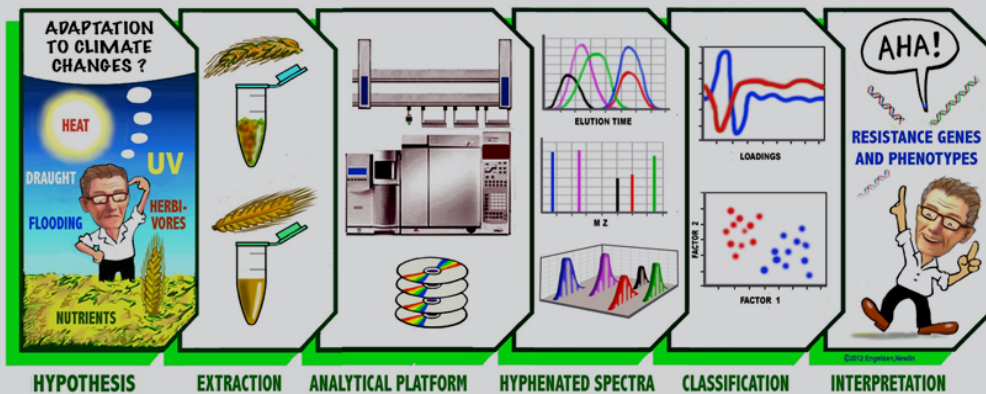




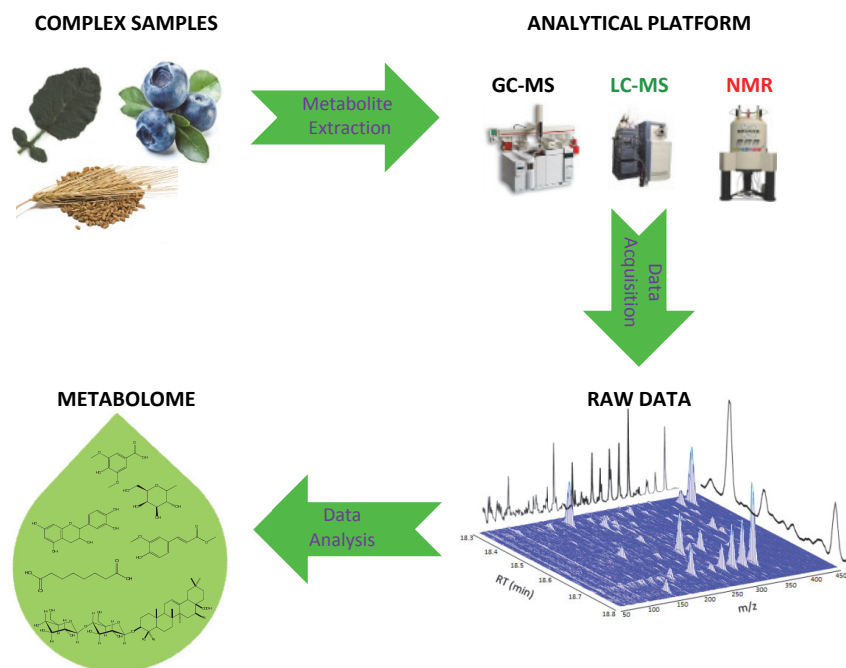
# Metabolomics and bioactive substances in plants

## PLANT METABOLOMICS



PHD THESIS · 2013  
BEKZOD KHAKIMOV BAHROMOVICH

# Metabolomics and bioactive substances in plants



PhD thesis  
Bekzod Khakimow Bahromovich

Spectroscopy & Chemometrics, Department of Food Science  
University of Copenhagen  
Denmark

## **Title**

Metabolomics and bioactive substances in plants

## **Submission**

September 2013

## **Supervisors**

Main supervisor:

Professor Søren Balling Engelsen, Spectroscopy & Chemometrics, Department of Food Science, Faculty of Science, University of Copenhagen

Co-supervisor:

Professor Søren Bak, Plant Biochemistry, Department of Plant and Environmental Sciences, Faculty of Science, University of Copenhagen

## **Opponents**

Assoc. Prof. Henrik Toft Simonsen (Chairman)  
Department of Plant and Environmental Sciences  
Faculty of Science, University of Copenhagen

Dr. Federico Marini  
Department of Chemistry  
University of Rome "La Sapienza"

Prof. Royston Goodacre  
School of Chemistry  
Manchester Institute of Biotechnology  
University of Manchester

ISBN 978-87-7611-654-5

Printed by SL grafik, Frederiksberg C, Denmark ([www.slgrafik.dk](http://www.slgrafik.dk))

*"The living cell does not necessarily follow the programme of the laboratory. Indeed it might be doubted whether any substances as such ever appear except at the end. The actual process might well be like a shuffling of cards, whereby the order of the cards is altered and the order or relative position is the important thing. On the anabolic side there is always the face; on the katabolic side there is always the back."*

Interpretation from the chapter 93 of *THE CANON OF MEDICINE*

Avicenna (Ibn Sina) 980 – 1037 A.D.



## ACKNOWLEDGEMENTS

This PhD project was financed by the Faculty of Science, University of Copenhagen for support to the elite-research area “Metabolomics and bioactive compounds” and conducted between the Quality & Technology (currently Spectroscopy & Chemometrics), Department of Food Science and Plant Biochemistry, Department of Plant and Environmental Sciences. I am grateful to all my colleagues and friends at Q&T and Plant Biochemistry who supported me during these three years. Most importantly, I would like to thank my supervisors Prof. Søren Balling Engelsen and Prof. Søren Bak thanks to whom I found myself in a warm and friendly scientific environment. Their endless support throughout the project, useful advices and openness to new ideas positively reflected in my work. I am thankful to Assoc. Prof. Frans van den Berg for his willingness to assist in any scientific and technical questions, to Assoc. Prof. Mohammed S. Motawia for useful discussions in organic chemistry, to José M. Amigo for his friendly advices in chemometrics, to Assoc. Prof. Flemming Hofmann for his assistance in NMR laboratory. I also would like to thank Assoc. Prof. Mikael A. Petersen for advices in GC-MS lab and Prof. Rasmus Bro for teaching me multi-way methods. Moreover, I appreciate help of Dr. Morten A. Rasmussen, Assoc. Prof. Birthe M. Jespersen and Prof. Lars Munck for opening the doors of wonderful barley world. Thanks to regular Barbarea group meetings I have learned a lot about plant-insect interactions and wish to thank all of the members of the group.

In addition, I would like to mention a great support of my friends whom I met in Copenhagen and became part of my life in Denmark. My special thanks to my friends, Daniela Rago, Gözde Gürdenize, Hamid Babamoradi, José M. Amigo, Maider Vidal, Sanni Matero, Francesco Savorani, Carl Emil Aae Eskildsen, Lotte Sørensen, Vera Kuzina, Anna Petersen and Signe Hoff. Despite being far away, my family was always the main support and I am thankful to my parents and to my sisters and two very important people in my life, Komronbek Khakimov and Shahina Khakimova. Moreover, I would like to thank my friend Otabek Suvonov who was always there for me, to share good moments and to overcome challenges. I am also grateful to my friends Suhrob, Zafar, Ulugbek and Jamol for their emails and skype calls that were important for me during the last three years. Finally, I thank Maria Pushkareva for her support and designing figures of this thesis.

## ABSTRACT

Metabolomic analysis of plants broadens understanding of how plants may benefit humans, animals and the environment, provide sustainable food and energy, and improve current agricultural, pharmacological and medicinal practices in order to bring about healthier and longer life. The quality and amount of the extractable biological information is largely determined by data acquisition, data processing and analysis methodologies of the plant metabolomics studies. This PhD study focused mainly on the development and implementation of new metabolomics methodologies for improved data acquisition and data processing. The study mainly concerned the three most commonly applied analytical techniques in plant metabolomics, GC-MS, LC-MS and NMR. In addition, advanced chemometrics methods e.g. PARAFAC2 and ASCA have been extensively used for development of complex metabolomics data processing and analysis methods. The first study (*Journal of Chromatography A*, 1266 (2012) 84–94) demonstrated how the application of a multi-way decomposition method, PARAFAC2, can help in providing maximum extraction of metabolite features from the raw LC-MS data obtained from complex plant extracts. The second study (*Analytical and Bioanalytical Chemistry*, In Press, DOI: 10.1007/s00216-013-7341-z) outlines a novel GC-MS derivatization method using TMSCN for trimethylsilylation for improved analysis of complex biological mixtures. A review paper (*Journal of Cereal Science*, Accepted, DOI: 10.1016/j.jcs.2013.10.002) written for the Journal of Cereal Science comprises current analytical challenges and perspectives of cereal metabolomics with emphasis on new development in the use of multivariate data analysis methods for exploitation of the full information level in the analytical platforms. The fourth study (*Journal of Experimental Botany*, Submitted) combined the knowledge gained from the first and second studies and applied cutting-edge chemometric methods in a real case biological question related to barley breeding. This study revealed several biological questions associated with plant- environment, plant-gene mutation relationships and alterations of the plants' physiology during their development stages.

# LIST OF PUBLICATIONS

## Paper 1

**Khakimov, B.**, Amigo, J. M., Bak, S. and Engelsen, S. B. (2012). Plant metabolomics: Resolution and quantification of elusive peaks in liquid chromatography-mass spectrometry profiles of complex plant extracts using multi-way decomposition methods. *Journal of Chromatography. A* 1266, 84-94.

## Paper 2

**Khakimov, B.**, Mohammed, S. M., Bak, S. and Engelsen, S. B. (2013). The use of trimethylsilyl cyanide derivatization for robust and broad spectrum high-throughput gas-chromatography-mass spectrometry based metabolomics. *Analytical and Bioanalytical Chemistry*. In press, DOI: 10.1007/s00216-013-7341-z

## Paper 3

**Bekzod Khakimov**, Søren Bak, Søren Balling Engelsen. The art of high-throughput cereal metabolomics: Current analytical technologies, challenges and perspectives. *Journal of Cereal Science*. Accepted, DOI: 10.1016/j.jcs.2013.10.002

## Paper 4

**Bekzod Khakimov**, Morten Arendt Rasmussen, Birthe Møller Jespersen, Lars Munck, Søren Balling Engelsen. The emerging barley seed metabolome studied by mutant analysis and advanced GC-MS: evaluation of the effects of development stage, genotype and growth temperature by ASCA. *Journal of Experimental Botany*. Submitted.

## Supplementary Paper

### Paper 5

Augustin, J. M., Drok, S., Shinoda, T., Sanmiya, K., Nielsen, J. K., **Khakimov, B.**, Olsen, C. E., Hansen, E. H., Kuzina, V., Ekstrøm, C. T. et al. (2012). UDP-Glycosyltransferases from the UGT73C Subfamily in *Barbarea vulgaris* Catalyze Sapogenin 3-O-Glucosylation in Saponin-Mediated Insect Resistance. *Plant Physiology* 160, 1881-1895.

## Popular Science Paper

**Bekzod Khakimov**, Søren Balling Engelsen, Rasmus Bro, Lars Nørgaard. (2012). Plante-metabolomics: opdagelse af nye bioaktive stoffer med PARAFAC2. *Dansk Kemi*, 93, nr. 12



## ABBREVIATIONS

AMDIS – Automatic Mass Spectral Deconvolution and Identification System

ANOVA – Analysis of Variance

ASCA – ANOVA-Simultaneous Component Analysis

BPC – Base Peak Chromatogram

BSTFA - N,O-Bistrifluoroacetamide

CE – Capillary Electrophoresis

DAD – Diode Array Detector

DNA – DeoxyriboNucleic Acid

DoE – Design of Experiment

ECVA - Extended Canonical Variates Analysis

EI – Electron Ionization

FT - Fourier Transform

FTICR - Fourier Transform Ion Cyclotron Resonance

GC – Gas Chromatography

HCN – Hydrogen Cyanide

iCoshift - Interval-Correlation-shifting

iECVA – Interval Extended Canonical Variates Analysis

IR – Infrared spectroscopy

LC – Liquid Chromatography

MANOVA - multivariate-ANOVA

MCR – Multivariate Curve Resolution

MS – Mass Spectrometry

MSTFA - N-Methyl-N-(trimethylsilyl) trifluoroacetamide

MTBSTFA - N-Methyl-N-tert-butyl-dimethylsilyl trifluoroacetamide

NIR – Near Infrared spectroscopy

NIST - National Institute of Standards and Technology

NMR – Nuclear Magnetic Resonance spectroscopy

OSC - Oxidosqualene Cyclase

PARAFAC – PARAllel FACtor Analysis

PARAFAC2 - PARAllel FACtor Analysis2

PCA – Principal Component Analysis

PLS – Partial Least Squares regression

RNA - Ribonucleic Acid

SIM – Selected Ion Monitoring

TIC – Total Ion Current

TLC – Thin Layer Chromatography

TMS – Trimethylsilyl

TMSCN – Trimethylsilyl Cyanide

TOF – Time-of-Flight

UGT - Uridine Diphosphate (UDP)-Glycosyltransferases

UV-VIS - Ultraviolet–Visible spectroscopy

## CONTENTS

Acknowledgements.....	- 5 -
Abstract .....	- 6 -
List of Publications .....	- 7 -
Abbreviations .....	- 8 -
1 Introduction.....	- 10 -
1.1 Metabolomics towards improved health, food and environment.....	- 10 -
1.2 The aim of the project.....	- 12 -
1.3 Brief description of the thesis and publications .....	- 12 -
2 Plant metabolomics.....	- 15 -
2.1 Background and state of the art .....	- 15 -
2.2 Plant bioactive substances.....	- 25 -
2.3 High-throughput metabolomics in development and improvement of crop plant cultivars.....	- 27 -
3 Advanced GC-MS and chemometrics for metabolome analysis .....	- 31 -
3.1 GC-MS .....	- 31 -
3.2 Trimethylsilyl cyanide (TMSCN) based derivatization .....	- 35 -
3.3 Design of experiment (DoE).....	- 39 -
3.4 Minimization of non-sample related variations in metabolomics.....	- 44 -
3.5 PARAllel FACtor Analysis 2 (PARAFAC2).....	- 46 -
3.6 ANOVA-simultaneous component analysis (ASCA) .....	- 52 -
4 Unpublished studies .....	- 55 -
4.1 Optimization of comprehensive metabolomic protocol for GC-MS, LC-MS and NMR analysis of <i>Barbarea vulgaris</i> leaves .....	- 55 -
4.1.1 GC-MS method optimization.....	- 56 -
4.1.2 1D H <sup>1</sup> NMR analysis of G and P type <i>Barbarea vulgaris</i> plant leaves .....	- 63 -
4.1.3 Tandem LC-MS analysis of G and P type <i>Barbarea vulgaris</i> plant leaves .....	- 67 -
4.2 TMSCN based derivatization and GC-MS detection of triterpenes produced by combinatorial biochemistry in tobacco leaves.....	- 71 -
4.3 Structure elucidation of triterpenoid saponins of the insect resistant and susceptible <i>Barbarea vulgaris</i> plants.....	- 78 -
5 Outreach.....	- 81 -
5.1 Will plants save the planet and can plant metabolomics play a key role?.....	- 81 -
5.2 Perspectives .....	- 82 -
6 References .....	- 87 -

# 1 INTRODUCTION

## 1.1 Metabolomics towards improved health, food and environment

*"Nature is the best Artist, Architect, Engineer and Doctor."* Cells are probably the most sophisticated creations of nature and they are considered to be the building blocks of the surrounding life. Cells are the smallest unit of life in which several hundreds of different bio-chemical processes occur simultaneously. The information that is necessary to regulate all these processes is preserved in the nucleus of the cells, which mainly consist of DNA and RNA. DNA and RNA are the two important components of the cells that regulate their living and development. Genes are small regions of DNA and RNA that carry information on one or more quality traits of the corresponding cells and phenotype(s). The life cycle of the living cells is all about metabolite and protein synthesis, their interactions and degradations. All the physiological processes of the living cells involve chemical and biochemical reactions that determine the metabolomic status of cells. The metabolome of cells comprise all metabolites present in the given status of the cell, which to some extent continuously changes depending on internal and external factors. Cells are able to synthesize several hundred metabolites within a short time period. These are difficult or not possible to reproduce in modern laboratories. When exposed to pathogens or any other internal and external stress, cells are able to produce defense metabolites that are absent or present in very small amounts under normal conditions (stress-free conditions) in cells. This, in turn, gives rise to several questions such as, where do these metabolites come from, how are they synthesized, which genes possess information about these response reactions and how are these metabolites able to assist the cells to cope with stress?

The cell metabolome greatly differs between various tissues, organs and organisms and is strongly influenced by the surrounding environment and genotype. For instance, mammalian and plant cells have different structures; therefore, their metabolomes differ significantly. The metabolomes of both cell types are complex; however, in the thesis I will mainly focus on the plant cell metabolome. Plant cell metabolism is a complex system where several hundreds of metabolites are synthesized simultaneously, involving several biosynthetic pathways. Recent studies have established an understanding of the biosynthetic pathways of a few classes of metabolites and alterations of cell metabolomic equilibriums in response to some internal and external factors. Although the genome of some plants e.g. *Arabidopsis thaliana* and *Oryza sativa*, are fully sequenced and a great amount of transcriptomic and proteomic data are available, only a very small portion of the biochemistry present in the plant cells is understood. This is mainly due to the complexity of the systems; for example, several genes might influence one factor or several factors might be the function of one gene. The current state of plant science is far away from the point where one can visualize the plant cell metabolism as one whole system and answer all the related questions such as: 1. Which metabolites are present in cells, 2. How do the metabolites differ in various tissues, organs and organisms, 3.

## 1. INTRODUCTION

Which factors can alter the cell metabolome, 4. Where, when and how are these metabolites synthesized in cells, and 5. What are the roles of these metabolites in cell functions? In order to answer all these important questions one must be able to detect all these metabolites in the cells in a quantitative manner and elucidate their structures. This field of the research has become one of the biggest areas of "omics" technology in the post-genomic era and today it is called METABOLOMICS.

Plant metabolomics is a rapidly developing scientific field that focuses on quantitative and qualitative analysis of all the metabolites of plant cells. Analysis of the plant metabolome allows for understanding of the influence of the plant genes on the quality traits of the phenotypes (Fiehn, 2002). Today, more and more plant studies are applying metabolomics approaches to uncover systems biology, to understand natural defense mechanisms of plants against external stresses, to identify the main bioactive compounds that enhance the health-beneficial properties of crop plants and to evaluate effects of the environment and genetic modifications. In combination with proteomics and transcriptomics, metabolomic analysis of plants has become a powerful approach for identification of functions of genes, discovery of biomarkers and for elucidation of biosynthetic pathways. Therefore, today, plant metabolomics has become a key tool of plant science to understand plants' physiology, develop and improve new crop plants towards higher yield and resistance to the continuous climate challenges as well as to improve their health-beneficial values. Thus, *"gene-metabolome-phenotype analysis (GMP) can be considered as a train, which is taking us towards improved health-food-environment (HFE) much faster and safer than any other vehicle."*

However, due to the complexity of the plant cell metabolome and the limited capabilities of the current analytical technologies, plant metabolomics has not yet reached the stage where it can provide quantitative data for the whole metabolome. This in fact shields a small, but very important portion of the information that may result in extended and/or new knowledge in systems biology. Therefore, today, studies that involve metabolomic method developments for increasing the range of the detectable metabolites, sensitivity and reproducibility to establish the high-throughput protocols constitute a substantial part of the research in metabolomics.

Despite the recent advances in chromatographic and electrophoretic separation and spectrometric and spectroscopic detection of a wide range of metabolites of the complex biological mixtures, current metabolomics cannot provide a complete picture of metabolomes of phenotypes. Main challenges of the current metabolomics can be divided into two different classes: 1. Problems arising from the qualitative analysis and 2. Problems arising from the quantitative analysis. Although modern analytical platforms such as GC-MS, GC x GC-MS, LC-MS, LC x LC-MS, CE-MS and LC-NMR provide a vast amount of metabolomic data, it is not always possible to identify all the detected metabolites. This is mainly due to the high complexity of the metabolome of the different species and limitations of the comprehensive metabolomic databases. However, it is worth mentioning that global metabolomic databases e.g. MassBank, NIST, Wiley, as well as species specific metabolomic libraries are developing rapidly and some of them have become very useful in the identification of unknowns. Most of these

metabolite libraries are based on mass spectrometric data, e.g. EI-MS data from GC-MS, exact mass of metabolites from LC-MS experiments. Recently, much effort has been put into the development of metabolite libraries based on tandem MS and NMR data (Cui et al., 2008; Ludwig et al., 2012; Wishart, 2007; Wishart et al., 2013). However, these libraries are still to be developed and enriched before metabolomics labs can utilize them.

The second biggest challenge of metabolomics is associated with quantitative analysis. Obtaining quantitative metabolomic data is probably the most challenging step of comprehensive metabolomic studies. The simplest requirement of quantitative metabolomic analysis is that the analyst must make sure that concentrations of metabolites present in the sample mixture and their recorded responses have a one-to-one relationship and that we are able to estimate the level of error when this relationship is disturbed. High quantitative value of the metabolomic data requires great effort in all the steps of the metabolomic workflow, starting from the plant growing or animal living conditions over the sampling methodology to the analysis and interpretation of the data.

### 1.2 The aim of the project

*The main goal of this PhD study was to implement and develop new methodologies for improving metabolomic data acquisition and analysis in plant metabolomics studies.* The project mainly involved GC-MS, LC-MS and NMR based metabolomics and application of advanced chemometrics e.g. DoE, PARAFAC2 and ASCA for improved metabolomic data acquisition and data analysis.

### 1.3 Brief description of the thesis and publications

This PhD thesis comprises most of the work performed within the project. Section 1.1 provides a brief introduction to plant metabolomics and highlights where and how it is used and the main expected outcomes. In Section 2, the pros and cons of the main elements of plant metabolomics are described in more detail. The section covers aims of different metabolomic studies, e.g. targeted and untargeted analysis, demonstrates the main challenges of the quantitative data acquisition and briefly discusses the bioactive substances of the plants. In addition, Section 2 demonstrates the role of metabolomics in development and improvement of crop plants. The section also describe the most commonly used commercial and free metabolomic data processing software, metabolite databases available to date as well as other useful web sources that assist newcomers in the field. Section 3 describes advanced GC-MS and chemometrics-based plant metabolomics methodologies. The section demonstrates important considerations in high quality quantitative metabolomic data acquisition by GC-MS and the novel derivatization methodology based on trimethylsilyl cyanide (TMSCN) (Khakimov et al., 2013). The

section also demonstrates the use of the design of experiment (DoE) in optimization of plant metabolomics protocols and addresses the main sources of experimental errors, which introduces non-sample related variations. Section 3 outlines the use of advanced chemometric techniques for extracting hidden information from the LC-MS and GC-MS type of the metabolomics data by using the multi-way decomposition technique, PARAllel FACtor Analysis 2 (PARAFAC2), and holistic evaluation of the designed metabolomics data by using ANOVA-simultaneous component analysis (ASCA). Section 4 describes three separate studies performed within the project that are not yet condensed and submitted for publication. The first study demonstrates a development of the metabolomics protocol for profiling the saponin content as well as untargeted analysis of the F2 population of *B. vulgaris* plant leaves from the limited amount of the plant material (~ 10 mg). The second study depicts GC-MS analysis of the new triterpenoids produced by combinatorial biochemistry in tobacco leaves by using the constrains developed from the oxidosqualene cyclases and P450s of the *B. vulgaris* plants. Finally, the third study demonstrates purification of the saponin content of the P and G type *B. vulgaris* plants, prior to structure elucidation and tentative characterization of the major saponins detected from a LC-MS/MS analysis.

**Paper 1** (published) demonstrates the first application of the multi-way decomposition method, PARAllel FACtor Analysis 2 (PARAFAC2), to LC-MS based metabolomics data. This paper demonstrates resolution and quantification of the elusive peaks of the possible bioactive triterpenoid saponins of the *Barbarea vulgaris* plants against an insect herbivore, *Phyllotreta nemorum*. Moreover, it provides a tutorial on the use and validation of the PARAFAC2 method in conjunction with LC-MS data.

**Paper 2** (accepted) describes the development of a novel derivatization technique for GC-MS based metabolomics of complex biological mixtures. In this work, for the first time we have used trimethylsilyl cyanide (TMSCN) as a trimethylsilylation reagent for the comprehensive GC-MS analysis of a wide range of polar and non-volatile metabolites. The silylation capabilities of TMSCN are compared to the most commonly used silylation reagent MSTFA. The results of the analysis showed that TMSCN-based derivatization outperforms MSTFA-based silylation in terms of reaction speed, sensitivity and repeatability of GC-MS profiles. Moreover, the paper highlights and discusses some of the crucial aspects of the comprehensive GC-MS analysis including automation, importance of consistent derivatization time, sample preparation and injection.

**Paper 3** (accepted) is a comprehensive review of cereal metabolomics. The paper describes state-of-the-art analytical technologies, current challenges and perspectives of high-throughput cereal metabolomics. The review is mainly written with an aim to assist scientists who are not specialists in metabolomics, chemometrics and analytical chemistry in gaining an overview of the current state of the art in metabolomics. The paper describes all the steps of the metabolomic workflow from sample harvesting to the data analysis and interpretation. Each analytical platform is described separately and advantages and limitations highlighted. The main sources of bias introduced into the metabolomic

## 1. INTRODUCTION

data are discussed and possible solutions are provided. Moreover, the review describes a raw metabolomic data preprocessing tools, supervised and unsupervised chemometric methods of data analysis and provides selected examples of applications.

**Paper 4** (Under review) demonstrates GC-MS metabolomic profiling of flour from whole-grain barley seeds of three genetically different cultivars and revealed dynamics of barley metabolome during the grain-filling period and effects of growing temperature and genotype. The study revealed detection of 247 metabolites, 89 of which were identified based on their EI-MS and RIs. In this work all the above mentioned methodologies developed within the project, e.g. TMSCN based derivatization, comprehensive GC-MS profiling and PARAFAC2 based chromatographic data processing were employed on the real biological samples. The study also shows the power of the ANOVA-simultaneous component analysis (ASCA) for exploration of the metabolomics data derived from the designed experiment.

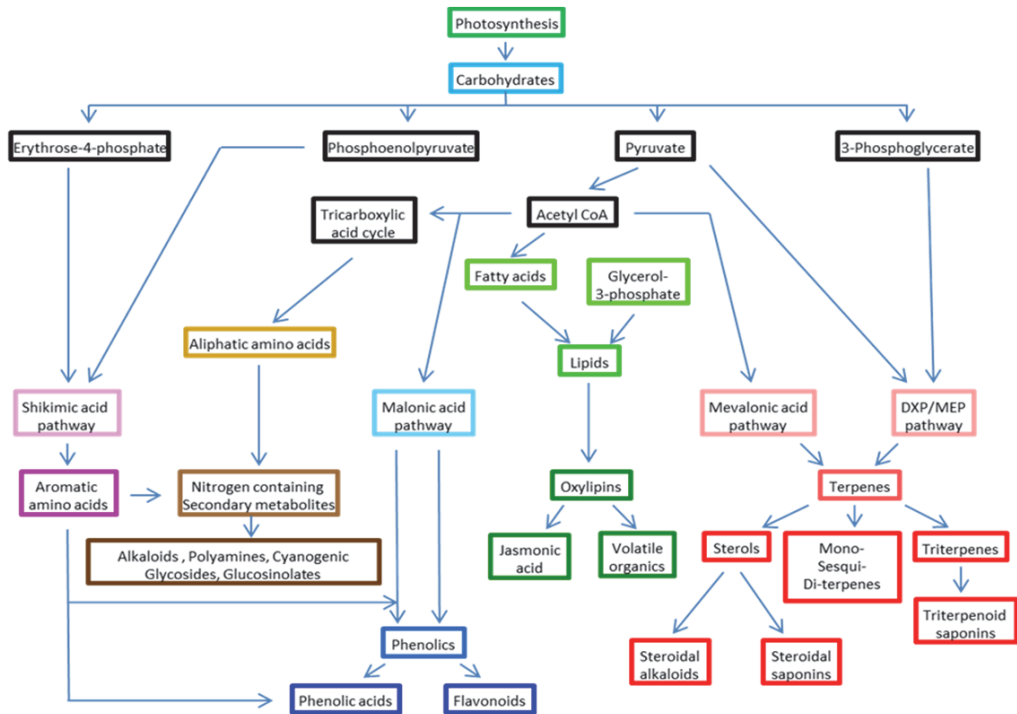
## 2 PLANT METABOLOMICS

### 2.1 Background and state of the art

Plant metabolites usually refers to small molecules with MW of up to 1500 Da that are intermediate and/or final products of the plant cell metabolism (Veber et al., 2002; Wishart, 2007). Plant cell metabolism is the complex physico-chemical event comprising photosynthesis, respiration, and biosynthesis and degradations of a broad range of organic molecules. Plant cell metabolism covers all the biochemical transformations occurring within the plant organism to sustain life, promote growth, reproduction, defense and response to the surrounding environment. These metabolomic transformations are usually divided into two main categories: catabolism (produce energy by breaking molecules into smaller units) and anabolism (use energy for constructing cell molecules from smaller units) reactions that are catalyzed by the several different enzymes. Plant metabolites are also divided in two classes: primary (metabolites that are directly involved in growth and development) and secondary (metabolites that are not directly involved in plant's growth, but possess important functions such as defense against various stresses) metabolites. As the whole systems, collections of all these metabolites within the cells, tissues and organisms are called metabolomes of the biological systems (Oliver et al., 1998). Thus, the metabolome of a plant is complex and it continuously changes throughout the lifetime of plants. Several biosynthetic pathways work simultaneously within each cell and one metabolite might be the part of several pathways. Hundreds of different enzymes that are functions of different genes alter the plant cell metabolome in different ways within seconds. Different types of plant cells have their own unique metabolome and accordingly, the metabolomes of different tissues and organs differ and make distinct contributions to the metabolome of the organism as a whole. The main biosynthetic pathways start with the carbohydrates synthesized during photosynthesis and involve shikimic acid pathway, pentose-phosphate pathway, mevalonic acid pathway, malonic acid pathway, malonyl CoA and other pathways (Bentley, 1999; Ganem, 1978; Koffas et al., 1999; Lynene, 1967) (Figure. 1). Each of these biosynthetic pathways shares some common metabolites and even enzymes, though a whole metabolomic profile and the end products are unique to each pathway. However, due to pleiotropy, metabolites of different pathways that are believed to be unrelated to each other may also be altered simultaneously. Information about regulations of these complex biochemical routes are encoded in the form of nucleotide sequences of the messenger ribonucleic acids (mRNAs) that are encoding the corresponding biocatalysts (enzymes) when it is demanded. Thus, plant metabolism is a well-controlled, complex, but flexible system.



## 2. PLANT METABOLOMICS



**Figure 1.** Simplified overview of major pathways of secondary-metabolite biosynthesis and their interrelationships with primary metabolism (modified from Schmidt (Schmidt et al., 2005)).

During plant development, reproduction as well as genetic modifications, internal and external stress cause significant alteration in plant cell metabolome. Observation of all these biochemical transformations is not possible yet. However, quenching the system and observing the given state of the cell metabolome is becoming a common approach to understand the complex biochemical processes that are occurring within cells. Although it has long been known that dereplication of chemical composition of plants may provide more insight into their biology and chemistry, very little was done until the Russian scientist Mikhail Tsvet invented chromatography in the early 20<sup>th</sup> century. He documented the first method for separation of plant pigments, chlorophyll and carotenoids, which brought a huge resonance in the field and initiated further research on decomposition of plants and animal-derived samples. The field developed relatively slower until the early 1950's when the gas chromatography-mass spectrometry era began (Gohlke and Mcclafferty, 1993). In the late 1960's, after improvements in GC-MS technology for performing comprehensive analyses, the measure of biological systems' responses to various external factors significantly improved. The first attempts to understand the complex metabolism of cells were performed by using GC-MS based detection of metabolites and the term "metabolomic profile" was introduced (Horning and Horning, 1971; Horning et al., 1968; Horning, 1971). Later, by the development of new analytical platforms and data analysis statistical

## 2. PLANT METABOLOMICS

methods, analysis of biological systems' metabolomes became even more common for revealing functions of genes, systems biology, mechanisms of diseases, and a new field was born, namely METABOLOMICS. Metabolomics covers systematic analysis of cell metabolomes in a quantitative manner that facilitates understanding of cell responses to pathophysiological stimuli or genetic modification. The approach was pioneered in late 1990's in the field of toxicology (Nicholson et al., 1999).

In contrast to genomics, proteomics and transcriptomics, metabolomics provides a rapid and closer view to systems biology and enables evaluation of genes and proteins to the level of phenotypes (Fiehn, 2002). In its simplest term, plant metabolomics can be described as the approach based on quantitative and/or qualitative measurements of low molecular metabolites of plants as the function of genetic modifications and various biotic and abiotic stresses. Metabolomics itself has been divided into different approaches based on the purpose of studies and the type of information obtained from the metabolomic analysis (Dunn, 2008; Fiehn, 2002). Three main metabolomic analysis approaches include targeted analysis, metabolomic profiling, and metabolomic fingerprinting. Targeted analysis is usually employed in screening studies where metabolite(s) of interest are known in advance. For example, quantitative analysis of mycotoxins from food and/or food raw materials (Rahmani et al., 2009; Royer et al., 2004) or quantification of defense secondary metabolites of plants induced during the herbivore attacks. Targeted metabolomics entails extensive prior sample preparation in order to improve metabolite recovery and quantification. In addition, many targeted analysis studies employ selected screening by observing characteristic m/z ions (selected ion monitoring (SIM) experiments can be performed using LC- and GC-MS) and/or investigating only the retention time region where target metabolite(s) elute. In turn, this may result in enhanced quantification and reduce experimental time and cost. Metabolomic profiling is a probably the mostly applied approach, which is based on semi-quantitative (only few profiling studies use absolute quantification) analysis of pre-defined class or classes of metabolites. Profiling usually focuses on one or two classes of metabolites such as phenolics, organic acids (these two classes can be analyzed using one protocol, see Paper 4), amino acids, isoprenoids or carbohydrates. Most metabolomic profiling studies target to observe the alterations occurred in one or more biosynthetic pathways. Therefore, metabolite extraction and sample preparation protocols are usually focused on desired classes of compounds. This approach applies various analytical platforms such as LC-MS (Kuzina et al., 2009), GC-MS (Roessner et al., 2000), NMR (Savorani et al., 2010a). Metabolomic fingerprinting is a rapid measurement of systems' metabolome showing their characteristic patterns. Usually, this approach does not involve extensive sample preparation or purification, but focuses on rapid snapshots of phenotypes. Metabolomic fingerprinting has found a wide application in rapid classification of biological samples as well as in screening and disease diagnosis (Chen et al., 2007; Choi et al., 2005; Kruger et al., 2008; Mattoli et al., 2006). This approach mainly applies spectroscopic techniques such as NMR, NIR, FT-IR, fluorescence, direct fusion mass spectrometry, GC-MS and LC-MS.

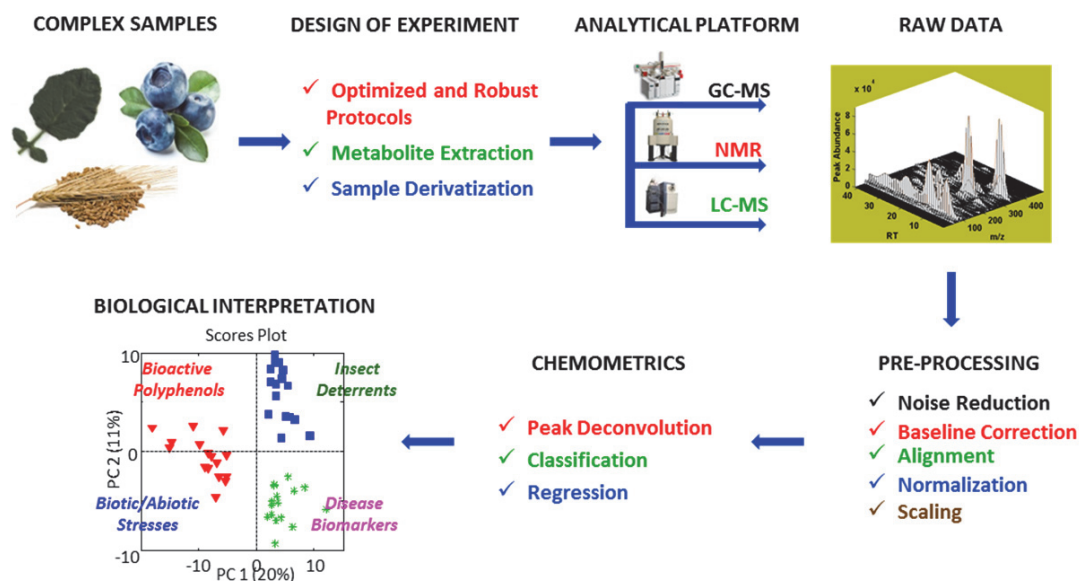
## 2. PLANT METABOLOMICS

Recent advances of analytical platforms and developments of metabolomic protocols and data analysis methods have enabled more insight into the metabolome of plants (Allwood and Goodacre, 2010; Fiehn, 2008; Lindon and Nicholson, 2008; Lisec et al., 2006; Okazaki and Saito, 2012). Today, more than 200,000 metabolites belonging to the plants' kingdom are known, including both primary and secondary metabolites (Fiehn, 2002). Among these, secondary metabolites, such as terpenes, alkaloids and phenolics, comprise 43,000, 12,000 and 8,000 metabolites, respectively (Bernhoft, 2013; Chen et al., 2011). Each plant cell contains several thousand distinct metabolites. Some of these metabolites possess biological activity, serve as disease biomarkers, and assist in elucidation of new biosynthetic pathways and functions of genes. However, current state-of-the-art metabolomic technologies are not capable of detecting all metabolites of plant cells in a single method. Most comprehensive and relatively high-throughput analytical techniques that are frequently applied in various metabolomics studies comprise NMR, GC-MS, LC-MS, CE-MS, FT-IR and LC-DAD (see Paper 3). In the case of metabolomic analysis of complex samples by applying optimized metabolomic protocols, these techniques allow detection of up to 50 (NMR) or several hundred (e.g. up to 500 in GC-MS, LC-MS, CE-MS) most abundant metabolites of the investigated sample mixtures. These numbers, however, are increasing by continuous developments of analytical platforms e.g. GC × GC-MS, LC × LC-MS, FTICR-MS, LC-NMR, raw data processing methods e.g. PARAFAC2, AMDIS, MCR and improvements of unbiased and global metabolite extraction and derivatization protocols. Nevertheless, a detectible part of the metabolome in a single method (a single metabolomic protocol and one analytical platform) still remains significantly lower compared to the actual metabolome of the cells, at the given status. In fact, identification of unknown plant metabolites and broadening of the detectible part of the metabolome is one of the main challenges of current metabolomics. The ultimate goal of several studies focuses on developments of metabolomic technologies to increase the detectible part of the metabolome, which in turn leads to improve the capabilities of current metabolomics and may further assist in identification of several functions of genes simultaneously or allow better understanding of the pleiotropic effects and biological systems in general. However, it is worth mentioning that the analytical techniques such as NMR, LC-MS and GC-MS are capable of detecting a very broad range of metabolites; therefore, the physico-chemical diversity of the metabolome is not any more the primary limitation of metabolomics. In fact, the separation power and the sensitivity of the techniques are the main limiting factors, since the concentration of cell metabolites at any given time may significantly vary and the techniques are able to detect only the first few hundred most abundant metabolites, while the low concentration metabolites remain unseen and/or hindered by the more abundant metabolites. *Thus, the ultimate goal of metabolomics method development studies can be reached when the detectible part of the metabolome will be equal or close to the actual metabolome of the investigated sample matrix.*

An overview of the general plant metabolomic workflow for the comprehensive metabolomics studies aimed at uncovering targeted or untargeted biochemical phenomena is demonstrated in Figure 2. Hypothetically, we can image this as one integrated system employing several separation and

## 2. PLANT METABOLOMICS

detection techniques simultaneously, and the complex sample matrix, e.g. plant leaf, seed or drop of plasma, can be directly introduced into the system and automated robots perform data acquisition. Similar systems such as LC-NMR, LC-NMR-MS have already shown a high potential for automated separation and sample up-concentration, followed by detection (Jaroszewski, 2005a; Jaroszewski, 2005b; Jaroszewski, 2007). However, due to technical complications and high cost, these analytical platforms are not yet commonly used in comprehensive metabolomics studies.



**Figure 2.** Overview of the Plant Metabolomics Workflow

Today, the majority of metabolomic studies are purpose-oriented and deal with absolute or semi-quantitative detection of prior known and/or unknown metabolites from various sample matrices. In plant metabolomics, the main determining factor is the type of the investigated metabolites, e.g. polar, semi-polar, non-polar metabolites or conjugated metabolites, to other cell components, such as phenolics and saponins. The choice of metabolite extraction method, e.g. solvent, mechanical stirring, time and temperature, is probably the most important factor for obtaining the desired metabolomic data. Three main features of the investigated metabolites, concentration (approximate), range of the molecular masses and polarity of metabolites, play a key role in optimization of metabolite extraction protocols and determine which analytical platform to use. It is worth mentioning that not many comprehensive plant metabolomic studies apply appropriate optimization of metabolomic protocols by considering the most important factors and testing the protocols to improve their robustness, since

## 2. PLANT METABOLOMICS

only the reliable and robust protocols may provide meaningful data, thus true biological information. Table 1 comprises free and commercial metabolomic databases, data preprocessing/analysis software available to date, and lists some of the useful websites and outstanding metabolomics-related review papers to assist newcomers in the field.

<b>Metabolomic Data Bases</b>	
I.	<p><b>Wiley &amp; NIST</b>            Combined Wiley 10<sup>th</sup> edition + NIST 11/12 possess 870,000 spectra obtained from GC-EI-TOF, GC-EI-Q and LC-MS/MS  <a href="http://www.sisweb.com/software/ms/wiley.htm#registrynistcombined">http://www.sisweb.com/software/ms/wiley.htm#registrynistcombined</a>  <a href="http://webbook.nist.gov/chemistry/">http://webbook.nist.gov/chemistry/</a></p>
II.	<p><b>Golm Metabolome Database (GMD)</b>            GC-EI-TOF and GC-EI-Q Mass Spectral (MS) and Retention Time Index (RI) Libraries (MSRI) of metabolites from various biological systems (Kopka et al., 2005)  <a href="http://gmd.mpimp-golm.mpg.de/">http://gmd.mpimp-golm.mpg.de/</a></p>
III.	<p><b>MassBank</b>            GC-EI-TOF, GC-EI-QQ, LC-ESI-IT, LC-ESI-Q, LC-ESI-QQ, LC-ESI-QIT, LC-ESI-ITFT, LC-ESI-QTOF, LC-ESI-ITTOF, LC-APPI-QQ, LC-APCI-QTOF, MALDI-TOF and CE-ESI-TOF libraries of metabolites (Horai et al., 2010)  <a href="http://www.massbank.jp/">http://www.massbank.jp/</a></p>
IV.	<p><b>Fiehn GC-MS Database</b>            Contain GC-EI-Q and GC-EI-TOF data of metabolites (Kind et al., 2009)  <a href="http://fiehnlab.ucdavis.edu/Metabolite-Library-2007">http://fiehnlab.ucdavis.edu/Metabolite-Library-2007</a></p>
V.	<p><b>METLIN</b>            Comprises Mass Spectral data of more than 64,000 metabolites obtained using LC-ESI-QTOF (Smith et al., 2005)  <a href="http://metlin.scripps.edu/index.php">http://metlin.scripps.edu/index.php</a></p>
VI.	<p><b>The Human Metabolome Database (HMDB)</b>            Contain chemical, clinical and biochemical data of 41,519 metabolites found in the human body (Wishart et al., 2013)  <a href="http://www.hmdb.ca/">http://www.hmdb.ca/</a></p>
VII.	<p><b>The Madison Metabolomics Consortium Database (MMCD)</b>            Contain chemical formula, names and synonyms, structure, physical and chemical properties of more than 10,00 metabolites reported in the literature and combine 1D, 2D NMR and MS data for 500 metabolites (Cui et al., 2008)  <a href="http://mmcd.nmrfam.wisc.edu/">http://mmcd.nmrfam.wisc.edu/</a></p>
VIII.	<p><b>The Birmingham Metabolite Library Nuclear Magnetic Resonance database (BML-NMR)</b>            Contain 3328 experimental 1D and 2D J-resolved NMR spectra of 208 metabolite standards (Ludwig et al., 2012)  <a href="http://www.bml-nmr.org/">http://www.bml-nmr.org/</a></p>
IX.	<p><b>Kyoto encyclopedia of genes and genomes (KEGG)</b>            Large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies  <a href="http://www.genome.jp/kegg/">http://www.genome.jp/kegg/</a></p>

## 2. PLANT METABOLOMICS

X.	<p><b>MetaCyc (encyclopedia of metabolic pathways)</b>  MetaCyc contains more than 2042 pathways from more than 2414 different organisms involved in both primary and secondary metabolism, as well as associated compounds, enzymes, and genes  (Caspi et al., 2012)  <a href="http://metacyc.org/">http://metacyc.org/</a></p>
XI.	<p><b>Reactome</b>  Reactome is an open-source, open access, manually curated and peer-reviewed pathway database. It includes biological pathways, signaling, innate and acquired immune function, transcriptional regulation, translation, apoptosis and classical intermediary metabolism and possess several Entities (nucleic acids, proteins, complexes and small molecules) participating in these reactions (Croft et al., 2011)  <a href="http://www.reactome.org/">http://www.reactome.org/</a></p>
XII.	<p><b>Chemical Entities of Biological Interest (ChEBI)</b>  ChEBI is one of the most comprehensive chemical compounds database comprising data from four different databases (the Integrated relational Enzyme (IntEnz), KEGG, Chemical Component Dictionary of the Protein Data Bank (PDBeChem), chemical database of bioactive molecules with drug-like properties (ChEMBL) (Hastings et al., 2013)  <a href="http://www.ebi.ac.uk/chebi/">http://www.ebi.ac.uk/chebi/</a></p>
XIII.	<p><b>PubChem</b>  PubChem consist of three components, PubChem Substance, PubChem Compound, and PubChem BioAssay and provide information about small molecules' biological activity, physico-chemical properties, chemical structure similarity search and links to biological properties of the metabolites via PubMed scientific literature (Wang et al., 2009b; Wang et al., 2010; Wang et al., 2012)  <a href="http://pubchem.ncbi.nlm.nih.gov/">http://pubchem.ncbi.nlm.nih.gov/</a></p>
XIV.	<p><b>ChemSpider</b>  Free database with access to over 29 million structures, properties and associated information by integrating and linking compounds from more than 440 data sources (Ekins, 2009; Williams and Tkachenko, 2010; Williams and Tkachenko, 2011)  <a href="http://www.chemspider.com/">http://www.chemspider.com/</a></p>
XV.	<p><b>Biological Magnetic Resonance Data Bank (BMRB)</b>  BMRB contains 1D and 2D NMR data for wide range of pathway and/or species specific biopolymers and metabolites  <a href="http://www.bmrwisc.edu/metabolomics/">http://www.bmrwisc.edu/metabolomics/</a></p>
XVI.	<p><b>Spectral Data Base for Organic Compounds (SDBS)</b>  SDBS contains EI-MS, 1H and 13C NMR, FT-IR, Raman and Electron Spin Resonance (ESR) data 34,000 compounds  <a href="http://sdbs.riodb.aist.go.jp/sdbs/cgi-bin/cre_index.cgi">http://sdbs.riodb.aist.go.jp/sdbs/cgi-bin/cre_index.cgi</a></p>
XVII.	<p><b>MetaboLights</b>  MetaboLights is an open-access database which is cross-species, cross-technique and covers metabolite structures and their reference spectra as well as their biological roles, locations and concentrations, and experimental data from metabolic experiments (Haug et al., 2013)  <a href="http://www.ebi.ac.uk/metabolights/">http://www.ebi.ac.uk/metabolights/</a></p>
XVIII.	<p><b>BiGG</b>  A Biochemical Genetic and Genomic knowledgebase of large scale metabolic</p>

## 2. PLANT METABOLOMICS

	<p>reconstructions</p> <p>The database accounts for the functions of 1,496 ORFs, 2,004 proteins, 2,766 metabolites, and 3,311 metabolic and transport reactions (Schellenberger et al., 2010)</p> <p><a href="http://bigg.ucsd.edu/">http://bigg.ucsd.edu/</a></p>
XIX.	<p><b>KNAPSAck Metabolomics</b></p> <p>This is a plant metabolomic database comprising 20,741 species and 50,048 metabolites and allow metabolites search from MS peak, molecular weight and molecular formula, and species (Afendi et al., 2012)</p> <p><a href="http://kanaya.naist.jp/KNAPSAck/">http://kanaya.naist.jp/KNAPSAck/</a></p>
<p><b>Metabolomic Data Processing Software</b></p>	
I.	<p><b>PARAFAC2</b></p> <p>Processing GC-MS, LC-MS, LC-DAD and CE-MS type of three-way data (Bro et al., 1999)</p> <p><a href="http://models.life.ku.dk/algorithms">http://models.life.ku.dk/algorithms</a></p>
II.	<p><b>iCoshift</b></p> <p>Alignment of NMR, LC-UV, GC-FID and CE-UV type of two dimensional data (Savorani et al., 2010b)</p> <p><a href="http://models.life.ku.dk/algorithms">http://models.life.ku.dk/algorithms</a></p>
III.	<p><b>Correlation Optimized Warping (COW)</b></p> <p>Alignment of NMR, LC-UV, GC-FID and CE-UV type of two dimensional data (Nielsen et al., 1998)</p> <p><a href="http://models.life.ku.dk/algorithms">http://models.life.ku.dk/algorithms</a></p>
IV.	<p><b>DOSY Toolbox</b></p> <p>Processing Diffusion-Ordered Spectroscopy NMR data (Nilsson and Morris, 2008)</p>
V.	<p><b>FastChrom</b></p> <p>Matlab-based method for baseline correction, peak detection, and assignment (grouping) of similar peaks across samples from single-channel data (e.g. GC-FID) and multi-channel data (e.g. total ion chromatogram from GC-MS) (Johnsen et al., 2013)</p> <p><a href="http://models.life.ku.dk/algorithms">http://models.life.ku.dk/algorithms</a></p>
VI.	<p><b>Load2Chrom</b></p> <p>Performs automatic peak assignment using PARAFAC2, PARAFAC and MCR based deconvoluted mass spectra and comparing against metabolite data bases e.g., NIST (Murphy et al., 2012)</p> <p><a href="http://models.life.ku.dk/algorithms">http://models.life.ku.dk/algorithms</a></p>
VII.	<p><b>AMDIS</b></p> <p>The Automated Mass Spectral Deconvolution and Identification System (AMDIS) extracts spectra for individual components in a GC/MS data file and identifies compounds by matching these spectra against a reference library (Stein, 1999)</p> <p><a href="http://chemdata.nist.gov/mass-spc/amdis/">http://chemdata.nist.gov/mass-spc/amdis/</a></p>
VIII.	<p><b>GAVIN</b></p> <p>Matlab based free software complement to AMDIS for processing GC-MS metabolomic data (Behrends et al., 2011)</p>
IX.	<p><b>MetAlign</b></p> <p>Metabolomic data preprocessing software for LC-MS and GC-MS type of data (Lommen, 2009; Lommen and Kools, 2012)</p> <p><a href="http://www.wageningenur.nl/en/show/MetAlign.htm">http://www.wageningenur.nl/en/show/MetAlign.htm</a></p>

## 2. PLANT METABOLOMICS

X.	<p><b>MZmine</b> Metabolomic data preprocessing software for LC-MS and GC-MS type of data (Katajamaa et al., 2006; Pluskal et al., 2010) <a href="http://mzmine.sourceforge.net/">http://mzmine.sourceforge.net/</a></p>
XI.	<p><b>XCMS</b> Metabolomic data preprocessing software for LC-MS and GC-MS type of data (Smith et al., 2006; Tautenhahn et al., 2012) <a href="http://metlin.scripps.edu/xcms/">http://metlin.scripps.edu/xcms/</a></p>
XII.	<p><b>MarkerLynx</b> Metabolomic data preprocessing software for LC-MS and GC-MS type of data (Waters, UK) <a href="http://www.waters.com/waters/de_DE/MarkerLynx-/nav.htm?locale=de_DE&amp;cid=513801">http://www.waters.com/waters/de_DE/MarkerLynx-/nav.htm?locale=de_DE&amp;cid=513801</a></p>
XIII.	<p><b>MetaboliteDetector</b> Deconvolution and analysis of GC-MS based metabolomics data (Hiller et al., 2009) <a href="http://md.tu-bs.de/">http://md.tu-bs.de/</a></p>
XIV.	<p><b>MetabolomeExpress</b> Web-based processing and analysis of GC-MS based metabolomics data (Carroll et al., 2010) <a href="https://www.metabolome-express.org/">https://www.metabolome-express.org/</a></p>
XV.	<p><b>MetaboloAnalyst</b> Web-based analytical pipeline for high-throughput metabolomics studies, which provide processing e.g. baseline correction, alignment, peaks detection, normalization and multivariate statistical analysis, as well as data annotation for various MS and NMR based metabolomics data (Xia et al., 2009; Xia et al., 2012) <a href="http://www.metaboanalyst.ca/">http://www.metaboanalyst.ca/</a></p>
XVI.	<p><b>MetaboMiner</b> This is a Java based software that automatically or semi-automatically identifies metabolites in complex bio-fluids from 2D NMR spectra, including 1H-1H total correlation spectroscopy (TOCSY) and 1H-13C heteronuclear single quantum correlation (HSQC) data using libraries such as HMDB, HMCD, MMRB and other databases (Xia et al., 2008) <a href="http://wishart.biology.ualberta.ca/metabominer/">http://wishart.biology.ualberta.ca/metabominer/</a></p>
XVII.	<p><b>OpenMS</b> This is an open-source software C++ library for LC/MS data management and analyses. It offers an infrastructure for the development of mass spectrometry related software and powerful 2D and 3D visualization (Lange et al., 2005; Sturm et al., 2008) <a href="http://open-ms.sourceforge.net/">http://open-ms.sourceforge.net/</a></p>
XVIII.	<p><b>Seven Golden Rules Software</b> Calculates molecular formulas from high resolution mass spectrometry data by considering seven heuristic rules: (1) restrictions for the number of elements, (2) LEWIS and SENIOR chemical rules, (3) isotopic patterns, (4) hydrogen/carbon ratios, (5) element ratio of nitrogen, oxygen, phosphor, and sulphur versus carbon, (6) element ratio probabilities and (7) presence of trimethylsilylated compounds (Kind and Fiehn, 2007) <a href="http://fiehnlab.ucdavis.edu/projects/Seven_Golden_Rules/">http://fiehnlab.ucdavis.edu/projects/Seven_Golden_Rules/</a></p>
XIX.	<p><b>PolySearch</b> This is a web-based tool that supports more than 50 different classes of queries (e.g., diseases, tissues, cell compartments, gene/protein names, SNPs, mutations, drugs and metabolites) against nearly a dozen different types of text, scientific abstract or bioinformatics databases (Cheng et al., 2008) <a href="http://wishart.biology.ualberta.ca/polysearch/index.htm">http://wishart.biology.ualberta.ca/polysearch/index.htm</a></p>



## 2. PLANT METABOLOMICS

XX.	<b>COLMAR</b> Complex Mixture Analysis by NMR is a web-based identification of metabolites based on their chemical shifts and J-coupling constants (COLMAR) (Robinette et al., 2008) <a href="http://spinportal.magnet.fsu.edu/">http://spinportal.magnet.fsu.edu/</a>
XXI.	<b>FiD</b> Fragment iDentificator (FiD) is a software that allows structural identification of product ions produced with tandem mass spectrometric measurement of low molecular weight organic compounds by search over all possible fragmentation paths and outputs a ranked list of alternative structures (Heinonen et al., 2008) <a href="http://www.cs.helsinki.fi/group/sysfys/software/fragid/">http://www.cs.helsinki.fi/group/sysfys/software/fragid/</a>
XXII.	<b>MetATT</b> This is a web-based metabolomics tool for analyzing two-factor and time-series data and offer PCA, ANOVA, ASCA, and other multivariate methods of analysis and data interpretation (Xia et al., 2011) <a href="http://metatt.metabolomics.ca/">http://metatt.metabolomics.ca/</a>
XXIII.	<b>MSFACTs</b> It is a standard Java/Swing application that imports, aligns, and reformats spectral and chromatographic data e.g. GC-MS, UV, IR and NMR (Duran et al., 2003) <a href="http://www.noble.org/plantbio/sumner/msfacts/">http://www.noble.org/plantbio/sumner/msfacts/</a>
XXIV.	<b>msInspect</b> This is an open-source software for rapid inspection and processing of LC-MS data (May et al., 2007) <a href="http://proteomics.fhcrc.org/CPL/home.html">http://proteomics.fhcrc.org/CPL/home.html</a>
XXV.	<b>MathDAMP</b> This software allows preprocessing, normalization and visualization of GC-MS, LC-MS and CE-MS type of raw datasets on a datapoint-by-datapoint basis (Baran et al., 2006) <a href="http://mathdamp.iab.keio.ac.jp/">http://mathdamp.iab.keio.ac.jp/</a>
XXVI.	<b>GASP</b> Free software for GC-MS metabolomics data alignment and visualization <a href="http://www.flintbox.com/public/project/1210">http://www.flintbox.com/public/project/1210</a>
XXVII.	<b>apLCMS</b> Adaptive processing of LC-MS metabolomics data (R package) (Yu et al., 2009) <a href="http://web1.sph.emory.edu/apLCMS/">http://web1.sph.emory.edu/apLCMS/</a>
XXVIII.	<b>Maltcms, ChromA and ChromA4D</b> These are three comprehensive chromatography and mass spectral data e.g. LC-MS, GC-MS; GC x GC-MS processing software from Bielefeld University that performs alignment, generate peak tables, mass spectral search and allow data visualization (Hoffmann et al., 2012; Hoffmann and Stoye, 2009) <a href="http://maltcms.sourceforge.net/">http://maltcms.sourceforge.net/</a>
XXIX.	<b>Further MS based Structure Elucidation Software can be found at</b> <a href="http://fiehnlab.ucdavis.edu/staff/kind/Metabolomics/Structure_Elucidation/">http://fiehnlab.ucdavis.edu/staff/kind/Metabolomics/Structure_Elucidation/</a>

**Table 1.** Commonly used metabolomics databases and data processing software. More information about useful metabolomics databases, software and other useful resources can be accessed via the homepages of Metabolomics Society (<http://www.metabolomicssociety.org/>), Fiehn Laboratory (<http://fiehnlab.ucdavis.edu/>), Biological Magnetic Resonance Data Bank (<http://www.bmrwisc.edu/>), Arita Laboratory, University of Tokyo (<http://metabolomics.jp/>) and Spectroscopy and Chemometrics Research Group, University of Copenhagen (<http://models.life.ku.dk/>).

## 2.2 Plant bioactive substances

Plant bioactive substances are referred to as the chemical compounds that have direct and/or indirect biological effects. In fact, plant primary metabolites, nutrients as well as plant secondary metabolites elicit some biological activities when their concentration is high enough. However, bioactive substances of plants are usually referred to plant secondary metabolites that possess pharmacological or toxicological effects in man and animals. Plant secondary metabolites are not directly involved in the growth and development of plants and their absence does not cause an immediate stress effect in plants, but lack of these metabolites in the longer term might cause serious injuries of plants. Bioactive metabolites are synthesized in plant cells along with the primary metabolites (carbohydrates, amino acids, proteins and lipids) that may involve several different biosynthetic pathways. During evolution, plants have developed such biosynthetic routes to produce secondary metabolites that are necessary to survive in the surrounding environment. Most common plant secondary metabolites can be divided into three main classes: terpenes, phenolics and alkaloids. Plant phenolics, including phenolic acids and flavonoids, possess free radical scavenging activity, while terpenes may attract pollinators or seed dispersers and alkaloids as well as saponins are the main feeding deterrents against herbivores (Crozier et al., 2006). However, it is worth mentioning that different plant species may contain different types and concentrations of the plant secondary metabolites and the bioactive metabolite profile of genetically very close plant species may significantly differ (Faizal and Geelen, 2013).

One of the main classes of bioactive substances is the glycosides. Glycosides cover a broad range of metabolites that possess a mono-, di- or oligosaccharide moieties (glycone part) bound to the other non-sugar part, namely aglycone. These include saponins, glucosinolates, cyanogenic glycosides and glycosides of flavonoids, proanthocyanidins and tannins. Saponins are one of the most common plant glycosides that consist of aglycones, triterpenoids or sterols (modified triterpenoids that have three methyl groups less at position C-4 and C-14) attached to the sugar moieties. Saponins are usually referred as a soap forming compounds that are amphipathic due to their hydrophilic glycone (sugar) moieties and the hydrophobic aglycone. Based on the number of positions of the aglycones where sugar moieties attached, saponins are classified as monodesmosidic, bidesmosidic and etc. Moreover they are further classified based on the structures of the aglycones, e.g., dammaranes, tirucallanes, lupanes, hopanes, oleananes, 23-nor oleananes, taraxasteranes, ursanes, cycloartanes, lanostanes, cucurbitanes, and steroids (Faizal and Geelen, 2013; Vincken et al., 2007). It is worth to mention that the saponin profile of plants are often complex and different species possess different types of saponins. Several studies showed that saponins produced in plants are mainly stored in leaves and roots (Li and Hu, 2009; Tang et al., 2009). Biological activities of saponins are very broad and include insecticidal, fungicidal, molluscicidal, pesticidal activities as well as broad range of pharmacological activities (De Geyter et al., 2012; Diab et al., 2012; Kuzina et al., 2011; Kuzina et al., 2009; Osbourn et al., 2011; Sun et al., 2009; Takahashi et al., 2010). In addition to these, saponins play a key role in plants' defense and development. For example, triterpenoid saponins derived from oleanolic acid are

## 2. PLANT METABOLOMICS

found to be associated with the resistance of the glabrous type winter cress, *B. vulgaris* against insect herbivores such as *Phyllotreta nemorum* and *Plutella xylostella* (Agerbirk et al., 2003b; Kuzina et al., 2009; Nielsen et al., 2010; Shinoda et al., 2002). Whereas, triterpenoid avenacins were found to be associated with the defense mechanism of the oat plant against fungi and bacteria (Mugford et al., 2009; Osbourn, 1996; Osbourn et al., 2003). Other studies demonstrated stimulating effect of saponins e.g. chromo-saponin 1 derived from  $\beta$ -amyrin, in the development of plant roots (Rahman et al., 2001) and germination (Evidente et al., 2011; Zambou et al., 1993). Moreover, overexpression of  $\beta$ -amyrin and lupeol based saponins depicted an indirect effects on the improvement of plant nodulation (Confalonieri et al., 2009).

Another class of plant bioactive substances is the polyphenols that are the major defense metabolites against radicals formed due to the ultraviolet radiations and during photosynthesis (Manach et al., 2004). Moreover, polyphenols constitute to the major plant defense metabolites against other pathogenesis and they are considered as the main bioactive compounds that possibly have a great health promoting properties in human and animals (Kishimoto et al., 2013; Mayer, 2006; Pandey and Rizvi, 2009; Petti and Scully, 2009; Pourcel et al., 2007; Xia et al., 2010). Structures of polyphenols are very diverse and based on the number of phenolic rings and the way they are bound to each other, polyphenols can be divided into phenolic acids, lignans, flavonoids and stilbenes. Among these polyphenols, phenolic acids, flavonoids and their derivatives, e.g. esters and glycosides are the most commonly distributed in the plant kingdom. Phenolic acids can be classified as derivatives of benzoic e.g. gallic acid and cinnamic acids e.g. protocatechuic acid. These phenolic acids are present in high amounts in red fruits and in tea, as well as in vegetables such as onions (Manach et al., 2004; Tomas-Barberan and Clifford, 2000). However, chemical diversity and the concentrations of cinnamic acid based phenolic acids are much higher in fruits, vegetables and in cereal plant than compared to benzoic acid derived phenolics. Among these, phenolics, ferulic, caffeic, p-coumaric, sinapinic and syringic acids are the most common metabolites often detected in various plants. These phenolic acids are mainly present in form of glycosides and/or bounded to organic acids and other cell membrane components. Caffeic acid is found to be the most abundant phenolic acid of most fruits, while ferulic acid dominates in cereals. It is worth to mentioning, that the concentrations of these phenolics vary greatly depending on plant organ and their developmental stage (Clifford and Scalbert, 2000; Hatcher and Kruger, 1997). It is also worth mentioning that phenolics are mainly present in the outer parts of the fruits, cereals and vegetables (skin and leaves) because their biosynthesis is stimulated by light.

Flavonoids including flavones, flavanones, isoflavones and flavonols are also common polyphenols of different plant food products such as fruits and vegetables. These mainly occur in the form of glycosides possessing mainly glucose, rhamnose and less frequently galactose and other sugar moieties. These polyphenols are the major metabolites that constitute the color of fruits and their accumulation is also highly depend on light (Macheix et al., 1990). Studies have shown that some fruits of the same three and even between the different sides of the single piece of fruit might have significantly different flavonoid profile depending on exposure to light (Herrmann, 1976; Price et al.,

1995). Biological activity of these polyphenols might be significantly different in fresh products compared to processed products. Since they easily degrade and, thus lose and/or alter their biological activity. This is also important in polyphenols bioavailability (Manach et al., 2004; Scalbert and Williamson, 2000).

### **2.3 High-throughput metabolomics in development and improvement of crop plant cultivars**

Metabolomics generate vast amount of data that contain valuable information about the physiological status of the biological systems, up or down regulations of genes and influence of biotic and abiotic factors. Recently, metabolomics was highly appreciated as a powerful tool for the improvement and development of crop plants such as cereals (Ferne and Schauer, 2009; Schauer and Ferne, 2006). This is due to the high sensitivity and capacity of metabolomics to screen large number of samples in a high-throughput manner, thus providing a rapid and relatively cheap approach to measuring the responses of cereal plants towards internal and external factors. In addition, metabolomics allow detection of various responses at different levels, including primary and secondary metabolites that are crucial for plants in order to survive and adapt to the surrounding environment. Therefore, by measuring the metabolome of cereals it is possible to evaluate the phenotypes at the molecular level and analyze the quality traits of plants in a broad range (Figure 2 of Paper 3). The high complexity of most metabolic data sets does not allow for the identification of all detected metabolites. Nevertheless, metabolomics is still limited to provide an analysis of a small portion of the actual metabolome as it only allows for detection of a few hundred of the most abundant metabolites. However, metabolomics enables extraction of much more valuable biological information about the systems responses to the environment and the genetic modifications than any other omics platforms. Currently metabolomics is widely utilized for evaluation of cereal plants' resistance against various pathogens (Bollina et al., 2011), salt stress (Widodo et al., 2009), drought (Bowne et al., 2012) and other biotic/abiotic stresses (Andersson et al., 2010; Balmer et al., 2013). Capabilities of the comprehensive metabolomic approaches such as GC-MS based metabolomic profiling allow evaluation of the effects of single gene mutation, pleiotropy and growth temperature (see Paper 4) (Bino et al., 2004).

High-throughput analytical platforms applied in crop plant metabolomics studies differ by their sensitivity, reproducibility and by the range of the detectable metabolites (Allwood et al., 2011; Okazaki and Saito, 2012). The most commonly applied techniques include chromatography or electrophoresis based separation followed by spectroscopic or spectrometry based detection systems. NMR based metabolomics plays a key role in crop development as it allows rapid and quantitative detection of the most abundant metabolomic pool such as carbohydrates, amino acids, and

## 2. PLANT METABOLOMICS

nucleotides (Baker et al., 2006; Barros et al., 2010; Curtis et al., 2009; Gavaghan et al., 2011; Graham et al., 2009; Manetti et al., 2006; Wu et al., 2013). One of the main advantages of NMR based metabolomics is that it provides rapid, highly reproducible and quantitative detection of the broad range of metabolites that possess hydrogen atoms or any other atom with NMR active nucleus such as phosphorous, nitrogen and carbon. However, most high-throughput NMR metabolomic studies are performed by measuring the 1D  $^1\text{H}$  NMR spectra. Sample preparation and pretreatment of the NMR based metabolomics is relatively simpler than in GC-MS or LC-MS based metabolomics and metabolites are directly detected from the mixture samples without prior separation steps. Therefore, NMR spectra of complex biological samples can be regarded as an unique fingerprint of the whole system. These fingerprints possess both qualitative and quantitative metabolomic information that easily can be turned into biological information with the help of multivariate data analysis (Engelsen et al., 2013). The main drawback of NMR is its sensitivity and in many cases it only allows for the detection of the most abundant metabolites (e.g. first 50 metabolites) while, the low concentration metabolites remain undetected and/or hidden behind the peaks of more abundant metabolites (Paper 3).

Other high-throughput detection techniques commonly used in the development of the crop plants are based on vibrational and electronic spectroscopy methods that are capable of measuring various physico-chemical properties of the samples as one whole system, and often less specific to the individual metabolites. These techniques include, NIR, IR, FT-IR, UV-VIS, Raman and Fluorescence spectroscopy. However, these spectroscopic methods allow rapid and quantitative analysis of the bulk primary metabolites such as total starch content, fat content, protein content, dietary fibers and sugars, and facilitate rapid non-destructive evaluation of the desired quality traits of the cultivars. In this term NIR spectroscopy in combination with multivariate methods of analysis, serve as a powerful tool for the rapid proxy evaluation of cereal cultivars grown under different environmental conditions and/or effects of genetic modifications (Jacobsen et al., 2005; Jensen et al., 1982; Munck Lars, 1992; Munck et al., 2001; Munck et al., 2010; Nørgaard et al., 2000). As an example, FT-Raman and NIR has been demonstrated to be an efficient method for rapid screening of rice seeds for protein and amylose content (Sohn et al., 2004). Raman, FT-IR, UV-VIS and Fluorescence methods showed a capability for high-throughput analysis of cereal plants for their chemical composition in various different studies focused on crop development, assessment and rapid screening studies (Barron and Rouau, 2008; Greene and Bain, 2005; Manolache et al., 2013; Mikkelsen et al., 2013; Siuda et al., 2006; Zekovic et al., 2012).

Separation followed by detection techniques such as GC-MS, LC-MS, CE-MS and LC-DAD facilitate more informative metabolomics data that allow quantification and identification of the individual metabolites.

These analytical platforms allow detection of low concentration metabolites that are difficult or impossible to observe by NMR and other spectroscopic methods, thus more metabolites can be detected in a semi-quantitative manner that allow metabolomic profile comparison of the samples.

## 2. PLANT METABOLOMICS

However, absolute quantification of all these metabolites is laborious or impossible when several hundred metabolites are detected, and the identities of most metabolites are unknown. Only few comprehensive metabolomic studies have performed absolute quantification, since the biological questions behind the majority metabolomics studies can be answered by relative quantification. However, it is worth mentioning that both metabolomic analyses, the semi-quantification (or relative quantification) and absolute quantification, require an optimized and robust metabolomics protocol, since slight alterations of the analysis may have significant impact on the quantitative nature of the data and may cause over or under estimation of the observed results.

Metabolomic data obtained from GC-MS and LC-MS techniques are usually very complex and not all detected metabolites can be identified. For example, paper 4 demonstrates the detection of 247 metabolites from barley flour in a single GC-MS metabolomic profiling experiment, but only 89 metabolites could be identified based on their EI-MS and RIs. However, this is the current status of the GC-MS metabolomics based on the quadrupole MS and EI-MS library search. Although, the method has been used for a more than two decades and has been applied to various plant and animal tissues, the databases are still not rich enough to allow efficient identification. In the case of LC-MS or GC-MS that are based of accurate mass detection, metabolites are identified by their accurate mass rather than fragmentation pattern and these approaches also allow identification of less than half of the metabolites that could be detected from the complex plant samples. Tandem MS is one of the solutions for enhancing metabolite identification, however most of the time it allows only tentative identification. Table 1 lists the up-to-date metabolomic data bases originated from EI-MS, accurate mass and tandem MS libraries.

The separation followed by detection techniques are less high-throughput than the direct detection techniques like, NMR or NIR. This is due to the complexity of the sample preparation steps in chromatographic methods such as GC-MS analysis which require derivatization. In GC-MS and LC-MS analysis, the metabolites of the mixture samples are separated prior to detection. Separation of complex samples occur in a reproducible manner to a certain extent (e.g. retention time shifts, sensitivity loss, baseline drifts are commonly faced problems) and for the limited number of samples, since the instruments have limitations with the number of samples that can be analyzed in a single sequence without further maintenance (e.g. cleaning and replacement of the parts). Moreover, in LC-MS systems the complexity of the metabolites may cause ion suppression for the metabolites eluting closely, which may result in uncertain quantification. Separation of metabolites depends on many instrumental parameters as well as the nature of the mobile phase-stationary phase interactions. Slight alterations and/or inconsistencies of these parameters during the analysis may cause a high level of error in the metabolomics data and mask the original variation. Usually, the level of error is relatively higher in separation followed by detection techniques than in direct detection methods such as NMR, and NIR. Appropriately, conducted metabolomics studies involve replicates as well as internal standards in which to some extent help to minimize the experimental variations. Thus, GC-MS, LC-MS and CE-MS remain one of the most powerful analytical platforms of metabolomics and provide the

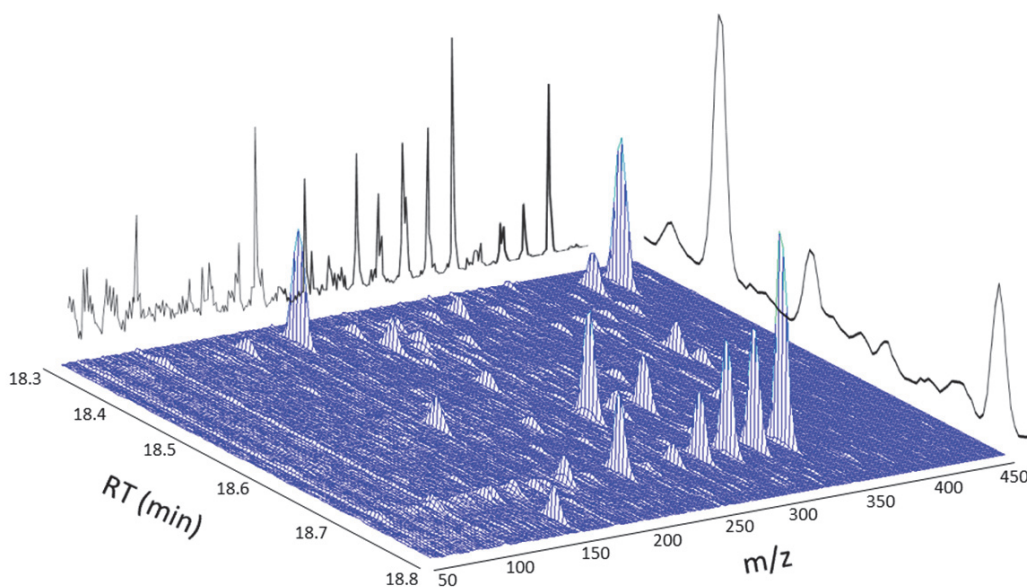
## 2. *PLANT METABOLOMICS*

most comprehensive metabolic analysis of the systems. In crop plant development and improvement, these techniques play a key role and allow simultaneous identification of alterations of the several biosynthetic pathways. Up to date, more than hundred original researches papers have been published in metabolomics in crop plant cultivars applications (Table 1 and Figure 6 in Paper 3).

## 3 ADVANCED GC-MS AND CHEMOMETRICS FOR METABOLOME ANALYSIS

### 3.1 GC-MS

Gas chromatography coupled to mass spectrometry (GC-MS) is based on separation of metabolites by applying heat until they reach their boiling point, fly through the GC column, and reach the MS. Metabolites are separated in a GC column based on their boiling point and partitioning coefficients between mobile phase (GC carrier gas) and stationary phase (GC column). The outlet of the GC column opens in an ionization chamber of a mass spectrometer where eluted metabolite will be ionized by a beam of the highly charged electrons (in the case of electron ionization (EI)). Resulted ions further travel through the mass analyzer where ions are separated based on their mass to charge ratio ( $m/z$ ) and finally reaches the detector. This process is repeated several times at the fixed scan speed e.g.  $20 \text{ s sec}^{-1}$  throughout the analysis. Thus, resolution of the metabolites depends on both, GC separation and scan speed of the mass spectrometer. At each scan point, full e.g. 50-500  $m/z$  or selected ion monitoring (SIM) e.g. one specific  $m/z$  ion, mass spectra can be recorded and obtained chromatographic data for many scans over the investigated range of the elution time will have three or two dimensional structure. Figure 3 demonstrates an interval of the raw GC-MS data recorded at  $m/z$  range of 50-450.



**Figure 3.** Interval of raw GC-MS data.



### 3. ADVANCED GC-MS AND CHEMOMETRICS FOR METABOLOME ANALYSIS

High chromatographic resolution power is the main advantage of GC-MS over other hyphenated techniques. It is the most established platform within the hyphenated techniques mainly due to its relatively simpler operation, lower cost and higher reproducibility. However, the range of the detectable metabolites is limited to those that are thermally stable and volatile under the GC-MS conditions. To date, this covers metabolites with molecular weight of up to 1400 Da and covers various classes such as amino acids, carbohydrates, fatty acids, small molecular organic acids, phenolics, terpenes and sterols. Most of the comprehensive GC-MS metabolomic analysis of complex biological samples requires sample derivatization where non-volatile and/or thermally unstable metabolites are chemically altered for increasing their detection. Depending on the type of the mass spectrometers, GC-MS allows identification of metabolites based on their EI-MS fragmentation patterns (quadrupole based mass analyzers) and accurate mass measurements (TOF based mass analyzers). Due to the long traditional use of electron ionization (EI), EI-MS based GC-MS metabolomic data bases are the richest pools of the metabolites available to date e.g. NIST and Wiley libraries. Another type of ionization method, which is less frequently used in GC-MS, is chemical ionization (CI). CI is considered to be a soft form of ionization and may provide information about the mass of the molecular ion but less fragmentation ions.

GC-MS has become the mostly utilized hyphenated analytical platform in many different metabolomic analyses, such as targeted analysis of pesticides, adulterants, pharmaceuticals and their byproducts from urine and plasma samples. Moreover, it found a wide use in identification of pollutant levels in water and air, aroma compounds of food and food raw materials, and finally in comprehensive metabolomic fingerprinting and profiling of various complex samples derived from microorganisms, plants and mammalian systems. In plant metabolomics, GC-MS plays a key role and current literature contains substantial amounts of GC-MS based plant metabolomic studies aimed at uncovering effects of genetic modifications, biotic/abiotic stresses, elucidation of plants' natural defense mechanisms, and identification of health promoting substances of medicinal plants (see Paper 3). To date, the majority of the plant derived primary and secondary metabolites can be quantitatively detected by GC-MS, in a relatively high-throughput manner. Optimized metabolomics protocols employing high-throughput metabolite extraction and use of autosamplers for sample derivatization and injection, allow quantitative GC-MS analysis of up to several hundred samples. However, high quality quantitative GC-MS data of complex plant samples require a lot of effort in all the steps of the analysis, starting from the sample harvesting to the data acquisition. This will be further discussed in section 3.4.

GC-MS based metabolomics can be divided in three distinctive steps: sample preparation, data acquisition, and data analysis. The first two steps play an important role in quality of the obtained data, while the latter facilitates extraction of biological knowledge from the data. Several studies has been published on optimization of GC-MS protocols for large scale metabolomic profiling of complex

### 3. ADVANCED GC-MS AND CHEMOMETRICS FOR METABOLOME ANALYSIS

samples (Danielsson et al., 2012; Fiehn et al., 2000; Gullberg et al., 2004; Jiye et al., 2005; Lisec et al., 2006). Main sources of experimental errors introduced during the sample preparation, derivatization and analysis are highlighted in (Kanani et al., 2008; Kanani and Klapa, 2007; Khakimov et al., 2013; Little, 1999). Additional precautions in acquisition of quantitative GC-MS data from large sample sets and development of the novel derivatization method are described in the following section and in Paper 2.

One of the main challenges of GC-MS plant metabolomics is associated with identification of unknown metabolites. In GC-MS, metabolites can be identified by comparing their retention indices (RIs) and characteristic fragmentation pattern of their mass spectra (EI-MS) against databases. However, not all the metabolites detected from the complex plant extract are present in databases. Identification of some of these unknown metabolites can be confirmed by using authentic standards if one has prior knowledge what these metabolites might be. The systems allowing hyphenation of GC with more informative detection systems such as NMR and preparative GC are not mature enough to analyze large sample sets and/or identify several hundred metabolites within complex samples. However, GC-MS databases are becoming richer e.g. the 10<sup>th</sup> edition of the Wiley GC-MS library contain EI-MS mass spectra of 638 thousand distinct compounds. To date several commercial and free GC-MS databases comprising metabolites identified from various species are available (Table 1). Moreover, identification of unknowns based on accurate mass measurements (1-3 ppm) have been improved by using general chemistry rules, elemental composition and isotopic patterns of the metabolites (Kind and Fiehn, 2007). This is one of the promising trends in identification of new metabolites in a high-throughput manner and requires further research in the field to develop a validated method that would allow metabolomics labs to apply the method in a daily routine identification.

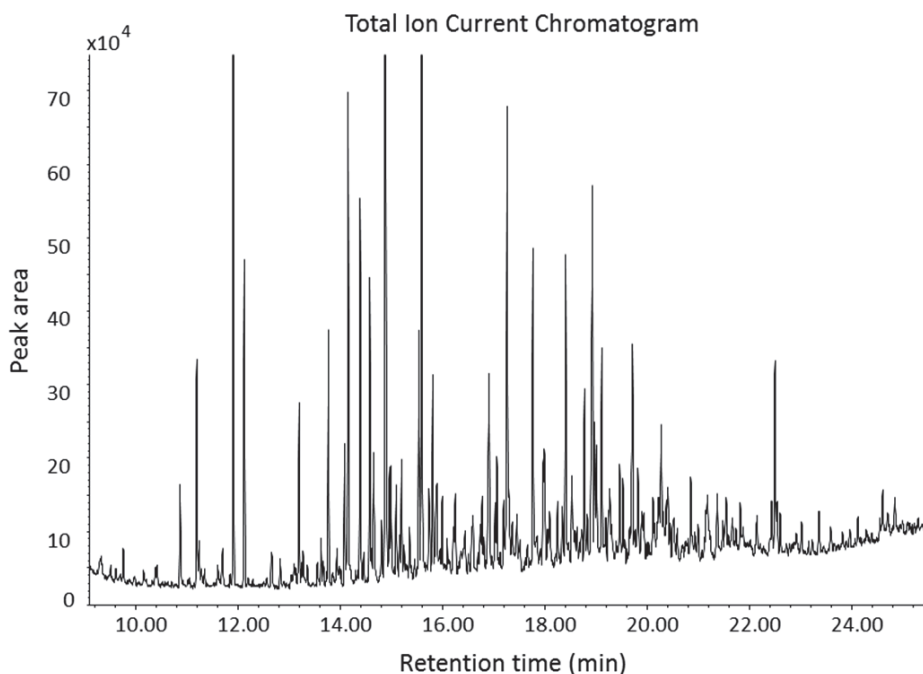
The next challenging issue of GC-MS metabolomics, which is partly solved, is processing of the complex data and extraction of relevant information. Non-specialist analysts need user-friendly methods for rapid quantification of resolved or partially resolved peaks from the complex GC-MS chromatograms and annotate detected peaks. This is partially possible by using commercial chromatographic data processing software such as ChemStation (Agilent), DataAnalysis (Bruker), Chromeleon (Dionex), ChromaTOF (Leco) and Empower (Waters). However, is it not straightforward when dealing with complex profiles where several peaks are overlapped and/or closely eluted, thus hampering their quantification and identification. These software may allow deconvolution of such peaks by using their mass spectra. However, it requires manual work for each peak and for each sample, which is not an optimal solution for two reasons, firstly, it is laborious and secondly, such a quantification will be biased. It is also worth mentioning that automatic quantification of resolved and overlapped peaks by using this software may not always be reliable due to the inconsistency of the retention times of the peaks over the samples, peak shape alterations and finally, peaks having a low s/n ratio might be completely ignored. However, this approach remains common among many labs due to easier

### 3. ADVANCED GC-MS AND CHEMOMETRICS FOR METABOLOME ANALYSIS

operation of commercial software and in fact, they are well suitable when the chromatographic system is well optimized and detection primarily focuses on few targeted metabolites.

In comprehensive plant metabolomics that apply GC-MS for profiling and/or untargeted analysis, conventional chromatographic data analysis software can assist in a limited extend e.g. initial data visualization, library search. Up to date, several GC-MS data processing software have been developed for manual, automated and/or semi-automated deconvolution of mass spectra, baseline correction, metabolite quantification and identification (see Paper 3). Unfortunately, all of these processing approaches have some disadvantages related to their capabilities and use. One of the mostly utilized stand-alone GC-MS data processing software, Automated Mass Spectral Deconvolution and Identification System (AMDIS) was developed in late 1990s and it is commonly used in GC-MS metabolomics (Stein, 1999). However, the technique requires manual processing and validation. While AMDIS (that is mostly used up to date) can handle only one sample at a time, the cutting edge technology based on multi-way decomposition modeling, PARAllel FACtor Analysis 2 (PARAFAC2) performs the same task in a more efficient manner and facilitates extraction of more information. However, this approach also has some disadvantages related to its use by non-scientists. PARAFAC2 based chromatographic data processing requires division of the data into smaller (less complex) intervals in elution time dimension and requires validation of the models (Amigo et al., 2010a; Amigo et al., 2010b; Bro et al., 1999; Khakimov et al., 2012). Detailed features of PARAFAC2 and current research on its development will be discussed in section 3.5.

GC-MS based plant metabolomics can be considered as a mature field that possess well established protocols related to the data acquisition and analysis (Fiehn, 2008; Liseč et al., 2006; Lytovchenko et al., 2009; Shuman et al., 2011; t'Kindt et al., 2009). Modern GC-MS labs are able to analyze several hundreds of complex plant samples within one sequence and detect up to 500 distinct metabolites from each sample. Figure 4 shows complex TIC chromatogram of the raw GC-MS data obtained from the trimethylsilylated barley seed extract (data from Paper 4). As mentioned earlier, the handling of such complex data is not easier than its acquisition. Indeed, today, not all the metabolomic studies perform comprehensive analysis of the raw data, but rather limited deductive analysis to the target of the study. Typically, hundreds of unassigned peaks remain unpublished and no further efforts are made to identify these metabolites, elucidate their biological origin and evaluate their relationships with assigned metabolites. In fact, publication of these unassigned metabolites by providing as much information as possible about their identities e.g. RI, EI-MS, accurate mass, origin, along with the data may significantly benefit future metabolomics studies.



**Figure 4.** TIC chromatogram of GC-MS data from a methanol extract of whole-grain flour of barley seeds.

### 3.2 Trimethylsilyl cyanide (TMSCN) based derivatization

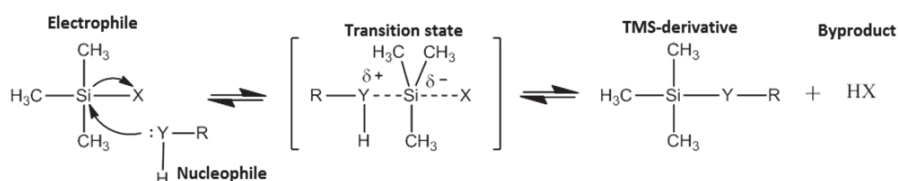
In order to detect metabolites in GC-MS, they must be volatile and thermally stable under the given instrumental conditions. It is important to meet both requirements, to be volatile and thermally stable, for the reliable detection, since volatile, but thermally unstable metabolite, can only be analyzed for qualitative purposes, and not quantitative because they may be partially degraded due to high temperature inside the GC column or mass spectrometer. Moreover, thermally stable, but nonvolatile metabolites cannot pass through the GC column and remain in the injection port. Since the development of gas chromatography, one of the compromises for enhancing metabolite detection involved chemical derivatization of metabolites towards increasing their thermal stability and lowering their boiling points. These metabolites mainly include those that possess polar and reactive functional groups such as carboxylic acid (RCOOH), alcohol (ROH), aldehyde (RCHO), amines (RNH<sub>2</sub>, R<sub>2</sub>NH), amides (RCONH<sub>2</sub>) and thiols (RSH). These functional groups result in significant intermolecular interactions such as hydrogen bonding, coulombic forces and Van der Waals forces that increase their boiling points. In addition, they increase the reactivity of metabolites at elevated temperature. Chemical derivatization is based on replacing the active hydrogen atoms of these polar and reactive functional groups of the metabolites with more inert and non-polar groups, which eliminate strong

intermolecular interactions. Common chemical derivatization methods applied in GC-MS include silylation, alkylation and acylation. This section focus on silylation techniques only and more specifically on the novel applications of the trimethylsilylation reagent, trimethylsilyl cyanide (TMSCN) for derivatization of complex biological samples.

Silylation is the most commonly used derivatization method in GC-MS analysis and it utilizes trimethylsilylation, where active hydrogen atom is replaced with a trimethylsilyl (TMS:  $[-\text{Si}(\text{CH}_3)_3]$ ) group, while tert-butyldimethylsilylation (MTBS:  $[-\text{Si}(\text{CH}_3)_2(\text{C}(\text{CH}_3)_3)]$ ) and chloromethyl dimethylsilylation ( $[-\text{SiCH}_2\text{Cl}(\text{CH}_3)_2]$ ) are less often employed. Among these methods, tert-butyldimethylsilylation exhibit significantly higher stability of the derivatized products against hydrolysis and provide more stable derivatization (Blau and Halket, 1994; Poole and Zlatkis, 1979). Moreover, the MTBS silylation reagent, MTBSTFA, provides characteristic fragmentation patterns and fragment ions representing mainly  $[\text{M}-57]^+$  and  $[\text{M}-131]^+$ , though it is less sensitive towards sterically hindered functional groups due to its relatively larger molecular size (Schummer et al., 2009). However, due to the higher number of available reagents and higher silylation reactivity, historically, trimethylsilylation is the most commonly applied method. Trimethylsilylation of metabolites depend on the following factors: **1.** Reaction temperature, **2.** Reaction time, **3.** Applied solvent (if any) and its ratio to the reagent, **4.** The nature of the leaving group of the reagent (its size and basicity, the reaction will proceed most efficiently if the leaving group is a weaker base than the silylating substrate), **5.** Chemistry of the substrate (its basicity and steric hindrance), **6.** Byproduct (its volatility and interference in the reaction equilibrium, as well as ability to increase the electrophilicity of the substrate will further catalyze the reaction towards formation of TMS-derivatives), **7.** Dryness of the reaction mixture (presence of minute amount of moisture may significantly destruct an equilibrium towards hydrolysis of derivatives). However, even the general laboratory practices suggest that a higher silylation temperature favor the formations of the TMS-derivatives, the derivatization protocols must consider a compromise to avoid degradation of metabolites and/or their derivatives due to the elevated temperature. The other important compromise is the silylation reaction time, which must be optimized prior to achieving an acceptable s/n ratio and to avoid degradation of metabolites and/or their derivatives due to a long contact with the aggressive silylation reagents. Derivatization of complex biological mixtures involves several competing silylation reactions that occur simultaneously, where each reaction requires silylation reagent. Therefore, excess amount of the reagent will provide best silylation yield. Depending on the physico-chemical properties of the silylating substrates of the complex mixture and their concentrations, each silylation reaction may be completed at different times for different metabolites. Moreover, the complete silylation time of a metabolite depend on the complexity of the sample matrix (number of competing reactions). Thus, it is not straightforward to find an optimal silylation time when all metabolites are fully converted to their TMS-derivatives. Alternatively, an optimal silylation time of the complex mixture can be defined as the time when the maximum number of metabolites reaches their highest abundance. Most importantly, the silylation

time of the samples that are subjected to the quantitative comparison must be consistent, although it may not be an optimal time for some of the metabolites of the mixtures.

In addition, the solvent used in silylation reactions significantly influences the reaction yield and the stability of the produced TMS-derivatives. Silylation reactions performed in the solvents with a moderate polarity follow  $S_N2$ -Si mechanism (Figure 5) (Blau and Halket, 1994; Pierce, 1968; Schummer et al., 2009; van Look et al., 1995). The main requirement for the solvent is that it should be inert to the applied reagent and derivatives. A weakly basic solvent such as pyridine suits the silylation reactions and increase the solubility of the metabolites as well as increasing reaction speed by bonding the acidic protons present in the mixture. However, it is challenging to bring all the metabolites into a pyridine/reagent solvent system, when dealing with complex mixtures with a wide range of polarity. In fact, some protocols use solvent free derivatization where the reagent itself plays the role of solvent. This is a good approach, since it eliminates artifacts related to the solvent and increases the s/n ratio.



**Figure 5.** General scheme of the trimethylsilylation reaction, which occur via a  $S_N2$  mechanism.

Today, more than ten different silylation reagents are available and they differ by their reactivity towards different functional groups, byproducts, stability and prices (Little, 1999; Poole, 2013). Among them, N-Methyl-N-(trimethylsilyl) trifluoroacetamide (MSTFA) is the most commonly used reagent for derivatization of various functional groups in GC-MS metabolomic profiling and fingerprinting (Fiehn et al., 2000; Lisec et al., 2006; Roessner et al., 2000). Today, the majority of the complex mixture derivatization protocols use 20-40  $\text{mg ml}^{-1}$  solution of the methoxiamine hydrochloride ( $\text{CH}_3\text{ONH}_2 \cdot \text{HCl}$ ) in pyridine for methoxiamination of aldehyde and ketone functional groups by incubating at 30-40°C for 60-1020 min, followed by MSTFA based silylation by addition of 40-100  $\mu\text{l}$  reagent and incubation at 30°C for 20-90 min. These protocols demonstrated powerful derivatization, covering a wide range of primary and secondary metabolites. However, experimental variation related to the derivatization still constitute a substantial amount of the total variation that may hide the biological information present in the data (Kanani et al., 2008; Kanani and Klapa, 2007). These variations are mainly related to the non-reproducible profiles e.g. significant fluctuations of the metabolite peaks between replicates, interference of artifact and byproduct peaks.

In the course of the PhD study, a new trimethylsilylation reagent, trimethylsilyl cyanide (TMSCN) was evaluated for its silylation capabilities towards various molecular families in GC-MS metabolomics. Today, the use of TMSCN for silylation purposes is limited to the targeted silylation of few compounds (Riggio et al., 1992), though some studies have showed its clear advantages for silylation of alcohols, phenols, organic acids, amino acids and carbohydrates under mild and solvent free conditions (Mai and Patil, 1986). The silylation capability of TMSCN has been investigated towards metabolite standard mixture containing amino acids, carbohydrates, phenolic acids, flavonoids, organic acids, and sterols as well as for phenolic extracts of the blueberry fruits. The trend of the derivatization reactions was monitored by GC-MS over the different silylation times varying from 5 min to 60 hours. In parallel with TMSCN, MSTFA reagent was used for silylation of the same mixture samples. The results of this method comparison study are comprised in Paper 2. Moreover, the use of TMSCN in GC-MS detection of plant triterpenes are described in sections 4.1 and 4.2.

For many metabolites, TMSCN based silylation outperformed MSTFA in terms of reaction speed, sensitivity, and repeatability. The main advantages of using TMSCN as an alternative trimethylsilylation reagent are comprised below:

**1. Small molecular size.** TMSCN is probably the smallest silylation reagent that has high reactivity both under neutral and basic conditions. This has been demonstrated by derivatization comparison of TMSCN and MSTFA towards sterically hindered labile protons of 2,6-diphenylphenol (Paper 2, Supp. Fig.1 ). Trimethylsilyl chloride (TMCS) also has small molecular size, but it is a very weak silyl donor in the absence of base.

**2. Inert byproduct.** The only byproduct formed during the TMSCN based silylation is hydrogen cyanide (HCN) that has boiling point of 26°C. It will not interfere with peaks of volatile metabolites and will be eluted even before the solvent. The other feature of the byproduct is that it can further enhance the reaction speed by protonating TMSCN that will lead to increase its electrophilicity and catalyze the reaction towards production of TMS-derivatives. In contrast to byproducts of some reagents such as TMCS that can hydrolyze the TMS-derivatives, HCN is too weak an acid to do this, which makes the produced TMS-derivatives more stable.

**3. As a solvent.** As mentioned earlier TMSCN has a smaller molecular size than most of the silylation reagents and the electron density of the molecular is shifted towards cyanide group, which makes it a polar, molecular and structurally it resembles acetonitrile, one of the most common organic solvents. In the course of the present study, we have observed greater solubility of carbohydrates, phenolic acids, triterpenes, and amino acids in TMSCN than in MSTFA (although exact solubility were not measured). This makes TMSCN even more suitable for silylation of complex biological mixtures by avoiding the use of an additional solvent. In paper 2, it has been shown that the application of pure

TMSCN result in a much greater *s/n* ratio of TMS-derivative peaks than when it is used in conjunction with methoximation step, which is performed in the solvent pyridine.

**4. Less artifacts.** As mentioned above, the TMSCN byproduct does not result in any artifacts. Compared to MSTFA, TMSCN based silylation resulted in more repeatable GC-MS profiles of both, artificial metabolite mixture and of blueberry fruit extracts.

**5. Mild conditions.** In paper 2, TMSCN and MSTFA were compared using the same temperature (37°C) that is most commonly applied in the literature, but TMSCN is able to provide greater silylation yield at room temperature. Moreover, Mai and Patil also showed that up to 80-97% reaction yield toward various metabolites is reached within 5 min at 25°C (Mai and Patil, 1986).

**6. Price.** TMSCN is cheaper than most of its alternative derivatization reagents. Based on the pricelist of the Sigma-Aldrich (on 4 Aug 2013), the price for 25 ml of highest purity reagents are as follows: TMSCN (950 dkk), MSTFA (3207 dkk), BSTFA (1403 dkk) and MTBSTFA (3380 dkk). However, as in the case of the all derivatization reagents, TMSCN also have some drawbacks. One its disadvantage is related to the formation of hydrogen cyanide (HCN) as a byproduct. However, the hazard caused by HCN can easily be minimized to a level that is safe for utilizing the reagent in routine GC-MS analysis. All the safety considerations and practical aspects of using TMSCN are provided in details in the Supporting Material of the paper 2. The other disadvantage of TMSCN is the stability of the TMS-derivatives when they are in contact with the reagent for more than a few hours. In this term, MSTFA is much more aggressive and product degradation occurs more rapidly. However, in high-throughput GC-MS analysis of several samples, derivatization time must be strictly controlled and the samples must be injected into GC at the optimal silylation time. This will minimize the experimental variations related to the product degradation.

### 3.3 Design of experiment (DoE)

#### *Objectives of DoE*

DoE is frequently used to optimize factors of metabolomics protocol, that lead to detection of a broader range of metabolites, increase *s/n* ratio, reproducibility, enable high-throughput analysis, reduce experimental cost, level of biased and artifact effects. DoE is about setting up a series of experiments representative to the investigated question. Usually, in plant metabolomics, DoE starts by identifying and specifying the number of experimental parameters e.g. factors of sample preparation, metabolite extraction steps and analytical instrument parameters, followed by identification of their ranges that must be investigated. The next step is to define the number of responses that will be measured in each experiment of DoE e.g. *s/n* ratio of one specific or several metabolites, the number of detectable metabolites and reproducibility of the profiles. Then DoE can be created depending on



### 3. ADVANCED GC-MS AND CHEMOMETRICS FOR METABOLOME ANALYSIS

the number of parameters, and their investigated ranges and experiments will be performed in a randomized order. Finally, the obtained data are analyzed by regression analysis, which estimates the relationships between changes in factors to the changes in response(s). The regression models provide knowledge on the importance of parameters to the response(s) and assist in identification of the parameter values that will lead to maximizing the desired response. However, when DoE has an aim to optimize several responses, identification of the best experimental condition, that would fulfill the goal of all the responses, will become difficult. Such situations require that the analysts make a decision on choosing an experiment that will reflect a compromise between the response variables.

DoE allows estimation of the factors and their interaction effects to the response(s), evaluation of the systematic (effect) and unsystematic (noise) variations of the experiments, and provide a reliable prediction of the optimal conditions. Mostly, DoE assists in reaching three objectives of experiments that will lead to the efficient, robust and high-throughput metabolomic protocols. These objectives are screening, optimization and robustness testing that are performed in the given order by using the DoE approach. Screening is the first step of the DoE and it attempts to answer the questions such as, 1) What are the factors mostly influencing the fluctuations of the measured responses? and 2) In which ranges these factors must be investigated for identification of the most efficient combination of the factors? Screening requires relatively fewer experiments in relation to the investigated factors than in optimization. In plant metabolomics, DoE based screening of factors are mainly focused on estimation of the range of the investigated factors, e.g. sample pretreatment, metabolite extraction, derivatization parameters and instrumental settings. However, choosing globally optimal factors is more challenging, due to the diversity of the metabolome and significantly different concentrations of the metabolites. Optimization is the second step of the DoE, which assists in finding the best combination of the investigated factors. Optimization involves series of experiments designed by varying the levels of factors and use the obtained data for predicting the desired response variable e.g. s/n ratio or limit of detection of metabolites etc. Robustness testing is the last step of the DoE that evaluates the sensitivity of the response to the small fluctuations of the factors. An optimal experimental condition is not necessarily the most robust. Therefore, it is crucial to evaluate the levels of factors that may cause non-stable experiment and adjust them in such a way that they will provide a reproducible protocol, even if it is slightly shifted from the optimal conditions. This is probably the most important requirement in quantitative metabolomics, since only the reproducible metabolomic protocols may provide reliable data, from which true biological variations can be evaluated.

#### *Experimental noise*

DoE mainly demonstrates two kinds of experimental variations. The represents variations related to the effects of the factors and the second represents variations related to the noise. Noise is the unsystematic part of the variation that is present due to inaccurate measurement, instrumental fluctuations and/or other types of bias. However, in order to estimate the effect of factors, it is crucial to know the noise level boundaries. For instance, if one attempts to optimize the extraction protocol of some low concentration metabolites by varying temperature, it is necessary to perform several (5-7)

replicate analyses in order to find the noise level, at the initial experimental conditions. The noise level can be calculated as the standard deviation of the concentrations measured for replicate samples. This will indicate the error level, which might be present in each measurement performed by using the same protocol. Then, this error level must be taken into account when evaluating the effect of the extraction temperature upon metabolite concentrations.

#### *DoE models*

After selection of the investigated experimental factors, determination of the experimental domain and the response(s), it is time to design the experiment. DoE is based on developing a model,  $y=f(x)$ , that would allow prediction of the  $y$  response variable from the polynomial function  $f(x)$ , which represents the relationship between the experimental factors and the response(s). Depending on the objectives of the DoE (screening, optimization and robustness test) different modeling strategies are applied. For example, screening and robustness testing are performed by linear ( $y=b_0 + b_1x_1 + b_2x_2$ ) and second order interaction ( $y=b_0 + b_1x_1 + b_2x_2 + b_{12}x_1x_2 + E$ ) models, while optimization requires the quadratic models ( $y=b_0 + b_1x_1 + b_2x_2 + b_{11}x_1^2 + b_{22}x_2^2 + b_{12}x_1x_2 + E$ ) that allow estimation of linear and non-linear relationships between the factors and response(s).

#### *Factorial design*

Factorial design employs models based on linear or interaction effects of the experiment factors and their influence on the response(s). Table 2 demonstrates a  $2^3$  full factorial design of the metabolomic experiment where three factors e.g. ( $x_1$ ) solvent concentration, ( $x_2$ ) extraction time and ( $x_3$ ) extraction temperature are varied in two different levels that result in total of 8 experiments. This model can be written as a third order interaction model ( $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 + b_{123}x_1x_2x_3$ ) and each effect, individual factor effects ( $b_1, b_2, b_3$ ) and their interaction effects ( $b_{12}, b_{13}, b_{23}$  and  $b_{123}$ ) can be calculated by using the signs of the corresponding factor columns. For example,  $b_1$  that represents the main effects of the solvent concentration (when it is changed from low level to the high level) can be determined as  $b_1 = 1/8 (-1 + 1.2 + 0.9 + 1.1 - 0.8 + 1 - 1.3 + 0.9) = 0.25$ . This indicates that the high level of  $x_1$  factor will increase the level of  $y$  response by 0.25 (in original unit of the  $y$ ). More detailed description of the model calculation is provided in (Lundstedt et al., 1998). Graphically this design can be illustrated as a cube where three dimensions are defined by  $x_1, x_2$  and  $x_3$  (Fig. 6). A middle point of the cube (blank circle), which is experiment number 9 (Table 2) represent a central point of the experiment. Usually, this is the starting point of the experiment and it will be performed in several replicates to estimate the level of noise. Moreover, these center point experiments will answer the question, whether or not the experimental factors and responses have a non-linear relationship (if measured response,  $y$ , of the experiment number 9 greatly differs to the  $b_0$ ), if so then quadratic models must be used.

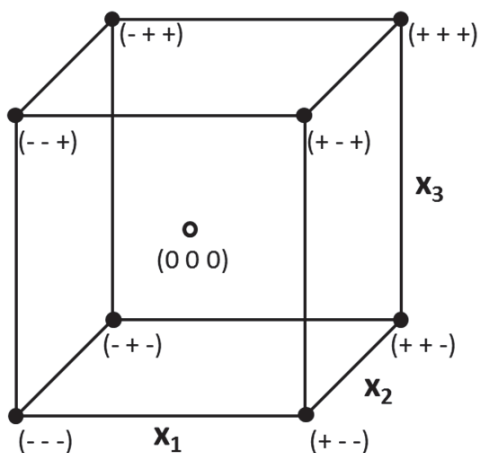


Figure 6.  $2^3$  full factorial design.

#### Fractional factorial design

In the case of the full factorial design, screening and robustness testing of the metabolomic protocol is defined by  $\mathbf{N} = 2^k$ , where  $k$  is the number of investigated factors. For example, 8 or 12 factors will require 256 and 4096 experiments to be performed, respectively. However, in most cases, this exceeds the limit of performable experiments and not all the interactions between 12 factors are significant. Therefore, a good screening and robustness testing start by evaluation of the main effects (Eq.1.), followed by addition of the factors' interaction experiments that will reveal the effects of two or more factors' alterations at a time (Eq.2.). Usually, additions of the interaction effect experiments are based on results of the main effects (interactions between the factors with high influence in response are the most interesting).

$$y = \beta_0 + \sum \beta_i X_i + \varepsilon \quad (\text{Eq. 1.})$$

$$y = \beta_0 + \sum \sum \beta_{ij} X_{ij} + \varepsilon \quad (\text{Eq. 2.})$$

For example, evaluation of the main effects of the protocol that is screened for 8 factors can be performed in 9 experiments, one for each factor and one center point where factors are set to their center. Normally, for estimation of the noise level, this center point is performed several times. In addition to the main effects, if it is necessary to evaluate second order interactions (interactions of all 8 variables by varying two factors at a time) the screening test will require 37 experiments. Full experimental domain of this protocol is spanned by 8 factors that can be exemplified as corners of the 8 dimensional hyper cube. Screening of this protocol that cover a complete experimental domain requires many experiments, which are expensive. This can be compensated by applying the fractional factorial design, which allows selection of only those "informative" experiments that cover a maximal volume of the experimental domain in a limited number of experiments.

### 3. ADVANCED GC-MS AND CHEMOMETRICS FOR METABOLOME ANALYSIS

Exp.No	x <sub>1</sub>	x <sub>2</sub>	x <sub>3</sub>	x <sub>1</sub> x <sub>2</sub>	x <sub>1</sub> x <sub>3</sub>	x <sub>2</sub> x <sub>3</sub>	x <sub>1</sub> x <sub>2</sub> x <sub>3</sub>	y
1	-	-	-	+	+	+	-	1
8	+	-	-	-	-	+	+	1.2
3	-	+	-	-	+	-	+	0.9
4	+	+	-	+	-	-	-	1.1
5	-	-	+	+	-	-	+	0.8
6	+	-	+	-	+	-	-	1
7	-	+	+	-	-	+	-	1.3
8	+	+	+	+	+	+	+	0.9
9	0	0	0	0	0	0	0	1

**Table 2.** 2<sup>3</sup> full factorial design. (-) x<sub>1</sub> is solvent concentration 50%, (+) x<sub>1</sub> is solvent concentration 90%, (-) x<sub>2</sub> extraction time 2 min, (+) x<sub>2</sub> extraction time 5 min, (-) x<sub>3</sub> extraction temperature 65°C and (+) x<sub>3</sub> extraction temperature 95°C. Experiment 9 is the center point where factors are set to 70%, 3.5 min and 80°C.

A number of experiments performed in the fractional factorial design is determined by  $N = 2^{k-p}$ , where  $p$  is the size of the fraction. For example 2<sup>7-4</sup> fractional factorial design will cover 1/16 fraction of the full 2<sup>7</sup> design (full design requires 128 experiments) and it is equal to the 2<sup>3</sup> full factorial designs (requires 8 experiments, Table 2). In this example, the main and interaction effects of three factors e.g. **x<sub>1</sub>, x<sub>2</sub> and x<sub>3</sub>** (the most important factors that influence measured response(s)) will be investigated, while other four factors remain constant. By performing eight experiments, it will be possible to estimate the main effects e.g. **x<sub>1</sub>** and **x<sub>3</sub>**, the second order interaction e.g. **x<sub>1</sub>x<sub>3</sub>** and the third order interaction (**x<sub>1</sub>x<sub>2</sub>x<sub>3</sub>**). Mostly, fractional factorial designs are performed to evaluate up to four or five order interactions. This may involve systems where changes in levels of five factors will have a significant interaction effects. However, the effect of the factors to the response will not be as accurate as in the case of the full factorial design. This is due to the confounding of the factors. In the case of the 2<sup>7-4</sup> fractional factorial design, the main effect of the **x<sub>1</sub>** factor is confounded by the interaction effects of the **x<sub>2</sub>** and **x<sub>3</sub>** factors, and the main effect of the **x<sub>2</sub>** factor is confounded by the interaction of the **x<sub>1</sub>** and **x<sub>3</sub>** factors. In the same manner, two factor interaction effects are contaminated by the three factor interactions. This leads to the conclusion that, fractional factorial based screening and robustness testing can provide holistic results when the interaction effects of the two factors are not significantly contaminating the main effect of the third factor. In practice, a chosen design generator(s) will control the confounding effects, resolution of the fractional factorial design and the fraction of the investigating experimental domain. A more detailed description on generations of a fractional factorial design, evaluation of the models and separation of confounded effects is provided in the literature (Eriksson et al., 2000; Lundstedt et al., 1998).

#### *D-optimal design*

As mentioned earlier, the D-optimal design can be applied for screening and optimization problems and it can handle a non-linear relationship between the factors and the response(s). D-optimal design is based on selecting the best subset of experiments, from all available combinations of experiments that will span the largest experimental domain as possible. This is based on calculation of the determinant of the relevant  $X'X$  of the experimental matrix  $X$ . The algorithm of the D-optimal design will select a user defined number of experiments from all possible combinations in such a way that the selected experiments will possess the maximum determinant of the  $X'X$  experimental matrix. Then selected experiments will be performed in a randomized order and evaluated in relation to the measured response(s). In the least squares manner the model can be written as  $\mathbf{y} = \mathbf{b}_0 + \mathbf{b}_1\mathbf{x}_1 + \mathbf{b}_2\mathbf{x}_2 + \mathbf{E}$  or  $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{E}$  and  $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ . By selecting a number of most representative experiments, D-optimization algorithm assumes that the regression model is correct. D-optimal design can be applied in many different situations that are not suitable for fractional factorial or other approaches. Moreover, it is one of the most commonly used optimization approaches, which considers interaction effects of the factors. Application of D-optimal design in conjunction with fractional factorial design has found a wide use in metabolomic studies (Danielsson et al., 2012; Gullberg et al., 2004).

### **3.4 Minimization of non-sample related variations in metabolomics**

Plant metabolomics deals with quantitative measurements of a wide range of metabolites with different concentrations and finds biological information based on the changes of these metabolite levels due to some effects e.g. genetic modifications, biotic/abiotic stresses, growing season and fertilizers. The variation of one metabolite level caused by one specific effect may vary from sample to sample and different metabolites might be altered in different levels. The main aim of the metabolomic study is to capture these variations as accurately as possible. However, these variations, to some extent, are always confounded by non-sample related variations. In order to extract meaningful information out of metabolomic data, analysts must be able to minimize and estimate the level of non-sample-related variations, e.g., to separate the real biological variations from the artificial data variations.

DoE is an advanced approach for minimization of the non-sample related variations in metabolomic studies and this approach has already demonstrated its potential in many studies. Prior to DoE, the main problems (factors that cause in introduction of errors) must be defined and their influence to the level of artifacts should be estimated. In fact, this requires prior knowledge about the system and sources of errors. In plant metabolomics, the main sources of non-sample-related variations are sampling (representative mass reduction, harvesting, storing, processing), metabolite extraction and instrumental variations. The level of the non-sample-related variations introduced by these sources depend on the complexity of the metabolomic protocol, number of investigated samples and the

### 3. ADVANCED GC-MS AND CHEMOMETRICS FOR METABOLOME ANALYSIS

capabilities of the analytical platforms (linear range of the detectors, stability, high-throughput screening). For example, untargeted GC-MS metabolomics of plant leaves requires careful evaluation of the following steps of the metabolomic workflow: **1.** Plant leaves must be harvested in the same way and with a minimum time interval, **2.** The time interval between harvesting and quenching must be constant throughout the samples, **3.** In order to minimize experimental error, metabolite extraction must be performed in smaller batches including 12-15 samples at a time, **4.** Use of readily volatile and/or other solvent systems with unstable composition must be minimized, **5.** Metabolite extraction of the sample set which is subject to the quantitative comparison must be performed by using the exact same protocol, **6.** Samples must be stored under cooled condition that is safe for the complex extract, **7.** The derivatization method must be as unbiased as possible toward the sample matrix and functional groups of the different metabolites, **8.** The derivatization time must be controlled and kept constant, **9.** The GC-MS sample introduction method must be optimized towards reduction of artifacts such as septum and injection port derived peaks, **10.** The split ratio (depend on the concentration of metabolites, injection volume and detection limit of the detector) between the column and vent flow rates must allow introduction of sufficient amount of the sample to detect a maximum number of metabolites and at the same time avoid detector saturation. **11.** In order to minimize the sample loss during the injection, splitless time, the injection port temperature and/or temperature ramp must be optimized, **12.** The oven temperature program and the carrier gas flow rate must provide acceptable chromatographic resolution, and at the same time, to avoid metabolite degradation due to the high temperature and/or interactions with a stationary phase, **13.** Mass spectrometer scan speed (at the specified  $m/z$  range) must allow the best possible resolution and provide a reasonable  $s/n$  ratio (too high a scan speed may result in a low  $s/n$  ratio and noise peaks, while a slow scan rate can suffer from unresolved peaks). Thus, substantial part of the non-sample-related variations can be reduced by appropriate optimization of the above-mentioned steps of the GC-MS metabolomic workflow. However, the reproducibility of the GC-MS profiles of the biological and technical replicates can be further improved by detailed study of one or more steps (factors) of the workflow by DoE approach.

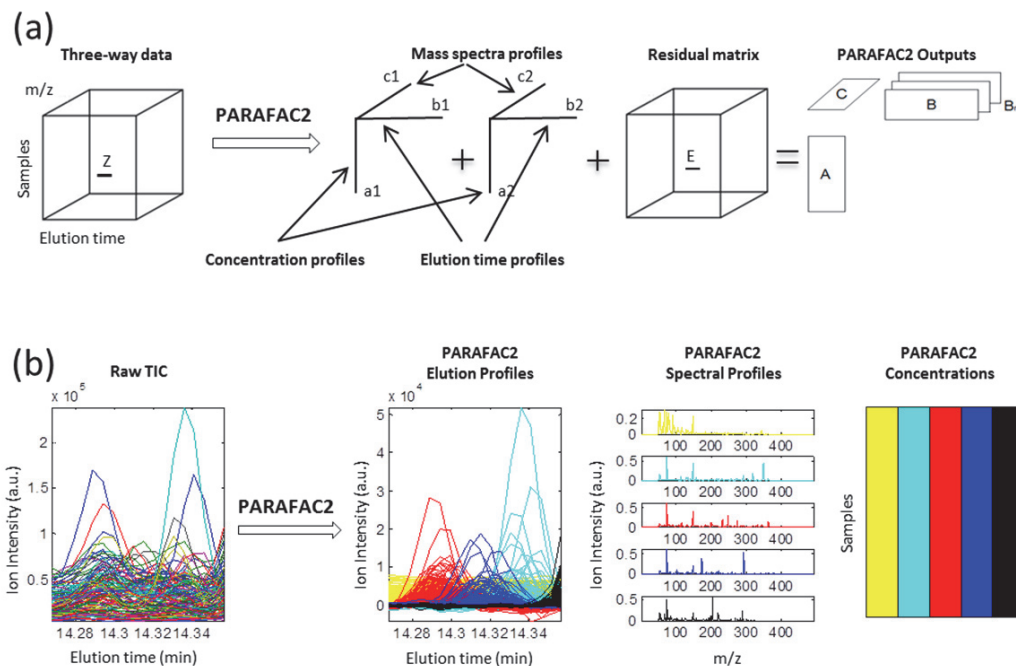
In the given example, DoE based optimization can be applied for increasing a number of detectible metabolites, the  $s/n$  ratio and/or decrease specific artifact effects. While robustness testing e.g. fractional factorial design may assist in evaluation of the minute fluctuations of factors such as metabolite extraction temperature or time. Moreover, sample derivatization (methoxiamination and/or trimethylsilylation) conditions such as, derivatization time, temperature, reagent concentration as well as gas chromatography conditions like, injection temperature, initial temperature, heating rate and gas flow rate can be optimized to obtain best profiles. The best GC-MS profiles are usually determined by their reproducibility and the amount of information. However, by development of systems biology and metabolomics, the need for extraction of a significantly smaller size of biological variations from the complex metabolomic data is increasing. This urges to minimize non-sample related variation to a level that it will be several fold lower than the smallest biological variation

caused by the effects in question. Thus, both minimum error and maximum information are the main targets of most studies aimed at development of metabolomic protocols.

### 3.5 PARAllel FACtor Analysis 2 (PARAFAC2)

PARAFAC2 is a multi-way decomposition method, which is commonly applied to model three-way data sets (Bro et al., 1999; Harshman, 1972; Kiers et al., 1999). It is originated from the PARAllel FACtor Analysis that was developed in late 1960s by Richard Harshman (Harshman, 1970). Both methods, PARAFAC and PARAFAC2 can be considered as the generalization of the principal component analysis (PCA) to higher order arrays. In contrast to PCA, PARAFAC2 does not suffer from rotational problems and is able to model three-way data sets by decomposing it into a smaller number of components that will be represented by scores and loadings (Fig. 7 (a)). In order to condense the three-way data in such a way, PARAFAC2 applies some constrains that restricts the degree of freedom and provides simpler and more robust models. Therefore, PARAFAC2 may not be able to capture the variation that PCA could explain (this is often the case when dealing with a complex data sets with higher order of variations). Any three-way data can be unfolded to two-way matrix (in any mode) and modeled by PCA, which may result in explanation of the substantial part of the data variation, however interpretation of such a model is challenging. Instead, PARAFAC2 offers several advantages for exploration of three-way arrays. Firstly, PARAFAC2 solutions are unique, which means its scores and loadings directly represent the modes of the investigated data array. Moreover, PARAFAC2 models are robust and easily interpretable. However, the model validation might be challenging, if data is complex. Although, recent studies performed for improvement of PARAFAC (Bro and Kiers, 2003) and PARAFAC2 (Kamstrup-Nielsen et al., 2013) model validation provide more reliable and easier way of deciding the number of components that would describe the data best.

### 3. ADVANCED GC-MS AND CHEMOMETRICS FOR METABOLOME ANALYSIS



**Figure 7.** (a) PARALLEL FACTor Analysis based decomposition of the three-way raw GC-MS data interval. (b) Example of the PARAFAC2 based processing of the raw GC-MS metabolomic data interval.

PARAFAC can decompose a data cube if the correct number of components are fitted and describe the data. However, some three-way data sets may suffer from disturbance of its trilinear structure. For example, retention time shifts in chromatography. In such a situation, PARAFAC will fail to provide reliable models. The main difference between PARAFAC and PARAFAC2 is that PARAFAC2 is less restrictive to the trilinear structure of the data and it is able to cope with data shifts in some extent. For example, in chromatography, retention time shifted peaks of the same metabolites over the different samples can still be modeled as the same chemical, because PARAFAC2 uses not only retention time but also the mass spectral information (since mass spectra of these shifted peaks will be identical if they are derived from the same metabolite). All these features of PARAFAC2 e.g. uniqueness, shift and noise handling, and easier interpretation made the method very useful for processing raw metabolomic three-way data sets derived from hyphenated platforms such as GC-MS (Amigo et al., 2008; Amigo et al., 2010a; Amigo et al., 2010b), LC-DAD (Garcia et al., 2007; Marini et al., 2011) and LC-MS (Khakimov et al., 2012). By PARAFAC2 processing of such hyphenated metabolomic data, it is possible to extract all the quantitative and qualitative information (Fig.7 (b)). The PARAFAC2 model of the three-way raw GC-MS data defined by elution times  $\times$  mass spectra  $\times$  samples provide three outputs: 1. PARAFAC2 elution time profiles that represent the elution profiles of the resolved



### 3. ADVANCED GC-MS AND CHEMOMETRICS FOR METABOLOME ANALYSIS

peaks, 2. PARAFAC2 mass spectral profiles that correspond to the actual mass spectra of the resolved peaks and 3. PARAFAC2 concentration profiles, which represents the areas of the resolved peaks.

Today PARAFAC2 is probably one of the most comprehensive methods for processing raw metabolomic data that allows extraction of vast amount of information in a high-throughput manner (several hundred samples can be processed simultaneously depending on the capabilities of the PC). Paper 1 demonstrates the first application of PARAFAC2 to LC-MS metabolomic data and discusses model validation and other considerations related to model interpretation. The main advantages of PARAFAC2 for processing chromatographic data are briefly described below:

#### *Baseline elimination*

Baseline is one of the common challenges in chromatography and it mainly arises from the stationary phase, temperature ramp program in GC, gradient elution program in LC, inconsistent column pressure, pH, contaminations, reagent and solvent. By modeling raw chromatographic data without any pre-processing steps, it is possible to eliminate the baseline as a separate component of the PARAFAC2 model.

#### *Resolution of overlapped chromatographic peaks*

Overlapping of chromatographic peaks is the most commonly faced challenge in GC-MS, LC-MS and CE-MS based metabolomic studies of complex mixtures. This is due to the insufficient separation power of the techniques for chemically similar metabolites. PARAFAC2 can resolve such overlapped peaks using their mass spectra. The resolution power of PARAFAC2 depends on the overlapping level (significance of overlapped peak shoulder) and the mass spectral difference between the overlapped peaks. Although, some chromatographic data analysis software allow resolution of overlapped peaks that requires manual resolution of one sample at a time, while PARAFAC2 can handle several hundred GC-MS profiles simultaneously.

#### *Retention time shifts*

Retention time inconsistency of metabolites over samples is also a significant drawback of chromatographic systems. This might be due to inconsistent gradient program, column degradation and contamination, temperature and pH fluctuations, contamination of injection system and other mechanical and/or electronic problems arising during the runs. As mentioned earlier, PARAFAC2 is able to model retention time shifted peaks of the same metabolites as the same component if their mass spectra are identical. This eliminates any need for prior alignment of the data.

#### *Low s/n ratio*

If the investigated chromatographic data complexity allows the development of a valid PARAFAC2 model (with the correct number of components that is equal to the number of variation present in the data), it facilitates detection of low s/n peaks. Baseline elimination and chromatographic peak

### 3. ADVANCED GC-MS AND CHEMOMETRICS FOR METABOLOME ANALYSIS

resolution further enhances the chance to detect very low s/n peaks that are severely hidden by the baseline and/or abundant peaks.

#### *Accurate quantification*

PARAFAC2 allows automatic peak quantification with a minimum interference of the analysts. PARAFAC2 concentration profiles represent the actual areas of the resolved peaks that is not always possible by using conventional chromatographic data analysis software due to the overlapping and high complexity of the data. For quantification of peaks based on their area, the method considers the shape of the peak individually for each sample, thus width, heights and retention time differences of the peaks do not influence their quantification.

#### *Mass spectral deconvolution*

PARAFAC2 performs automatic mass spectral deconvolution of resolved, overlapped, low s/n ratio and other elusive chromatographic peaks. Mass spectral profiles of PARAFAC2 models represent experimental mass spectra of metabolites. PARAFAC2 mass spectra displayed 95-99% similarity with the actual mass spectra of the resolved saponin peaks (Khakimov et al., 2012). In terms of mass spectral deconvolution, PARAFAC2 outperforms its alternatives such as AMDIS and ChromaTOF by its high-throughput nature that allows processing of several samples at a time. Most importantly, PARAFAC2 facilitates the deconvolution of mass spectra of severely hindered peaks that are hidden by the baseline, artifact peaks and peaks that are more abundant. PARAFAC2 has a high potential for deconvolution of mass spectra obtained from the quadrupole and ion trap mass analyzers, while processing of UPLC-QTOF-MS and GC-QTOF-MS are not reported yet.

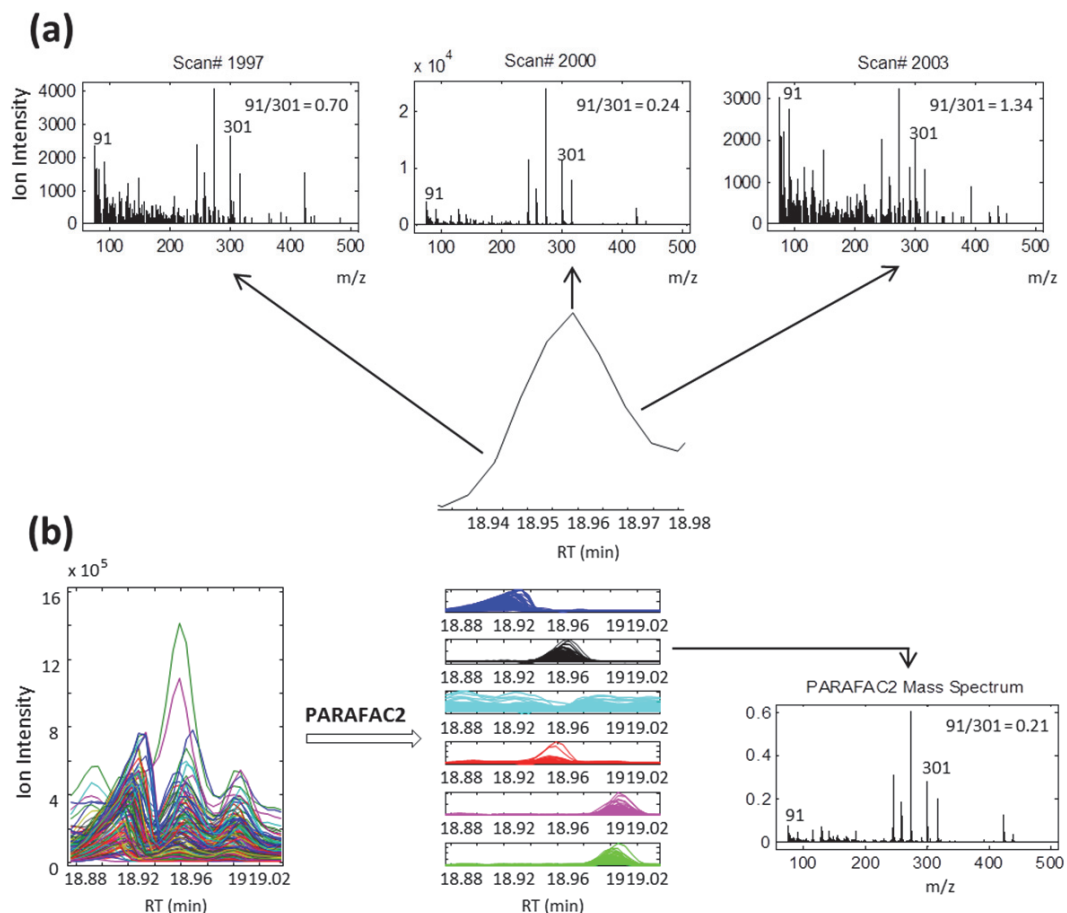
#### *Elimination of mass spectral skewing*

Mass spectral skewing (distortion of relative mass spectral peak intensities) is one of the problems that hampers qualitative analysis of metabolomic data obtained by using relative low scan rate mass spectrometers (Watson and Sparkman, 2007b). Spectral skewing occurs due to the transient changes in the partial pressure of the metabolite inside the ionization chamber, as they elute from the column. In other words, it can be described as a disagreement between the gas chromatographic separation systems and the mass spectrometry detection system. The mass spectrometer requires a constant pressure of the metabolite during the data acquisition. However, it is not possible when the sample concentration is changing in the GC eluate with time. Unfortunately, modern GC-MS systems cannot provide a constant pressure of the sample in ionization chamber while its full mass spectra are recorded. This causes alterations of the original ratios of the m/z peaks of metabolites and makes their identification difficult. Spectral skewing is more pronounced when the peaks' become broader. In this case, a reliable mass spectrum of the peak must be evaluated when the partial pressure of the metabolite is more stable (the top of the peak).

### 3. ADVANCED GC-MS AND CHEMOMETRICS FOR METABOLOME ANALYSIS

Figure 8 (a) shows an example of the EI-MS mass spectral skewing when the GC-MS peak width becomes broader. In this example the mass spectrum of a peak is demonstrated at three different scan points, when the peak started to elute (scan # 1997), when the peak reached its maximum (scan # 2000) and when the peak is decreasing (scan # 2003). These three mass spectra are different, although they represent the same metabolite. At the earlier scan point (scan # 1997),  $m/z$  peak intensities tend to increase from low  $m/z$  to high  $m/z$ . This is because quadrupole mass analyzer scans through the investigated  $m/z$  range from low  $m/z$  to high  $m/z$  that requires some time to record whole spectrum, in which partial pressure of the metabolite gradually increases. Therefore, by the time higher  $m/z$  ions are flying through the quadrupole, the concentration of the metabolite will be higher in the ionization chamber, which in fact results in higher ion intensities. In contrast to this, at the later scan point (scan # 2003), the partial pressure of the metabolite is relatively lower than at the earlier scan point, therefore intensities of the lower  $m/z$  peaks are higher than the intensities of the higher  $m/z$  peaks. While, at the maximum of the peak (scan# 2000) the relative ratios of the  $m/z$  peaks are closer to the original mass spectrum of this metabolite. The ratios of 91  $m/z$  to 301  $m/z$  are 0.7023, 0.2410 and 1.3488 in scan points 1997, 2000 and 2003, respectively.

### 3. ADVANCED GC-MS AND CHEMOMETRICS FOR METABOLOME ANALYSIS



**Figure 8.** (a) Illustration of mass spectral skewing due to dynamics in partial pressure of the analyte in an instrument (GC-Quadrupole-MS) that is scanned from a low  $m/z$  value to a high  $m/z$  value during GC elution (b) Elimination of mass spectral skewing by PARAFAC2 model based mass spectral deconvolution.

Another advantage of PARAFAC2 modeling of raw GC-MS data, which is not mentioned in its earlier applications, is that it can solve above-mentioned mass spectral skewing of broad peaks. This is because PARAFAC2 averages the mass spectrum of the resolved peak across all its scan points, and the obtained mass spectrum, which will better reflect the original spectrum of the metabolite (Fig. 8 (b)). The PARAFAC2 mass spectral profile of the broad GC-MS peak demonstrated in figure 8 (a) shows that the ratio of the 91  $m/z$  to 301  $m/z$  is 0.2103 which is very close to the ratio of these  $m/z$  peaks in the original mass spectrum of this peak (0.2410). This proves that, the PARAFAC2 based mass spectral deconvolution is accurate, even in the presence of the spectral skewing.

However, the PARAFAC2 approach also has some drawbacks primarily related to its use by non-specialists. Firstly, because it has not implemented into a graphical user interface, yet, that can simplify its use in everyday analysis. Although, an automatic model validation approach is suggested, it is not yet mature therefore, choosing an optimal number of components to describe the data best still requires some chemometric knowledge. Despite its comprehensiveness, the method provides fruitful results when the data is less complex. Therefore, PARAFAC2 based processing of raw chromatographic data is mainly performed in intervals, where the data is divided into smaller intervals in retention time dimension (Amigo et al., 2008; Khakimov et al., 2012). This, however, cannot be considered as the method's disadvantage, since division of the complex chromatographic data into smaller intervals facilitates better understanding of the data and it can be performed in few minutes even if the data is as complex as GC-MS profiles of crude extracts that may contain up to 500 peaks. Moreover, interval based PARAFAC2 modeling of chromatographic data enables detection of low s/n peaks and even under the noise peaks, which are usually ignored in the shadow of much abundant peaks and baseline drifts. The only case when the resolution power of PARAFAC2 method is limited is when isomer peaks are severely overlapped. It is, however, difficult to resolve such peaks because their mass spectra are usually very similar or even identical.

## 3.6 ANOVA-simultaneous component analysis (ASCA)

The significance of the variation between two groups of samples, based on one measured variable is usually estimated by using *student's t-test*, which was introduced by W. S. Gosset in 1908 (Box, 1987). The significance of variation present among several groups of samples can be evaluated by analysis of variance (ANOVA) (Searle, 1971). ANOVA is a statistical hypothesis testing method and it is based on the null hypothesis assumption (no difference between investigated groups based on the measured variable). In order to estimate the significance of the variations of groups, ANOVA calculates the variation present within each group and the means of these variations between the groups. The mostly utilized value for evaluation of the significance of the treatment is the *p-value*, which will indicate the chance of such a group differentiations due to the treatment, under the null hypothesis. For multivariate data where several responses are measured, significance of variance among groups can be evaluated by using multivariate-ANOVA (MANOVA) (Mardia et al., 1979). However, when multivariate data become complex, like in the case of metabolomics, and variables co-vary, the MANOVA fails to provide a reliable significance test due to assumptions that are not valid. In addition, several other approaches based on PCA (Bratchell, 1989) and PLS (Ståhle and Wold, 1990) were developed for the analysis of variance in multivariate data sets.

Analysis of variance (ANOVA)-simultaneous component analysis (ASCA) is the most recent and advanced method for analysis of variance that is suitable for the various kinds of multivariate data

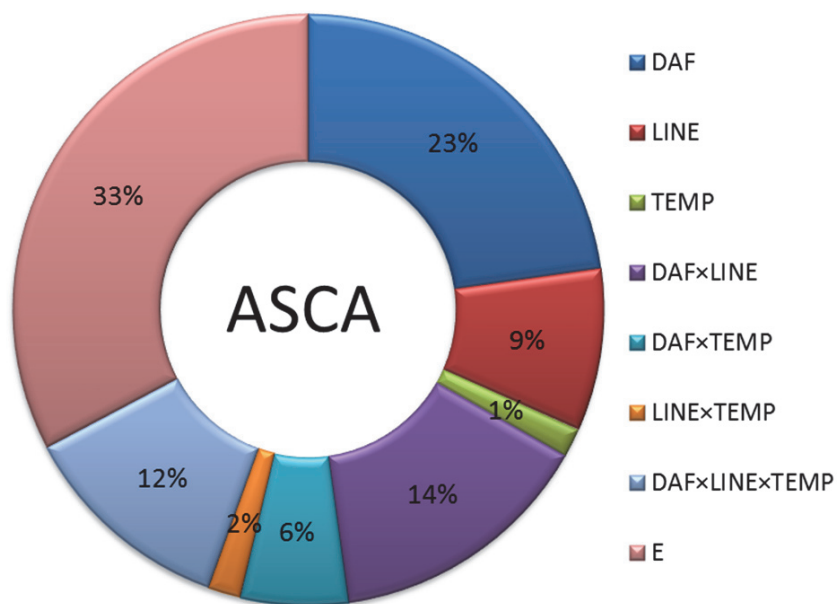
### 3. ADVANCED GC-MS AND CHEMOMETRICS FOR METABOLOME ANALYSIS

sets, including metabolomic data (Smilde et al., 2005). ASCA is developed based on the multilevel component analysis and can be applied to various datasets obtained from designed experiments. Initially, ASCA was developed for separating the underlying structures present in the complex metabolomic data. The ASCA model can be described as a direct multivariate generalization of the ANOVA model (Equation 4 in Smilde et al., 2005) and just like PCA analysis, it is based on solving the least squares problem (Equation 6 in Smilde et al., 2005). In the example of the GC-MS metabolomic data set described in Paper 4, metabolomic variations of whole-grains of three barley genotypes associated with the grain filling time (days after flowering (DAF)), genotype differences (LINE) and growth temperature (TEMP) can be evaluated for all detected metabolites by applying ASCA modeling. This model will comprise three main and four interactions effects like as demonstrated in equation 3:

$$\mathbf{X} = \mathbf{X}_{\text{DAF}} + \mathbf{X}_{\text{LINE}} + \mathbf{X}_{\text{TEMP}} + \mathbf{X}_{\text{DAF} \times \text{LINE}} + \mathbf{X}_{\text{DAF} \times \text{TEMP}} + \mathbf{X}_{\text{LINE} \times \text{TEMP}} + \mathbf{X}_{\text{DAF} \times \text{TEMP} \times \text{LINE}} + \mathbf{E} \quad (\text{Eq. 3})$$

The metabolome of a system is sensitive to external perturbations, thus designed metabolomic studies usually causes introduce of substantial amount of variation reflecting the design rather than pure biology. Separation of such an interfering effect allows extraction of hindered information (Fig.9.). Thus, ASCA partitions the total variation of the dataset into separate parts that represent variations corresponding to the different factors and allows estimation of the significance of these factor effects. Detailed theory of ASCA and its comparison with other alternatives can be found in (Jansen et al., 2005; Smilde et al., 2008; Zwanenburg et al., 2011). To date, ASCA has demonstrated a high potential for extracting information from the designed metabolomic datasets obtained from NMR (van Velzen et al., 2008), GC-MS (Chang et al., 2006) and LC-MS (Wang et al., 2009a).

### 3. ADVANCED GC-MS AND CHEMOMETRICS FOR METABOLOME ANALYSIS



**Figure 9.** ANOVA-simultaneous component analysis (ASCA) based separation of the variance present in the designed metabolomic data obtained from barley seed GC-MS metabolomic profiling (from paper 4). DAF - days after flowering or grain filling effect, LINE - metabolomic differences present between investigated three different barley lines, TEMP - growing temperature effect and E - random variation.

## 4 UNPUBLISHED STUDIES

This section comprises three studies performed within this PhD project that are not published yet. The first study concerns development of the metabolomics protocol by applying three different analytical platforms, GC-MS, LC-MS and NMR. The aim of this study was to perform untargeted analysis as well as metabolomic profiling of saponins in plant-insect interaction study between F2 population generated from the resistant glabrous (G) type and susceptible pubescent (P) types of the *Barbarea vulgaris* plants and the insect, *Phyllotreta nemorum*. The second study involved targeted GC-MS metabolomic analysis of the triterpenes produced by combinatorial biochemistry in tobacco plant leaves. In this study the new trimethylsilylation method by using TMSCN was applied (Khakimov et al., 2013) and demonstrated its power for quantitative analysis of minute amounts of the triterpenes produced in tobacco leaves. The third study demonstrates purification of the P and G type *B. vulgaris* plants saponins and comprises LC-MS/MS results. The aim of this study was to uncover the structures of the most abundant triterpenoid saponins of the two different plants and provide better understanding of the plant-insect interactions and the relationships between saponin structures in *B. vulgaris* and their toxicity to *P. nemorum*.

### 4.1 Optimization of comprehensive metabolomic protocol for GC-MS, LC-MS and NMR analysis of *Barbarea vulgaris* leaves

One of the aims of the PhD project was to conduct a comprehensive metabolomic study of the *B. vulgaris* F2 population derived from a cross of the parental glabrous (G) and pubescent (P) type. These two morphologically different phenotypes of *B. vulgaris* also differ by their resistance to an insect herbivore *P. nemorum* (Agerbirk et al., 2003a; Shinoda et al., 2002). Resistance of G-type plants to herbivory by the flea beetle larvae of *P. nemorum* was evaluated by targeted metabolomics based on feeding deterrent activity bioassays. This enabled identification of feeding deterrents, triterpenoid saponins, such as hederagenin cellobioside (Shinoda et al., 2002) and oleanolic acid cellobioside (Agerbirk et al., 2003a). A more comprehensive metabolomic analysis for identification of *B. vulgaris* bioactive metabolites was performed by Kuzina et al., 2009 where they conducted LC-MS based metabolomic profiling of saponins of both, G- (resistant), and P- (susceptible) parents, and the segregating F2 population (Kuzina et al., 2009). This approach confirmed previous findings and uncovered two additional saponin like metabolites that depicted high correlation with plants' resistance against insect larvae for parental plants as well as the F2 population. The 160 F2 plants represented the whole range from full susceptibility to full resistance to flea beetle larvae and varied by their content of saponins. Later, two unknown saponin like metabolites of the resistant plants found in Kuzina et al., 2009 were identified as gypsogenin and 4-epihederagenin cellobioside (Nielsen



#### 4. UNPUBLISHED STUDIES

et al., 2010). The first application of the multi-way decomposition method, PARAFAC2, on LC-MS type of data (Khakimov et al., 2012) employed raw LC-MS metabolomic data obtained from leaves of the F2 populations, which is also used in Kuzina et al., 2009. This study showed that metabolomic data treatment is an important step for using all the information present in the data set and to extract as much biological information as possible. PARAFAC2 based processing of this LC-MS data enabled tentative identification of five more saponin like metabolites, in addition to the four previously found saponins, to be associated with F2 plants' resistance against *P. nemorum* larvae based on PLS regression and correlation analysis between PARAFAC2 scores of resolved peaks and F2 plants' resistance level. It is worth to mention that all these studies involved either targeted metabolomics for identification of insect deterrent bioactive saponins or attempted to find metabolomic differences of parental plants and F2 population primary by focusing on saponins. However, it is more likely that, apart from saponins, the F2 population generated from the cross of two significantly different phenotypes of *B. vulgaris* may result in alterations of other metabolites derived from different biosynthetic pathways due to pleiotropy. Therefore, comprehensive metabolomics of the F2 population may provide more insight into metabolome-insect resistance relationship and evaluation of other than mevalonate pathways by covering a broader range of metabolites.

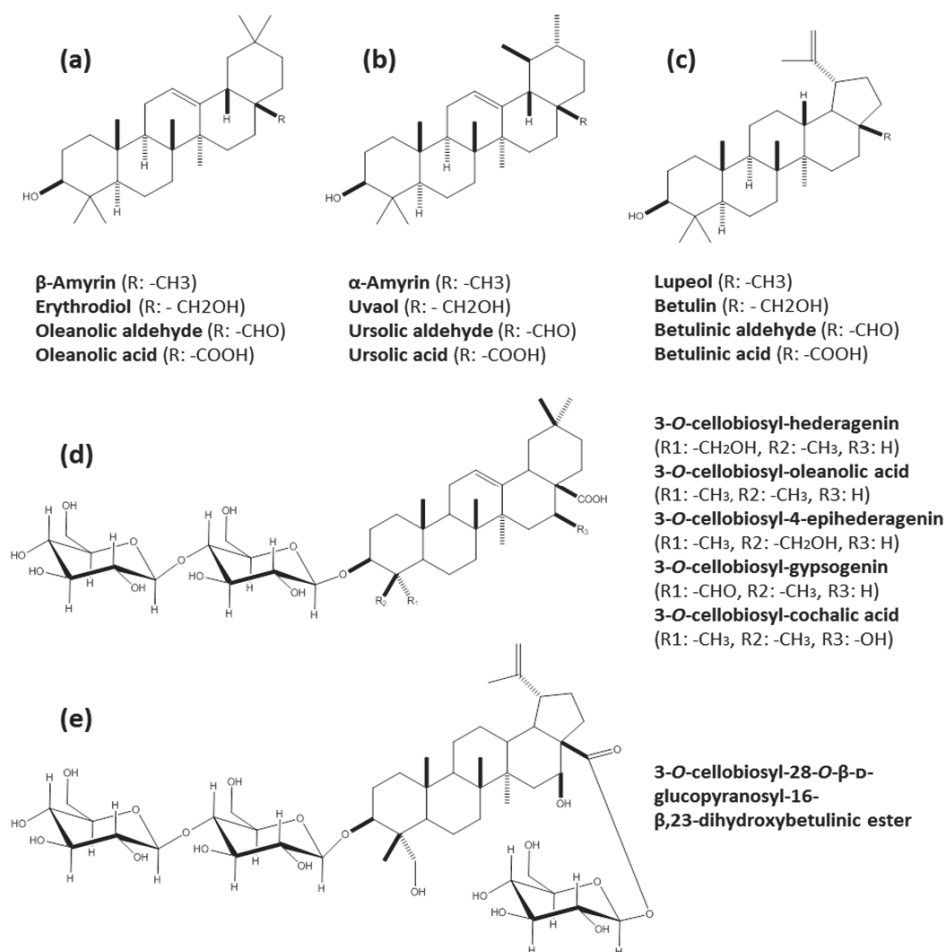
In this study, we attempted to develop a single metabolomic protocol covering as many metabolites as possible, including semi-polar metabolites such as saponins. For this reason, we aimed to employ three analytical platforms, GC-MS, LC-MS and NMR in parallel, since they complement each other and allow detection of a broader range of metabolites than using a single platform. One of main limitations for performing such a comprehensive metabolomics was the limited amount of the sample material. Approximately 10 mg of the fresh leaves of F2 population used in Kuzina et al., 2009 were kept in a -80°C freezer. Our aim was to develop a single metabolomic protocol to extract maximum metabolome in an untargeted way, split this extract into three parts, and use them for GC-MS, LC-MS and NMR analysis. All the metabolomic protocol optimization works were performed by using 4-12 weeks old G- and P-type plants grown in a climate chamber. Performance of metabolomic protocol was evaluated for each analytical platform separately.

##### 4.1.1 GC-MS method optimization

Due to the high chromatographic resolution and sensitivity of GC-MS, we decided to use it for detection of primary e.g. amino acids, carbohydrates, organic acids and secondary metabolites e.g. terpenes and phenolics of *B. vulgaris* leaves. Our aim was to detect as wider range of metabolites as possible from the 10 mg of plant leaves in an untargeted approach. Nevertheless, detection of triterpenoid profiles of the F2 population remains crucial, since this fraction of the metabolites may hold valuable biological information present among the F2 plants. However, GC-MS detection of *B.*

#### 4. UNPUBLISHED STUDIES

*vulgaris* plant triterpenes is not straightforward, since they are linked to one, two or more sugar moieties and form saponins. The boiling points of saponins usually exceed the maximum allowed temperature of the GC-MS e.g. oleanolic acid itself (without any sugar bonded to it) has a melting point of > 300°C, which makes them non-volatile. Therefore, we started GC-MS method optimization by establishing a method for quantitative detection of triterpenes of plants by using standard compounds, aglycones, such as hederagenin, oleanolic acid, betulinic acid and a saponin,  $\alpha$ -hederin, which is structurally similar to hederagenin cellobioside (difference between  $\alpha$ -hederin and hederagenin cellobioside is that sugar moieties of  $\alpha$ -hederin consist of arabinopyranosyl and mannopyranosyl, while the latter possess two glucopyranosyles) (Fig.10).



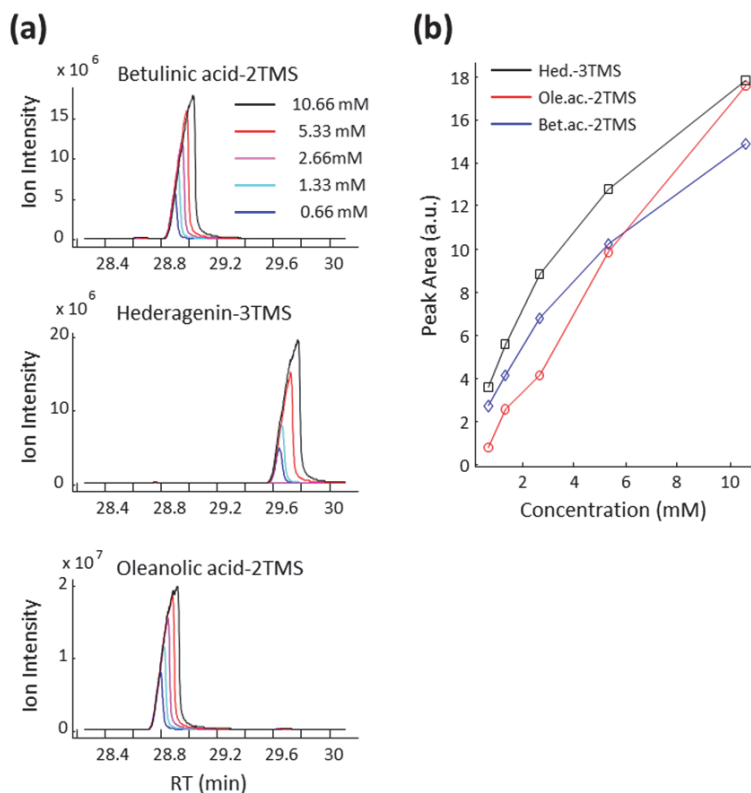
**Figure 10.** Structures of triterpenoid aglycones derived from  $\beta$ -Amyrin **(a)**,  $\alpha$ -Amyrin **(b)**, lupeol **(c)** and five oleanolic acid derived triterpenoid saponins **(d)** and one lupane type triterpenoid saponin identified from the G-type *B. vulgaris* **(e)**.

#### 4. UNPUBLISHED STUDIES

In the first step, we attempted to detect standard aglycones, hederagenin, oleanolic acid and betulinic acids by direct injection of 1  $\mu\text{l}$  aliquot of standard mixture of 0.2  $\text{mg ml}^{-1}$  solution of aglycones in methanol. However, it was not possible to detect aglycones from the GC-MS analysis, though injection port and GC oven program was set to the maximum allowed temperatures, 320°C and 330°C, respectively. This suggest that the applied temperature was not enough to vaporize the aglycones and fly through the GC column and/or were degraded inside the column or injection port, since they possess polar functional groups such as  $-\text{COOH}$  and  $-\text{OH}$ . In a second attempt, we tried trimethylsilylation of complete dried 50  $\mu\text{l}$  aliquot of the initially used standard mixture with 50  $\mu\text{l}$  MSTFA and TMSCN separately by incubating at room temperature for 2 hours and injected 1  $\mu\text{l}$  aliquot into GC-MS. This enabled detection of three peaks at different retention times that each corresponded to the TMS-derivatives of the three standard aglycones (Fig.11 (a)). However, the abundance of aglycone peaks were up to 4 fold greater when they were derivatized by using TMSCN. This depicted higher silylation capabilities of TMSCN over the most commonly used reagent MSTFA. Therefore, we have performed a comprehensive trimethylsilylation method comparison study by using these two reagents towards silylation of various primary and secondary metabolites from a standard mixture and blueberry fruit extracts (paper 2). Thus, TMSCN outperformed MSTFA in terms of silylation reaction rate, sensitivity, unbiased silylation of various functional groups and profile reproducibility. Thus, further derivatization reactions employed TMSCN only. Moreover, the calibration curve of the TMSCN derivatization of triterpenoid aglycones also showed a high quantitative power of the method (Fig.11 (b)).

Thus, GC-MS detection of triterpenoid aglycones of saponins were established, however, detection of saponins remains challenging, since their molecular size and boiling point is even greater than their aglycones. GC-MS detection of silylated  $\alpha$ -hederin was not possible, even by using splitless mode injection and by varying silylation conditions (24-70°C and 5 min – 8 hour). TLC analysis of the derivatized and not derivatized  $\alpha$ -hederin depicted two different spots and suggested that the trimethylsilylation reaction did occur, though it is impossible to know how many hydroxyl functional groups of  $\alpha$ -hederin are derivatized. This led to the fact that the boiling point of  $\alpha$ -hederin-nTMS is higher than the GC-MS temperature settings and/or it is not stable under high temperature. Thus, it was concluded that detection of the saponin fraction of metabolites of *B. vulgaris* plant leaves was not possible by direct derivatization of the plant complex extracts and GC-MS analysis. Therefore, it was decided to apply acidic hydrolysis of the plant extracts prior to derivatization. Since acidic hydrolysis cleaves the glycosidic and ester bonds through which sugar moieties are attached to the aglycones. However, hydrolyzation followed by derivatization based GC-MS analysis will result in detection of triterpene aglycone pool rather than each triterpenoid saponin separately. It is worth mentioning that acidic hydrolysis of complex plant extract will affect most of the metabolites that possess ester, glycosidic and even ether bonds. Therefore, in order to evaluate the intact metabolites of the extract, it is also crucial to analyze the samples without hydrolysis. According to the previous studies and the latest LC-MS/MS analysis of the parental G and P type of *B. vulgaris* (Fig. 16 and Table 4) there are

more than ten different aglycones that make up the total saponin profile of the plants (Khakimov et al., 2012; Kuzina et al., 2009). Thus, GC-MS analysis of hydrolyzed F2 plant extracts will provide a valuable triterpenoid profile as well as other secondary metabolites such as phenolics, which are also mostly present in conjugated forms with the carbohydrates and other cell membrane components.



**Figure 11. (a)** Superimposed TIC chromatograms of GC-MS data obtained on trimethylsilylated standard triterpenoids, betulinic acid, hederagenin, oleanolic acid at five different concentrations. **(b)** Calibrations curves of the triterpenoids.

In order to develop an optimal GC-MS metabolomic protocol for analysis of hydrolyzed and not hydrolyzed extracts, GC-MS instrumental parameters, including injection, GC oven program and MS settings were optimized to gain high sensitivity, accuracy and reproducibility. Optimal GC-MS settings were established by taking advantages of previously published GC-MS instrumental protocols for analysis of complex sample matrices (Engewald et al., 1999; Fialkov et al., 2007; Fiehn et al., 2000; Heiden et al., 2001; Horning and Horning, 1971; Liseč et al., 2006; Roessner et al., 2000). This required a compromise between the chromatographic resolutions of metabolites and their sensitivity. Moreover, the mode of injection e.g. split and splitless modes greatly determined the metabolomic data quality, since split mode injection methods were not able to detect low s/n ration peaks, while in

#### 4. UNPUBLISHED STUDIES

splitless mode injection several high abundant peaks were overloaded, which make their quantification impossible. Therefore, a compromise was necessary to detect as many metabolites as possible in a quantitative and high-throughput manner. An optimized final GC-MS protocol is presented in box 1.

<p><b>OVEN PROGRAM: ON</b> 40 °C for 3 min then 12 °C/min to 300 °C for 8 min Run Time: 32.667 min 5 min (Post Run): 40 °C</p> <p><b>BACK PTV INLET H2</b> Mode: PTV Solvent Vent Heater: Off Pressure: On 9.3896 kPa Total Flow: On 36.2 mL/min Septum Purge Flow: On 20 mL/min Gas Saver: Off Purge Flow to Split Vent: 15 mL/min at 2.5 min Vent Flow: 200 mL/min Vent Pressure: 7 kPa Until 0.3 min Cryo: Off</p> <p><b>THERMAL AUX 2 {MSD TRANSFER LINE}</b> Heater: On Temperature Program: On 290 °C for 0 min Run Time: 32.667 min</p> <p><b>COLUMN #1</b> Phenomen 7HG-G018-11Zebron ZB 5MSi 5%Phe 95%DiMe p 370 °C: 30 m x 250 µm x 0.25 µm In: Back PTV Inlet H2 Out: Vacuum (Initial): 40 °C Pressure: 9.3896 kPa Flow: 1.2 mL/min Average Velocity: 58.982 cm/sec Holdup Time: 0.84772 min Flow Program: Off 1.2 mL/min for 0 min Run Time: 32.667 min 5 min (Post Run): 0.99842 mL/min</p>	<p><b>GERSTEL MAESTRO SYSTEM SETTINGS</b> Maestro Runtime: 37.67 min GC Cool Down Time: 5.00 min Cryo Timeout: 10.00 min</p> <p><b>GERSTEL CIS</b> CIS: used Cryo Cooling: used Heater Mode: Standard Initial Temperature: 120 °C Equilibration Time: 0.30 min Initial Time: 0.30 min Ramp 1 Rate: 5.00 °C/s End Temp: 320 °C Hold Time: 10.00 min</p> <p><b>GERSTEL MPS Liquid Injection</b> Syringe: 10ul</p> <p><b>SAMPLE PARAMETERS</b> Sandwich: used with sample above Top Air Volume: 2.0 uL Inj. Volume: 1.0 uL Air Volume above: 1.0 uL Solvent Plug Volume: 0.0 uL Sandwich Solvent: Wash1 Air Volume below: 1.0 uL Inj. Speed: 2.50 uL/s Fill Volume: 5.0 uL Fill Strokes: 2 Fill Speed: 0.20 uL/s Viscosity Delay: 2 s Eject Speed: 100.00 uL/s Pre Inj. Delay: 1 s Post Inj. Delay: 1 s Inj. Penetration: 41.00 mm Sample Tray Type: VT98 Vial Penetration: 29.00 mm</p>	<p><b>CLEANING PARAMETERS</b> Preclean Sample: 1 Wash Station 1: Wash1 Preclean Solv.1: 2 Postclean Solv.1: 2 Fill Speed Solv.1: 1.00 uL/s Viscosity Delay Solv.1: 1 s Eject Speed Solv.1: 70.00 uL/s Information Solv.1: Acetone Wash Station 2: Wash2 Preclean Solv.2: 2 Postclean Solv.2: 2 Fill Speed Solv.2: 1.00 uL/s Viscosity Delay Solv.2: 1 s Eject Speed Solv.2: 70.00 uL/s Information Solv.2: Hexane</p> <p><b>MS ACQUISITION PARAMETERS</b> Acquisition Mode: Scan Solvent Delay: 8.50 min EMV Mode: Gain Factor Gain Factor: 2.00 Resulting EM Voltage: 1847 Low Mass: 50.0 High Mass: 500.0 Threshold: 150 MS Source: 230 C maximum 250 C MS Quad: 150 C maximum 200 C Timed Events Time (min): State (MS On/Off) 8.50 On 25.50 Off</p>
--	---	---

**Box 1.** Optimized GC-MS instrumental settings for untargeted GC-MS metabolomics of *B. Vulgaris* plant leaves.

After establishment of an optimal GC-MS protocol for quantitative detection of broad range of metabolites of hydrolyzed and not hydrolyzed extracts of the plant leaves, it was also necessary to optimize the metabolite extraction, hydrolysis (for saponification) and sample derivatization steps of the protocol. Pilot screening experiments demonstrated that metabolomic protocol factors such as extraction solvent composition (1), extraction time (2) and temperature (3), hydrolyzation time (4) and temperature (5) as well as trimethylsilylation time (6) and temperature (7) were the most important for the number of metabolites detected from the GC-MS analysis and their s/n ratios. Based on these

#### 4. UNPUBLISHED STUDIES

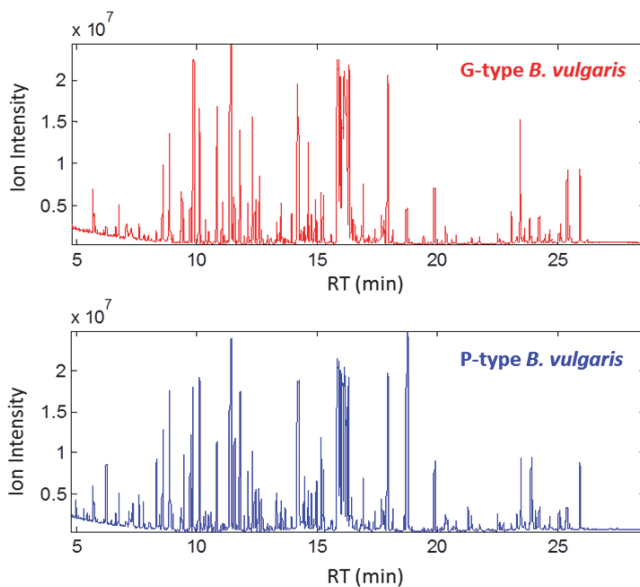
screening studies and previously published work, it was decided to use methanol as an extraction solvent (Kuzina et al., 2009; Liseć et al., 2006) and hydrochloric acid at a final concentration of 1M for hydrolysis (Arranz and Calixto, 2010; Zadernowski et al., 2005). Thus,  $2^{7-3}$  fractional factorial experiment was designed for the identification of an optimal combination of the above-mentioned seven factors of the GC-MS metabolomic protocol (Table 3).

		1	2	3	4	5	6	7
		Solvent (% of methanol)	Extraction time (min)	Extraction temperature (°C)	Hydrolyzation time (hour)	Hydrolyzation Temperature (°C)	Silylation time (hour)	Silylation temperature (°C)
1.	---+++	50	5	100	1	99	16	24
2.	++---+	100	15	70	1	99	16	24
3.	---+++	50	5	100	16	60	0,5	40
4.	+---+-	100	5	70	16	99	0,5	24
5.	+++++	100	5	100	16	60	16	24
6.	++---+	100	15	70	16	60	0,5	40
7.	0000000	75	10	85	8,5	79,5	8,25	32
8.	+++++-	50	15	100	16	99	0,5	24
9.	+-----	100	5	70	1	60	16	40
10.	+++---	100	15	100	1	60	0,5	24
11.	0000000	75	10	85	8,5	79,5	8,25	32
12.	---+++	50	15	100	1	60	16	40
13.	0000000	75	10	85	8,5	79,5	8,25	32
14.	-----	50	5	70	16	99	16	40
15.	-+++-+	50	15	70	16	60	16	24
16.	-----	50	5	70	1	60	0,5	24
17.	+++++	100- (85%)	5	100	1	99	0,5	40
18.	+---+-	50	15	70	1	99	0,5	40
19.	+++++	100	15	100	16	99	16	40

Table 3.  $2^{7-3}$  fractional factorial design (16 experiments + 3 center points) developed for optimization of the metabolite extraction and derivatization conditions for GC-MS analysis of the *B. vulgaris* plant leaves. \* Experiment number 17 performed best in terms of s/n ratio on detected hederagenin-3TMS and oleanolic acid-2TMS peaks. However, model suggests that reduction of methanol concentration from 100% to 85% will slightly improve the protocol.

This enabled identification of the optimal conditions by performing only 19 experiments varying seven factors in the specified ranges and three replicates of center points where all the factors were set to the center points. All the experiments were evaluated by measuring the response variables, peak abundances of hederagenin-3TMS and oleanolic acid-2TMS, since the main aim from this optimization was to enhance the detection of the triterpenes. In addition to this, three replicates of the center point experiments demonstrated the method reproducibility. Total ion current (TIC) chromatogram of the raw GC-MS data obtained from the non-hydrolyzed leaf extracts of G and P type plants are demonstrated in figure 12. However, these chromatographic data have not yet been comprehensively investigated, since the data of the whole F2 population has not been recorded. Nevertheless, analysis of the data by using ChemStation software suggested the presence of more than 200 hundred metabolites, though their mass spectrum has not been compared against libraries.

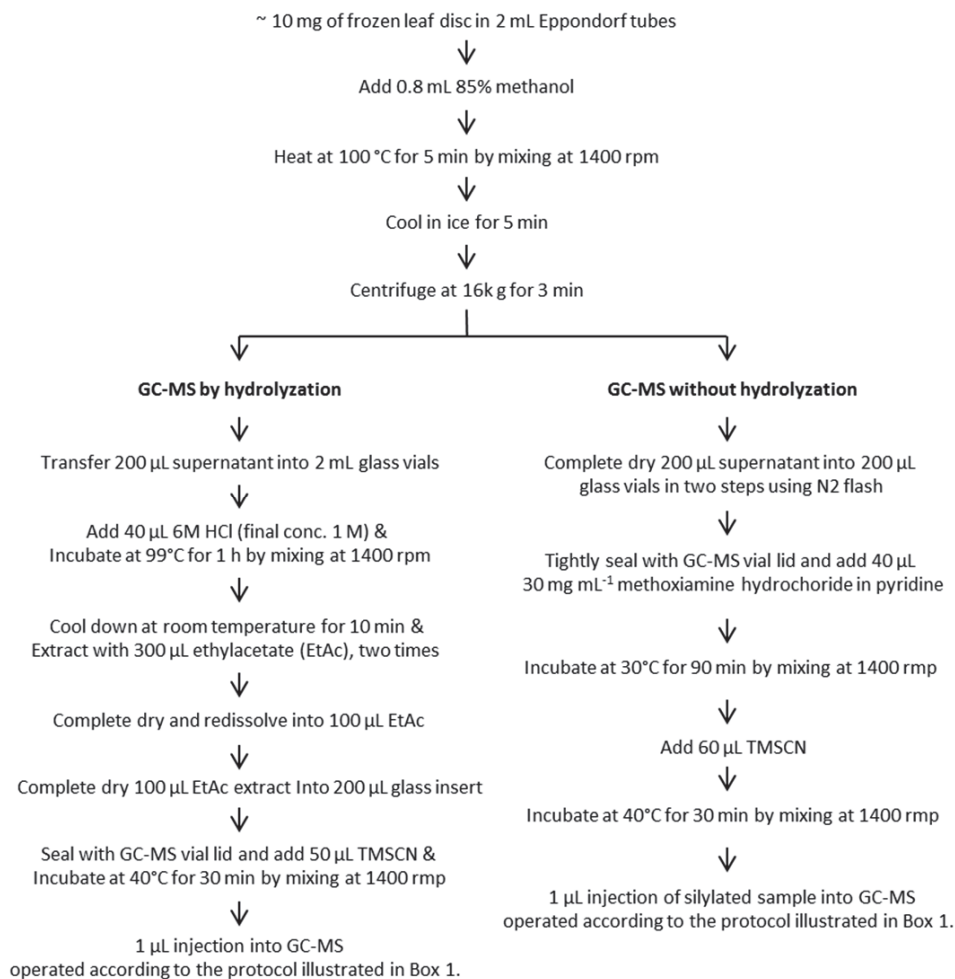
#### 4. UNPUBLISHED STUDIES



**Figure 12.** Total ion current (TIC) chromatograms of the raw GC-MS data obtained from not hydrolyzed leaf extracts of G and P type *B. vulgaris* plants, based on the metabolomic protocols illustrated in figure 12.

Thus, the final protocol was established (Fig.13) and its robustness was tested by performing GC-MS analysis of four biological replicates of G type plant leaves, which resulted in >92% similarities of the obtained GC-MS profiles.

#### 4. UNPUBLISHED STUDIES



**Figure 13.** Scheme of the metabolomic protocol developed for GC-MS, LC-MS and NMR analysis of the *B. vulgaris* plant leaves by using design of experiment. For LC-MS analysis 50 µL aliquot of the final extract was transferred into 2 mL glass vials, complete dry under reduced pressure and store in -20 °C until analysis. Likewise, for the NMR analysis 300 µL of aliquot was transferred into 2 mL glass vials, complete dry under reduced pressure and store in -20 °C until analysis.

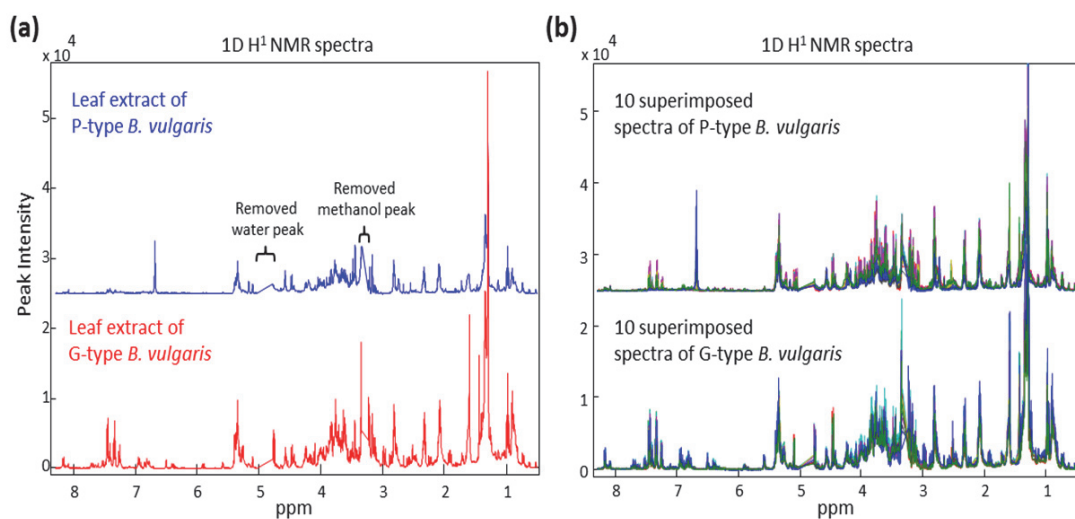
#### 4.1.2 1D $^1\text{H}$ NMR analysis of G and P type *Barbarea vulgaris* plant leaves

Metabolite extraction protocol (Fig.13), described above, has been used to extract 10 mg leaf discs of 10 replicate samples (harvested from 10 different leaves of the same age plants grown in the same climate chamber) from G and P-type plants. Moreover, three mixture samples were introduced, where half of the leaf disc was from G and the other half was from P type plants. 300 µL aliquot of obtained



#### 4. UNPUBLISHED STUDIES

metabolite extracts were completely dried in 1.5 ml glass vials and re-dissolved in 500  $\mu$ l methanol-d<sub>4</sub> (99.8%) containing 5  $\mu$ l of 20 mg ml<sup>-1</sup> TSP in deuterated water. 1D <sup>1</sup>H NMR spectra were recorded using a Bruker Avance DSX 500 NMR spectrometer (11.7 T) operating at 500.13 MHz, and equipped with a BBI probe for 5 mm (o.d.) sample tubes. Data acquisition for all the samples was automated from IconNMR automation software and each sample was automatically, tuned, matched and shimmed. A total of 512 number of scans were recorded at room temperature and the obtained spectra were referenced towards TSP (3-(Trimethylsilyl)-Propionic acid-d<sub>4</sub>) peak at 0.00 ppm. Raw NMR data of 23 samples were imported into Matlab after baseline and phase correction using TopSpin (version 13.1, Bruker BioSpin). The NMR spectra of P and G type of plants were similar, although significant differences were pronounced in the anomeric and aromatic regions (Fig.14).



**Figure 14.** 1D <sup>1</sup>H NMR spectra of the leaf extracts from G and P type *B. vulgaris* plants, based on the metabolomic protocol illustrated in figure 13. **(a)** NMR spectrum of P- and G- type plants. **(b)** Superimposed NMR spectra of 10 P- and G- type plants.

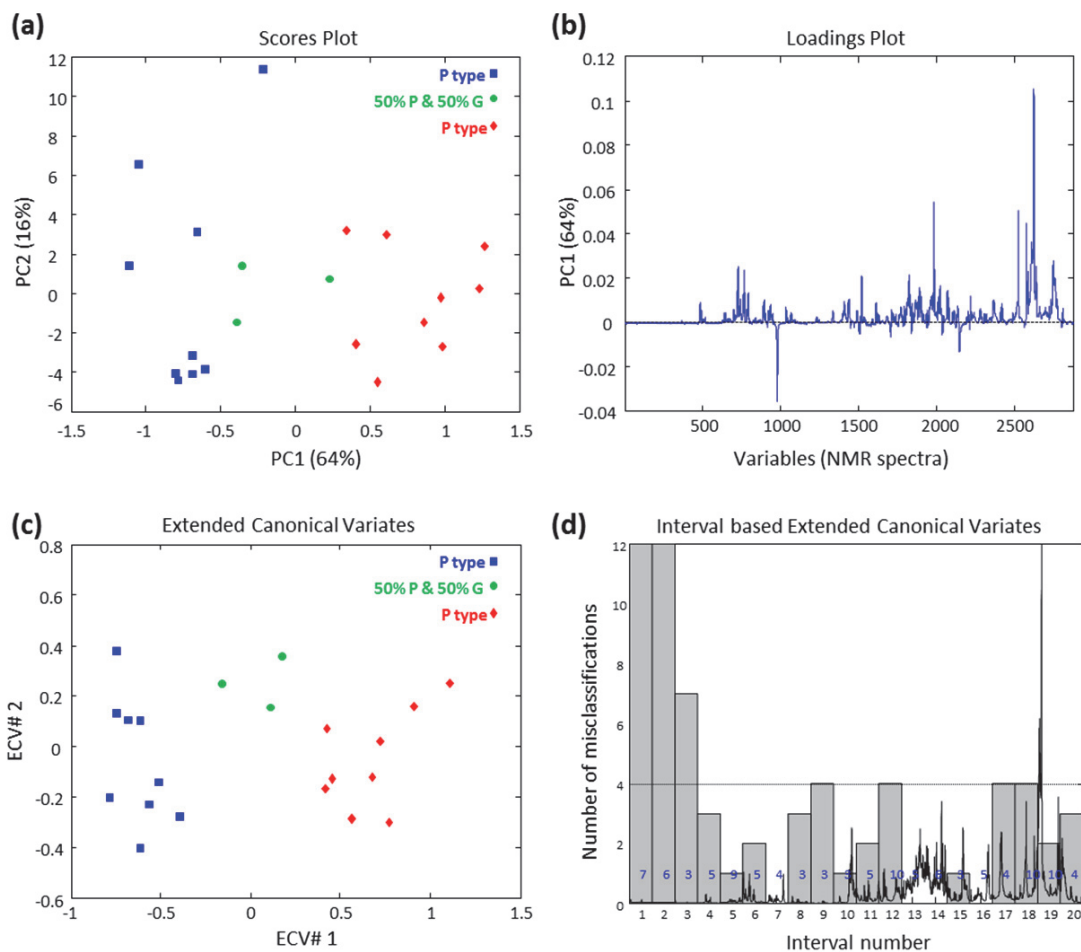
Comparison of NMR spectra of pure hederagenin and oleanolic acid mono-glycosides (see paper 5) with NMR spectra of G and P type plants' complex extracts did not show significant signals of these saponins. This is primarily due to the low concentration of saponins. As mentioned earlier, the sensitivity of NMR is inferior to GC-MS and LC-MS, however it provides reproducible detection of the most abundant e.g. first 50 metabolites of the investigated complex mixtures. Since, the concentration of saponins are significantly lower than the amount of other proton containing metabolites present in

#### 4. UNPUBLISHED STUDIES

the 85% methanol extract of the leaf discs, their peaks are not observed in the spectra and/or greatly masked by the peaks of more abundant metabolites.

Prior to multivariate data analysis, obtained NMR data was preprocessed to minimize non-sample related variations by removing non-informative regions of the spectra, and aligning the chemical shifts changes of the same metabolites over different samples. First, the data was aligned by using iCoshift (Savorani et al., 2010c) where all samples were aligned towards the median spectra. Then, non-informative regions of the spectra, including residual water and methanol peaks were removed. Since the investigated plant leaf discs are thought to have the same weight and the metabolite extraction protocol was identical for all samples, ideally, there is no need for normalizing this kind of data. However, due to the minute differences in sample weight and some experimental variations e.g. extraction, introduction of non-sample related variations is unavoidable. Therefore, we have attempted to apply four different normalization techniques and evaluated their effects by observing PCA based separation of G and P type plants and compared the results with the non-normalized data. These included: the probabilistic quotient normalization method (Dieterle et al., 2006), ref norm method by using the area of the triplet peak in region 2.79 – 2.84 ppm, the sum of the absolute values of all variables for the given sample (1-Norm) and the sum of the squared value of all variables for the given sample (2-Norm). Obtained results suggest that normalization of the data by the area of the triplet peak between 2.79-2.84 ppm was best for separating G type plants from P type (Fig.15 (a)). Separation of G and P type plants was due to the PC1 that captured 64% of the total variance present in the data, which suggest significant difference between the two aspects of plants. The loading plot of this PCA model demonstrates a broad range of variables (NMR peaks) that are responsible for the separation (Fig.15 (b)). In addition, ECVA based classification of G and P type *B. vulgaris* plants revealed slightly better separation than in PCA (Figure 15 (c)). Whereas interval based ECVA modeling enabled identification of the most informative NMR regions for the separation (Fig.15 (d)). iECVA modeling was performed by dividing the NMR data into 20 intervals with the same size and developing individual ECVA models for each interval. This approach showed that some regions of the NMR spectra were more distinctive for the two different plants than other regions and allowed separation of three classes (class 1: G type, class 2: P type and class 3: mixture of G and P) with a zero misclassification.

#### 4. UNPUBLISHED STUDIES



**Figure 15.** (a) Scores plot of the PCA analysis developed on 1D  $^1\text{H}$  NMR data demonstrated in figure 14 (b). Preprocessing of the data included, alignment of the slight chemical shift changes using iCoshift and sample wise normalization of the data to minimize the non-sample-related variation. (b) Loadings of PC1 that is responsible for the separation. (c) Extended canonical variates of the global ECVA model of the same NMR data. (d) Interval based ECVA results where data was divided into 20 intervals and ECVA models were computed for each interval separately. Dotted line is number of misclassifications (4 for 8 LV's) for global model and italic numbers (blue) are optimal LVs in interval model.

These findings suggest that, despite the differences present in the low concentration secondary metabolites of the G and P plants e.g. saponins, there are other obvious differences in more abundant metabolites of these two different phenotypes of *B. vulgaris*. Thus, metabolite extraction protocol (Fig.13), which is developed by using fractional factorial design, can be used for NMR based metabolomic fingerprinting of the F2 population. Therefore, the NMR metabolomic approach can be applied to capture the metabolomic differences between the F2 plants varying by their resistance to

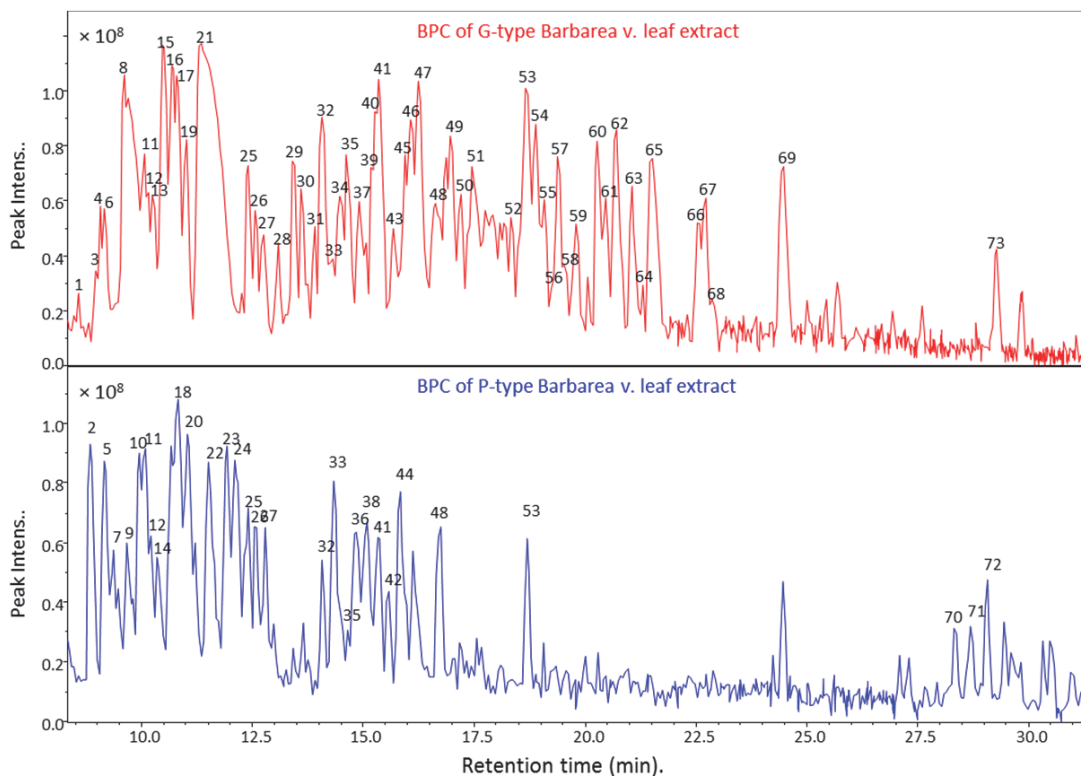
the flea beetle larvae (*Phyllotreta nemorum*). However, 1D  $^1\text{H}$  NMR spectra of the saponin-enriched fraction of the G and P type plant extracts obtained by SPE based fractionation (see section 4.3) demonstrated significantly improved NMR peaks of saponins, like hederagenin and oleanolic acid cellobioside. This was especially more pronounced in the aliphatic (singlet peaks of C24, C26, C27 and C30 between 0.7-1.16 ppm) and anomeric regions ( $\text{H1}'$  and  $\text{H1}''$  of hexoses between 4.2-4.5 ppm).

### 4.1.3 Tandem LC-MS analysis of G and P type *Barbarea vulgaris* plant leaves

As described in the NMR analysis of the parental P and G type *B. vulgaris* plants, 10 biological replicates from each type were harvested and metabolites were extracted according to the protocol described in figure 13. 200  $\mu\text{l}$  aliquot of the resulted metabolite extract in 85% methanol was completely dried under reduced pressure, at 30°C, and re-dissolved in 50  $\mu\text{l}$  of 50% methanol. Obtained extracts were used for LC-MS/MS analysis. LC-MS consisted of a Agilent 1100 Series LC (Agilent Technologies), equipped with a Gemini NX column (35°C; 2.0  $\times$  150 mm, 3.5 mm; Phenomenex) and coupled to a Bruker HCTUltra ion-trap mass spectrometer (Bruker Daltonics). Mobile phases were eluent A, water with 0.1% (v/v) formic acid, and eluent B, acetonitrile with 0.1% (v/v) formic acid. The gradient program was as follows: 0 to 1 min, isocratic 12% B; 1 to 33 min, linear gradient 12% to 80% B; 33 to 35 min, linear gradient 80% to 99% B; 35 to 38 min, isocratic 99% B; 38 to 45 min, isocratic 12% B at a constant flow rate of 0.2  $\text{mL min}^{-1}$ . The detector was operated in negative mode and included tandem mass spectrometry at two stages MS/MS and three stages MS/MS/MS. LC-MS profiles of G and P type plants were different, where G plant contained more peaks than P type (Fig.16) and based on the fragmentation patterns of the resolved peaks from both types of the plants, the majority of the metabolites were glycosides such as saponins. All six previously identified saponins (Figure 10) were present in the G type LC-MS profiles, while they were not detected in the P type plant (Table 4). Molecular masses of aglycones [Aglycone-H]<sup>-</sup>, oleanolic acid (455.3), 4-epihederagenin (471.3), gypsogenin (469.3), hederagenin (471.7) and cochalic acid (471.5) were detected in MS/MS and MS/MS/MS fragmentation patterns of the corresponding saponins eluted at retention times 24.5, 22.9, 22.5, 21.5 and 20.7 min, respectively. Apart from the previously known triterpenes of the *B. vulgaris*, LC-MS/MS data allowed tentative characterization of several other triterpenoid saponin-like metabolites of P and G type plants. This included eight new molecular masses (441.3, 443.3, 447.1, 457.5, 472.5, 473.3, 485.3, 487.5) that might possibly be derived from the saponins of *B. vulgaris* with triterpenoid backbones. Among these molecular masses, 457.5 match with the molecular mass of the triterpenoid saponin soyasapogenol B, 473.3 soyasapogenol A and 487.5 match with the mass of bayogenin. Although to date, there are no reports in the literature on *B. vulgaris* saponins with such aglycones. However, metabolite profiling of triterpenoid saponins of *Medicago truncatula* based on

#### 4. UNPUBLISHED STUDIES

accurate mass measurement by using UPLC ESI FT-ICR MS demonstrated tentative identification of 79 different triterpenoid saponins from hairy roots (Pollier et al., 2011). Tandem mass spectral data suggested that all these saponins were derived from ten different triterpenoid backbones. These triterpenoids included, 455 (soyasapogenol E), 457 (soyasapogenol B), 469 (tentatively identified as the hederagenin with the hydroxyl group at 28 position being oxidized to aldehyde group), 471 (hederagenin), 473 (soyasapogenol A), 485 (tentatively identified as the bayogenin with the hydroxyl group at 28 position being oxidized to aldehyde group) and 487 (bayogenin).



**Figure 16.** Base Peak Chromatogram (BPC) of the 85% methanol extracts of G and P type *B. vulgaris* leaves. Metabolites are numbered according to their order in Table 4.

Thus, it can be hypothesized that *B. vulgaris* possess various triterpenes that are mainly derived from  $\beta$ -amyrin. Tentative characterization of P and G type saponins derived from the all above mentioned 12 different aglycones were based on MS/MS data that showed characteristic fragmentation patterns of glycosides with loss of hexose (162), pentose (132) and methyl-pentose (146) (Table 4). It is worth mentioning that the intensities of P type metabolites eluting at the earlier retention times were

#### 4. UNPUBLISHED STUDIES

comparable to the intensities of the G type metabolites. The fragmentation patterns suggested that these glycosides are saponins with three and more sugar moieties. However, at the later elution times, the P type profile is rather simpler than the profile of G type plant. This suggests that P type plants contain less semi-polar metabolites than the G type does. Thus, semi-polar saponins with two and one sugar moieties were present in significantly low amounts in P plant than in G plant.

No	RT (min)	MS (m/z)	MS2 (m/z)	MS3 (m/z)
1.g	8.6	887.4	741.3 (MS-146)	561.1 (MS2-18-162)
2.p	8.8	755.4	609.3 (MS-146)	285 (MS2-324)
3.g	9.0	1241.5	933.4 (MS-162-146)	591.2 (MS2-18-324)
4.g	9.1	1079.4	933.5 (MS-162)	591.2 (MS2-18-324)
5.p	9.1	755.4	609.3 (MS-146)	285 (MS2-324)
6.g	9.2	1095.4	787.5 (MS-162-146)	607.1 (MS2-18-162)
7.p	9.4	933.5	625.3 (MS-162-162)	300.9 (MS2-324)
<b>8.g</b>	<b>9.6</b>	<b>1079.6</b>	<b>771.5 (MS-146-162)</b>	<b>447.1 (MS2-324)</b>
<b>9.p</b>	<b>9.7</b>	<b>593.3</b>	<b>447.1 (MS-146)</b>	<b>447.1</b>
<b>10.p</b>	<b>9.9</b>	<b>917.5</b>	<b>771.3 (MS-146)</b>	<b>485.3 (MS2-124-162)</b>
<b>11.b</b>	<b>10.05</b>	<b>1079.5</b>	<b>593.3 (MS-324-162)</b>	<b>447.1 (MS2-146)</b>
12.b	10.15	1079.4	933.5 (MS-162)	591.2 (MS2-18-324)
13.g	10.25	1109.4	787.4 (MS-322)	607.1 (MS2-18-162)
<b>14.p</b>	<b>10.4</b>	<b>917.5</b>	<b>771.3 (MS-146)</b>	<b>485.3 (MS2-124-162)</b>
15.g	10.5	1063.8	917.5 (MS-146)	771.5 (MS2-146)
16.g	10.7	1093.7	947.6 (MS-146)	771.2 (MS2-14-162)
17.g	10.8	1063.8	917.5 (MS-146)	771.5 (MS2-146)
18.p	10.8	931.5	785.4 (MS-146)	609.3 (MS2-176)
19.g	11.0	1093.7	947.6 (MS-146)	771.5 (MS2-14-162)
20.p	11.0	917.5	771.4 (MS-146)	609.4 (MS2-162)
21.g	11.3	868.2	422.1 (MS-446)	422.1
22.p	11.5	917.5	771.4 (MS-146)	609.4 (MS2-162)
23.p	11.9	901.6	755.4 (MS-146)	609.4 (MS2-146)
24.p	12.1	931.5	785.4 (MS-146)	609.4 (MS2-176)
<b>25.b</b>	<b>12.4</b>	<b>901.5</b>	<b>755.3 (MS-146)</b>	<b>447.2 (MS2-146-162)</b>
26.b	12.5	931.4	785.4 (MS-146)	609.1 (MS2-14-162)
27.b	12.7	931.4	785.4 (MS-146)	609.1 (MS2-14-162)
<b>28.g</b>	<b>13.1</b>	<b>959.7</b>	<b>797.6 (MS-162)</b>	<b>473.4 (MS2-162-162)</b>
<b>29.g</b>	<b>13.4</b>	<b>1135.8</b>	<b>811.8 (MS-324)</b>	<b>487.5 (MS2-162-162)</b>

4. UNPUBLISHED STUDIES

<b>30.g</b>	<b>13.6</b>	<b>974.7</b>	<b>767.8 (MS-45-162)</b>	<b>443.3 (MS2-162-162)</b>
31.g	13.9	857.7	811.5 (MS-46)	605.4 (MS2-44-162)
<b>32.b</b>	<b>14.1</b>	<b>1087.9</b>	<b>879.6 (MS-46-162)</b>	<b>487.5 (MS2-68-324)</b>
<b>33.b</b>	<b>14.3</b>	<b>959.7</b>	<b>797.6 (MS-162)</b>	<b>473.4 (MS2-162-162)</b>
34.g	14.5	697.5	633.3 (MS-64)	327.0 (MS2-306)
<b>35.b</b>	<b>14.6</b>	<b>1059.9</b>	<b>811.8 (MS-248)</b>	<b>487.5 (MS2-162-162)</b>
<b>36.p</b>	<b>14.8</b>	<b>1120.7</b>	<b>795.8 (MS-324)</b>	<b>472.4 (MS2-162-162)</b>
37.g	14.9	757.5	693.2 (MS-64)	357.0 (MS2-336)
<b>38.p</b>	<b>15.1</b>	<b>1002.9</b>	<b>959.8 (MS-44)</b>	<b>473.4 (MS2-324-162)</b>
<b>39.g</b>	<b>15.2</b>	<b>959.7</b>	<b>797.6 (MS-162)</b>	<b>473.4 (MS2-162-162)</b>
40.g	15.25	1071.8	863.5 (MS-46-162)	795.5 (MS2-68)
<b>41.b</b>	<b>15.35</b>	<b>943.1</b>	<b>619.4 (MS-162-162)</b>	<b>457.4 (MS2-162)</b>
<b>42.p</b>	<b>15.6</b>	<b>959.7</b>	<b>797.6 (MS-162)</b>	<b>473.4 (MS2-162-162)</b>
<b>43.g</b>	<b>15.7</b>	<b>1234.6</b>	<b>811.8 (MS-292-132)</b>	<b>487.5 (MS2-162-162)</b>
<b>44.p</b>	<b>15.8</b>	<b>1002.9</b>	<b>959.8 (MS-44)</b>	<b>473.4 (MS2-324-162)</b>
<b>45.g</b>	<b>15.9</b>	<b>1044.0</b>	<b>796.0 (MS-248)</b>	<b>471.4 (MS2-162-162)</b>
<b>46.g</b>	<b>16.1</b>	<b>943.1</b>	<b>619.4 (MS-162-162)</b>	<b>457.4 (MS2-162)</b>
<b>47.g</b>	<b>16.2</b>	<b>943.1</b>	<b>619.4 (MS-162-162)</b>	<b>457.4 (MS2-162)</b>
<b>48.b</b>	<b>16.6</b>	<b>943.1</b>	<b>619.4 (MS-162-162)</b>	<b>457.4 (MS2-162)</b>
<b>49.g</b>	<b>17.0</b>	<b>943.1</b>	<b>619.4 (MS-162-162)</b>	<b>457.4 (MS2-162)</b>
<b>50.g</b>	<b>17.2</b>	<b>943.1</b>	<b>619.4 (MS-162-162)</b>	<b>457.4 (MS2-162)</b>
51.g	17.4	829.8	783.6 (MS-46)	621.4 (MS2-162)
<b>52.g</b>	<b>18.3</b>	<b>943.1</b>	<b>619.4 (MS-162-162)</b>	<b>457.4 (MS2-162)</b>
<b>53.b</b>	<b>18.6</b>	<b>811.8</b>	<b>649.5 (MS2-162)</b>	<b>487.5 (MS2-162)</b>
<b>54.g</b>	<b>18.9</b>	<b>811.8</b>	<b>649.5 (MS2-162)</b>	<b>487.5 (MS2-162)</b>
<b>55.g</b>	<b>19.1</b>	<b>809.7</b>	<b>647.5 (MS-162)</b>	<b>485.4 (MS2-162)</b>
<b>56.g</b>	<b>19.2</b>	<b>811.8</b>	<b>649.5 (MS2-162)</b>	<b>487.5 (MS2-162)</b>
<b>57.g</b>	<b>19.4</b>	<b>809.7</b>	<b>647.5 (MS-162)</b>	<b>485.4 (MS2-162)</b>
<b>58.g</b>	<b>19.6</b>	<b>1027.7</b>	<b>779.6 (MS-248)</b>	<b>455.3 (MS2-162-162)</b>
<b>59.g</b>	<b>19.8</b>	<b>649.7</b>	<b>487.3 (MS-162)</b>	<b>487.3</b>
<b>60.g</b>	<b>20.2</b>	<b>795.7</b>	<b>633.5 (MS-162)</b>	<b>471.5 (MS2-162)</b>
<b>61.g</b>	<b>20.5</b>	<b>795.7</b>	<b>633.5 (MS-162)</b>	<b>471.5 (MS2-162)</b>
<b>62.g</b>	<b>20.7</b>	<b>795.7</b>	<b>633.5 (MS-162)</b>	<b>471.5 (MS2-162)</b>
<b>63.g</b>	<b>21.0</b>	<b>827.9</b>	<b>781.8 (MS-46)</b>	<b>457.2 (MS2-162-162)</b>
<b>64.g</b>	<b>21.3</b>	<b>795.7</b>	<b>633.5 (MS-162)</b>	<b>471.5 (MS2-162)</b>
<b>65.g</b>	<b>21.5</b>	<b>841.8</b>	<b>795.7 (MS-46)</b>	<b>471.7 (MS2-162-162)</b>
<b>66.g</b>	<b>22.5</b>	<b>839.9</b>	<b>793.5 (MS-46)</b>	<b>469.3 (MS2-162-162)</b>

#### 4. UNPUBLISHED STUDIES

<b>67.g</b>	<b>22.7</b>	<b>825.7</b>	<b>779.6 (MS-46)</b>	<b>455.3 (MS2-162-162)</b>
<b>68.g</b>	<b>22.9</b>	<b>679.7</b>	<b>633.5 (MS-46)</b>	<b>471.3 (MS2-162)</b>
<b>69.g</b>	<b>24.5</b>	<b>825.7</b>	<b>779.6 (MS-46)</b>	<b>455.3 (MS2-162-162)</b>
70.p	28.3	885.7	867.5 (MS-18)	657.4 (MS2-210)
71.p	28.7	885.7	867.5 (MS-18)	657.4 (MS2-210)
<b>72.p</b>	<b>29.1</b>	<b>695.7</b>	<b>677.4 (MS-18)</b>	<b>441.4 (MS2-236)</b>
73.g	29.2	809.8	763.5 (MS-46)	601 (MS2-162)

Table 4. LC-MS/MS based fragmentation patterns of major metabolites detected from the G-type (g), P-type (p) and in both types (b) of *Barbarea v.* 85% methanol extract. Metabolite number 62 represent cochalic acid cell., 65 hederagenin cell., 66 gypsogenin cell., 68 4-epihederagenin cell., and 69 oleanolic acid cell. \* Loss of 162 correspond to cleavage of hexose, 146 methyl-pentose, 324 double hexose, 292 double methyl-pentose, 132 pentose, 176 hexuronic acids, 18 water, 46 formic acid, 44 carbon dioxide, while the loss of m/z ion 14, 64, 68, 248, 276, 306 and 336 remain unknown. Possible saponins with triterpenoid backbones are highlighted in bold.

## 4.2 TMSCN based derivatization and GC-MS detection of triterpenes produced by combinatorial biochemistry in tobacco leaves

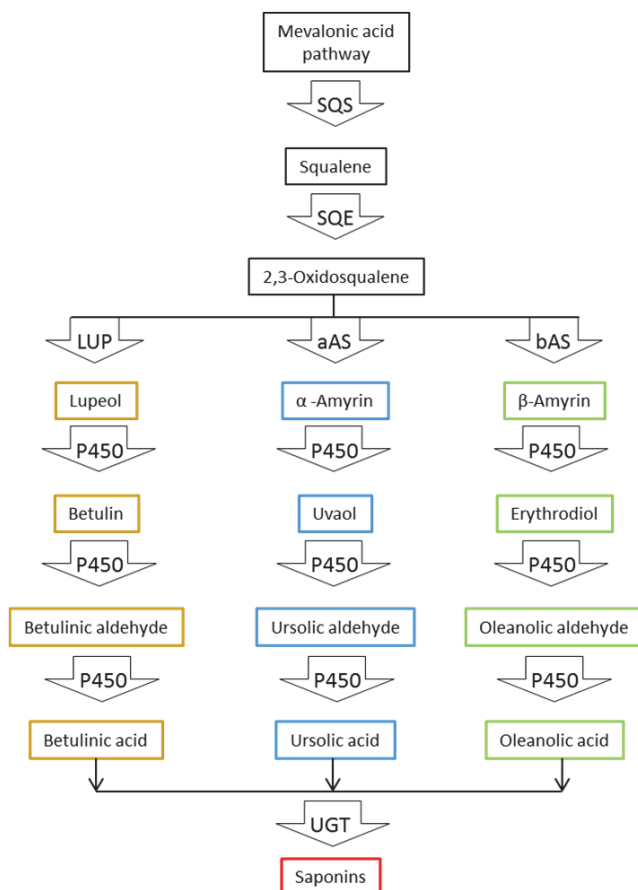
This section demonstrates a part of the GC-MS results of the unpublished work conducted on saponin biosynthetic pathway elucidation in *B. vulgaris*. The first committed step in saponin biosynthesis is the 2,3-oxidosqualene, which is a major product of the mevalonic acid pathways (Augustin et al., 2011). 2,3-oxidosqualene is the substrate for several enzymes that belong to the oxidosqualene cyclase (OSCs) family and all these enzymes use it in different ways to cyclase triterpenes and sterols. Depending on the plant species, the ratios between these cyclization products are different (Vincken et al., 2007). Some plants e.g. *B. vulgaris* tend to produce more triterpenes and use them for defense against biotic stresses. There are mainly three kinds of triterpene backbones that are used for synthesis of triterpenoid saponins. These are  $\beta$ -amyrin,  $\alpha$ -amyrin and lupeol that are the products of  $\beta$ -amyrin synthase (bAS),  $\alpha$ -amyrin synthase (aAS) and lupeol synthase (LUP), respectively (Fukushima et al., 2011) (Fig.17). Then, these triterpenoid backbones undergo various modifications e.g. oxidation that lead to introduction of hydroxyl, ketone, aldehyde and carboxylic acid functional groups, mediated by cytochrome P450-dependent monooxygenases (P450s) (Schuler and Werck-Reichhart, 2003). These result in production of erythrodiol, oleanolic aldehyde, followed by oleanolic acid from  $\beta$ -amyrin and uvaol, ursolic aldehyde, followed by ursolic acid from  $\alpha$ -amyrin and betulin, betulinic aldehyde, followed by betulinic acid derived from lupeol. Recent studies demonstrated the roles of some members of the cytochrome P450 enzymes in biosynthesis of triterpenoid saponins in plants (Fukushima et al., 2011; Geisler et al., 2013; Hamberger and Bak, 2013). Then, oxidation products of the triterpenes undergo further decorations by uridine diphosphate (UDP)-glycosyltransferases (UGTs) that add sugar moieties at the different positions of the aglycone (in different extend e.g. there are



#### 4. UNPUBLISHED STUDIES

saponins with two, three, four and more sugars) and result in diverse saponin pools (Augustin et al., 2012).

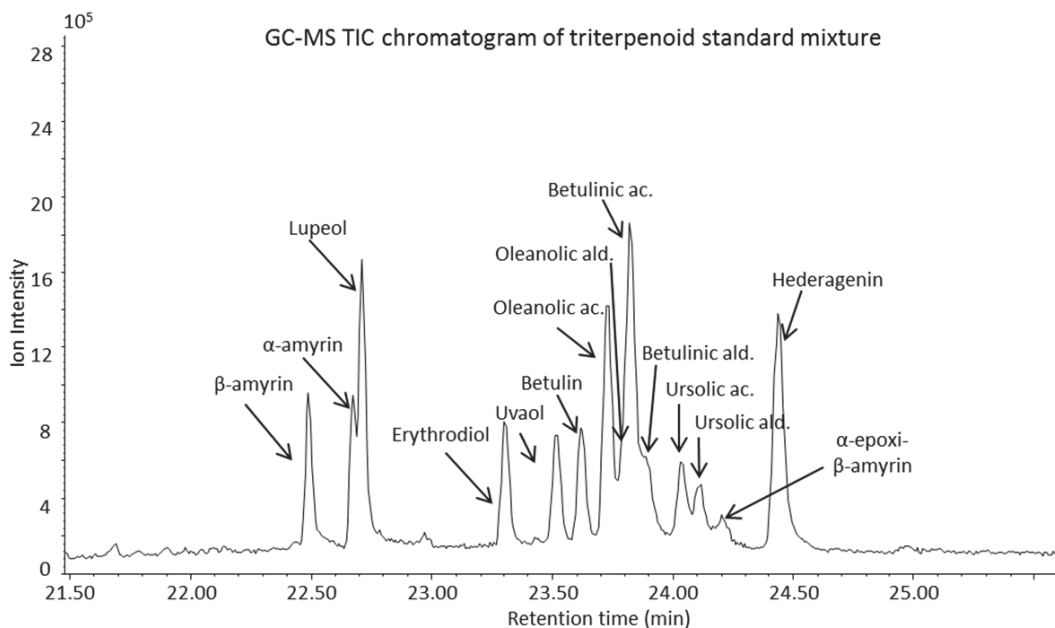
This study dealt with the functional characterization of *B. vulgaris* genes that are believed to be involved in saponin biosynthesis. A total of eight *B. vulgaris* genes (6 OSCs and 2 P450s) were investigated by transient expression in *Nicotiana benthamiana* leaves by using a transient plant expression system (CPMV-HT) based on cowpea mosaic virus (CPMV) (Sainsbury et al., 2009). The oxidosqualene cyclases of *B. vulgaris* were BvLUP2, BvLUP5 and PEN1 from G and P type plants, while the P450s were BvCYP716A from G and P type plants. Young leaves of tobacco plants were infiltrated with *A. tumefaciens* containing CPMV-HT constructs developed by using OSCs either alone or in combination with P450s. The resulted triterpenoids produced in tobacco leaves were evaluated by GC-MS analysis of plant leaf extracts.



**Figure 17.** Simplified overview of the biosynthetic pathways of triterpenoid saponins. SQS: squalene synthase, SQE: squalene epoxidase, LUP: lupeol synthase, aAS:  $\alpha$ -Amyrin synthase, bAS:  $\beta$ -Amyrin synthase, P450: cytochrome P450-dependent monooxygenases, UGT: uridine diphosphate (UDP)-glycosyltransferases.

#### 4. UNPUBLISHED STUDIES

A targeted GC-MS metabolomic protocol for detection of triterpenes was developed based on trimethylsilylation of plant crude extracts using novel derivatization reagent, trimethylsilyl cyanide (TMSCN) (Khakimov et al., 2013). Derivatization time, temperature, reagent amount and injection method were optimized toward increasing s/n ratios of the triterpene TMS-derivatives peaks. The metabolomic protocol was evaluated by GC-MS analysis of the standard mixture containing 14 different triterpenoid aglycones, at which some were expected to occur in the tobacco leaves infiltrated with the CPMV-HT constructed (Figure 18).

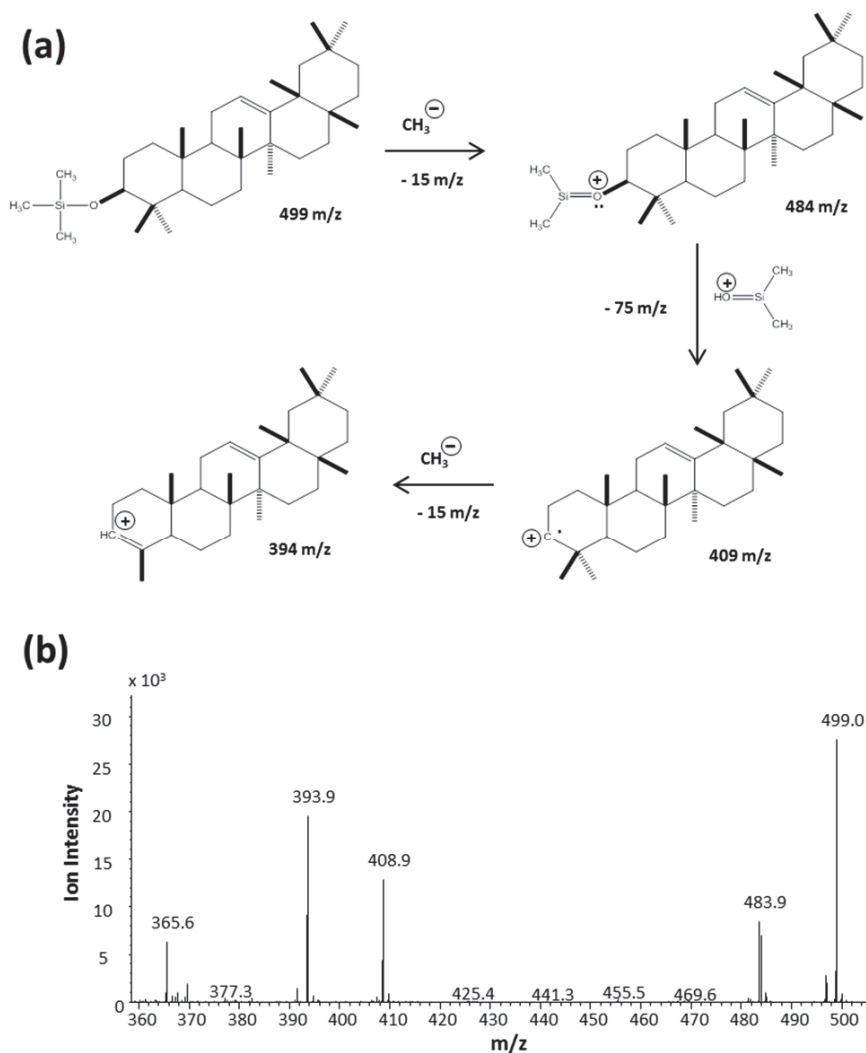


**Figure 18.** GC-MS profile of the triterpenoid standard mixture containing 14 different aglycones. 50 mL aliquot of the standard mixture solution where each aglycone was dissolved in methanol at the final concentration of  $0.25 \text{ mg mL}^{-1}$  was completely dried under reduced pressure and trimethylsilylated using TMSCN reagent. More detailed GC-MS protocol is given in section 4.2. The peaks of all triterpenoids correspond to their trimethylsilylated derivatives.

Fingerprint electron ionization mass spectra (EI-MS) patterns of all the triterpenes standard obtained from the GC-MS analysis were added in a database developed in-house. EI-MS fragmentation patterns of these triterpenes possess a general pattern, which was characteristic for the trimethylsilylated derivatives (Watson and Sparkman, 2007a). These general fragmentation patterns showed loss of a methyl ( $15 \text{ m/z}$ ) functional group followed by loss of dimethylsilyl oxonium ( $75 \text{ m/z}$ ) and the second methyl group. Figure 19 demonstrates an example of such a fragmentation pattern for trimethylsilylated  $\beta$ -Amyrin. This pattern was common for all the investigated triterpene standards,

#### 4. UNPUBLISHED STUDIES

and allowed estimation of the molecular masses of the aglycones. Although the ratios of these characteristic  $m/z$  ions generated before or after the loss of methyl and dimethylsilyl oxonium fragments were different for different triterpenes. Thus, this laid a foundation for the estimation of the molecular masses of the new triterpenoids.



**Figure 19. (a)** Common EI-MS fragmentation pattern observed for all trimethylsilylated triterpenes in the example of  $\beta$ -amyrin. These characteristic patterns are represented by loss of methyl ( $-15 m/z$ ), followed by loss of dimethylsilyl oxonium ( $-75 m/z$ ) and loss of the second methyl ( $-15 m/z$ ) group from the final silylated products of the triterpenes. **(b)** EI-MS of  $\beta$ -amyrin, in range of 360-500  $m/z$ , that shows characteristic pattern.

#### 4. UNPUBLISHED STUDIES

Metabolite extraction protocol from the tobacco leaves was as follows: single leaf (with a fresh weight of 1.5 g) upon 6 days post-infiltration was ground to a fine powder under liquid N<sub>2</sub> and metabolites were extracted by using 1.5 mL ethyl acetate (EtAc) and vortexed at room temperature for 30 min. Then, 200 µl aliquot of EtAc extract was completely dried inside the GC-MS glass insert under reduced pressure, at 40°C and tightly sealed with golden magnetic GC-MS vial lids with a silicone septum. All the samples were derivatized and analyzed in GC-MS in a single sequence automated by using GERSTEL MultiPurpose Sampler (MPS) with DualRait WorkStation integrated to a GC-MS system from Agilent. Samples were trimethylsilylated at a constant derivatization time and analyzed in a random order. Each sample was trimethylsilylated by addition of 50 µl pure TMSCN reagent using a 100 µl syringe installed in the autosampler and incubated at 40°C for 40 min by shaking at 750 rpm by using the agitator of the autosampler. After derivatization, 1 µl aliquot of the derivatized sample (note: prior test showed that 200 µl of dried plant extract was fully soluble in 50 µl pure TMSCN) was injected in splitless mode at the splitless time of 3 min (purge flow to split vent, septum purge flow and column flow were 10, 3 and 1.7 mL min<sup>-1</sup>, respectively) into the GC-MS cooled injection system (CIS) by using a 10 µL syringe of the autosampler. Injection parameters and considerations regarding TMSCN based derivatization can be found in the Supporting Information A of the paper 2 (Khakimov et al., 2013). The GC-MS consisted of an Agilent 7890A GC and an Agilent 5975C series MSD. GC separation was performed on an Agilent HP-5MS column (30 m x 250 µm x 0.25 µm) by using hydrogen as a carrier gas. The GC oven temperature program was as follows: initial temperature 40°C, heating rate 12.0°C min<sup>-1</sup>, end temperature 310°C, hold time 8.0 min and post run time 5 min at 40°C. Mass spectra were recorded in the range of 50-700 *m/z* with a scanning frequency of 3.2 scans s<sup>-1</sup>, and the MS detector was switched off during the 20 min solvent delay time, since the analysis was mainly targeted to capture the triterpenoids that elute later than 21 min in this method. The transfer line, ion source and quadrupole temperatures were set to 280, 230 and 150°C, respectively. The mass spectrometer was tuned according to the manufacturer recommendations by using perfluorotributylamine (PFTBA). The obtained GC-MS chromatographic data was analyzed using Agilent Technologies' ChemStation software (version: E.02.02.1431) and DataAnalysis software (version 4.0) from Bruker Daltonics.

The GC-MS results demonstrated formation of novel products for all the constructs transformed in *A. tumefaciens* and infiltrated into *N. benthamiana* leaves for transient expression, when OSCs (BvLUP2, BvLUP5 and PEN1) of *B. vulgaris* were expressed alone or in combination with P450s (BvCYP716A) (Figure 20). Infiltration of tobacco leaves with G or P type BvLUP2, revealed formation of comparable amounts of lupeol. Infiltration of tobacco leaves with P and G type BvLUP2 in combination with P450 (BvCYP716A) resulted in significant reduction of the lupeol peak and formation of the lupeol oxidation product, betulinic acids and two other unknown peaks (**unk1** at RT 23.64 and **unk2** at RT 24.18 min). It is worth mentioning that the abundances of the two unknown peaks were higher than the abundance of the produced betulinic acid. Although the RT of the **unk1** was the same as the RT of the betulin, which is the intermediate products form during the oxidation of lupeol to betulinic acid, its EI-MS fragmentation pattern was different. Likewise, RT of the **unk2** matched with the RT of α-epoxi-β-

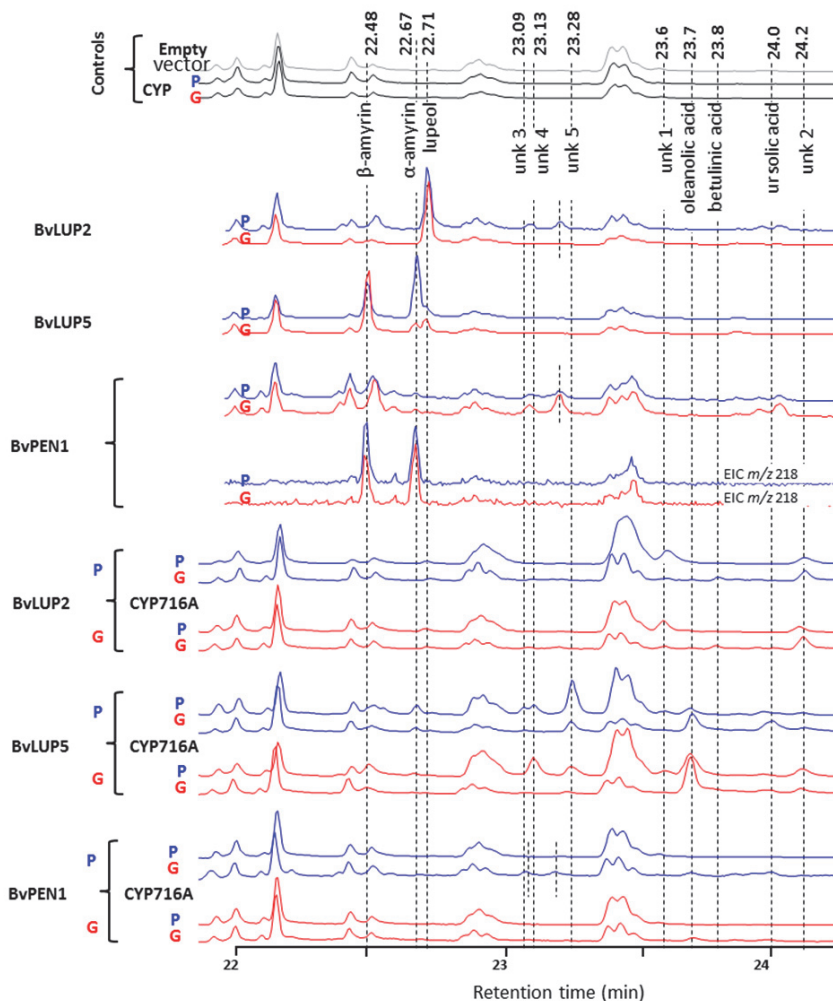
#### 4. UNPUBLISHED STUDIES

amyrin, though their mass spectra were different. Thus, these two unknown metabolites peaks might originate from other products of lupeol oxidation catalyzed by endogenous *N. benthamiana* P450s. None of these compounds were detected in control leaves (infiltration buffer, empty-vector-infiltrated plants) or leaves that were infiltrated with the P450 (CYP716A) construct alone and plants showed no changes in the triterpene profile compared to control plants.

Infiltration of tobacco leaves with OSCs, BvLUP5 from P and G type plants resulted in formation of three different triterpenoid backbones,  $\beta$ -amyrin,  $\alpha$ -amyrin and lupeol. In the case of BvLUP5 from G type *B. vulgaris* the major product was  $\beta$ -amyrin, while peaks of  $\alpha$ -amyrin and lupeol were significantly lower than the  $\beta$ -amyrin peak. In contrast to this, the BvLUP5 from P type showed comparable amounts of  $\beta$ -amyrin and  $\alpha$ -amyrin and the peak of lupeol was significantly lower. Formation of three different triterpene backbones suggest greater diversity of BvLUP5 gene compared to the BvLUP2 gene. Infiltration of BvLUP5 genes from G and P type plants together with the cytochrome P450 (CYP716A) from G and P type plants revealed further modifications of the above mentioned cyclization products and resulted in formation of oleanolic, ursolic and betulinic acids that are the P450 based oxidation products of the  $\beta$ -amyrin,  $\alpha$ -amyrin and lupeol, respectively. Moreover, these constructs revealed two unknown peaks at retention times 23.64 (**unk1**) and 24.18 (**unk2**) min that were previously found in BvLUP2 based constructs and additionally, three more unknown peaks at retention times 23.09 (**unk3**), 23.13 (**unk4**) and 23.28 (**unk5**) min. Co-expression of BvCYP716A from the G type plant with BvLUP5 of the G type plant showed accumulation of mainly oleanolic acid, while the product range increased when the BvLUP5 (G type) was co-expressed with BvCYP716A from the P type plant (Fig. 20). Similarly, more novel products were detected when BvLUP5 (P type) was co-expressed with BvCYP716A (P-type). This indicates that BvCYP716A (P type) originated from the flea-beetle-susceptible plant is less specific than BvCYP716A (G type), which originates from the insect-resistant G-type plant. It is worth mentioning that unknown peaks 3-5 were highly expressed from the OSCs and P450s originated from the P type *B. vulgaris* plant than in G type plant. In the constructs developed from BvLUP5 (P type) and CYP716A (P type), both, **unk3** and **unk4** were present in comparable amounts and the highest abundance of **unk5** was observed. In contrast to this, replacement of P type CYP716A with the G type gene resulted in disappearance of **unk3** and **unk4** and reduction of **unk5** approximately two fold (Fig.20). Likewise, combinational constructs of BvLUP5 (G type) with CYP716A (G type) showed no peaks of these unknown metabolites, while the combination of BvLUP5 (G type) with CYP716A (P type) showed peaks of **unk4** and **unk5**. RTs of **unk5** (23.28 min) and erythrodiol (23.31 min), which is the intermediate product formed during the oxidation of  $\beta$ -amyrin to produce oleanolic acid, are very close to each other, though their EI-MS patterns significantly differ. Thus, **unk3**, **unk4** and **unk5** remain unknown P450 catalyzed oxidation products of triterpenes that might correspond to intermediate products that were not present in the standard mixture containing 14 triterpenes, towards formation of oleanolic, ursolic or betulinic acids or other unexpected metabolites. However, EI-MS patterns of **unk3** and **unk4** matched with the two unidentified metabolites observed in the similar study that dealt

4. UNPUBLISHED STUDIES

with characterization of P450s' role in the biosynthesis of triterpenoids in *Medicago truncatula* (Fukushima et al., 2011).



**Figure 20.** GC-MS profiles of *N. benthamiana* extracts infiltrated with six different OSCs or OSCs co-expressed with BvCYP716A-G and BvCYP716A-P. The positions of novel compounds are indicated with vertical lines, together with the names of the novel compounds when standards were available, and their fragmentation pattern and retention time were identical with those from the novel compounds. Empty vector construct pEAQ-HT, pEAQ-HT-CYP716A-G and pEAQ-HT-CYP716A-P were used as the control.

Tobacco leaves infiltrated with BvPEN1 (P type) and BvPEN1 (G type) also revealed cyclization of 2,3-oxidosqualene into  $\beta$ -amyrin and  $\alpha$ -amyrin, however the amounts of these cyclization product were much lower compared to the BvLUP2 and BvLUP5 catalyzed cyclization. Co-infiltration of BvPEN1 with CYP716A led to formation of the oxidation products of  $\beta$ -amyrin and  $\alpha$ -amyrin, oleanolic and ursolic acids, as expected. Consequently, the products were also produced in a lower amount, and accordingly peaks were better observed by extraction ion chromatography using characteristic m/z ions (Figure 20).

### 4.3 Structure elucidation of triterpenoid saponins of the insect resistant and susceptible *Barbarea vulgaris* plants

Two different phenotypes of the *B. vulgaris* plant, glabrous (G)-resistant and pubescent (P)-susceptible to the insect herbivore, differ morphologically, cytologically (P-type plants are covered by hairs, while G-type plants are not) and biochemically (glucosinolate and saponin profile of the plants are different) (Agerbirk et al., 2003b; Kuzina et al., 2009). Despite all these differences, variations of plants' resistance towards insects were found to correlated with their saponin content (Kuzina et al., 2009). Up to date, structures of six triterpenoid saponins have been elucidated from the G-type plants, out of which four of them depicted high correlation with the resistance against herbivore. This include hederagenin cellobioside (Shinoda et al., 2002), oleanolic acid cellobioside (Agerbirk et al., 2003a), gypsogenin cellobioside and 4-epihederagenin cellobioside (Nielsen et al., 2010). The role of the other two triterpenoid saponins of G-type plants, cochalic acid cellobioside (Nielsen et al., 2010) and the first lupine type saponin, 3-O-cellobiosyl-28-O- $\beta$ -D-glucopyranosyl-16- $\beta$ ,23-dihydroxybetulinic ester, found in *Brassicaceae* family (unpublished work by Kristensen et al), remain unknown. Structures of these saponins are demonstrated in figure 10. However, saponins present in the P-type plants, that are favorable food source for various insects including, *P. nemorum* and *P. xylostella*, remain unidentified.

LC-MS and LC-MS/MS data acquired on crude G- and P-type *B. vulgaris* plant extracts suggest a presence of a much broader range of saponins from both plant types (Augustin et al., 2012; Kuzina et al., 2009). In addition to this, PARAFAC2 based comprehensive analysis of LC-MS data recorded on parental G- and P-types plants and the segregating F2 population demonstrated five additional saponins, which also showed significant correlation with plant's resistance level against flea beetle larvae of *Phyllotreta nemorum* (Khakimov et al., 2012). However, their structures were tentatively characterized being saponins of G- and P-type plants with two and three sugar moieties. LC-MS/MS data of the G- and P-type *B. vulgaris* plants demonstrated in section 4.1 (Figure 16 and Table 4). These studies show peaks of several saponin-like metabolites with different masses of saponin backbones and different decoration of the sugar moieties. The results showed presence of more than ten different molecular masses of possible triterpenoids with slight alterations in the structures of the

#### 4. UNPUBLISHED STUDIES

backbones. This included aglycone molecular masses of already known triterpenoid saponins of the G-type *B. vulgaris* such as 472.7 (hederagenin, 4-epihederagenin, cochalic acid), 470.7 (gypsogenin), 456.7 (oleanolic acid), 474.7 (new lupine type saponin) as well as unknown aglycone masses at  $m/z$  442.4, 444.3, 448.1, 458.4, 472.5, 473.4, 486.3, 488.5. Metabolites with these aglycone masses were repeatedly eluted at the different retention times and mostly possessed characteristic fragmentation patterns that demonstrated the loss of hexoses (162) and methyl-pentoses (146). It is worth mentioning that the same molecular mass of aglycones were part of the different types of saponin-like metabolites with different levels of glycosylation. The observed fragmentation patterns mostly suggested presence of saponins with two and three sugar moieties, where saponins with three sugar moieties were in relatively comparable amounts in both types of plants, while the saponins with two sugar moieties were mainly present in G-type plants. This raises a question, whether the resistance of the G-type *B. vulgaris* is associated with saponins possessing two sugars only (as it was found previously) and the saponins with three sugars do not have influence on the resistance of the plants.

Thus, it is interesting to study the structures of the saponins of G- and P-type *B. vulgaris* plants as that may bring new knowledge into the plant-insect interactions, triterpenoid biosynthesis and Quantitative Structure-Activity Relationship (QSAR). One of the most important questions is how the saponins of G- and P-type plants differ structurally and how this is associated with their resistance and/or susceptibility towards insects or is it only the matter of concentration. It is also crucial to know, if the saponin backbone structure or the decoration of the sugar moieties or both have influence on toxicity of the saponins to the insects. Previously, similar study was performed with oleanolic acid mono-glycoside and hederagenin mono-glycoside produced *in vitro* and feeding assays were applied by using flea beetle (*Phyllotreta nemorum*) (Augustin et al., 2012). The study demonstrated higher feeding deterrent properties of hederagenin mono-glycoside than oleanolic acid mono-glycoside. However, our aim from this study is not to isolate individual saponins from G and P type plants and evaluate their activity by performing bioassay tests, but structurally characterize the most abundant saponins of the G- and P-type *B. vulgaris*.

A semi-targeted metabolomic protocol for extraction of saponins from the G- and P-type *B. vulgaris* plant leaves was developed using methanol extraction and enrichment of saponin fraction by using solid phase extraction (SPE) cartridges for purification. The metabolomic protocol was as follows: freeze dried leaves of the plants were soaked into the 55% ethanol in wt/vol of 1:10 and boiled in a water bath for 5 min, followed by 10 min of extraction in ultrasonic bath at 50°C. The obtained extracts were filtered, dried under reduced pressure and re-dissolved in 30% methanol in a vol/vol of 3:1, between initial extract and methanol. Then, 60 mL of the 30% methanol extract was run through a Strata C18 SPE cartridge (10g, 60 mL), which was pre-conditioned with 30% methanol in advance and fraction 1 (30% methanol fraction) was obtained. Then, the same cartridge was flushed with 60 mL of 90% methanol which resulted in fraction 2 (90% methanol fraction), and finally 60 mL of 100% methanol (fraction 3) recovered the most non-polar metabolites of the plant extract. Fractions 1-3



#### 4. UNPUBLISHED STUDIES

were up concentrated by drying and re-solubilizing in 50% methanol, in a vol/vol of 4:1 between the initial extracts and 50% methanol. Then, all fractions were analyzed by thin layer chromatography (TLC) and LC-MS/MS, which showed that fraction 2 (90% methanol) contained the majority of saponins present in the 55% ethanol extract of the *B. vulgaris* leaves. While, fractions 1 and 3 possessed significantly lower amounts of saponins, based on the intensities of metabolite peaks observed from the LC-MS/MS profiles. LC-MS/MS experiment performed in this study was identical to the protocol demonstrated in section 4.1. Moreover, the results of these LC-MS/MS experiments of saponin enriched fraction 2 were in agreement with the earlier LC-MS analysis performed on the G- and P-type plant leaves (section 4.1 and Table 4). Most of the saponins of both types of plants possessed 2-3 sugar moieties and G-type plants contained more peaks and with higher abundances. More detailed investigation of the LC-MS/MS profiles of the 90% methanol fraction (fraction 2) of both plants revealed more than 30 saponin-like metabolites, including the peaks of the already known six saponins from the G-type plant, and 20 metabolites from the P-type plant (data not shown). Thus, we have tentatively characterized the saponins of P-type *B. vulgaris*, in addition to the G-type plant, and further structure elucidation of saponins of both types of plants will be performed on fraction 2, by using hyphenated methods of analysis such as LC-NMR.

## 5 OUTREACH

### 5.1 Will plants save the planet and can plant metabolomics play a key role?

Plants are important in every aspect of life on planet earth and they contribute with an essential effect in maintaining the habitable environment on earth for human and animals. The life we live today is achieved mainly due to the scientific and technological developments. Plant science is an important part of modern science, which has been conducted for several centuries, and that today deals with several different topics e.g. organic food production, sustainable food and energy, pharmacology, medicine and the problems related to climate changes. One of the most powerful current tools of plant science is metabolomics. Analysis of plant metabolome facilitates an understanding of metabolome-environment, metabolome-gene, metabolome-health and metabolome-phenotype relationships. These, in turn allow us to gain new insight into plant cell regulations, how to utilize plants in an optimal way and harvest benefits. However, plant metabolomics is a relatively new area of plant science and its capabilities are still expanding by developments of analytical platforms, metabolomic protocols and data processing methods.

This PhD study was conducted with the aim of expanding the limits of plant metabolomics by development, improvement and implementation of the new metabolomics protocols. During this PhD project, seven separate works were conducted that are published either submitted for publication or not yet submitted. The first study (Paper 1) implemented a new method for processing of raw LC-MS metabolomics data by using a multi-way decomposition method PARAFAC2. This is the first application of PARAFAC2 on LC-MS data sets and it includes a detailed tutorial in use of PARAFAC2 as well as description of its advantages over existing methods and drawbacks. During the second study (Paper 2), we have developed a new GC-MS metabolomics protocol for derivatization of complex biological samples based on TMSCN trimethylsilylation. The method was demonstrated to outperform existing methodologies commonly used in the literature in terms of sensitivity, speed, repeatability and finally yet importantly, the method provides an unbiased detection of broad range of metabolites. In the third study (Paper 3) we have reviewed the current challenges and perspectives of analytical technologies and high-throughput metabolomics protocols in cereal science. The review addresses most important aspects of quantitative metabolomics and highlights cutting-edge methods in metabolomic data acquisition, data preprocessing and data analysis. In the fourth study (Paper 4) we have compiled our knowledge gained from the 1<sup>st</sup> and the 2<sup>nd</sup> studies and applied them to a real biological question related to barley plants. In this study we have demonstrated the power of PARAFAC2 based metabolomics data processing, TMSCN derivatization and cutting-edge metabolomics data analysis approaches by using a chemometric method, namely, ANOVA-simultaneous component analysis (ASCA). The study revealed several biological mechanisms

associated with plant-environment, plant-gene mutation relationships and alterations of the plants' physiology during their development stages. Thus, the study proved the efficiency of the newly developed metabolomics methods to explore as much information as possible.

Three other studies that are in preparation are described in section 4. Study 5 demonstrates the use of Design of Experiment (DoE) for optimization of the metabolomic profiling protocol by applying three different analytical platforms, GC-MS, LC-MS and NMR. The study resulted in establishment of the complete metabolomic protocol, including metabolite extraction, derivatization (for GC) and data acquisition for the comprehensive metabolomic analysis of *B. vulgaris* plant leaves. Study 6 demonstrates development of targeted GC-MS metabolomic analysis of plant triterpenoids produced by combinatorial biochemistry in tobacco plant leaves. This study was focused on elucidation of the biosynthetic pathway of triterpenes. A new derivatization methodology developed in the second study, once again proved its high quantitative power and allowed detection of triterpenes produced in tobacco leaves by transient experiment of OSCs and P450s. Study 7 concerns the incomplete work performed on structure elucidation of unknown triterpenoid saponins from the insect resistant (glabrous or G type) and susceptible (pubescent or P type) *B. vulgaris* plants.

## 5.2 Perspectives

During the PhD project, I faced several questions for which it was difficult or impossible to find an answer from the literature. Several aspects of the plant metabolomics workflow (Figure 2) still require improved and validated methodologies for data acquisition and processing. This mainly concerns the detection of metabolites in a quantitative manner and the data processing of obtained complex data in order to enhance the extractable biological information. In the following I describe some of these research ideas that I find it useful to perform.

### *Tutorial for Optimization of Plant Metabolomics Protocols*

To date, there is no single comprehensive tutorial paper, which covers the pros and cons of the most frequently used protocols in plant metabolomics. A single protocol cannot meet all the requirements of the different metabolomics studies. Very few studies describe optimization of plant metabolomics protocols (Gullberg et al., 2004) and there is a need for a detailed tutorial for new comers in the field. Such a tutorial should cover design of experiment (DoE) for optimization of the plant metabolomics protocols and explain all the involved steps in detail. One of the most frequently arising issues in plant metabolomics is how to optimize protocols to achieve the best results in a short period of time by a limited number of pilot experiments. In my view, this tutorial paper should explain what the analyst must consider when performing targeted/untargeted analysis, metabolite profiling or fingerprinting. It should also address that in quantitative metabolomics, the protocols must be optimized towards three most important response variables: 1<sup>st</sup> robustness, 2<sup>nd</sup> s/n ratio and 3<sup>rd</sup> relevant information (number

## 5. OUTREACH

of detectable metabolites). Moreover the influence of the several important factors of the protocols (e.g. sample harvesting, mass reduction, quenching, storage, metabolite extraction, derivatization and instrumental conditions) must be highlighted in the examples of the outstanding metabolomics protocol optimization studies performed in human and animal bio-fluids and tissue extracts (Danielsson et al., 2012; Jiye et al., 2005).

### *Small volume-high-throughput derivatization for GC-MS metabolomics of biological mixtures*

On-line sample derivatization is an attractive method for GC-MS analysis, which is able to provide high-throughput and less expensive derivatization than the conventional derivatization methods. Several studies have shown a high-potential of on-line sample derivatization, including acylation and silylation in targeted and metabolomic profiling studies (Cheng et al., 2011; Ho and Ding, 2012; Lin et al., 2005; Liu et al., 2002; Tzing et al., 2006). However, very few studies have shown application of on-line derivatization for comprehensive GC-MS metabolomics of biological samples. The derivatization methodology described in Paper 2 promises an efficient on-line derivatization methodology for quantitative detection of a broad range of metabolites from a complex mixtures by using the novel trimethylsilylation reagent TMSCN. This is due to the high silylation reactivity of the TMSCN. Paper 2 and (Mai and Patil, 1986; Riggio et al., 1992) demonstrate the high silylation rate of TMSCN towards various functional groups, which can be used to develop a small-volume-high-throughput derivatization method for the GC-MS analysis of plant and animal tissue. TMSCN can be tested to develop a small volume in-needle derivatization methodology by using 10-30  $\mu\text{L}$  of the reagent and mixing it with 5-20  $\mu\text{L}$  of the complex sample extract (e.g. blood, urine, plant extract) within the needle used for injection. Since, TMSCN is able to provide rapid silylation at room temperature, the reaction mixture can be directly injected in GC-MS in solvent vent mode by slowly evaporating excess amount of TMSCN and the solvent. The same procedure can be performed in relatively high amounts of sample volume by using bigger syringes. These methodologies allow minimization of incubation time and the use of GC-MS vials and expensive lids that need to be used for automation of the whole analysis. Instead, GC-MS injection syringes will serve as a reaction incubator and it can be rinsed with solvents between each analysis. To date, we have tried TMSCN mostly with plant derived samples mixtures, however, it is expected to perform equally good with animal fluids or tissue extracts.

### *High-throughput GC-MS metabolomic data processing methods: a case study*

In metabolomics, one of the basic problems is to process the raw data to extract the quantitative information as accurate as possible. In GC-MS (same in LC-MS) analysis, peaks of the different metabolites might be overlapped or have low s/n ratio that hampers their quantification and the extraction of the mass spectra of the metabolites. In the literature, AMDIS is the mostly used method for processing and deconvolution of the mass spectrum of each analyte. However, in the last decade several alternative methods to AMDIS were developed. In addition, several instrument manufacturers also provide user friendly software packages that allow mass spectral deconvolution and peak

## 5. OUTREACH

quantification. It is worth mentioning that these methods may possess disadvantages when the data becomes too complex or when several hundred samples must be analyzed. Most of these software packages require manual processing of each sample, in which often leads to an increased bias due to analyst interference. The PARAFAC2 method demonstrated in Papers 2 and Paper 4 is an efficient method for processing raw GC-MS and LC-MS data and allow high-throughput analysis, accurate quantification and mass spectral deconvolution. I would be interesting to compare the performances of all these methods, including AMDIS, PARAFAC2, MCR, MetaboliteDetector, ChromaTOF (LECO) and ChemStation (Agilent) with respect to (1) resolution power, (2) sensitivity, (3) mass spectral deconvolution, (4) speed and (5) quantitative power. This can be performed by processing the GC-MS data obtained from the standard mixture samples that contain chemically very similar stereoisomers. Such a study would allow to evaluate different data processing methods for resolution of the severely overlapped peaks of stereoisomers as well as for accuracy of the deconvoluted mass spectra (what is the similarity between the deconvoluted mass spectra of the metabolite and its original mass spectra) and quantification.

### *Automated Chromatographic data Processing System (ACPS)*

Several advantages and drawbacks of the PARAFAC2 method are demonstrated within this thesis and in many other research papers (Amigo et al., 2010a; Amigo et al., 2010b; Bro et al., 1999; Khakimov et al., 2012). The method is becoming more and more common among metabolomics labs and people are starting to benefit from the advantages of PARAFAC2 in different research fields. However, one of the main drawbacks of the method originates from its use, as knowledge of chemometrics and coding skills in Matlab is still required. Since, PARAFAC2 based raw chromatographic data processing is becoming more common, it is important to develop a user friendly, preferably graphical user interface based software that can assist non-specialists to use the method. Similar software already been developed for the MCR method (Jaumot et al., 2005). Preliminary workflow of such software may include following steps:

- 1) Input: Three-way data (e.g. GC-MS, LC-MS, CE-MS, LC-DAD)
- 2) Determination of interval(s) in retention time dimension to be modeled individually (automatic or manually)
- 3) Choice of pre-processing method
- 4) Choice of method for modeling: PARAFAC2 or PARAFAC
- 5) Choice of the maximum number of factors to be fitted (automated validation algorithm might be implemented inside the software (Kamstrup-Nielsen et al., 2013))
- 6) Execution
- 7) Validation of the obtained models by exploring each model separately e.g. plot elution, mass spectral and concentration profiles of the resolved peaks, and if necessary repeat steps 5 and 6
- 8) Calculation of retention indices of the resolved peaks based on their retention time and compare it with an selected library (an RI library will be linked to the software)

## 5. OUTREACH

- 9) Use of PARAFAC2 resolved mass spectral profiles of peaks to search the most similar EI-MS match by using public databases such as NIST, Wiley that will be linked to the software
- 10) Import of concentration profiles into Excel or Matlab

In fact, such software is needed in many metabolomics labs that deal with big data sets, in order to properly process the obtained data and efficiently utilize gained information. However, development of this software requires a lot of effort, but it will pay back by changing currently used time consuming, laborious and in many cases black-box style data processing work, making it faster, more fun and informative.

Development of high-throughput, unbiased, quantitative and broad range metabolomics tools including analytical platforms and data processing as well as optimized metabolomics protocols may lead to significant improvements in many fields of scientific research and industrial advances. This includes today's top scientific topics such as synthetic biology, natural product discovery, sustainable food, energy, organic food production and solar energy. I believe, the role of metabolomics, particularly plant metabolomics will be a significant player in solving the above-mentioned top issues that our world is facing today. Since the metabolome of biological systems contains vast amount of information that is only partly analyzable, future research in metabolomics and method development will focus on the extraction of maximum quantitative and qualitative information and use it towards improving crops, medicine, food and life style in general.

**5. OUTREACH**

## 6 REFERENCES

- Afendi, F. M., Okada, T., Yamazaki, M., Hirai-Morita, A., Nakamura, Y., Nakamura, K., Ikeda, S., Takahashi, H., Altaf-Ul-Amin, M., Darusman, L. K. et al., 2012. KNAPSAcK Family Databases: Integrated Metabolite-Plant Species Databases for Multifaceted Plant Research. *Plant and Cell Physiology* 53.
- Agerbirk, N., Olsen, C. E., Bibby, B. M., Frandsen, H. O., Brown, L. D., Nielsen, J. K. and Renwick, J. A. A., 2003a. A saponin correlated with variable resistance of *Barbarea vulgaris* to the diamondback moth *Plutella xylostella*. *Journal of Chemical Ecology* 29, 1417-1433.
- Agerbirk, N., Orgaard, M. and Nielsen, J. K., 2003b. Glucosinolates, flea beetle resistance, and leaf pubescence as taxonomic characters in the genus *Barbarea* (Brassicaceae). *Phytochemistry* 63, 69-80.
- Allwood, J. W., De Vos, R. C. H., Moing, A., Deborde, C., Erban, A., Kopka, J., Goodacre, R. and Hall, R. D., 2011. *Plant Metabolomics and Its Potential for Systems Biology Research: Background Concepts, Technology, and Methodology*.
- Allwood, J. W. and Goodacre, R., 2010. An Introduction to Liquid Chromatography-Mass Spectrometry Instrumentation Applied in Plant Metabolomic Analyses. *Phytochemical Analysis* 21, 33-47.
- Amigo, J. M., Popielarz, M. J., Callejon, R. M., Morales, M. L., Troncoso, A. M., Petersen, M. A. and Toldam-Andersen, T. B., 2010a. Comprehensive analysis of chromatographic data by using PARAFAC2 and principal components analysis. *Journal of Chromatography A* 1217, 4422-4429.
- Amigo, J. M., Skov, T. and Bro, R., 2010b. ChromATHography: Solving Chromatographic Issues with Mathematical Models and Intuitive Graphics. *Chemical Reviews* 110, 4582-4605.
- Amigo, J. M., Skov, T., Coello, J., Maspoch, S. and Bro, R., 2008. Solving GC-MS problems with PARAFAC2. *Trac-Trends in Analytical Chemistry* 27, 714-725.
- Andersson, A. A., Kamal-Eldin, A. and Aman, P., 2010. Effects of Environment and Variety on Alkylresorcinols in Wheat in the HEALTHGRAIN Diversity Screen. *Journal of Agricultural and Food Chemistry* 58, 9299-9305.
- Arranz, S. and Calixto, F. S., 2010. Analysis of polyphenols in cereals may be improved performing acidic hydrolysis: A study in wheat flour and wheat bran and cereals of the diet. *Journal of Cereal Science* 51, 313-318.
- Augustin, J. M., Drok, S., Shinoda, T., Sanmiya, K., Nielsen, J. K., Khakimov, B., Olsen, C. E., Hansen, E. H., Kuzina, V., Ekstrom, C. T. et al., 2012. UDP-Glycosyltransferases from the UGT73C Subfamily in *Barbarea vulgaris* Catalyze Saponin 3-O-Glucosylation in Saponin-Mediated Insect Resistance. *Plant Physiology* 160, 1881-1895.
- Augustin, J. M., Kuzina, V., Andersen, S. B. and Bak, S., 2011. Molecular activities, biosynthesis and evolution of triterpenoid saponins. *Phytochemistry* 72, 435-457.
- Baker, J. M., Hawkins, N. D., Ward, J. L., Lovegrove, A., Napier, J. A., Shewry, P. R. and Beale, M. H., 2006. A metabolomic study of substantial equivalence of field-grown genetically modified wheat. *Plant Biotechnology Journal* 4, 381-392.



## 6. REFERENCES

- Balmer, D., Flors, V., Glauser, G. and Mauch-Mani, B., 2013. Metabolomics of cereals under biotic stress: current knowledge and techniques. *Frontiers in Plant Science* 4, 82.
- Baran, R., Kochi, H., Saito, N., Suematsu, M., Soga, T., Nishioka, T., Robert, M. and Tomita, M., 2006. MathDAMP: a package for differential analysis of metabolite profiles. *BMC Bioinformatics* 7.
- Barron, C. and Rouau, X., 2008. FTIR and Raman signatures of wheat grain peripheral tissues. *Cereal Chemistry* 85, 619-625.
- Barros, E., Lezar, S., Anttonen, M. J., van Dijk, J. P., Rohlig, R. M., Kok, E. J. and Engel, K. H., 2010. Comparison of two GM maize varieties with a near-isogenic non-GM variety using transcriptomics, proteomics and metabolomics. *Plant Biotechnology Journal* 8, 436-451.
- Behrends, V., Tredwell, G. D. and Bundy, J. G., 2011. A software complement to AMDIS for processing GC-MS metabolomic data. *Analytical Biochemistry* 415, 206-208.
- Bentley, R., 1999. Secondary metabolite biosynthesis: The first century. *Critical Reviews in Biotechnology* 19, 1-40.
- Bernhoft, A., 2013. Bioactive compounds in plants-benefits and risks for man and animals. Novus forlag, Oslo 2010: The Norwegian Academy of Science and Letters.
- 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. et al., 2004. Potential of metabolomics as a functional genomics tool. *Trends in Plant Science* 9, 418-425.
- Blau, K. and Halket, J. E., 1994. Handbook of Derivatives for Chromatography. Wiley, Chichester, UK.
- Bollina, V., Kushalappa, A. C., Choo, T. M., Dion, Y. and Rioux, S., 2011. Identification of metabolites related to mechanisms of resistance in barley against *Fusarium graminearum*, based on mass spectrometry. *Plant Molecular Biology* 77, 355-370.
- Bowne, J. B., Erwin, T. A., Juttner, J., Schnurbusch, T., Langridge, P., Bacic, A. and Roessner, U., 2012. Drought Responses of Leaf Tissues from Wheat Cultivars of Differing Drought Tolerance at the Metabolite Level. *Molecular Plant* 5, 418-429.
- Box, J. F., 1987. Guinness, Gosset, Fisher, and Small Samples. *Statistical Science* 2, 45-52.
- Bratchell, N., 1989. Multivariate response surface modelling by principal components analysis. *Journal of Chemometrics* 3, 579-588.
- Bro, R., Andersson, C. A. and Kiers, H. A. L., 1999. PARAFAC2 - Part II. Modeling chromatographic data with retention time shifts. *Journal of Chemometrics* 13, 295-309.
- Bro, R. and Kiers, H. A. L., 2003. A new efficient method for determining the number of components in PARAFAC models. *Journal of Chemometrics* 17, 274-286.
- Carroll, A. J., Badger, M. R. and Millar, A. H., 2010. The MetabolomeExpress Project: enabling web-based processing, analysis and transparent dissemination of GC/MS metabolomics datasets. *BMC Bioinformatics* 11.

## 6. REFERENCES

- Caspi, R., Altman, T., Dreher, K., Fulcher, C. A., Subhraveti, P., Keseler, I. M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L. A. et al., 2012. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research* 40, D742-D753.
- Chang, W. T., Thissen, U., Ehlert, K. A., Koek, M. M., Jellema, R. H., Hankemeier, T., van der Greef, J. and Wang, M., 2006. Effects of growth conditions and processing on *Rehmannia glutinosa* using fingerprint strategy. *Planta Medica* 72, 458-467.
- Chen, F., Tholl, D., Bohlmann, J. and Pichersky, E., 2011. The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant Journal* 66, 212-229.
- Chen, H. W., Wortmann, A. and Zenobi, R., 2007. Neutral desorption sampling coupled to extractive electrospray ionization mass spectrometry for rapid differentiation of bilosamples by metabolomic fingerprinting. *Journal of Mass Spectrometry* 42, 1123-1135.
- Cheng, C. Y., Wang, Y. C. and Ding, W. H., 2011. Determination of Triclosan in Aqueous Samples Using Solid-phase Extraction Followed by On-line Derivatization Gas Chromatography-Mass Spectrometry. *Analytical Sciences* 27, 197-202.
- Cheng, D., Knox, C., Young, N., Stothard, P., Damaraju, S. and Wishart, D. S., 2008. PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Research* 36, W399-W405.
- Choi, Y. H., Sertic, S., Kim, H. K., Wilson, E. G., Michopoulos, F., Lefeber, A. W. M., Erkelens, C., Kricun, S. D. P. and Verpoorte, R., 2005. Classification of *Ilex* species based on metabolomic fingerprinting using nuclear magnetic resonance and multivariate data analysis. *Journal of Agricultural and Food Chemistry* 53, 1237-1245.
- Clifford, M. N. and Scalbert, A., 2000. Ellagitannins - nature, occurrence and dietary burden. *Journal of the Science of Food and Agriculture* 80, 1118-1125.
- Confalonieri, M., Cammareri, M., Biazzi, E., Pecchia, P., Fevereiro, M. P. S., Balestrazzi, A., Tava, A. and Conicella, C., 2009. Enhanced triterpene saponin biosynthesis and root nodulation in transgenic barrel medic (*Medicago truncatula* Gaertn.) expressing a novel beta-amyrin synthase (*AsOXA1*) gene. *Plant Biotechnology Journal* 7, 172-182.
- Croft, D., O'Kelly, G., Wu, G. M., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B. et al., 2011. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research* 39, D691-D697.
- Crozier, A., Clifford, M. and Ashihara, H., 2006. *Plant Secondary Metabolites. Occurrence, Structure and Role in the Human Diet.* Blackwell Publishing.
- Cui, Q., Lewis, I. A., Hegeman, A. D., Anderson, M. E., Li, J., Schulte, C. F., Westler, W. M., Eghbalian, H. R., Sussman, M. R. and Markley, J. L., 2008. Metabolite identification via the Madison Metabolomics Consortium Database. *Nature Biotechnology* 26, 162-164.
- Curtis, T. Y., Muttucumar, N., Shewry, P. R., Parry, M. A. J., Powers, S. J., Elmore, J. S., Mottram, D. S., Hook, S. and Halford, N. G., 2009. Effects of Genotype and Environment on Free Amino Acid Levels

## 6. REFERENCES

- in Wheat Grain: Implications for Acrylamide Formation during Processing. *Journal of Agricultural and Food Chemistry* 57, 1013-1021.
- Danielsson, A. P. H., Moritz, T., Mulder, H. and Spegel, P., 2012. Development of a gas chromatography/mass spectrometry based metabolomics protocol by means of statistical experimental design. *Metabolomics* 8, 50-63.
- De Geyter, E., Smagghe, G., Rahbe, Y. and Geelen, D., 2012. Triterpene saponins of *Quillaja saponaria* show strong aphicidal and deterrent activity against the pea aphid *Acyrtosiphon pisum*. *Pest Management Science* 68, 164-169.
- Diab, Y., Ioannou, E., Emam, A., Vagias, C. and Roussis, V., 2012. Desmettianosides A and B, bisdesmosidic furostanol saponins with molluscicidal activity from *Yucca desmettiana*. *Steroids* 77, 686-690.
- Dieterle, F., Ross, A., Schlotterbeck, G. and Senn, H., 2006. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in H-1 NMR metabolomics. *Analytical Chemistry* 78, 4281-4290.
- Dunn, W. B., 2008. Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. *Physical Biology* 5.
- Duran, A. L., Yang, J., Wang, L. and Sumner, L. W., 2003. Metabolomics spectral formatting, alignment and conversion tools (MSFACTs). *Bioinformatics* 19, 2283-2293.
- Ekins, S., 2009. ChemSpider. *Chemistry World* 6, 70.
- Engelsen, S. B., Savorani, F. and Rasmussen, M. A., 2013. Chemometric Exploration of Quantitative NMR Data. *eMagRes* 2, 267-278.
- Engewald, W., Teske, J. and Efer, J., 1999. Programmed temperature vaporiser-based injection in capillary gas chromatography. *Journal of Chromatography A* 856, 259-278.
- Eriksson, L., Johansson, E., Kettaneh-Wold, N., Wikstrom, C. and Wold, L., 2000. Design of Experiments: Principles and Applications. Imetrics Academy, Umeå, Sweden.
- Evidente, A., Cimmino, A., Fernandez-Aparicio, M., Rubiales, D., Andolfi, A. and Melck, D., 2011. Soyasapogenol B and trans-22-dehydrocampesterol from common vetch (*Vicia sativa* L.) root exudates stimulate broomrape seed germination. *Pest Management Science* 67, 1015-1022.
- Faizal, A. and Geelen, D., 2013. Saponins and their role in biological processes in plants. *Phytochemical Reviews*. DOI 10.1007/s11101-013-9322-4
- Fernie, A. R. and Schauer, N., 2009. Metabolomics-assisted breeding: a viable option for crop improvement? *Trends in Genetics* 25, 39-48.
- Fialkov, A. B., Steiner, U., Lehotay, S. J. and Amirav, A., 2007. Sensitivity and noise in GC-MS: Achieving low limits of detection for difficult analytes. *International Journal of Mass Spectrometry* 260, 31-48.
- Fiehn, O., 2002. Metabolomics - the link between genotypes and phenotypes. *Plant Molecular Biology* 48, 155-171.

## 6. REFERENCES

- Fiehn, O., Kopka, J., Trethewey, R. N. and Willmitzer, L., 2000. Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Analytical Chemistry* 72, 3573-3580.
- Fiehn, O., 2008. Extending the breadth of metabolite profiling by gas chromatography coupled to mass spectrometry. *Trac-Trends in Analytical Chemistry* 27, 261-269.
- Fukushima, E. O., Seki, H., Ohyama, K., Ono, E., Umemoto, N., Mizutani, M., Saito, K. and Muranaka, T., 2011. CYP716A Subfamily Members are Multifunctional Oxidases in Triterpenoid Biosynthesis. *Plant and Cell Physiology* 52, 2050-2061.
- Ganem, B., 1978. From glucose to aromatics: recent developments in natural products of the shikimic acid pathway. *Tetrahedron* 34, 3353-3383.
- Garcia, I., Ortiz, M. C., Sarabia, L. and Aldama, J. M., 2007. Validation of an analytical method to determine sulfamides in kidney by HPLC-DAD and PARAFAC2 with first-order derivative chromatograms. *Analytica Chimica Acta* 587, 222-234.
- Gavaghan, C. L., Li, J. V., Hadfield, S. T., Hole, S., Nicholson, J. K., Wilson, I. D., Howe, P. W., Stanley, P. D. and Holmes, E., 2011. Application of NMR-based Metabolomics to the Investigation of Salt Stress in Maize (*Zea mays*). *Phytochemical Analysis* 22, 214-224.
- Geisler, K., Hughes, R. K., Sainsbury, F., Lomonosoff, G. P., Rejzek, M., Fairhurst, S., Olsen, C. E., Motawia, M. S., Melton, R. E., Hemmings, A. M. et al., 2013. Biochemical analysis of a multifunctional cytochrome P450 (CYP51) enzyme required for synthesis of antimicrobial triterpenes in plants. *Proceedings of the National Academy of Sciences*
- Gohlke, R. S. and McLafferty, F. W., 1993. Early Gas-Chromatography Mass-Spectrometry. *Journal of the American Society for Mass Spectrometry* 4, 367-371.
- Graham, S., Amigues, E., Migaud, M. and Browne, R., 2009. Application of NMR based metabolomics for mapping metabolite variation in European wheat. *Metabolomics* 5, 302-306.
- Greene, P. R. and Bain, C. D., 2005. Total internal reflection Raman spectroscopy of barley leaf epicuticular waxes in vivo. *Colloids and Surfaces B-Biointerfaces* 45, 174-180.
- Gullberg, J., Jonsson, P., Nordstrom, A., Sjoström, M. and 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, 283-295.
- Hamberger, B. and Bak, S., 2013. Plant P450s as versatile drivers for evolution of species-specific chemical diversity. *Philosophical Transactions of the Royal Society B-Biological Sciences* 368, 1612.
- Harshman, R. A., 1970. Foundations of the PARAFAC procedure: Model and conditions for an 'explanatory' multi-mode factor analysis. *UCLA Working Papers in Phonetics* 16, 1-80.
- Harshman, R. A., 1972. PARAFAC2: mathematical and technical notes. *UCLA Working Papers in Phonetics* 22, 30-44.

## 6. REFERENCES

- Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M. et al., 2013. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Research* 41, D456-D463.
- Hatcher, D. W. and Kruger, J. E., 1997. Simple phenolic acids in flours prepared from Canadian wheat: Relationship to ash content, color, and polyphenol oxidase activity. *Cereal Chemistry* 74, 337-343.
- Haug, K., Salek, R. M., Conesa, P., Hastings, J., de Matos, P., Rijnbeek, M., Mahendrakar, T., Williams, M., Neumann, S., Rocca-Serra, P. et al., 2013. MetaboLights-an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Research* 41, D781-D786.
- Heiden, A. C., Kolahgar, B. and Pfannkoch, E., 2001. Benefits of Using Programmed Temperature Vaporizers (PVTs) instead of Hot Split/Splitless Inlet for Measurements of Volatile by Liquid, Headspace, and Solid Phase microExtraction (SPME) Techniques. Application Note 7/2001, GERSTEL GmbH & Co.KG, Germany.
- Heinonen, M., Rantanen, A., Mielikainen, T., Kokkonen, J., Kiuru, J., Ketola, R. A. and Rousu, J., 2008. FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data. *Rapid Communications in Mass Spectrometry* 22, 3043-3052.
- Herrmann, K., 1976. Flavonols and flavones in food plants: a review. *Journal of Food Technology* - 11, 433-448.
- Hiller, K., Hangebrauk, J., Jaeger, C., Spura, J., Schreiber, K. and Schomburg, D., 2009. MetaboliteDetector: Comprehensive Analysis Tool for Targeted and Nontargeted GC/MS Based Metabolome Analysis. *Analytical Chemistry* 81, 3429-3439.
- Ho, Y. C. and Ding, W. H., 2012. Solid-phase Extraction Coupled Simple On-line Derivatization Gas Chromatography - Tandem Mass Spectrometry for the Determination of Benzophenone-type UV Filters in Aqueous Samples. *Journal of the Chinese Chemical Society* 59, 107-113.
- Hoffmann, N., Keck, M., Neuweger, H., Wilhelm, M., Hogy, P., Niehaus, K. and Stoye, J., 2012. Combining peak- and chromatogram-based retention time alignment algorithms for multiple chromatography-mass spectrometry datasets. *BMC Bioinformatics* 13.
- Hoffmann, N. and Stoye, J., 2009. ChromA: signal-based retention time alignment for chromatography-mass spectrometry data. *Bioinformatics* 25, 2080-2081.
- Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K. et al., 2010. MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry* 45, 703-714.
- Horning, E. C. and Horning, M. G., 1971. Human Metabolic Profiles Obtained by Gc and Gc/Ms. *Journal of Chromatographic Science* 9, 129-&.
- Horning, M. G., 1971. Drug Metabolism Studies in Newborn Using Gas Chromatography Mass Spectrometry. *Applied Spectroscopy* 25, 140.
- Horning, M. G., Moss, A. M. and Horning, E. C., 1968. Formation and gas-liquid chromatographic behavior of isometric steroid ketone methoxime derivatives. *Analytical Biochemistry* 22, 284-294.

## 6. REFERENCES

- Jacobsen, S., Sondergaard, I., Moller, B., Desler, T. and Munck, L., 2005. A chemometric evaluation of the underlying physical and chemical patterns that support near infrared spectroscopy of barley seeds as a tool for explorative classification of endosperm, genes and gene combinations. *Journal of Cereal Science* 42, 281-299.
- Jansen, J. J., Hoefsloot, H. C. J., van der Greef, J., Timmerman, M. E., Westerhuis, J. A. and Smilde, A. K., 2005. ASCA: analysis of multivariate data obtained from an experimental design. *Journal of Chemometrics* 19, 469-481.
- Jaroszewski, J. W., 2005a. Hyphenated NMR methods in natural products research, Part 1: Direct hyphenation. *Planta Medica* 71, 691-700.
- Jaroszewski, J. W., 2005b. Hyphenated NMR methods in natural products research, Part 2: HPLC-SPE-NMR and other new trends in NMR hyphenation. *Planta Medica* 71, 795-802.
- Jaroszewski, J. W., 2007. Hyphenated MMR techniques and MMR-based metabolomics in studies of medicinal plants. *Planta Medica* 73, 802.
- Jaumot, J., Gargallo, R., de Juan, A. and Tauler, R., 2005. A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB. *Chemometrics and Intelligent Laboratory Systems* 76, 101-110.
- Jensen, S. A., Munck, L. and Martens, H., 1982. The Botanical Constituents of Wheat and Wheat Milling Fractions .1. Quantification by Autofluorescence. *Cereal Chemistry* 59, 477-484.
- Jiye, A., Trygg, J., Gullberg, J., Johansson, A. I., Jonsson, P., Antti, H., Marklund, S. L. and Moritz, T., 2005. Extraction and GC/MS analysis of the human blood plasma metabolome. *Analytical Chemistry* 77, 8086-8094.
- Johnsen, L. G., Skov, T., Houlberg, U. and Bro, R., 2013. An automated method for baseline correction, peak finding and peak grouping in chromatographic data. *Analyst* 138, 3502-3511.
- Kamstrup-Nielsen, M. H., Johnsen, L. G. and Bro, R., 2013. Core consistency diagnostic in PARAFAC2. *Journal of Chemometrics* 27, 99-105.
- Kanani, H., Chrysanthopoulos, P. K. and Klapa, M. I., 2008. Standardizing GC-MS metabolomics. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences* 871, 191-201.
- Kanani, H. H. and Klapa, M. I., 2007. Data correction strategy for metabolomics analysis using gas chromatography-mass spectrometry. *Metabolic Engineering* 9, 39-51.
- Katajamaa, M., Miettinen, J. and Oresic, M., 2006. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* 22, 634-636.
- Khakimov, B., Amigo, J. M., Bak, S. and Engelsen, S. B., 2012. Plant metabolomics: Resolution and quantification of elusive peaks in liquid chromatography-mass spectrometry profiles of complex plant extracts using multi-way decomposition methods. *Journal of Chromatography. A* 1266, 84-94.
- Khakimov, B., Motawia, M. S., Bak, S. and Engelsen, S. B., 2013. The use of trimethylsilyl cyanide derivatization for robust and broad spectrum high-throughput gas-chromatography-mass

## 6. REFERENCES

- spectrometry based metabolomics. *Analytical and Bioanalytical Chemistry* DOI: 10.1007/s00216-013-7341-z.
- Kiers, H. A. L., Ten Berge, J. M. F. and Bro, R., 1999. PARAFAC2 - Part I. A direct fitting algorithm for the PARAFAC2 model. *Journal of Chemometrics* 13, 275-294.
- Kind, T. and Fiehn, O., 2007. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics* 8.
- Kind, T., Wohlgemuth, G., Lee, D. Y., Lu, Y., Palazoglu, M., Shahbaz, S. and Fiehn, O., 2009. FiehnLib: Mass Spectral and Retention Index Libraries for Metabolomics Based on Quadrupole and Time-of-Flight Gas Chromatography/Mass Spectrometry. *Analytical Chemistry* 81, 10038-10048.
- Kishimoto, Y., Tani, M. and Kondo, K., 2013. Pleiotropic preventive effects of dietary polyphenols in cardiovascular diseases. *European Journal of Clinical Nutrition* 67, 532-535.
- Koffas, M., Roberge, C., Lee, K. and Stephanopoulos, G., 1999. Metabolic engineering. *Annual Review of Biomedical Engineering* 1, 535-557.
- Kopka, J., Schauer, N., Krueger, S., Birkemeyer, C., Usadel, B., Bergmuller, E., Dormann, P., Weckwerth, W., Gibon, Y., Stitt, M. et al., 2005. GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics* 21, 1635-1638.
- Kruger, N. J., Troncoso-Ponce, M. A. and Ratcliffe, R. G., 2008. H-1 NMR metabolite fingerprinting and metabolomic analysis of perchloric acid extracts from plant tissues. *Nature Protocols* 3, 1001-1012.
- Kuzina, V., Nielsen, J. K., Augustin, J. M., Torp, A. M., Bak, S. and Andersen, S. B., 2011. Barbarea vulgaris linkage map and quantitative trait loci for saponins, glucosinolates, hairiness and resistance to the herbivore Phyllotreta nemorum. *Phytochemistry* 72, 188-198.
- Kuzina, V., Ekstrom, C. T., Andersen, S. B., Nielsen, J. K., Olsen, C. E. and Bak, S., 2009. Identification of Defense Compounds in Barbarea vulgaris against the Herbivore Phyllotreta nemorum by an Ecometabolomic Approach. *Plant Physiology* 151, 1977-1990.
- Lange, E., Reinert, K., Gropl, C., Kohlbacher, O., Sturm, M. and Hildebrandt, A., 2005. OPENMS; a generic open source framework for chromatography/MS-based proteomics. *Molecular & Cellular Proteomics* 4, S25.
- Li, J. T. and Hu, Z. H., 2009. Accumulation and Dynamic Trends of Triterpenoid Saponin in Vegetative Organs of Achyranthus bidentata. *Journal of Integrative Plant Biology* 51, 122-129.
- Lin, W. C., Chen, H. C. and Ding, W. H., 2005. Determination of pharmaceutical residues in waters by solid-phase extraction and large-volume on-line derivatization with gas chromatography-mass spectrometry. *Journal of Chromatography A* 1065, 279-285.
- Lindon, J. C. and Nicholson, J. K., 2008. Spectroscopic and Statistical Techniques for Information Recovery in Metabonomics and Metabolomics. *Annual Reviews of Analytical Chemistry* 1, 45-69.
- Lisec, J., Schauer, N., Kopka, J., Willmitzer, L. and Fernie, A. R., 2006. Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nature Protocols* 1, 387-396.

## 6. REFERENCES

- Little, J. L., 1999. Artifacts in trimethylsilyl derivatization reactions and ways to avoid them. *Journal of Chromatography A* 844, 1-22.
- Liu, Y. Q., Cho, S. R. and Danielson, N. D., 2002. Solid-phase microextraction and on-line methylation gas chromatography for aliphatic carboxylic acids. *Analytical and Bioanalytical Chemistry* 373, 64-69.
- Lommen, A., 2009. MetAlign: Interface-Driven, Versatile Metabolomics Tool for Hyphenated Full-Scan Mass Spectrometry Data Preprocessing. *Analytical Chemistry* 81, 3079-3086.
- Lommen, A. and Kools, H. J., 2012. MetAlign 3.0: performance enhancement by efficient use of advances in computer hardware. *Metabolomics* 8, 719-726.
- Ludwig, C., Easton, J. M., Lodi, A., Tiziani, S., Manzoor, S. E., Southam, A. D., Byrne, J. J., Bishop, L. M., He, S., Arvanitis, T. N. et al., 2012. Birmingham Metabolite Library: a publicly accessible database of 1-D H-1 and 2-D H-1 J-resolved NMR spectra of authentic metabolite standards (BML-NMR). *Metabolomics* 8, 8-18.
- Lundstedt, T., Seifert, E., Abramo, L., Thelin, B., Nystrom, A., Pettersen, J. and Bergman, R., 1998. Experimental design and optimization. *Chemometrics and Intelligent Laboratory Systems* 42, 3-40.
- Lynene, F., 1967. Biosynthetic pathways from acetate to natural products. *Pure and Applied Chemistry*, 137-168
- Lytovchenko, A., Beleggia, R., Schauer, N., Isaacson, T., Leuendorf, J. E., Hellmann, H., Rose, J. K. C. and Fernie, A. R., 2009. Application of GC-MS for the detection of lipophilic compounds in diverse plant tissues. *Plant Methods* 5.
- Macheix, J.-J., Fleuriet, A. and Billot, J., 1990. Fruit phenolics. Boca Raton, FL: CRC Press.
- Mai, K. and Patil, G., 1986. Alkylsilyl Cyanides As Silylating Agents. *Journal of Organic Chemistry* 51, 3545-3548.
- Manach, C., Scalbert, A., Morand, C., Remesy, C. and Jimenez, L., 2004. Polyphenols: food sources and bioavailability. *American Journal of Clinical Nutrition* 79, 727-747.
- Manetti, C., Bianchetti, C., Casciani, L., Castro, C., Di Cocco, M. E., Miccheli, A., Motto, M. and Conti, F., 2006. A metabolomic study of transgenic maize (*Zea mays*) seeds revealed variations in osmolytes and branched amino acids. *Journal of Experimental Botany* 57, 2613-2625.
- Manolache, F. A., Hanganu, A., Duta, D. E., Belc, N. and Marin, D. I., 2013. The Physico-chemical and Spectroscopic Composition Characterization of Oat Grains and Oat Oil Samples. *Revista de Chimie* 64, 45-48.
- Mardia, K. V., Kent, J. T. and Bibby, J. M., 1979. Multivariate Analysis. London: Academic Press.
- Marini, F., D'Aloise, A., Bucci, R., Buiarelli, F., Magri, A. L. and Magri, A. D., 2011. Fast analysis of 4 phenolic acids in olive oil by HPLC-DAD and chemometrics. *Chemometrics and Intelligent Laboratory Systems* 106, 142-149.
- Mattoli, L., Cangi, F., Maidecchi, A., Ghiara, C., Ragazzi, E., Tubaro, M., Stella, L., Tisato, F. and Traldi, P., 2006. Metabolomic fingerprinting of plant extracts. *Journal of Mass Spectrometry* 41, 1534-1545.



## 6. REFERENCES

- May, D., Fitzgibbon, M., Liu, Y., Holzman, T., Eng, J., Kemp, C. J., Whiteaker, J., Paulovich, A. and McIntosh, M., 2007. A platform for accurate mass and time analyses of mass spectrometry data. *Journal of Proteome Research* 6, 2685-2694.
- Mayer, A. M., 2006. Polyphenol oxidases in plants and fungi: Going places? A review. *Phytochemistry* 67, 2318-2331.
- Mikkelsen, M. S., Jespersen, B. M., Larsen, F. H., Blennow, A. and Engelsen, S. B., 2013. Molecular structure of large-scale extracted beta-glucan from barley and oat: Identification of a significantly changed block structure in a high beta-glucan barley mutant. *Food Chemistry* 136, 130-138.
- Mugford, S. T., Qi, X. Q., Bakht, S., Hill, L., Wegel, E., Hughes, R. K., Papadopoulou, K., Melton, R., Philo, M., Sainsbury, F. et al., 2009. A Serine Carboxypeptidase-Like Acyltransferase Is Required for Synthesis of Antimicrobial Compounds and Disease Resistance in Oats. *Plant Cell* 21, 2473-2484.
- Munck Lars., 1992. The case of high lysine barley breeding, Barley: Genetics, Biochemistry, Molecular Biology and Biotechnology. pp. 573-603. Wallingford, Oxon, UK: C.A.B: International.
- Munck, L., Jespersen, B. M., Rinnan, Å., Seefeldt, H. F., Engelsen, M. M., Norgaard, L. and Engelsen, S. B., 2010. A physiochemical theory on the applicability of soft mathematical models-experimentally interpreted. *Journal of Chemometrics* 24, 481-495.
- Munck, L., Nielsen, J. P., Moller, B., Jacobsen, S., Sondergaard, I., Engelsen, S. B., Norgaard, L. and Bro, R., 2001. Exploring the phenotypic expression of a regulatory proteome-altering gene by spectroscopy and chemometrics. *Analytica Chimica Acta* 446, 171-186.
- Murphy, K. R., Wenig, P., Parcsi, G., Skov, T. and Stuetz, R. M., 2012. Characterizing odorous emissions using new software for identifying peaks in chemometric models of gas chromatography-mass spectrometry datasets. *Chemometrics and Intelligent Laboratory Systems* 118, 41-50.
- Nicholson, J. K., Lindon, J. C. and Holmes, E., 1999. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* 29, 1181-1189.
- Nielsen, N. J., Nielsen, J. and Staerk, D., 2010. New Resistance-Related Saponins from the Insect-Resistant Crucifer *Barbarea vulgaris*. *Journal of Agricultural and Food Chemistry* 58, 5509-5514.
- Nielsen, N. P. V., Carstensen, J. M. and Smedsgaard, J., 1998. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A* 805, 17-35.
- Nilsson, M. and Morris, G. A., 2008. Speedy component resolution: An improved tool for processing diffusion-ordered spectroscopy data. *Analytical Chemistry* 80, 3777-3782.
- Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J. P., Munck, L. and Engelsen, S. B., 2000. Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. *Applied Spectroscopy* 54, 413-419.
- Okazaki, Y. and Saito, K., 2012. Recent advances of metabolomics in plant biotechnology. *Plant Biotechnology Reports* 6, 1-15.

## 6. REFERENCES

- Oliver, S. G., Winson, M. K., Kell, D. B. and Baganz, F., 1998. Systematic functional analysis of the yeast genome. *Trends in Biotechnology* 16, 373-378.
- Osbourn, A., Goss, R. J. M. and Field, R. A., 2011. The saponins - polar isoprenoids with important and diverse biological activities. *Natural Product Reports* 28, 1261-1268.
- Osbourn, A. E., 1996. Preformed antimicrobial compounds and plant defense against fungal attack. *Plant Cell* 8, 1821-1831.
- Osbourn, A. E., Qi, X. Q., Townsend, B. and Qin, B., 2003. Dissecting plant secondary metabolism - constitutive chemical defences in cereals. *New Phytologist* 159, 101-108.
- Pandey, K. B. and Rizvi, S. I., 2009. Plant polyphenols as dietary antioxidants in human health and disease. *Oxidative Medicine and Cellular Longevity* 2, 270-278.
- Petti, S. and Scully, C., 2009. Polyphenols, oral health and disease: A review. *Journal of Dentistry* 37, 413-423.
- Pierce, A. E., 1968. Silylation of Organic Compounds. Pierce Chemical Company: Rockford, IL.
- Pluskal, T., Castillo, S., Villar-Briones, A. and Oresic, M., 2010. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* 11.
- Pollier, J., Morreel, K., Geelen, D. and Goossens, A., 2011. Metabolite Profiling of Triterpene Saponins in *Medicago truncatula* Hairy Roots by Liquid Chromatography Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Journal of Natural Products* 74, 1462-1476.
- Poole, C. F. and Zlatkis, A., 1979. Trialkylsilyl Ether Derivatives (Other Than Tms) for Gas-Chromatography and Mass-Spectrometry. *Journal of Chromatographic Science* 17, 115-123.
- Poole, C. F., 2013. Alkylsilyl derivatives for gas chromatography. *Journal of Chromatography A* 1296, 2-14.
- Pourcel, L., Routaboul, J. M., Cheynier, V., Lepiniec, L. and Debeaujon, I., 2007. Flavonoid oxidation in plants: from biochemical properties to physiological functions. *Trends in Plant Science* 12, 29-36.
- Price, S. F., Breen, P. J., Valladao, M. and Watson, B. T., 1995. Cluster Sun Exposure and Quercetin in Pinot-Noir Grapes and Wine. *American Journal of Enology and Viticulture* 46, 187-194.
- Rahman, A., Ahamed, A., Amakawa, T., Goto, N. and Tsurumi, S., 2001. Chromosaponin I specifically interacts with AUX1 protein in regulating the gravitropic response of arabidopsis roots. *Plant Physiology* 125, 990-1000.
- Rahmani, A., Jinap, S. and Soleimany, F., 2009. Qualitative and Quantitative Analysis of Mycotoxins. *Comprehensive Reviews in Food Science and Food Safety* 8, 202-251.
- Riggio, P. P., Karasiewicz, R. J., Rosen, P. and Toome, V., 1992. The Use of Trimethylsilyl Cyanide in Gas-Chromatographic Analysis. *Journal of Chromatographic Science* 30, 29-31.
- Robinette, S. L., Zhang, F. L., Bruschweiler-Li, L. and Bruschweiler, R., 2008. Web server based complex mixture analysis by NMR. *Analytical Chemistry* 80, 3606-3611.

## 6. REFERENCES

- Roessner, U., Wagner, C., Kopka, J., Trethewey, R. N. and Willmitzer, L., 2000. Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant Journal* 23, 131-142.
- Royer, D., Humpf, H. U. and Guy, P. A., 2004. Quantitative analysis of Fusarium mycotoxins in maize using accelerated solvent extraction before liquid chromatography atmospheric pressure chemical ionization tandem mass spectrometry. *Food Additives and Contaminants* 21, 678-692.
- Sainsbury, F., Thuenemann, E. C. and Lomonossoff, G. P., 2009. pEAQ: versatile expression vectors for easy and quick transient expression of heterologous proteins in plants. *Plant Biotechnology Journal* 7, 682-693.
- Savorani, F., Picone, G., Badiani, A., Fagioli, P., Capozzi, F. and Engelsen, S. B., 2010a. Metabolic profiling and aquaculture differentiation of gilthead sea bream by H-1 NMR metabonomics. *Food Chemistry* 120, 907-914.
- Savorani, F., Tomasi, G. and Engelsen, S. B., 2010b. icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal of Magnetic Resonance* 202, 190-202.
- Savorani, F., Tomasi, G. and Engelsen, S. B., 2010c. icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal of Magnetic Resonance* 202, 190-202.
- Scalbert, A. and Williamson, G., 2000. Dietary intake and bioavailability of polyphenols. *Journal of Nutrition* 130, 2073S-2085S.
- Schauer, N. and Fernie, A. R., 2006. Plant metabolomics: towards biological function and mechanism. *Trends in Plant Science* 11, 508-516.
- Schellenberger, J., Park, J. O., Conrad, T. M. and Palsson, B. O., 2010. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics* 11.
- Schmidt, D. D., Voelckel, C., Hartl, M., Schmidt, S. and Baldwin, I. T., 2005. Specificity in ecological interactions. Attack from the same lepidopteran herbivore results in species-specific transcriptional responses in two solanaceous host plants. *Plant Physiology* 138, 1763-1773.
- Schuler, M. A. and Werck-Reichhart, D., 2003. Functional genomics of P450s. *Annual Review of Plant Biology* 54, 629-667.
- Schummer, C., Delhomme, O., Appenzeller, B. M. R., Wennig, R. and Millet, M., 2009. Comparison of MTBSTFA and BSTFA in derivatization reactions of polar compounds prior to GC/MS analysis. *Talanta* 77, 1473-1482.
- Searle, S. R., 1971. Introduction to Variance Components. In *Linear Models*: John Wiley & Sons, Inc. pp. 376-420.
- Shinoda, T., Nagao, T., Nakayama, M., Serizawa, H., Koshioka, M., Okabe, H. and Kawai, A., 2002. Identification of a triterpenoid saponin from a crucifer, *Barbarea vulgaris*, as a feeding deterrent to the diamondback moth, *Plutella xylostella*. *Journal of Chemical Ecology* 28, 587-599.
- Shuman, J., Cortes, D., Armenta, J., Pokrzywa, R., Mendes, P. and Shulaev, V., 2011. Plant Metabolomics by GC-MS and Differential Analysis. In *Plant Reverse Genetics* (ed. A. Pereira): Humana Press. pp. 229-246.

## 6. REFERENCES

- Siuda, R., Balcerowska, G. and Sadowski, C., 2006. Comparison of the usability of different spectral ranges within the near ultraviolet, visible and near infrared ranges (UV-VIS-NIR) region for the determination of the content of scab-damaged component in blended samples of ground wheat. *Food Additives and Contaminants* 23, 1201-1207.
- Smilde, A. K., Hoefsloot, H. C. J. and Westerhuis, J. A., 2008. The geometry of ASCA. *Journal of Chemometrics* 22, 464-471.
- Smilde, A. K., Jansen, J. J., Hoefsloot, H. C. J., Lamers, R. J. A. N., van der Greef, J. and Timmerman, M. E., 2005. ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics* 21, 3043-3048.
- Smith, C. A., O'Maille, G., Want, E. J., Qin, C., Trauger, S. A., Brandon, T. R., Custodio, D. E., Abagyan, R. and Siuzdak, G., 2005. METLIN - A metabolite mass spectral database. *Therapeutic Drug Monitoring* 27, 747-751.
- Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. and Siuzdak, G., 2006. XCMS: Processing mass spectrometry data for metabolite profiling using Nonlinear peak alignment, matching, and identification. *Analytical Chemistry* 78, 779-787.
- Sohn, M., Himmelsbach, D. S. and Barton, F. E., 2004. A comparative study of Fourier transform Raman and NIR spectroscopic methods for assessment of protein and apparent amylose in rice. *Cereal Chemistry* 81, 429-433.
- Stähle, L. and Wold, S., 1990. Multivariate-Analysis of Variance (Manova). *Chemometrics and Intelligent Laboratory Systems* 9, 127-141.
- Stein, S. E., 1999. An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *Journal of the American Society for Mass Spectrometry* 10, 770-781.
- Sturm, M., Bertsch, A., Gropl, C., Hildebrandt, A., Hussong, R., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., Zerck, A., Reinert, K. et al., 2008. OpenMS-An open-source software framework for mass spectrometry. *BMC Bioinformatics* 9.
- Sun, H. X., Xie, Y. and Ye, Y. P., 2009. Advances in saponin-based adjuvants. *Vaccine* 27, 1787-1796.
- t'Kindt, R., Morreel, K., Deforce, D., Boerjan, W. and Van Bocxlaer, J., 2009. Joint GC-MS and LC-MS platforms for comprehensive plant metabolomics: Repeatability and sample pre-treatment. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences* 877, 3572-3580.
- Takahashi, S., Hori, K., Hokari, M., Gotoh, T. and Sugiyama, T., 2010. Inhibition of human renin activity by saponins. *Biomedical Research-Tokyo* 31, 155-159.
- Tang, H. F., Cheng, G., Wu, J., Chen, X. L., Zhang, S. Y., Wen, A. D. and Lin, H. W., 2009. Cytotoxic Asterosaponins Capable of Promoting Polymerization of Tubulin from the Starfish *Culcita novaeguineae*. *Journal of Natural Products* 72, 284-289.
- Tautenhahn, R., Patti, G. J., Rinehart, D. and Siuzdak, G., 2012. XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data. *Analytical Chemistry* 84, 5035-5039.

## 6. REFERENCES

- Tomas-Barberan, F. A. and Clifford, M. N., 2000. Dietary hydroxybenzoic acid derivatives - nature, occurrence and dietary burden. *Journal of the Science of Food and Agriculture* 80, 1024-1032.
- Tzing, S. H., Ghule, A., Liu, J. Y. and Ling, Y. C., 2006. On-line derivatization gas chromatography with furan chemical ionization tandem mass spectrometry for screening of amphetamines in urine. *Journal of Chromatography A* 1137, 76-83.
- van Look, G., Simchen, G. and Heberle, J., 1995. Silylating Agents, Fluka Chemie AG. Buchs, Switzerland.
- van Velzen, E. J. J., Westerhuis, J. A., van Duynhoven, J. P. M., van Dorsten, F. A., Hoefsloot, H. C. J., Jacobs, D. M., Smit, S., Draijer, R., Kroner, C. I. and Smilde, A. K., 2008. Multilevel data analysis of a crossover designed human nutritional intervention study. *Journal of Proteome Research* 7, 4483-4491.
- Veber, D. F., Johnson, S. R., Cheng, H. Y., Smith, B. R., Ward, K. W. and Kopple, K. D., 2002. Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry* 45, 2615-2623.
- Vincken, J. P., Heng, L., de Groot, A. and Gruppen, H., 2007. Saponins, classification and occurrence in the plant kingdom. *Phytochemistry* 68, 275-297.
- Wang, J. S., Reijmers, T., Chen, L. J., Van der Heijden, R., Wang, M., Peng, S. Q., Hankemeier, T., Xu, G. W. and van der Greef, J., 2009a. Systems toxicology study of doxorubicin on rats using ultra performance liquid chromatography coupled with mass spectrometry based metabolomics. *Metabolomics* 5, 407-418.
- Wang, Y. L., Bolton, E., Dracheva, S., Karapetyan, K., Shoemaker, B. A., Suzek, T. O., Wang, J. Y., Xiao, J. W., Zhang, J. and Bryant, S. H., 2010. An overview of the PubChem BioAssay resource. *Nucleic Acids Research* 38, D255-D266.
- Wang, Y. L., Xiao, J. W., Suzek, T. O., Zhang, J., Wang, J. Y. and Bryant, S. H., 2009b. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research* 37, W623-W633.
- Wang, Y. L., Xiao, J. W., Suzek, T. O., Zhang, J., Wang, J. Y., Zhou, Z. G., Han, L. Y., Karapetyan, K., Dracheva, S., Shoemaker, B. A. et al., 2012. PubChem's BioAssay Database. *Nucleic Acids Research* 40, D400-D412.
- Watson, J. T. and Sparkman, O. D., 2007a. Electron Ionization, Representative Fragmentations (Spectra) of Classes of Compounds (6. Trimethylsilyl Derivative). In *Introduction to Mass Spectrometry*: John Wiley & Sons, Ltd. pp. 315-448.
- Watson, J. T. and Sparkman, O. D., 2007b. Gas Chromatography/Mass Spectrometry. In *Introduction to Mass Spectrometry*: John Wiley & Sons, Ltd. pp. 571-638.
- Widodo, Patterson, J. H., Newbiggin, E., Tester, M., Bacic, A. and Roessner, U., 2009. Metabolic responses to salt stress of barley (*Hordeum vulgare* L.) cultivars, Sahara and Clipper, which differ in salinity tolerance. *Journal of Experimental Botany* 60, 4089-4103.

## 6. REFERENCES

- Williams, A. J. and Tkachenko, V., 2010. ChemSpider: How an online resource of chemical compounds, reaction syntheses, and property data can support green chemistry. *Abstracts of Papers of the American Chemical Society* 239.
- Williams, A. J. and Tkachenko, V., 2011. ChemSpider: Does community engagement work to build a quality online resource for chemists? *Abstracts of Papers of the American Chemical Society* 242.
- Wishart, D. S., 2007. Current Progress in computational metabolomics. *Briefings in Bioinformatics* 8, 279-293.
- Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y. F., Djombou, Y., Mandal, R., Aziat, F., Dong, E. et al., 2013. HMDB 3.0-The Human Metabolome Database in 2013. *Nucleic Acids Research* 41, D801-D807.
- Wu, D. Z., Cai, S. G., Chen, M. X., Ye, L. Z., Chen, Z. H., Zhang, H. T., Dai, F., Wu, F. B. and Zhang, G. P., 2013. Tissue Metabolic Responses to Salt Stress in Wild and Cultivated Barley. *Plos One* 8.
- Xia, E. Q., Deng, G. F., Guo, Y. J. and Li, H. B., 2010. Biological Activities of Polyphenols from Grapes. *International Journal of Molecular Sciences* 11, 622-646.
- Xia, J. G., Bjorndahl, T. C., Tang, P. and Wishart, D. S., 2008. MetaboMiner - semi-automated identification of metabolites from 2D NMR spectra of complex biofluids. *BMC Bioinformatics* 9.
- Xia, J. G., Mandal, R., Sinelnikov, I. V., Broadhurst, D. and Wishart, D. S., 2012. MetaboAnalyst 2.0-a comprehensive server for metabolomic data analysis. *Nucleic Acids Research* 40, W127-W133.
- Xia, J. G., Psychogios, N., Young, N. and Wishart, D. S., 2009. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Research* 37, W652-W660.
- Xia, J. G., Sinelnikov, I. V. and Wishart, D. S., 2011. MetATT: a web-based metabolomics tool for analyzing time-series and two-factor datasets. *Bioinformatics* 27, 2455-2456.
- Yu, T. W., Park, Y., Johnson, J. M. and Jones, D. P., 2009. apLCMS-adaptive processing of high-resolution LC/MS data. *Bioinformatics* 25, 1930-1936.
- Zadernowski, R., Naczek, M. and Nesterowicz, J., 2005. Phenolic acid profiles in some small berries. *Journal of Agricultural and Food Chemistry* 53, 2118-2124.
- Zambou, K., Spyropoulos, C. G., Chinou, I. and Kontos, F., 1993. Saponin-Like Substances Inhibit Alpha-Galactosidase Production in the Endosperm of Fenugreek Seeds - A Possible Regulatory Role in Endosperm Galactomannan Degradation. *Planta* 189, 207-212.
- Zekovic, I., Lenhardt, L., Dramicanin, T. and Dramicanin, M. D., 2012. Classification of Intact Cereal Flours by Front-Face Synchronous Fluorescence Spectroscopy. *Food Analytical Methods* 5, 1205-1213.
- Zwanenburg, G., Hoefsloot, H. C. J., Westerhuis, J. A., Jansen, J. J. and Smilde, A. K., 2011. ANOVA-principal component analysis and ANOVA-simultaneous component analysis: a comparison. *Journal of Chemometrics* 25, 561-567.



# Paper 1

**Bekzod Khakimov**, José Manuel Amigo, Søren Bak, Søren Balling Engelsen

Plant metabolomics: Resolution and quantification of elusive peaks in liquid chromatography-mass spectrometry profiles of complex plant extracts using multi-way decomposition methods

*Journal of Chromatography A*, 1266 (2012) 84–94







# Plant metabolomics: Resolution and quantification of elusive peaks in liquid chromatography–mass spectrometry profiles of complex plant extracts using multi-way decomposition methods

Bekzod Khakimov<sup>a,b,\*</sup>, José Manuel Amigo<sup>a</sup>, Søren Bak<sup>b</sup>, Søren Balling Engelsen<sup>a</sup>

<sup>a</sup> Quality & Technology, Department of Food Science, Faculty of Science, University of Copenhagen, Rolighedsvej 30, 1958 Frederiksberg C, Denmark

<sup>b</sup> Plant Biochemistry, Department of Plant and Environmental Sciences, Faculty of Science, University of Copenhagen, Rolighedsvej 30, 1958 Frederiksberg C, Denmark

## ARTICLE INFO

### Article history:

Received 5 September 2012

Received in revised form 7 October 2012

Accepted 8 October 2012

Available online 16 October 2012

### Keywords:

Plant metabolomics

Triterpenoid saponins

LC–MS

PARAFAC2

Automatic peak detection

*Barbarea vulgaris*

## ABSTRACT

Previous studies on LC–MS metabolomic profiling of 127 F2 *Barbarea vulgaris* plants derived from a cross of parental glabrous (G) and pubescent (P) type, revealed four triterpenoid saponins (hederagenin cellobioside, oleonic acid cellobioside, epihederagenin cellobioside, and gypsogenin cellobioside) that correlated with resistance of plants against the insect herbivore, *Phyllotreta nemorum*. In this study, for the first time, we demonstrate the efficiency of the multi-way decomposition method PARALLEL FACTOR analysis 2 (PARAFAC2) for exploring complex LC–MS data. PARAFAC2 enabled automated resolution and quantification of several elusive chromatographic peaks (e.g. overlapped, elution time shifted and low S/N ratio), which could not be detected and quantified by conventional chromatographic data analysis. Raw LC–MS data of 127 F2 *B. vulgaris* plants were arranged in a three-way array (elution time point × mass spectra × samples), divided into 17 different chromatographic intervals and each interval were individually modeled by PARAFAC2. Three main outputs of the PARAFAC2 models described: (1) elution time profile, (2) relative abundance, and (3) pure mass spectra of the resolved peaks modeled from each interval of the chromatographic data. PARAFAC2 scores corresponding to relative abundances of the resolved peaks were extracted and further used for correlation and partial least squares (PLS) analysis. A total of 71 PARAFAC2 components (which correspond to actual peaks, baselines and tails of neighboring peaks) were modeled from 17 different chromatographic retention time intervals of the LC–MS data. In addition to four previously known saponins, correlation- and PLS-analysis resolved five unknown saponin-like compounds that were significantly correlated with insect resistance. The method also enabled a good separation between resistant and susceptible F2 plants. PARAFAC2 spectral loadings corresponding to the pure mass spectra of chromatographic peaks matched well with experimentally recorded mass spectra (correlation based similarity >95%). This enabled to extract pure mass spectra of highly overlapped and low S/N ratio peaks.

© 2012 Elsevier B.V. All rights reserved.

## 1. Introduction

### 1.1. Plant metabolomics

Plant metabolomics deals with qualitative and quantitative analysis of metabolites from plant tissues and covers *targeted metabolomic analysis*, *metabolite profiling*, and *metabolic fingerprinting* [1]. Metabolomics may be viewed as the phenotypic endpoint of the sequence

genomics–transcriptomics–proteomics–metabolomics that reflects the dynamics of the plant. The metabolomic analysis of plant organisms has become a key technology for understanding the complexity of plant metabolism, its control and the link between genotypes and phenotypes [1,2]. Plant metabolomics approaches have successfully been applied in systems biology [3], biotechnology [4], functional genomics [5], environmental science [6], food chemistry [7] and medicinal chemistry [8].

### 1.2. Analytical techniques and metabolomic data

Advancements over the past decades in the capabilities of separation (e.g., GC, HPLC, UPLC, CE) and detection (e.g., MS, NMR, DAD and FID) techniques and in the development of new multivariate data analysis methods have led to a giant leap in knowledge about the plant metabolome and its response to internal and external

\* Corresponding author at: Quality & Technology, Department of Food Science, Faculty of Science, University of Copenhagen, Rolighedsvej 30, 1958 Frederiksberg C, Denmark. Tel.: +45 353 33239; fax: +45 353 33245.

E-mail addresses: [bzo@life.ku.dk](mailto:bzo@life.ku.dk) (B. Khakimov), [jmar@life.ku.dk](mailto:jmar@life.ku.dk) (J.M. Amigo), [bak@life.ku.dk](mailto:bak@life.ku.dk) (S. Bak), [se@life.ku.dk](mailto:se@life.ku.dk) (S.B. Engelsen).

perturbations. Application of hyphenated methods of analysis, for investigating the metabolome of plant organisms which contain a wide range of compounds with different chemical and physical properties is becoming more reliable and common approach. The choice of separation and detection methods is case specific and depends on the purpose of the metabolomic analysis. One of the most commonly applied techniques in plant metabolomics is liquid chromatography coupled to mass spectrometer (LC–MS) [9].

LC–MS can generate two, three and high-dimensional data sets, depending on the applied analytical technique and scan mode. The simplest example would be a data set obtained from a selected ion monitoring (SIM) experiment on LC–MS with one  $m/z$  channel (Fig. 1a). In this case, there would be only one measurement (the intensity of selected  $m/z$  channel) for each elution time point. For example, if a SIM experiment recorded for 25 min with a scanning frequency of 20 scans per minute, the obtained data would be a vector  $\mathbf{x}$  ( $1 \times 500$ ). If many samples are analyzed in the same way, the obtained data would be a two dimensional matrix  $\mathbf{X}(I \times J)$ , where  $I$  is the number of rows and it corresponds to the number of samples.  $J$  is the number of columns and it corresponds to the  $m/z$  signals' intensities measured at the selected  $m/z$  channel (Fig. 1b)). In general, when metabolomic data is acquired by LC–MS, a full mass spectrum (e.g. in the range of 200–1600  $m/z$ ) is obtained for each elution time scan point. In this case, the obtained data for each sample will become a two dimensional matrix  $\mathbf{Y}(I \times J)$ , where  $I$  is the number of rows corresponding to the elution time scan points.  $J$  is the number of columns corresponding to the number of selected  $m/z$  channels, respectively. Fig. 1c shows a landscape of the LC–MS data acquired for one sample (e.g., matrix  $\mathbf{Y}$ ) and Fig. 1d depicts such a data set obtained for many samples which forms three dimensional array  $\mathbf{Z}(I \times J \times K)$ , where  $K$  is the number of analyzed samples.

### 1.3. Common problems in chromatography

The traditional way of analyzing LC–MS data is to quantify well-resolved chromatographic peaks from total ion current (TIC) or base peak chromatograms (BPC) by calculating the peak area or peak height, followed by qualitative identification based on the mass spectra of the corresponding peaks. However, this approach is not always straightforward to perform. The analysis of the LC–MS metabolomic data, and the data obtained from other hyphenated methods can be significantly deteriorated by several different problems occurring during data acquisition (e.g. changes in elution time of peaks between the runs, overlapping, low  $s/n$  ratio of the peaks, non-Gaussian shape of the peaks and baseline drifts) [10]. Changes in the elution times of chromatographic peaks between the LC–MS runs are mainly caused by the instrumental uncertainties arising from small variations in pressure, temperature, pH, stationary phase and by wear of injection system or column. Moreover, the resolution power of the separation techniques is often not sufficient to obtain well-resolved peaks of isomers and/or chemically similar compounds. This will result in overlapping peaks and in turn more challenging peak quantification. Quantification of peaks by conventional chromatographic data analysis software may become difficult and laborious when baseline drifts are present and/or peaks have low  $s/n$  ratio.

Several methods have been proposed for solving the above-mentioned alignment problems in chromatographic data analysis. For example, correlation optimized warping (COW) [11,12] and interval correlation optimized shifting (*icos*shift) algorithms [13] have been proposed for alignment of the elution time shifted peaks and Boelens et al., 2004 [14] developed a method for removing the baseline drifts based on smoothing. For full evaluation of

problematic hyphenated LC–MS data, basically two different data analytical strategies can be followed:

- (1) Alignment of retention time shifts followed by PARAFAC analysis. This method has been applied to explore GC–MS [15] and LC–MS [16] metabolomic data.
- (2) Direct PARAFAC2 [17,18] modeling of chromatographic data with disturbed tri-linear structure (e.g. sample-to-sample retention time shifts of the peaks). If the elution time shifts are relatively confined to a certain extent, the PARAFAC2 algorithm is able to find and model the shifted peaks of the same chemical compounds, profiting the fact that they have the same mass spectral profile (or vice versa). This method has already demonstrated to work well for resolving chromatographic problems on GC–MS data [19]. Skov et al. [20] have demonstrated the application of these two methodologies in a comparative GC–MS study.

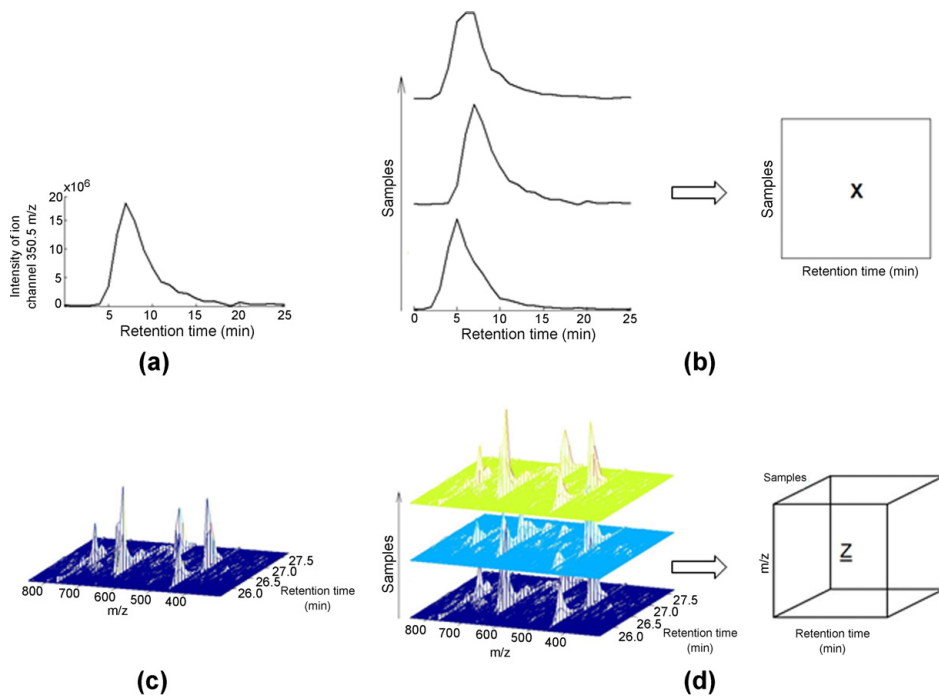
In this study, we illustrate an application of multi-way decomposition method PARAFAC2 for processing complex LC–MS data with minimum user interference. The methodology presented in this study is direct modeling of raw LC–MS data without any data-preprocessing step.

Moreover, some commercial software (e.g. Osiris) have been developed specifically for the prediction of the optimal chromatographic separation conditions for resolution of the overlapped peaks [21]. It is worth to mention that many LC–MS data processing have been performed using peak detection software (e.g., XCMS, MZmine, and MarkerLynx™), which can perform noise reduction, deconvolution, alignment and peak detection simultaneously [22]. The output of this processing is a table of detected peaks, where each peak is characterized by its own retention time and  $m/z$ . One of the main drawbacks of using such peak detection software is that, user must define several input parameters, which subsequently depend on investigated data set, and slight changes in one of these parameters may cause significant changes in obtained results.

### 1.4. Background and aim of the study

Defense compounds of *Barbarea vulgaris* against insect herbivores are well studied by metabolomic approaches. A targeted metabolomic analysis revealed the triterpenoid saponins hederagenin cellobioside [23] and oleanolic acid cellobioside [24] as the main defense compounds (Fig. 2). Kuzina et al. [25] performed untargeted LC–MS metabolomic profiling for investigating the metabolites correlated to the resistance level of 127 F2 *B. vulgaris* plants derived from parental resistant (G-type) and susceptible (P-type) *B. vulgaris* plants, against the flea beetle larvae, *Phyllotreta nemorum*. In the study, they processed the raw LC–MS data using MetAlign software, followed by correlation and principal component analysis (PCA). This approach revealed oleanolic acid cellobioside, hederagenin cellobioside and two unknown metabolites as the most correlated metabolites of F2 plants against herbivory. Later, two unknown metabolites that showed highly correlation to the insect resistance were identified as epihederagenin cellobioside and gypsogenin cellobioside (Fig. 1) [26].

The main aim of this research was to assess and illustrate the capabilities of PARAFAC2 for detection and quantification of elusive chromatographic peaks by solving common chromatographic problems occurring in LC–MS metabolomic data. In this study, we reevaluate the raw LC–MS data obtained from the untargeted metabolomic profiling of the 127 F2 plants [22] using PARAFAC2 and compare our findings with previous found results [22]. In addition to the latter,

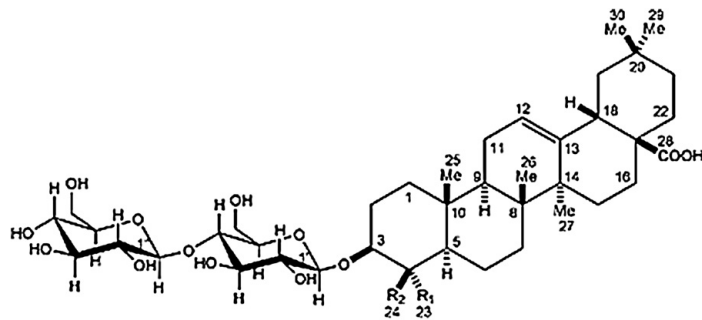


**Fig. 1.** Illustration of LC-MS metabolomic data structure: (a) a chromatogram obtained from the LC-MS analysis with one  $m/z$  channel (SIM experiment), (b) several samples analyzed by LC-MS with one  $m/z$  channel form a two dimensional data matrix, (c) landscape of raw LC-MS data obtained for one sample by recording several  $m/z$  signals (e.g. in the range of 300–900  $m/z$ ) for each elution time scan point, and (d) three-way structure of the raw LC-MS data obtained for three samples.

correlation analysis and PLS modeling were applied to the resolved component scores to investigate the correlation between PARAFAC2 resolved peaks and F2 plants' resistance against insects.

## 2. Experimental

F2 plants of *B. vulgaris* were generated by hybridization of the resistant G- (originated from Herlev, Denmark) and the susceptible



**Hederagenin cellobiosida** ( $R_1 = \text{CH}_2\text{OH}$ ,  $R_2 = \text{CH}_3$ )

**Oleanolic acid cellobiosida** ( $R_1 = \text{CH}_3$ ,  $R_2 = \text{CH}_3$ )

**Epihederagenin cellobiosida** ( $R_1 = \text{CH}_3$ ,  $R_2 = \text{CH}_2\text{OH}$ )

**Gypogenin cellobiosida** ( $R_1 = \text{CHO}$ ,  $R_2 = \text{CH}_3$ )

**Fig. 2.** Structures of known major triterpenoid saponins from *B. vulgaris* correlated with plants' defense against the insect herbivore, *P. nemorum*.

P- (originated from Tisø, Denmark) type *B. vulgaris* under greenhouse conditions. Young (3- to 12-week-old) leaves of F2 plants were used for determining resistance levels of the plants against flea beetle larvae (*Phyllotreta nemorum*) and metabolomic profiling by LC–MS. Further details about the plants and insect can be from [22].

### 2.1. Bioassays

Resistance levels of individual F2 plants were measured in bioassays using freshly harvested leaves of young plants and flea beetle larvae (less than 1-day-old). Five larvae were placed on each leaf and incubated at 24 °C for 72 h. After the incubation period, the number of survived larvae on the surface of the leaf was counted by stereomicroscope. For each F2 plant, six leaves were used in bioassay and the resistance levels of the plants were determined using 30 larvae. In this work, the average number of survived larvae per leaf disc was used as the resistance level of the F2 plants. The most resistant F2 plants have resistance level of 0, whereas the most susceptible F2 plants have resistance level of 5. Thus, 127 F2 plants were divided into three classes according to their resistance level: class 1 contained F2 plants with the resistance level of 0–1, class 2 samples contained partly resistant F2 plants with a resistance level of 1–4, and class 3 samples contained the susceptible F2 plants with a resistance level of higher than 4.

### 2.2. LC–MS metabolomic profiling

For acquisition of LC–MS metabolomic data, 8 mm leaf discs (about of 4 mg dry weight) were frozen in liquid nitrogen and kept at –80 °C. The frozen leaves were extracted with 500 µl of 85% methanol (60–70 °C) in a boiling water bath for 5 min and then cooled with ice. The extract was filtered through 45-µm Ultrafree-MC Durapore polyvinylidene difluoride filters (Millipore) before injecting into the LC–MS system. LC–MS analysis was performed on an Agilent 1100 Series liquid chromatograph (Agilent technologies) coupled to a Bruker Esquire 3000+ ion trap mass spectrometer (Bruker Daltonics). The column used was an XTerra MS C18 (3.5 µm, 2.1 mm × 100 mm; Waters). The mobile phases were solvent A (1 ml l<sup>-1</sup> formic acid and 50 µm NaCl) and solvent B (800 ml l<sup>-1</sup> acetonitrile and 1 ml l<sup>-1</sup> formic acid). The gradient program applied was as follows: 0–3 min, isocratic 18% B; 3–60 min, linear gradient 18–80% B; 60–65 min, linear gradient 80–100% B; 65–70 min, isocratic 100% B; 71–85 min, isocratic 18% B. The flow rate of the mobile phases was set to 0.2 ml min<sup>-1</sup>, the column temperature and the injection volume were 35 °C and 5 µl, respectively. The mass spectrometer was operated in a positive mode, and the ions were detected in the range of 300–1200 m/z.

### 2.3. Software

Chromatographic analysis of the LC–MS data was performed using the software, DataAnalysis, Version 4.0 (Bruker Daltonics). PARAFAC2, correlation and PLS regression analyses were performed using PLS Toolbox (Version 6.0.1, Eigenvector Research Inc. USA) working under MATLAB (Version 7.13.0.564, R2011b, The Mathworks, Inc., USA) environment. Raw LC–MS data was imported from netCDF files into MATLAB using available MATLAB codes.

## 3. Multi-way models

### 3.1. PARAFAC

Parallel Factor Analysis (PARAFAC) [27–29] is a generalization of principal component analysis (PCA) to the situation where a set of data matrices is to be analyzed. If two or more data matrices

with the same row and column units are combined, the resulting data become three-way and it can be modeled by the PARAFAC. Just like PCA, PARAFAC decomposes three-way data into sets of so-called scores and loadings matrices, which describes the data in a more condensed form than the original data. In practice, this means that by developing a PARAFAC model of a three-way LC–MS chromatographic data set (e.g. three-way array **Z** in Section 1.3) it is possible to resolve all the chromatographic peaks present in the data.

In order to develop a PARAFAC model, which describes the true underlying physicochemical variations of the chromatographic data, the model must be validated. A key consideration for developing a valid PARAFAC model is choosing the correct number of components. The number of components of the developed PARAFAC model for LC–MS chromatographic data sets correspond to the number of physicochemical variations (e.g., actual peaks, noise, baseline drifts) present in the data.

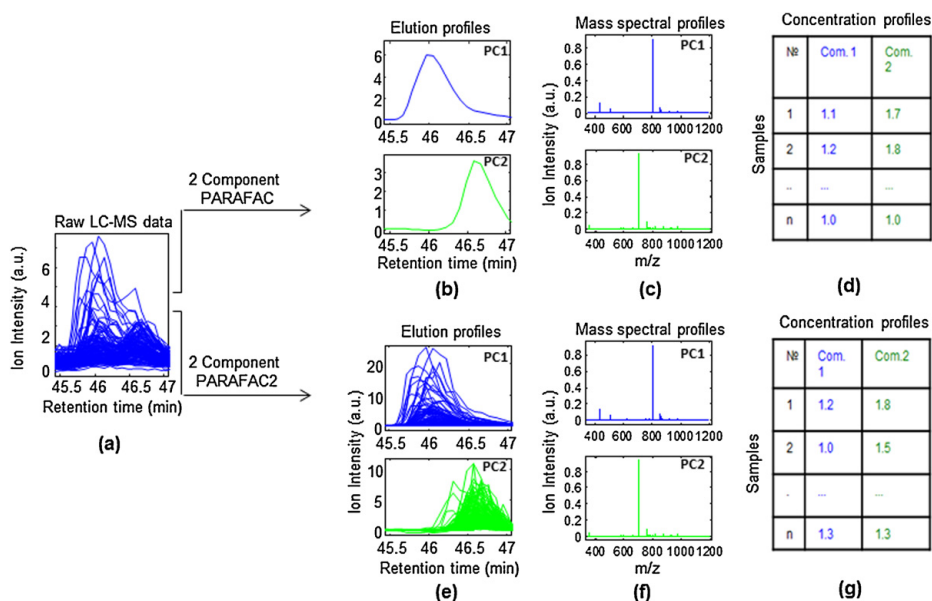
One of the advantages of applying PARAFAC for decomposing LC–MS data is that the original three-way structure of the data is maintained and the results can be directly interpreted. While PCA model cannot be developed for decomposing the three-way LC–MS data and requires data to be unfolded to form a table. The other advantage of PARAFAC models is that they provide a unique solution. The uniqueness of the PARAFAC models hold on that that the resolved elution profiles, spectral profiles and concentration profiles of the detected peaks reflect the original elution profiles (TIC chromatogram), mass spectral profiles and relative abundances of the peaks present in the data, respectively. One of the main prerequisite for the data to be modeled by PARAFAC is that the elution time of one chromatographic peak in a LC–MS data, should ideally be the same across the all samples. The problematic elution time shifts of the chromatographic peaks prevent a wider use of PARAFAC to model LC–MS data.

Fig. 3 shows the graphical illustration of the two-component PARAFAC and PARAFAC2 models of LC–MS data interval, which is investigated in this work. When the PARAFAC model of the raw LC–MS (Fig. 3a) data is developed, the model decomposes the data into one set of scores matrix and two sets of loadings matrices. Two sets of PARAFAC loadings correspond to the elution time profiles and the mass spectral profiles of resolved peaks, respectively (Fig. 3b and c). PARAFAC scores represent concentration profiles of resolved peaks, which correspond to the relative abundances of the peaks present in the chromatographic data (Fig. 3d). Elution profiles and the mass spectral profiles resolved for each component are characteristic for the individual metabolites and thus common to all samples, while the concentration profiles of the resolved peaks are sample-specific.

PARAFAC successfully been applied in many different areas of research. In chromatography it is used for detection of metabolites [30], in fluorescence spectroscopy PARAFAC modeling was performed for the analysis food samples [31]. Moreover, PARAFAC successfully been applied in combination with nuclear magnetic resonance (NMR) spectroscopy [32] and sensory data [33].

### 3.2. PARAFAC2

The PARAFAC2 model is less restrictive to the tri-linearity of the data [18,34] and allows successful modeling of LC–MS chromatographic data with elution time shifts of peaks across the samples. If the elution time shifts are relatively confined, the PARAFAC2 algorithm finds and models the shifted peaks of the same chemical compounds, profiting from the fact that they have identical mass spectral profile. In contrast to the PARAFAC model, the elution profiles resolved by PARAFAC2 (Fig. 3e) are sample-specific. PARAFAC2 resolves, as many elution profiles as there are samples, whereas PARAFAC only finds the common elution profile for all the



**Fig. 3.** Graphical illustration of two component PARAFAC and PARAFAC2 models of LC-MS chromatographic data interval investigated in this work. (a) Superimposed raw LC-MS chromatograms, (b) elution profiles resolved by PARAFAC, (c) spectral profiles resolved by PARAFAC and (d) concentration profiles (relative abundance of resolved peaks) calculated by PARAFAC model, (e) elution profiles resolved by PARAFAC2, (f) spectral profiles resolved by PARAFAC2 and (g) relative abundance calculated by PARAFAC2 model. \*Blue profiles describe the first component (PC1) of the PARAFAC and/or PARAFAC2 models and green profiles describe the second components (PC2) of the models. (For interpretation of references to color in this figure legend, the reader is referred to the web version of this article.)

samples. This feature of PARAFAC2 allows accurate quantification of the resolved peaks for each sample.

Thus, PARAFAC2 possess all the above-mentioned features of PARAFAC and, in addition, it can successfully model the three-way data, with a deteriorated tri-linear structure (e.g., elution time shifted peaks in LC-MS data). However, *calculating a global PARAFAC2 model of complex LC-MS data for resolving all the peaks present in the entire chromatogram is time consuming and requires computers with large operating memories.* In addition, such global PARAFAC2 model may not be representative for low *s/n* ratio peaks (minor metabolites), since the model will be mainly influenced by the major peaks. Moreover, due to the complexity of the data, it is difficult to find the optimal number of components for the model to describe the true underlying chemical information present in the data. Therefore, it is recommended to process complex chromatographic data by PARAFAC or PARAFAC2 in interval base [19]. This reduces the complexity of the data, and results in a parsimonious model that describes the raw data well. Finally, it is worth to mention that PARAFAC2 requires a minimum of preprocessing steps, and in most cases, the raw data directly modeled.

The PARAFAC2 can be considered as an alternative to the application of PARAFAC on data that has been thoroughly aligned [20]. PARAFAC2 successfully been applied in conjunction with GC-MS metabolomic data: resolution of elusive peaks from problematic regions of chromatograms [19,35–37], developing calibration models for precise quantification of estrogens [37], and identification of sources of oil spills [38]. In addition, PARAFAC2 is becoming increasingly used within different hyphenated analytical techniques such as HPLC-DAD [39–41], HPLC-IR [42] and HPLC-UV [43], and for monitoring pharmaceutical processes [44].

## 4. Results and discussion

### 4.1. Data analysis
















For reducing the complexity of the data, the LC-MS chromatographic profiles of the 127 F2 plant extracts were manually divided into 17 different chromatographic intervals. Each interval was confined by the presence of a baseline, and does not split peaks belonging to the same compound into different intervals (Fig. 4). Subsequently, each interval was modeled by PARAFAC2 individually.

In order to develop PARAFAC2 models, non-negativity constraints were applied in the spectral and sample modes of the three-way data, restricting the models to not find negative scores and/or loadings. Since in practice, mass spectral data and relative concentration values of the resolved peaks cannot be negative. This speed up the PARAFAC2 modeling and enabled obtaining unique PARAFAC2 models with spectral loadings reflecting the original mass spectral profiles of the raw data. Computation time of PARAFAC2 modeling depends on the complexity of the data, number of variations present in the data (components), settings of PARAFAC2 algorithm (e.g., non-negativity constraints, stop criteria, etc.), signal-to-noise ratio and capacity of the computer. As an example, computation time for developing a four component PARAFAC2 model of the biggest chromatographic data interval (interval number 4, Table 1) was 8.32 min when modeling was performed using a computer with Intel(R) Core(TM) i7 CPU processor, 12 GB RAM and 64-bit operating system. The performance of PARAFAC2 for modeling individually selected chromatographic intervals is presented in Table 1.

The concentration profiles obtained from PARAFAC2 models for each resolved peak were used for correlation and PLS regression

**Table 1**

The intervals of the chromatographic profiles modeled by PARAFAC2, their dimensions (the first mode is RT scan points, second mode is mass spectral dimension, and third mode is number of samples), optimal number of components and the explained variance of the PARAFAC2 models developed for each interval. <sup>1</sup>In intervals nos. 5, 7, 11, 12, 13, 14 and 15 some samples were identified as outliers and removed prior to the developing the final model.

Interval no.	Average elution profile	RT min	Dimensions of interval	Number of components	Explained variance (%)	Resolved peak numbers
1.		1.8–2.8	10 × 2123 × 127	4	89.7	1–4
2.		2.8–5.5	28 × 2123 × 127	5	90.4	5–9
3.		5.5–9.0	38 × 2123 × 127	3	91.9	10–12
4.		10.0–16.2	60 × 2123 × 127	4	88.8	13–16
5.		17.4–18.8	16 × 2123 × 113	2	86.9	17–18
6.		19.0–21.0	24 × 2123 × 127	2	96.5	19–20
7.		23.4–25.5	24 × 2123 × 121	4	92.2	21–24
8.		25.5–27.5	22 × 2123 × 127	7	90.1	25–31
9.		27.5–29.5	22 × 2123 × 127	7	90.3	32–38
10.		29.5–32.2	30 × 2123 × 127	6	87.7	39–44
11.		32.2–34.2	20 × 2123 × 123	4	83.7	45–48
12.		34.2–36.0	18 × 2123 × 122	5	79.2	49–53
13.		36.0–38.6	26 × 2123 × 114	6	67.4	54–59
14.		38.6–41.6	30 × 2123 × 115	3	94.8	60–62
15.		41.6–43.1	20 × 2123 × 115	4	76.8	63–66
16.		45.4–47.2	22 × 2123 × 127	3	91.8	67–69
17.		47.2–49.0	20 × 2123 × 127	2	81.4	70–71

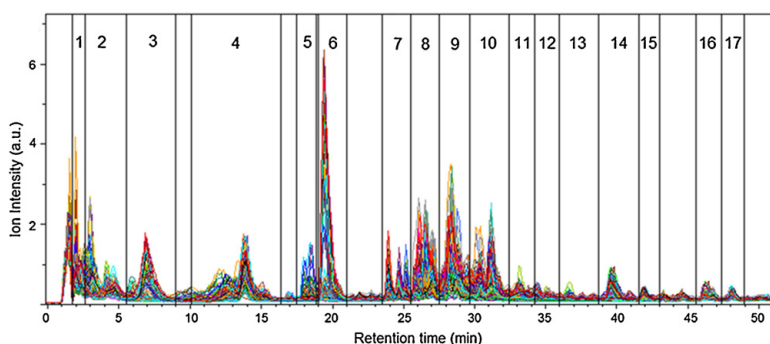


Fig. 4. The superimposed raw LC-MS chromatograms of the F2 plant samples and the 17 selected intervals.

analysis employing the resistance levels of 127 F2 plants and the relative abundances of the resolved peaks (metabolites). For better visual observation of the separation between the F2 plants based on the resolved LC-MS peaks, the samples were divided into three classes according to their resistance level (see Section 2.1).

#### 4.2. PARAFAC2 model validation

Model validation is an important step of PARAFAC2 analysis. Validation of PARAFAC2 models is illustrated in the example of the model developed for the chromatographic interval number 16. This interval (Table 1) corresponds to the region where one of the known saponin, oleanolic acid cellobioside, elutes. In order to develop a model which is able to describe the chemical information present in the data, one- to six-component PARAFAC2 models were fitted. The optimal number of components was determined based on: (1) the explained variance by the models and the residuals, (2) appearance of elution and spectral mode loadings, and (3) previous knowledge about the data (e.g. expected sources of main variations in the selected region of the chromatogram).

The resolved elution profiles of the PARAFAC2 models with two- and three-components closely reflect the raw chromatogram (Fig. 5). The PARAFAC2 model with three-components explained 91.8% of the variation in the data. The resolved MS spectral profiles for all three components were different and confirmed that they may correspond to three different chemical compounds (results not shown). The four-component model only explained additional 0.3%

variance (i.e. 92.1%) compared to the three-component model and the elution profiles differed from the original raw chromatogram (Fig. 5d). Moreover, the resolved elution profiles of component 1 and 2 (blue and green profiles, respectively Fig. 5d) strongly overlap and practically identical, and the score values obtained for these 2 components co-vary over samples. These results indicate that by extracting the fourth component, the PARAFAC2 model is forced to describe minor variations present in the mass spectra of few samples, and as a result, the same analyte variation is explained by two different components.

Thus, the three-component PARAFAC2 model was an optimal model for fitting this interval of the chromatogram. The first component of the model (blue profile, Fig. 5c) corresponded to oleanolic acid cellobioside, and its resolved mass spectral profile matched with experimentally measured mass spectrum of oleanolic acid cellobioside with a correlation based similarity of 98.5% (Fig. 6). As the LC gradient solvent contained sodium chloride (NaCl), and the mass spectra were recorded in positive mode, the base peak of 803.3  $m/z$  corresponds to the  $[M + Na]^+$  adduct ion of oleanolic acid cellobioside.

#### 4.3. Resolution of overlapped, elution time shifted and low $s/n$ peaks

The overlapping effect of chromatographic peaks is common in chromatography, and is due to the two or more analytes having the same or very close elution times at a given separation conditions.

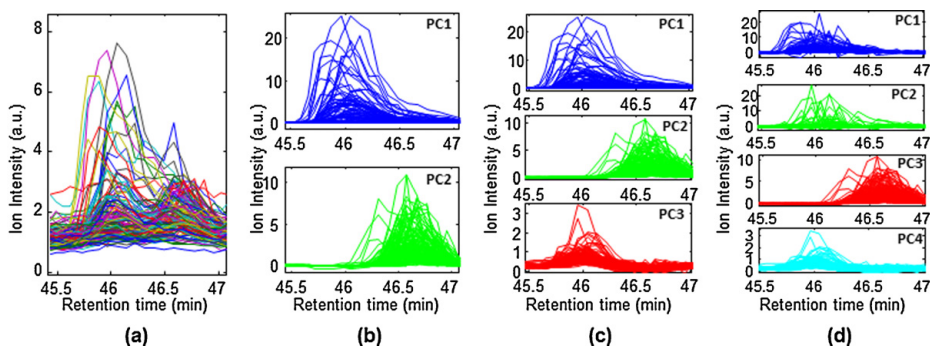
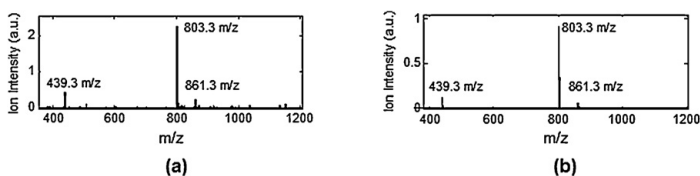


Fig. 5. (a) Superimposed raw LC-MS chromatograms of the F2 plant extracts in the chromatographic interval 16 (45.4–47.2 min), (b) resolved elution profiles of 2 component, (c) 3 component, and (d) 4 component PARAFAC2 models of the chromatographic interval. \*Blue, green, red and cyan color profiles correspond to the first (PC1), second (PC2), third (PC3), and fourth (PC4) components' elution profiles of PARAFAC2 models, respectively. (For interpretation of references to color in this figure legend, the reader is referred to the web version of this article.)





**Fig. 6.** (a) The mass spectrum of oleanolic acid cellobioside and (b) the mass spectral profile of oleanolic acid cellobioside peak resolved by PARAFAC2 (correlation based similarity >98%).

Resolution of overlapped peaks by PARAFAC2 is illustrated by the modeling the chromatographic interval number 15 (Fig. 7). Successive PARAFAC2 modeling of interval 15 revealed four component model to be an optimal for describing the data. The elution profiles of a four-component model (Fig. 7b) demonstrate the presence of overlapped peaks (blue, green, red and cyan elution profiles correspond to the component 1, 2, 3 and 4, respectively), which are cumbersome to visualize from the raw chromatogram (Fig. 7a).

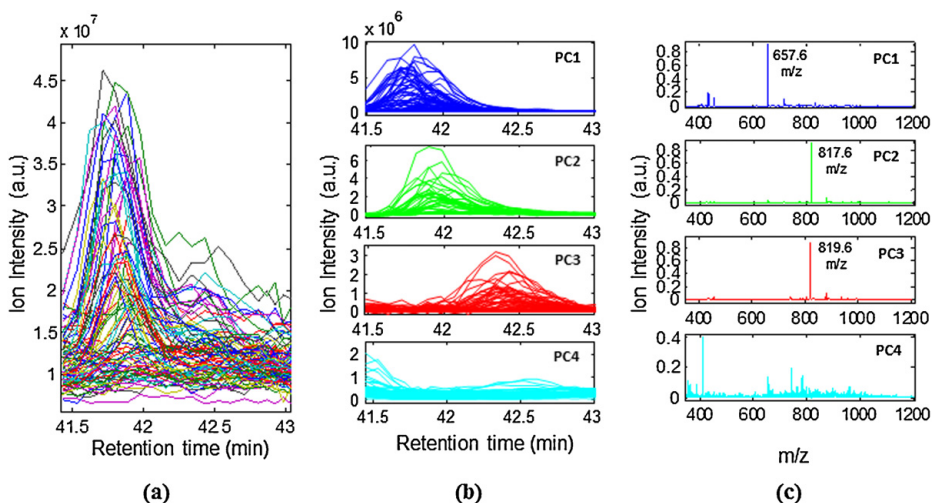
Fig. 7b shows the resolution of highly overlapped peaks (blue and green elution profiles), partly overlapped peaks (red elution profiles), as well as modeling of the baseline, which also partly describes the analyte eluted few seconds earlier (cyan elution profile). The second (green elution profile) and the third (red elution profile) components of the model correspond to known saponins found in *B. vulgaris*: gypsogenin cellobioside and epihederagenin cellobioside, respectively. The spectral profiles resolved by PARAFAC2 for all four components were different, and the two mass unit differences between gypsogenin cellobioside and epihederagenin cellobioside (gypsogenin cellobioside: 817.6  $m/z$ ,  $[M+Na]^+$  and epihederagenin cellobioside: 819.6  $m/z$ ,  $[M+Na]^+$ ) were also observed from the resolved mass spectral profiles (Fig. 7c).

Fig. 8a illustrates the raw chromatogram of interval number 12. By visual inspection of the raw chromatogram, it is difficult to assess the number of peaks present, and even more difficult to quantify them due to the low  $s/n$  ratio of the peaks. Five-component PARAFAC2 model fitted to this interval explained just below 80% of

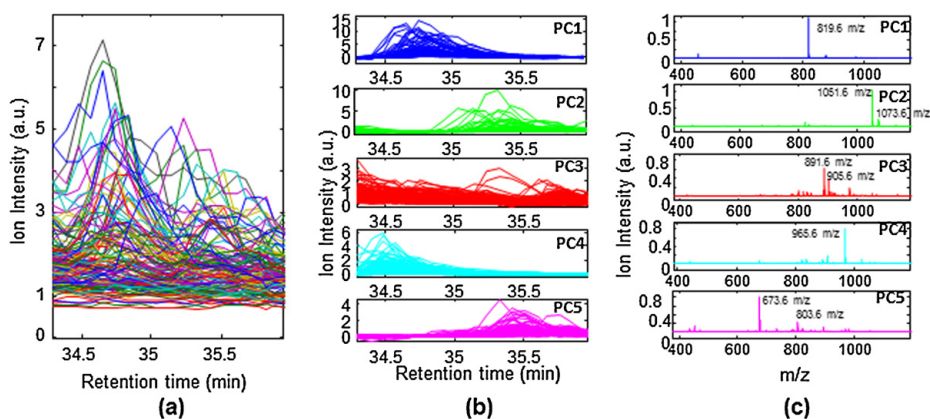
the variation present in the chromatographic data, and was able to resolve five distinct metabolite with different mass spectral profiles (Fig. 8c). In this case, the fourth and the fifth components of the model (cyan and magenta elution profiles, Fig. 8b) depict low  $s/n$  ratio peaks that were not visible and detectable from the raw data. In addition, Fig. 8 illustrates modeling of elution time shifted peaks (blue and green elution profiles, Fig. 8b), without any need for aligning the raw data. Though these peaks have significantly different elution times, they were modeled by PARAFAC2 as one component, since their mass spectral profiles were identical. Moreover, in the example of this chromatographic interval, it is possible to observe that the tail of the neighboring peaks, which eluted few seconds earlier, is modeled as the third component of the model (red elution profile, Fig. 8b).

#### 4.4. Resolved peaks of saponins and insect resistance

A total of 71 PARAFAC2 components (which correspond to actual peaks, including baselines and tails of neighboring peaks) were modeled from 17 different chromatographic intervals of LC–MS profiles (Table 1). In order to investigate the correlation between the resolved peaks from LC–MS chromatographic profiles of the F2 plants and their resistance against the flea beetle larvae, PARAFAC2 scores (relative abundance of metabolites) obtained for the resolved peaks were used for correlation and regression analyses (Fig. 9). In Fig. 9a, a bar plot shows correlation coefficients



**Fig. 7.** (a) Superimposed raw LC–MS chromatograms of the F2 plant extracts in the selected interval number 15, (b) resolved elution profiles and (c) corresponding mass spectral profiles of the four-component PARAFAC2 model developed for the chromatographic interval. \*Blue, green, red and cyan color profiles correspond to the first (PC1), second (PC2), third (PC3) and the fourth (PC4) components of the PARAFAC2 model. (For interpretation of references to color in this figure legend, the reader is referred to the web version of this article.)



**Fig. 8.** (a) Superimposed raw LC-MS chromatograms of the F2 plant extracts in the selected interval number 12, (b) resolved elution profiles and (c) corresponding mass spectral profiles of the five-component PARAFAC2 model developed for the chromatographic interval. \*Blue, green, red, cyan and magenta color profiles correspond to the first, second, third, fourth and the fifth components of the PARAFAC2 model. (For interpretation of references to color in this figure legend, the reader is referred to the web version of this article.)

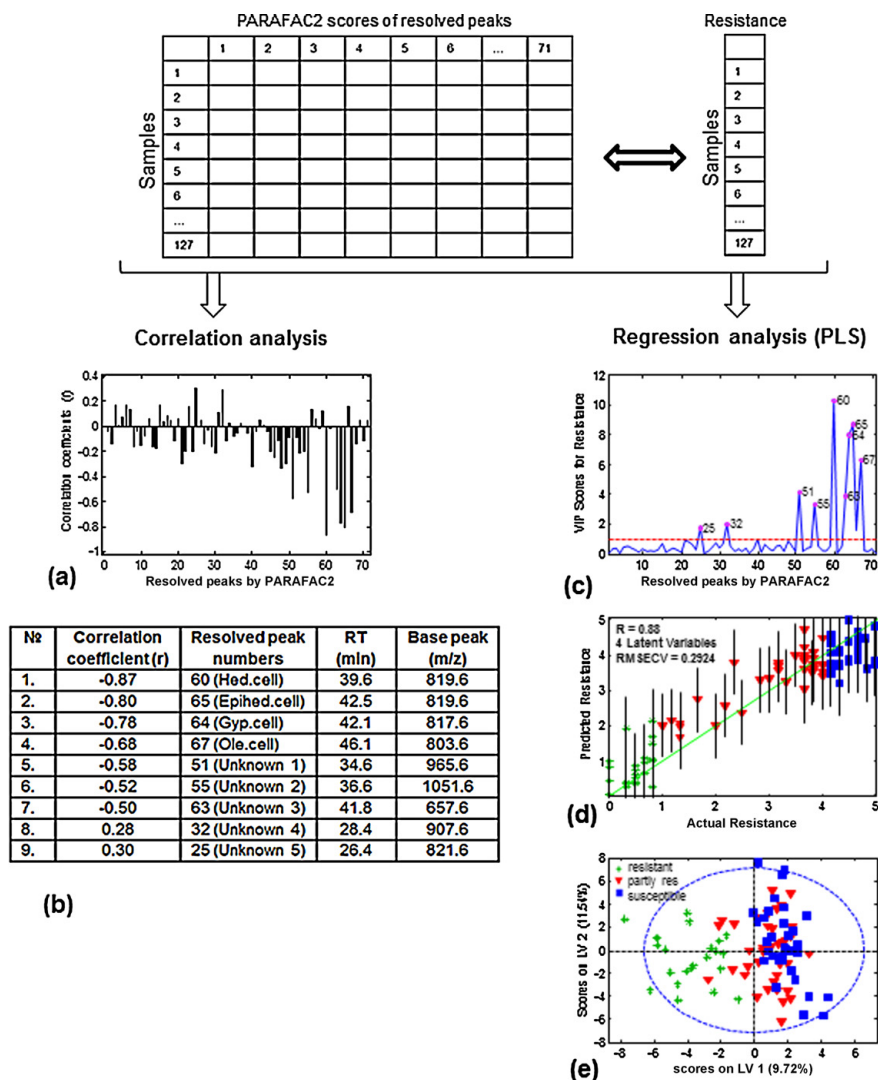
between the PARAFAC2 resolved peaks and the resistance level of F2 plants. Resistance level of F2 plants vary between zero and five, where most resistant plants are close to zero and susceptible plants are close to five (see Section 5.1). Therefore, negative correlation coefficients of the resolved peaks mean that they are positively correlated to the defense and vice versa. Correlation analysis showed that nine PARAFAC2 components (resolved peaks) out of 71 components were significantly correlated to the plants' resistance. The most positively correlated peaks to the plants' defense corresponded to the four previously known triterpenoid saponins of *B. vulgaris* (hederagenin cellobioside, oleanolic acid cellobioside, epihederagenin cellobioside, and gypsogenin cellobioside) (Fig. 9a, and b). One-way ANOVA (analysis of variance) test between PARAFAC2 scores of resolved peaks and resistance level of F2 plants confirmed significance of the correlations found from correlation and PLS analysis. *P*-values of nine mostly correlated peaks found for hypothesis test for null correlation were  $P < 0.0001$ .

The peak of *hederagenin cellobioside* modeled as the first component of the PARAFAC2 model of chromatographic interval number 14 (Table 1) and its complete mass spectral and elution profiles were successfully resolved (results not shown). Likewise, the peak of *oleanolic acid cellobioside* is modeled as the first component of the PARAFAC2 model of chromatographic interval number 16, and the resolved mass spectral and elution profiles are illustrated in Figs. 5 and 6. The second and the third components of the PARAFAC2 model of chromatographic interval number 15, represent the peaks of *gypsogenin cellobioside* and *epihederagenin cellobioside*, respectively (green and red elution profiles, Fig. 7b). Their resolved mass spectral profiles matched with the experimentally recorded mass spectra of these saponins (Fig. 7c).

Another five peaks that depict significant correlation were unknown compounds (Fig. 9-b). *Unknown 1*, which has a base peak of 965.6 *m/z* was found as the fourth component of the PARAFAC2 model of chromatographic interval 12 and it was the fifth most positively correlated peak to the resistance level. The resolution of *unknown 1* is shown in Fig. 8, where the cyan color profiles represent elution and mass spectral profiles (Fig. 8b and c). This peak was almost invisible from the raw LC-MS chromatogram (Fig. 8a), due to the overlapping effect. *Unknown 1* eluted 11.5 min earlier than oleanolic acid cellobioside. This indicates that it is a more polar compound than oleanolic acid cellobioside, since the LC gradient

program starts with a polar solvent, and the polarity of solvent gradually decreased over time (Section 5.2). Thus, *unknown 1* might be a new triterpenoid saponin of *B. vulgaris* with three sugar (hexose, 162 *m/z*) moieties attached to the aglycone with the same molecular mass as oleanolic acid (455.5 *m/z*). This assumption was further confirmed by its fragmentation pattern in LC-MS/MS experiment separately performed on parental G-type *B. vulgaris*. The LC-MS/MS approach assisted to fragment the precursor 965.6 *m/z* ions into its product ions and to record MS<sup>2</sup> and MS<sup>3</sup> spectra, which illustrated the loose of three hexose (162 *m/z*) moieties and formation of 455.5 *m/z* ion at MS<sup>3</sup> spectra (results not shown).

In contrast to the compounds listed in Fig. 9b, which are positively correlated to resistance, *unknown 4* and *unknown 5* depicted the highest positive correlation to the susceptibility level of F2 plants, and they were more abundant in susceptible plants compared to the resistant plants. These two peaks eluted much earlier (at RT 28.4 and 26.4 min., respectively) than the other highly correlated peaks resolved by PARAFAC2. This indicates that these compounds are more polar compared to the other unknown compounds. *Unknown 5* was modeled as the first component of the PARAFAC2 model of chromatographic interval 8, and its resolved mass spectral profile showed *m/z* signals at 821.6 *m/z* (base peak) and 983.6 *m/z*. The mass difference between these two *m/z* signals is 162 *m/z*, which corresponds to a sugar (hexose) moiety. The base peak of *unknown 5* (821.6 *m/z*) is 2 *m/z* unit higher than the base peak of hederagenin cellobioside (819.6 *m/z*). This indicates that *unknown 5* is most probably a triterpenoid saponin with three sugar (hexose) moieties attached to an aglycone, which is 2 *m/z* unit higher (473.5 *m/z*) than hederagenin (471.5 *m/z*). Thus, the aglycone structure of *unknown 5* may be identical to the structure of hederagenin, but the double bond present in position C12 is saturated (Fig. 2). This was further substantiated by LC-MS/MS experiments performed in negative mode on parental G- and P-type *B. vulgaris*. Visual observation of MS, MS<sup>2</sup> and MS<sup>3</sup> spectra revealed that the precursor ion with the molecular mass of 960 *m/z* loses three hexose moieties during fragmentation, and that the aglycone ion with the molecular mass of 473.5 *m/z* appeared in MS<sup>3</sup> spectra (results not shown). It is worth to mention that the intensity of this peak was much higher in P-type plants (the susceptible plant) than in G-type, which corroborates with our findings from PARAFAC2 modeling. The relative abundance of *unknown 5* determined by



**Fig. 9.** Correlation and PLS regression analysis between PARAFAC2 concentration profiles of 71 resolved peaks and the resistance level of F2 plants. (a) The bar plot shows the correlation coefficients of resolved peaks by PARAFAC2. (b) Table of nine most correlated resolved peaks, sorted according to their correlation coefficients ( $r$ ). (c) VIP scores of the 71 variables found by the PLS model. (d) Predicted versus measured plot of the PLS model shows the correlation between the actual and the predicted resistance of F2 plants found by the PLS regression model. (e) The score plot (LV1 vs LV2) of the PLS model shows separation between resistant (green) and susceptible (blue) F2 plants. (For interpretation of references to color in this figure legend, the reader is referred to the web version of this article.)

PARAFAC2 were 5 fold higher in susceptible F2 plants compared to the resistant F2 plants, which also reflects the positive correlation of *unknown 5* to flea beetle larvae susceptibility.

In order to investigate the relevance between the relative concentrations of the resolved metabolites and defense of F2 plants, a PLS regression model was constructed with PARAFAC2 determined relative abundance of resolved peaks as an X matrix, and resistance level of the F2 plants as a y vector (Fig. 9). The PLS regression analysis confirmed the importance of the nine peaks resolved by PARAFAC2 which were found in correlation analysis as the most significantly correlated to the F2 plants' resistance. Variable importance in

projection (VIP) scores obtained from the PLS model (Fig. 9c) for each variable (PARAFAC2 resolved peak) illustrate that only those nine variables have VIP scores higher than 1 (the threshold defined by the PLS model). This demonstrates that these nine variables are the most important variables for predicting the resistance level of the F2 plants. Predicted versus measured plot of the PLS model (Fig. 9d) shows the correlation ( $r$ ) of 0.88 between the actual resistance and the predicted resistance level found by the PLS model for each F2 plant. These findings confirm that the peaks resolved by PARAFAC2 from the LC–MS profiles were informative and correlated to the F2 plants' resistance against flea beetle

larvae. Moreover, a score plot of the PLS model illustrated a good separation between resistant and susceptible F2 plants (Fig. 9e). In addition, the loading plot of the PLS model showed a separation between the variables which are positively (left extreme of the loading plot, e.g. hederagenin cellobioside and oleanolic acid cellobioside) and negatively (right extreme of the loading plot, e.g. unknown 4 and unknown 5) correlated to the resistance level of F2 plants. This confirms the importance of the variables for predicting the resistance level and discriminating resistant and susceptible F2 plants (results not shown).

Finally, it is worth to mention that Kuzina et al. [22] detected 30 metabolites, which might be candidate metabolites for F2 plants' defense against the flea beetle larvae. In this study, PARAFAC2 resolved and quantified 28 out of those 30 metabolites, in addition to other 40 metabolites that did not show significant correlation to the plants' resistance against the insect. Moreover, PARAFAC2 resolved elusive peaks of unknown 4 and unknown 5 which have not been previously detected. The other triterpenoid saponin of *B. vulgaris* cochalic acid cellobioside [23] which did not correlate with defense, was also modeled by PARAFAC2 as the first component of the model of chromatographic interval 10.

## 5. Conclusions and remarks

The performance of PARAFAC2 modeling to resolve complex LC–MS profiles of plant metabolomic data was assessed and demonstrated. Complex and problematic chromatographic peaks with elution time shifts, strong overlaps, baseline drifts and low *s/n* peaks were successfully modeled by PARAFAC2, without any need for preprocessing the raw data. PARAFAC2 separately modeled all the peaks present in selected intervals of the LC–MS chromatograms, and enabled precise quantification of the relative abundance of the resolved peaks based on their area. This method of quantification is more accurate and unbiased than quantifying peaks based on the intensity of one or more marker *m/z* ion signals. A total of 71 PARAFAC2 components (which correspond to actual peaks, baselines and tails of neighboring peaks) were modeled from 17 different chromatographic intervals of the LC–MS data obtained from 127 F2 *B. vulgaris* plants. The concentration profiles of nine peaks exhibited a high correlation to the resistance level of F2 plants to flea beetle larvae. Four of these nine peaks had previously been identified as triterpenoid saponins of *B. vulgaris* and their resolved mass spectral profiles matched well with the experimentally obtained mass spectra. The remaining five peaks were saponin-like unknown compounds and based on the PARAFAC2 resolved mass spectral profiles, their structures could only be preliminarily assessed. All these features of PARAFAC2 finely illustrated its power for quantitative and qualitative analysis of complex LC–MS metabolomic data.

## Acknowledgements

We would like to thank our colleagues Miss. Gözde Gürdeniz (Department of Food Science, University of Copenhagen), Dr. Jörg M. Augustin (Department of Plant Biology and Biotechnology, University of Copenhagen) and Dr. Vera Kuzina (Department of Plant

Biology and Biotechnology, University of Copenhagen) for their helpful discussions and valuable comments on the manuscript. Faculty of Science is acknowledged for support to the elite-research area "Metabolomics and bioactive compounds" with a PhD stipendium to Bekzod Khakimov.

## References

- [1] O. Fiehn, *Plant Mol. Biol.* 48 (2002) 155.
- [2] J.W. Allwood, R.C.H. De Vos, A. Moing, C. Deborde, A. Erban, J. Kopka, R. Goodacre, R.D. Hall, *Methods Enzymol.* 500 (2011) 299.
- [3] V. Kuzina, J.K. Nielsen, J.M. Augustin, A.M. Torp, S. Bak, S.B. Andersen, *Phytochemistry* 72 (2011) 188.
- [4] H. Winning, N. Viereck, B. Wollenweber, F.H. Larsen, S. Jacobsen, I. Sondergaard, S.B. Engelsen, *J. Exp. Bot.* 60 (2009) 291.
- [5] K. Saito, M.Y. Hirai, K. Yonekura-Sakakibara, *Trends Plant Sci.* 13 (2008) 36.
- [6] J.M. Amigo, N. Ratola, A. Alves, *Atmos. Environ.* 45 (2011) 5988.
- [7] H.S. Son, G.S. Hwang, K.M. Kim, H.J. Ahn, W.M. Park, F. Van Den Berg, Y.S. Hong, C.H. Lee, *J. Agric. Food Chem.* 57 (2009) 1481.
- [8] J.H. Wu, X.H. Wang, Y.H. Yi, K.H. Lee, *Bioorg. Med. Chem. Lett.* 13 (2003) 1813.
- [9] J.W. Allwood, R. Goodacre, *Phytochem. Anal.* 21 (2010) 33.
- [10] J.M. Amigo, T. Skov, R. Bro, *Chem. Rev.* 110 (2010) 4582.
- [11] N.P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, *J. Chromatogr. A* 805 (1998) 17.
- [12] T. Skov, F. Van Den Berg, G. Tomasi, R. Bro, *J. Chromatogr. A* 1217 (2010) 484.
- [13] G. Tomasi, F. Savorani, S.B. Engelsen, *J. Chromatogr. A* 1218 (2011) 7832.
- [14] H.F.M. Boelens, R.J. Dijkstra, P.H.C. Eilers, F. Fitzpatrick, J.A. Westerhuis, *J. Chromatogr. A* 1057 (2004) 21.
- [15] S. Yang, J.S. Nadeau, E.M. Humston-Fulmer, J.C. Hoggard, M.E. Lidstrom, R.E. Synovec, *J. Chromatogr. A* 1240 (2012) 156.
- [16] D. Bylund, R. Danielsson, G. Malmquist, K.E. Markides, *J. Chromatogr. A* 961 (2002) 237.
- [17] R.A. Harshman, *UCLA Working Papers in Phonetics* 22 (1972) 30.
- [18] R. Bro, C.A. Andersson, H.A.L. Kiers, *J. Chemometr.* 13 (1999) 295.
- [19] J.M. Amigo, M.J. Popielarz, R.M. Callejon, M.L. Morales, A.M. Troncoso, M.A. Petersen, T.B. Toldam-Andersen, *J. Chromatogr. A* 1217 (2010) 4422.
- [20] T. Skov, J.C. Hoggard, R. Bro, R.E. Synovec, *J. Chromatogr. A* 1216 (2009) 4020.
- [21] S. Heinisch, E. Lesellier, C. Podevin, J.L. Rocca, A. Tchaplá, *Chromatographia* 44 (1997) 529.
- [22] G. Gürdeniz, M. Kristensen, T. Skov, L.O. Dragsted, *Metabolites* 2 (2012) 77.
- [23] T. Shinoda, T. Nagao, M. Nakayama, H. Serizawa, M. Koshioka, H. Okabe, A. Kawai, *J. Chem. Ecol.* 28 (2002) 587.
- [24] N. Agerbirk, C.E. Olsen, B.M. Bibby, H.O. Frandsen, L.D. Brown, J.K. Nielsen, J.A.A. Renwick, *J. Chem. Ecol.* 29 (2003) 1417.
- [25] V. Kuzina, C.T. Ekstrom, S.B. Andersen, J.K. Nielsen, C.E. Olsen, S. Bak, *J. Plant Physiol.* 151 (2009) 1977.
- [26] N.J. Nielsen, J. Nielsen, D. Staerk, *J. Agric. Food Chem.* 58 (2010) 5509.
- [27] R.A. Harshman, *UCLA Working Papers in Phonetics* 16 (1970) 1.
- [28] R. Bro, *Chemometr. Intell. Lab. Syst.* 38 (1997) 149.
- [29] R. Bro, *Multi-way analysis in the food industry: Models, Algorithms, and Applications*, PhD Thesis, University of Amsterdam, 1998.
- [30] D. Arroyo, M.C. Ortiz, L.A.S. Arabia, F. Palacios, *J. Chromatogr. A* 1187 (2008) 1.
- [31] J. Christensen, L. Norgaard, R. Bro, S.B. Engelsen, *Chem. Rev.* 106 (2006) 1979.
- [32] M. Dyrby, M. Petersen, A.K. Whittaker, L. Lambert, L. Norgaard, R. Bro, S.B. Engelsen, *Anal. Chim. Acta* 531 (2005) 209.
- [33] M. Cocchi, R. Bro, C. Durante, D. Manzini, A. Marchetti, F. Sacconi, S. Signinolfi, A. Ulrici, *Food Qual. Prefer.* 17 (2006) 419.
- [34] H.A.L. Kiers, J.M.F. Ten Berge, R. Bro, *J. Chemometr.* 13 (1999) 275.
- [35] J.M. Amigo, T. Skov, J. Coello, S. Maspoch, R. Bro, *Trends Anal. Chem.* 27 (2008) 714.
- [36] Y. Xu, W. Cheung, C.L. Winder, W.B. Dunn, R. Goodacre, *Analyst* 136 (2011) 508.
- [37] I. Garcia, L. Sarabia, M.C. Ortiz, J.M. Aldama, *Anal. Chim. Acta* 526 (2004) 139.
- [38] D. Ebrahimi, D.B. Hibbert, *J. Chromatogr. A* 1198 (2008) 181.
- [39] F. Marini, A. D'Aloise, R. Bucci, F. Buiarelli, A.L. Magri, A.D. Magri, *Chemometr. Intell. Lab. Syst.* 106 (2011) 142.
- [40] I. Garcia, M.C. Ortiz, L. Sarabia, J.M. Aldama, *Anal. Chim. Acta* 587 (2007) 222.
- [41] M. Vosough, A. Salemi, *Food Chem.* 127 (2011) 827.
- [42] K. Istvan, R. Rajko, G. Keresztury, *J. Chromatogr. A* 1104 (2006) 154.
- [43] J.M.M. Leitao, J.C.G.E. da Silva, *Chemometr. Intell. Lab. Syst.* 89 (2007) 90.
- [44] S. Matero, S. Poutiainen, J. Leskinen, S.P. Reinikainen, J. Ketolainen, K. Jarvinen, A. Poso, *Chemometr. Intell. Lab. Syst.* 96 (2009) 88.



# Paper 2

**Bekzod Khakimov**, Mohammed Saddik Motawie, Søren Bak, Søren Balling Engelsen

The use of trimethylsilyl cyanide derivatization for robust and broad spectrum high-throughput gas-chromatography-mass spectrometry based metabolomics

*Analytical and Bioanalytical Chemistry (2013) In press, DOI:  
10.1007/s00216-013-7341-z*



# The use of trimethylsilyl cyanide derivatization for robust and broad-spectrum high-throughput gas chromatography–mass spectrometry based metabolomics

Bekzod Khakimov · Mohammed Saddik Motawia ·  
Søren Bak · Søren Balling Engelsen

Received: 22 July 2013 / Revised: 22 August 2013 / Accepted: 2 September 2013  
© Springer-Verlag Berlin Heidelberg 2013

**Abstract** Reproducible and quantitative gas chromatography–mass spectrometry (GC-MS)-based metabolomics analysis of complex biological mixtures requires robust and broad-spectrum derivatization. We have evaluated derivatization of complex metabolite mixtures using trimethylsilyl cyanide (TMSCN) and the most commonly used silylation reagent *N*-methyl-*N*-(trimethylsilyl)trifluoroacetamide (MSTFA). For the comparative analysis, two metabolite mixtures, a standard complex mixture of 35 metabolites covering a range of amino acids, carbohydrates, small organic acids, phenolic acids, flavonoids and triterpenoids, and a phenolic extract of blueberry fruits were used. Four different derivatization methods, (1) direct silylation using TMSCN, (2) methoximation followed by TMSCN (M-TMSCN), (3) direct silylation using MSTFA, and (4) methoximation followed by MSTFA (M-MSTFA) were compared in terms of method sensitivity, repeatability, and derivatization reaction time. The derivatization methods were observed at 13 different derivatization times, 5 min to 60 h, for both metabolite mixtures. Fully automated sample derivatization and injection enabled excellent repeatability and precise

method comparisons. At the optimal silylation times, peak intensities of 34 out of 35 metabolites of the standard mixture were up to five times higher using M-TMSCN compared with M-MSTFA. For direct silylation of the complex standard mixture, the TMSCN method was up to 54 times more sensitive than MSTFA. Similarly, all the metabolites detected from the blueberry extract showed up to 8.8 times higher intensities when derivatized using TMSCN than with MSTFA. Moreover, TMSCN-based silylation showed fewer artifact peaks, robust profiles, and higher reaction speed as compared with MSTFA. A method repeatability test revealed the following robustness of the four methods: TMSCN>M-TMSCN>M-MSTFA>MSTFA.

**Keywords** Metabolomics · Gas chromatography–mass spectrometry · Trimethylsilyl derivatization · Methoximation · Trimethylsilyl cyanide

## Abbreviations

BSA	Bis(trimethylsilyl)acetamide
BSTFA	<i>N,O</i> -Bis(trimethylsilyl)trifluoroacetamide
EI	Electron impact
HCN	Hydrogen cyanide
MEOX	Methoxiamine
M-MSTFA	Methoximation followed by MSTFA-based silylation
M-TMSCN	Methoximation followed by TMSCN-based silylation
MPS	Multi-purpose sampler
MSTFA	<i>N</i> -methyl- <i>N</i> -(trimethylsilyl)trifluoroacetamide
PARAFAC2	Parallel Factor Analysis 2
PCA	Principal component analysis
RI	Retention index

**Electronic supplementary material** The online version of this article (doi:10.1007/s00216-013-7341-z) contains supplementary material, which is available to authorized users.

B. Khakimov · S. B. Engelsen  
Quality and Technology, Department of Food Science,  
Faculty of Science, University of Copenhagen, Rolighedsvej 30,  
Frederiksberg C, 1958 Copenhagen, Denmark

B. Khakimov (✉) · M. S. Motawia · S. Bak  
Plant Biochemistry, Department of Plant and Environmental  
Sciences, Faculty of Science, University of Copenhagen,  
Thorvaldsensvej 40, Frederiksberg C, 1871 Copenhagen, Denmark  
e-mail: bzo@life.ku.dk



TMCS	Trimethylchlorosilane
TMS	Trimethylsilyl
TMSCN	Trimethylsilyl cyanide

## Introduction

Gas chromatography–mass spectrometry (GC-MS) has become one of the favorite analytical platforms applied in metabolomics because of its high reproducibility and resolution power [1–3]. In contrast to NMR and LC-MS, GC-MS analysis requires metabolites to be thermally stable and volatile during analysis. To lower the boiling point of metabolites and increase volatility for GC-MS analysis, complex biological samples, such as plant and animal tissue extracts and biofluids need to be derivatized by chemical derivatization to improve or facilitate their detection. The most commonly used derivatization method involves methoximation followed by silylation [4–6]. During the methoximation step, metabolites with carbonyl ( $>C=O$ ) functional group react with the reagent (20–40 mg mL<sup>-1</sup> solutions of *O*-methylhydroxylamine hydrochloride in pyridine) and form oxime ( $>C=N-O-CH_3$ ) derivatives [7–9]. The main purpose of oximation is to form thermally stable derivatives that prevent cyclization of reducing sugars, formation of keto-enol tautomers of aldehydes and ketones with a proton in the  $\alpha$ -position, and to protect other carbonyl group containing metabolites from decarboxylation [10–12]. Silylation serves to substitute active hydrogen atoms of hydroxyl ( $-OH$ ), carboxylic acid ( $-COOH$ ), primary and secondary amines ( $R'NH_2$ ,  $R'R''NH$ ), and thiols ( $-SH$ ) with a trimethylsilyl ( $-Si(CH_3)_3$ ) group [13].

In GC-MS metabolomics, different silylation reagents have been applied and they differ by their reactivity, selectivity, side reactions, and byproducts [13–16]. The efficiency and speed of the silylation reaction depend on reaction temperature, time, and physicochemical properties of both the silylation reagent and the substrate. Among silylation reagents, *N*-methyl-*N*-(trimethylsilyl)trifluoroacetamide (MSTFA) has become the most commonly applied reagent because of its high reactivity towards a broad range of compound classes. Several studies on GC-MS method optimization suggest the use of MSTFA alone or together with 1 % of TMCS as a catalyst for silylation of complex mixtures [5, 17, 18]. Silylation reactions are reversible and have been shown to proceed via bimolecular nucleophilic substitution ( $S_N2$ -mechanism) at the silicon atom [19–22].

As silylation follows a  $S_N2$ -mechanism, the silylation depends on concentration of both the electrophile (silylation reagent) and the nucleophile (substrate). Besides concentration, reaction time, and temperature, silylation reaction yield

and rate depend on (1) nature of the leaving group of the silylation reagent, (2) chemistry of the substrate that reacts with silylation reagent, (3) steric effects, (4) influence of solvents, and (5) presence of catalysts. The general silylation activity order of the different functional groups is as follows: alcohols (primary > secondary > tertiary) > phenols > carboxylic acids > thiols > amines (primary > secondary) > amides. This is due to a number of factors. The silylation rate increases by increasing the affinity of the silicon atom to the nucleophile center of the substrate being silylated. The affinity of the silicon atom is highest to an oxygen atom [23, 24] and trimethylsilylation reagents with good leaving groups are more active in exchanging a trimethylsilyl group with an active hydrogen atom or a metal atom [19, 25]. In addition, the nucleophilic attack by a substrate on the electrophile silicon atom becomes easier when the covalent bond between the silicon atom and the leaving group is weak and easily dissociable [20, 23]. Solvents also influence the rate and the mechanism of silylation reactions. Pyridine is the mostly used solvent during methoximation and silylation reactions, and as it is a weak base, it can increase silylation reaction rate by scavenging active protons ( $H^+$ ) in the reaction mixture. Accessibility of the nucleophilic center of the substrate is crucial, and studies have demonstrated an influence of steric effects in silylation of branched secondary amines and primary amines [26]. To date, the most thoroughly studied silylation reagents are those with a Si–N bond, and silylated amides such as *N,O*-bis(trimethylsilyl)acetamide, *N,O*-bis(trimethylsilyl)trifluoroacetamide, and MSTFA have become common silylation reagents in GC-MS analysis. By contrast, very few studies have been conducted on silylation capabilities of compounds with a Si–C bond, such as trimethylsilyl cyanide (TMSCN). Two recent reviews on derivatization reagents and their reactions used for GC-MS analysis of a broad range of metabolites discuss the advantages and limitations of a variety of different derivatization reagents [13, 21]. However, TMSCN was not included as an alternative trimethylsilylation reagent. We show that TMSCN possesses several advantages over the majority of the reagents used up to date, including a high silylation reactivity and reproducibility.

In the present study, the use of TMSCN for derivatization of complex metabolite mixtures is demonstrated. In organic chemistry, TMSCN is mainly used as a source of cyanide group for various synthetic reactions [27]. Although Mai and Patil [28] have shown high silylation reactivity of TMSCN toward many functional groups, its potential as a derivatization reagent in comprehensive GC-MS analysis of metabolites mixtures has not been thoroughly studied. The only published, to our knowledge, example of the use of TMSCN for GC-MS analysis was published in the early 1990s, where it was used for the derivatization of the

prostacyclin analog I [29]. In this study, we compared GC-MS analysis of various classes of compounds silylated by using TMSCN and MSTFA.

Mai and Patil [28] have conducted a comprehensive study on silylation of alcohols, phenols, carboxylic acids, amines, and thiols using TMSCN. In the study, they showed that TMSCN outperformed many other reagents and provided mild and rapid derivatization. The study showed that under mild conditions: 5 min at 25 °C for most of alcohols, phenols, and carboxylic acids and 5–30 min at 25–100 °C for secondary amines, thiols, and carbohydrates, TMSCN-based silylation reaction yield reached up to 98 %. The study illustrated higher silylation efficiency and the reaction rate of TMSCN when compared with other alkyl cyanide derivatization reagents towards sterically hindered functional groups due to its relatively smaller molecular size. In addition, they showed high silylation reaction yield with neat (solvent free) TMSCN compared with the use of solvent since all metabolites were readily soluble and rather neutral pH of TMSCN enables non-destructive silylation of base sensitive compounds.

Byproduct formation is one of the limitations of most silylation reagents. Byproducts are formed during the silylation reactions and may result in formation of multiple artifact peaks that decrease profile reproducibility, hamper the detection of early eluting metabolites, and degradation of silylated metabolites. For example, one of the most commonly used silylation reagent in conjunction with MSTFA is TMCS that form hydrogen chloride as a byproduct, which is an aggressive acid towards TMS-derivatives. Hydrogen cyanide (HCN) is the only byproduct formed in TMSCN-based silylation. HCN is too weak an acid to hydrolyze the TMS-derivatized products, but by contrast, it can protonate TMSCN that lead to increased electrophilicity and thus serves to further increase silylation efficiency.

The purpose of this study is to assess the silylation capabilities of TMSCN towards various classes of metabolites that are often detected in GC-MS metabolomic studies of complex biological samples, and to compare with the silylation efficiency of the mostly used reagent MSTFA. Two metabolite mixtures, a standard mixture that compiled 35 different compounds including amino acids, carbohydrates, small organic acids, phenolic acids, flavonoids and triterpenoids, and a phenolic extract of blueberry fruits were used to evaluate the silylation capabilities of TMSCN and MSTFA. TMSCN and MSTFA silylation performances were evaluated in conjunctions with a methoximation step (using pyrimidine as a solvent) and without methoximation (direct silylation) where only the reagent itself acted as a solvent. Four different derivatization methods: (1) direct silylation using TMSCN, (2) methoximation followed by TMSCN (M-TMSCN), (3) direct silylation using MSTFA, and (4) methoximation followed by

MSTFA (M-MSTFA) were evaluated at 11–13 different silylation time points in the range of 5 min to 60 h.

## Materials and methods

### Preparation of metabolite standard mixture and extraction of blueberry fruits

Individual solutions of 35 standard compounds containing 6 amino acids (valine, serine, threonine, glycine, aspartic acid, and phenylalanine), 6 carbohydrates (ribitol, ribose, glucose, glucose-6-phosphate, maltose, and sucrose), 13 organic acids and phenolic compounds (benzoic acid, succinic acid, malic acid, palmitic acid, phenyllactic acid, 4-hydroxybenzoic acid, 2-hydroxy-2-methoxybenzoic acid, vanillic acid, 2-hydroxycinnamic acid, *p*-coumaric acid, caffeic acid, 4-hydroxiacetophenone, and vanillin), 2 polyphenols (naringenin and catechin), and 8 triterpenes (cholesterol,  $\beta$ -amyirin,  $\alpha$ -amyirin, lupeol, oleanolic acid, hederagenin, betulinic acid, and  $\alpha$ -epoxi- $\beta$ -amyirin) were prepared in the concentration of 2 or 0.25 mg mL<sup>-1</sup> (only for triterpenes). Most of the standard compounds were soluble in water, apart from polyphenols, caffeic acid, benzoic acid, 2-hydroxy-3-methoxybenzoic acid, palmitic acid, and triterpenes which were solubilized in dimethyl sulfoxide. Vanillic acid, *p*-coumaric acid and cholesterol were solubilized in 96 % ethanol. A metabolite standard mixture was prepared by combining 200- $\mu$ L aliquots of all the solutions of standard compounds, besides solutions of triterpenes, which were added in double amount (400  $\mu$ L). This standard mixture was later used for GC-MS analysis. Phenolic and organic acids of blueberry fruits were extracted essentially according to the protocol described by Zadernowski et al. [30]. To increase the extraction yield, slight modifications were introduced to the protocol. Blueberry fruits of low-bush blueberry (*Vaccinium myrtillus*) were purchased from the grocery shop Irma (Copenhagen, Denmark); 100 g of frozen blueberry fruits were extracted five times with 100 mL of 80 % (vol/vol) methanol at room temperature for 40 min by using an orbital shaker at 500 rpm. The extracts were centrifuged at 3,000 $\times$ g for 5 min, and the clear supernatants combined in 1,000 mL round bottom flask and dried using a rotary vacuum evaporator followed by freeze-drying; 1.5 g of freeze-dried extract was dissolved in 100 mL of 4 M sodium hydroxide, hydrolyzed under the nitrogen gas atmosphere for 4 h at room temperature while mixing at 300 rpm by using a magnetic stirrer. Then the solution was acidified with 6 M hydrochloric acid to pH 2, and extracted five times with diethyl ether (1:1, vol/vol) for 15 min while mixing at 300 rpm by using magnetic stirrer. To clean the ether extract from fatty acids and other nonpolar compounds, the combined ether fractions were dried using rotary vacuum evaporator, re-dissolved in

150 mL of 5 % (wt/vol) sodium bicarbonate solution, and then extracted five times with diethyl ether as described above. The remaining water phase was acidified with 6 M hydrochloric acid to pH 2, extracted five times with diethyl ether, the combined ether fractions dried finally re-dissolved in 17 mL of 80 % (vol/vol) methanol. The resulting extract referred as *phenolic extract A* mainly contains free phenolic and organic acids as well as phenolic acids derived from hydrolysis of ester bonds. The blueberry *phenolic extract B* was obtained in essential the same way as phenolic extract A, except that hydrolysis of 1.5 g of freeze-dried extract was performed using 100 mL of 2 M hydrochloric acid and stirring for 40 min at 94 °C. Beside the free phenolic acids, organic acids, and phenolic acids derived from hydrolysis of ester bonds, this extract also contain the phenolic acids derived from the hydrolysis of glycosidic bonds.

## Chemicals

All the compounds used for preparation of standard mixture and solvents were purchased from Sigma-Aldrich, except for 4-hydroxiacetophenone, hydrochloric acid (37 %), and sodium bicarbonate that were obtained from Merck, in best available quality. Dimethyl sulfoxide, diethyl ether, 2-hydroxy-3-methoxybenzoic acid, trimethylsilyl cyanide and *N*-methyl-*N*-(trimethylsilyl)trifluoroacetamide were purchased from Fluka in best available quality. Triterpenoid 12 $\alpha$ ,13 $\alpha$ -epoxy-3 $\beta$ -hydroxyoleanane was synthesized by M. S. Motawia. Water used throughout the study was purified using a Millipore Milli-Q lab water system equipped with 0.35  $\mu$ m filter membrane.

## Sample derivatization

Derivatization and injection of samples was fully automated by use of a GERSTEL MultiPurpose Sampler (MPS) with DualRait WorkStation integrated to a GC-MS system from Agilent. The MPS enabled reproducible sample derivatization in a high-throughput manner and was fully operated by MAESTRO software integrated with Agilent's ChemStation software. This enabled automation of individual and parallel sample derivatization steps from a single sequence developed for the analysis of several samples independently from chromatographic system and provided precise derivatization time control of each sample. Two types of syringes were installed, the left MPS was equipped with a 10- $\mu$ L syringe and used only for GC-MS injection, whereas the right MPS was equipped with a 100- $\mu$ L syringe and used in sample derivatization steps. Prior to derivatization, 100- $\mu$ L aliquot of the blueberry extract and 70- $\mu$ L aliquot of standard mixture samples were lyophilized in 150- $\mu$ L glass inserts under reduced pressure at room temperature by use of a vacuum centrifuge, transferred into 1.5 mL GC-MS vials, and sealed with magnetic caps with silicone septum under nitrogen gas to prevent

moisturization. Golden magnetic caps (ML 33032A from [www.mikrolab.dk](http://www.mikrolab.dk)) with silicon septum enabled MPS to move GC-MS vials and prevented solvent and/or reagent evaporation during the derivatization even after penetration of the septum by MPS needles. Lyophilized and sealed samples were placed on a MPS sample tray and further sample handling was fully automated.

Methoximation of samples was performed by addition of 40  $\mu$ L freshly prepared 20 mg mL<sup>-1</sup> methoxiamine hydrochloride (CH<sub>3</sub>ONH<sub>2</sub>·HCl) in pyridine and incubated for 90 min at 30 °C by agitation at 750 rpm. After methoximation, samples were silylated by addition of 40  $\mu$ L silylation reagent, and a total of 80  $\mu$ L reaction mixture was incubated at 37 °C by agitation at 750 rpm. To keep the volume of the reaction mixture constant in all four derivatization methods, direct silylation methods, TMSCN and MSTFA were performed by addition of 80- $\mu$ L pure silylation reagent and incubated at 37 °C by agitation at 750 rpm. Based on the reaction stoichiometry, the amount of applied silylation reagents exceeded at least 400 (TMSCN), 200 (M-TMSCN), 250 (MSTFA), and 125 (M-MSTFA) times the amount needed for silylation of all the available active hydrogen atoms present in the 70- $\mu$ L aliquot of standard mixture. GC-MS profiles of the standard mixture was evaluated at 11 different silylation times by using all four derivatization methods, whereas derivatization of blueberry extract was performed at 13 different silylation times by using three (TMSCN, M-TMSCN, and M-MSTFA) derivatization methods. Important practical considerations of the automated sample derivatization, GC-MS analysis are described in detail in the Electronic supplementary material (ESM; text).

## Data acquisition

An aliquot of 1.0  $\mu$ L derivatized sample was injected either in split (split ratio of 3:1 was used in blueberry extract analysis) or in splitless mode (for analysis of standard mixture) into a Gerstel cooled injection system (CIS) equipped with a glass wool packed liner. Detailed settings of left and right MPS syringes, sample incubating agitator, and CIS injection port parameters can be found in the ESM (text). The GC-MS consisted of an Agilent 7890A GC and an Agilent 5975C series MSD. GC separation was performed on an Agilent HP-5MS column (30 m  $\times$  250  $\mu$ m  $\times$  0.25  $\mu$ m) by using hydrogen carrier gas at the constant flow rate of 1.2 mL min<sup>-1</sup>. The GC oven temperature program was as follows: initial temperature, 60 °C; equilibration time, 1.0 min; heating rate, 12.0 °C min<sup>-1</sup>; end temperature, 310 °C; hold time, 6.0 min; and post-run time, 5 min at 60 °C. Mass spectra were recorded in the range of 50–750 *m/z* with a scanning frequency of 2.3 scans s<sup>-1</sup>, and the MS detector was switched off during the 3-min solvent delay time. The transfer line, ion source, and quadrupole temperatures were set to 280, 230, and 150 °C,

respectively. The mass spectrometer was tuned according to the manufacturer recommendations by using perfluorotributylamine.

#### Data analysis

Relative abundances of metabolites were calculated by Parallel Factor Analysis 2 (PARAFAC2) modeling of the raw GC-MS data [31, 32]. The method allowed precise quantification of well-resolved, co-eluted, overlapped, and low S/N ratio peaks using full mass spectra or marker  $m/z$  ions (for completely embedded peaks) of the metabolites. The outcome of PARAFAC2 models were (1) scores of each resolved peak that correspond to the relative areas of the detected peaks, (2) spectral loadings of each resolved peak which represent pure mass spectra of the detected peaks, and (3) elution time loadings that represent elution time profiles of peaks. For explorative analysis of the derivatization methods, principal component analysis (PCA) [33] was applied to a matrix containing relative abundances of detected metabolites (variables) in different derivatization methods (rows). Metabolites were identified based on their retention indices (RI) and EI-MS library match using commercial Wiley08 and NIST05 libraries as well as an in-house library of triterpenes. EI-MS library search was performed using original mass spectra or PARAFAC2 resolved mass spectra of chromatographic peaks. Retention indices of each metabolite were calculated using in-house MATLAB function based on the Van den Dool and Kratz equation [34] and retention times of C10–C40 alkanes that were analyzed using the same GC-MS method.

#### Software

GC-MS chromatographic data was analyzed using Agilent Technologies ChemStation software (version: E.02.02.1431). PARAFAC2 and PCA modeling were performed by using PLS Toolbox (Version 6.0.1, Eigenvector Research Inc. USA) working under MATLAB (Version 7.13.0.564, R2011b, the Mathworks Inc., USA) environment. CDF files of raw GC-MS data were imported into MATLAB using the function [35] which is available on [www.models.ku.dk](http://www.models.ku.dk).

#### Safety considerations

All derivatization reagents (including MSTFA and TMSCN) are highly toxic chemicals. Accordingly, derivatization reagents must be handled under the inert gas atmosphere in a fume hood. High attention must be paid to avoid contact with moisture and sun light. Appropriate vial lids must be tested for their ability to seal the vials and to prevent evaporation of the reagent during the derivatization. Derivatized samples and reagents must be removed from the autosampler shortly after injection and should be disposed according to the instructions provided by the suppliers. Users must be aware of the safety

precautions, prevent moisture that causes formation of byproducts in harmful concentrations. It is worth to mention that one of the possible byproducts of the TMSCN is hydrogen cyanide. Derivatization must be performed under inert gas if manual handling is necessary, if an autosampler is employed tightly sealed vials that prevent evaporation of both, the reagent and byproducts, should be used, and finally disposal of leftover silylation reagent and vials with silylated samples after analysis as a non-recyclable solutions. More detailed safety considerations of using TMSCN and MSTFA as well as reagent evaporations tests are described in detail in the [ESM \(text\)](#).

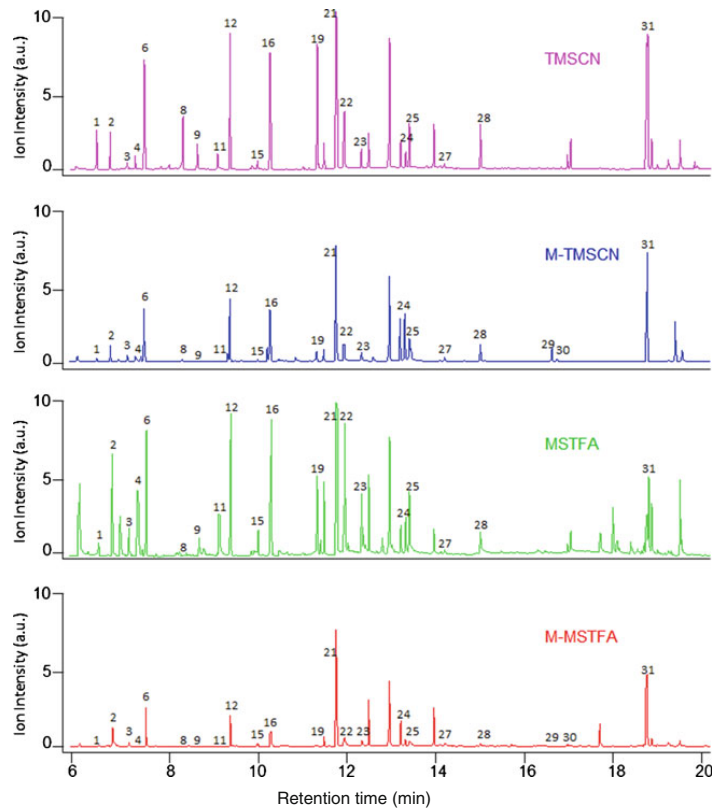
## Results and discussion

### Derivatization method comparison of standard mixture

#### Global analysis

The GC-MS profiles obtained from the four derivatization methods differed, both qualitatively and quantitatively, and they were significantly influenced by silylation time. The optimal silylation time of each of the derivatization method was defined as the time of incubation at which a maximum number of metabolites reached their highest peak intensity. Optimal silylation times of the derivatization methods were as follows: TMSCN, 40 min; M-TMSCN, 60 min; MSTFA, 30 min; and M-MSTFA, 60 min. Figure 1 illustrates total ion chromatograms of the standard mixture at the optimal silylation time of each method. Table 1 lists 41 derivatives that originated from 35 metabolites of the standard mixture and their relative ratios at the optimal silylation time of each metabolite in the four different derivatization methods. Derivatization products of all the 35 compounds used in the standard mixture, using both the methoxiamination followed by trimethylsilylation (M-TMSCN and M-MSTFA) and direct trimethylsilylation (TMSCN and MSTFA) methods, are highlighted in the ESM, Table S1. The repeatability of the derivatization methods was calculated from the relative standard deviations of abundances of 41 derivatives measured in four replicates, for each derivatization method at their optimal silylation times. The mean errors of the derivatization methods TMSCN, M-TMSCN, MSTFA, and M-MSTFA were 3.8 % (varying from 1.2 to 10.1 % for all metabolites), 6.4 % (1.2 to 17.8 %), 26.2 % (7.7 to 41.6 %), and 13.9 % (2.4 to 30.1 %), respectively. To evaluate the significance of the differences observed in the four different derivatization methods, PCA analysis was performed on the replicate data matrix (16×41) that include 16 samples, 4 replicates per method, and abundances of 41 TMS-derivatives listed in Table 1. Subsequently, ANOVA analysis was performed to evaluate the  $F$  statistics (variations between treatments/variations within treatments)

**Fig. 1** Total ion current chromatograms of GC-MS data obtained from the complex standard mixture using the four different derivatization methods at the optimal silylation times. Peaks are numbered in the same order as presented in the ESM, Table S1



and the  $p$  values by using the scores of PC1 and PC2 that explained more than 75 % of the variation. Figure S1 (ESM) shows the PC1 versus PC2 scores plot of the PCA analysis and the corresponding box plot of the ANOVA analysis performed on PC1 and PC2. This suggests rejection of the null hypothesis ( $p < 0.01$ ) with 95 % of the confidence and shows the significance of the differences observed between the four derivatization methods. For an exploratory evaluation of the four different derivatization methods at different silylation time points, two PCA models were developed. The first PCA model was developed on a  $X(56 \times 34)$  metabolomic data that compiled 34 metabolites detected in all 56 samples (four derivatization methods evaluated at eleven different silylation times, including replicates at the optimal derivatization times). The scores and loadings plot of the PCA model is displayed in Fig. 2. The PC1 versus PC2 scores plot of the PCA model reveal a partial separation of the derivatization methods (Fig. 2a), where samples of each method form its trajectory from the left to the right of the plot by increasing silylation time. Samples that are located to the very right side of the plot represent the silylation time points when TMS-

derivatives reached their highest intensities (optimal silylation time). As PC1 versus PC2 loadings plot of the model show that all metabolites are located in the right side of the plot having positive loadings in PC1 (Fig. 2b). However, further increase of silylation time resulted in decrease of metabolites' intensities, thus after optimal silylation time, samples are moved back to the left side of the scores plot. Around the optimal silylation times, the relative abundances of most of the metabolites were higher when using TMSCN and MSTFA methods compared with M-TMSCN and M-MSTFA, and accordingly, they showed higher scores in PC1 (Fig. 2a). The loadings plot also shows a partial separation of variables that assisted to compare different derivatization methods for detection of various metabolites by visual observation of scores and loadings plots. Most of the organic and phenolic acids are clustered on the upper right side of the plot having positive loadings in PC2, while triterpenes and metabolites with more than one exchangeable hydrogen atom (e.g., polyphenols, sucrose-8tms, ribitol-5tms, serine-3tms, glycine-3tms, and threonine-3tms) have negative loadings in PC2 and thus forms clusters on the lower right side of the plot. A

**Table 1** Trimethylsilyl (TMS) and methoxime-trimethylsilyl (MEOX-TMS) derivatives derived from 35 compounds of standard mixture sorted according to their retention time

No.	Substance	TMSCN <sup>a</sup>	M-TMSCN <sup>a</sup>	MSTFA <sup>a</sup>	M-MSTFA <sup>a</sup>	R <sub>I</sub> <sup>b</sup>	EI-MS <sup>c</sup>
1	Valine-2TMS	66.93 (40)	4.56 (60)	37.55 (30)	(150)	1,208	90
2	Benzoic acid-1TMS	3.50 (50)	2.05 (60)	5.31 (50)	(150)	1,242	94
3	Serine-2TMS	1.73 (10)	1.66 (40)	26.27 (10)	(30)	1,251	93
4	Threonine-2TMS	0.49 (20)	0.49 (40)	13.46 (20)	(60)	1,289	92
5	Glycine-3TMS	2.46 (150)	1.37 (30)	1.86 (20)	(150)	1,302	92
6	Succinic acid-2TMS	2.18 (40)	1.93 (150)	2.32 (40)	(300)	1,308	95
7	Serine-3TMS	29.34 (40)	1.10 (40)	0.54 (40)	(60)	1,353	95
8	Threonine-3TMS	12.86 (50)	1 (150)	0.42 (60)	No	1,375	95
9	Aspartic acid-2TMS	23.41 (50)	2.07 (150)	21.88 (40)	(150)	1,419	94
10	4-hydroxyacetophenone-1TMS	3.94 (40)	2.24 (150)	9.83 (20)	(150)	1,464	89
11	Malic acid-3TMS	1.94 (30)	1.59 (150)	2.25 (50 min)	(300)	1,489	91
12	Vanillin-1TMS	0.42 (150)	No	1 (30)	No	1,535	91
13	Phenylalanine-1TMS	34.51 (50)	0.69 (150)	58.55 (40)	(150)	1,548	86
14	Phenyllactic acid-2TMS	1.83 (50)	1.55 (150)	2.04 (20 min)	(150)	1,584	90
15	4-hydroxybenzoic acid-2TMS	1.92 (30)	1.48 (150)	3.43 (40)	(150)	1,626	94
16	Vanillin-MEOX-1TMS	No	2.23 (60)	No	(300)	1,648	91
17	2-hydroxy-3-methoxybenzoic acid-2TMS	5.36 (40)	4.48 (150)	4.32 (30)	(150)	1,692	94
18	( <i>trans</i> )-ribose-MEOX-4TMS	No	3.01 (60)	No	(150)	1,699	87
19	Ribitol-5TMS	1.41 (50)	1.16 (60)	1.35 (20)	(150)	1,745	98
20	Vanillic acid-2TMS	4.48 (50)	2.23 (150)	5.78 (30)	(150)	1,768	96
21	2-hydroxycinnamic acid-2TMS	2.35 (40)	2.16 (150)	7.62 (30)	(150)	1,811	95
22	( <i>trans</i> )-glucose-MEOX-5TMS	No	2.97(60)	No	(150)	1,925	89
23	<i>p</i> -coumaric acid-2TMS	4.23 (40)	1.77 (150)	5.74 (30)	(300)	1,940	88
24	( <i>cis</i> )-glucose-MEOX-5TMS	No	3.62 (60)	No	1 (60)	1,948	89
25	Palmitic acid-1TMS	1.31 (50)	1.86 (60)	1.80 (30)	(60)	2,044	91
26	( <i>trans</i> )-caffeic acid-3TMS	2.31 (30)	1.52 (150)	2.93 (30)	(150)	2,151	93
27	( <i>trans</i> )-glucose-6-phosphate-MEOX-4TMS	No	4.50 (150)	No	(150)	2,375	88
28	( <i>cis</i> )-glucose-6-phosphate-MEOX-4TMS	No	3.62 (150)	No	(150)	2,395	88
29	Sucrose-8TMS	2.32 (30)	2.15 (300)	1.60 (30)	(300)	2,709	90
30	( <i>trans</i> )-maltose-MEOX-8TMS	No	4.51 (300)	No	(300)	2,814	87
31	( <i>cis</i> )-maltose-MEOX-8TMS	No	4.61 (300)	No	(150)	2,843	87
32	Naringenin-3TMS <sup>d</sup>	7.85 (40)	2.86 (150)	6.11 (20)	(150)	2,905	85
33	Catechin-5TMS <sup>d</sup>	1.77 (150)	1.35 (30)	1.14 (30)	(40)	2,926	86
34	Cholesterol-1TMS	2.17 (40)	2.33 (150)	4.20 (30)	(60)	3,176	97
35	$\beta$ -amyrin-1TMS	1.74 (50)	1.92 (150)	1.81 (30)	(150)	3,402	98
36	$\alpha$ -amyrin-1TMS	2.04 (50)	2.07 (150)	1.94 (30)	(150)	3,443	99
37	Lupeol-1TMS	1.44 (60)	1.61 (150)	1.11 (50)	(150)	3,454	98
38	Oleanolic acid-2TMS	1.50 (40)	2.07 (150)	1.67 (30)	(300)	3,665	99
39	Betulonic acid-2TMS	2.47 (50)	2.16 (150)	2.04 (40)	(150)	3,687	99
40	$\alpha$ -epoxi- $\beta$ -amyrin-1TMS	4.02 (50)	3.43 (60)	2.58 (30)	(150)	3,755	97
41	Hederagenin-3TMS	1.83 (50)	1.93 (150)	1.29 (50)	(150)	3,790	99

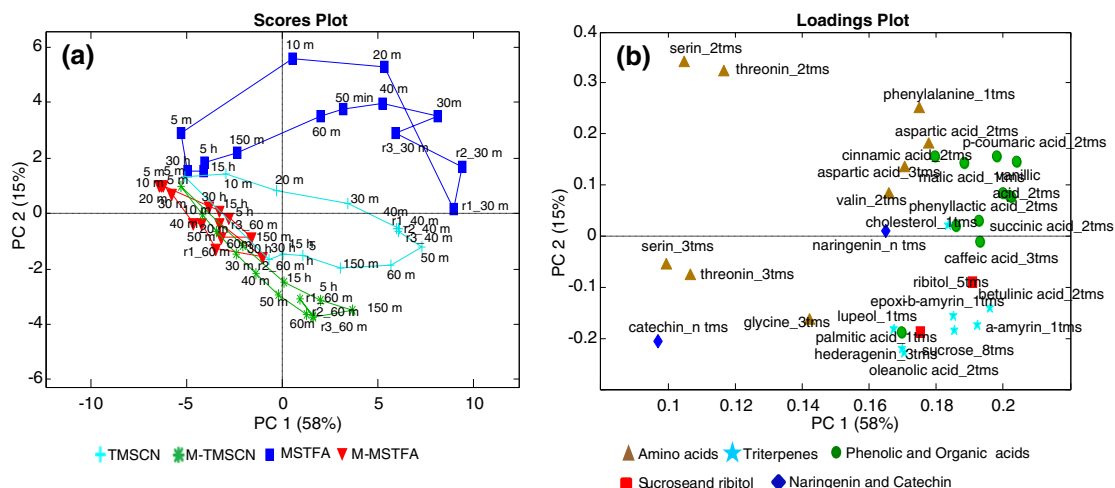
No no peak was observed

<sup>a</sup> PARAFAC2-based relative abundances of derivatives at their optimal silylation times in minutes (indicated in the brackets), for each derivatization method are illustrated as their ratio to the relative abundance of the corresponding derivatives in method M-MSTFA

<sup>b</sup> Retention indices of metabolite derivatives were calculated based on Van den Dool and Kratz equation by using retention times of C10–C40 alkanes

<sup>c</sup> EI-MS-based library match of the metabolites by using Wiley08 and NIST05

<sup>d</sup> PARAFAC2-based quantification of derivatives included only characteristic *m/z* ion but not full mass spectra



**Fig. 2** Scores (a) and loadings (b) plots of PCA model developed on a matrix containing PARAFAC2 scores (relative abundances) of TMS derivatives of the standard mixture detected by four derivatization methods, at 12 different silylation time points (including four replicates at the optimal silylation times)

visual observation of scores and loadings plot suggests that detection of triterpenes and metabolites with several exchangeable hydrogens is more efficient when using TMSCN and M-TMSCN than with MSTFA and M-MSTFA. This may be due to the relatively smaller molecular size of the TMSCN reagent than MSTFA, which enable a more rapid and efficient silylation of sterically hindered exchangeable hydrogen atoms of sucrose (hydroxyl groups at C2'), catechin, and amino acids (e.g., glycine-3tms and serine-3tms). By contrast, detection of some of the phenolic and organic acids was slightly more efficient when using MSTFA, suggesting that MSTFA samples are located on the upper right side of the scores plot. To compare the silylation capacity of TMSCN and MSTFA towards sterically hindered metabolites 2,6-diphenylphenol was analyzed at various silylation time points. This showed a rapid silylation with both reagents. Nevertheless, a considerable difference was observed in silylation time and efficiency. With TMSCN silylation, a maximum intensity of the TMS-product was observed at 5 s, whereas with MSTFA, a maximum was reached after 5 min and only reached an intensity of about 80 % of the TMSCN signal (see ESM, Fig. S2).

The scores plot of the first PCA model also assisted in evaluation of the derivatization methods' repeatability. The sums of the inner distances (Euclidean distances) of the four replicates to the center of a cluster that they form were calculated for each derivatization method. This parameter provided a measure of the compactness of a cluster of replicates. Euclidean distances of replicates of derivatization methods increased in following order: TMSCN (1.2) < M-TMSCN (4.4) < MSTFA (8.9) < M-MSTFA (18.4), and showed that the TMSCN method was the most reproducible while MSTFA was the least. This observation was in agreement with the

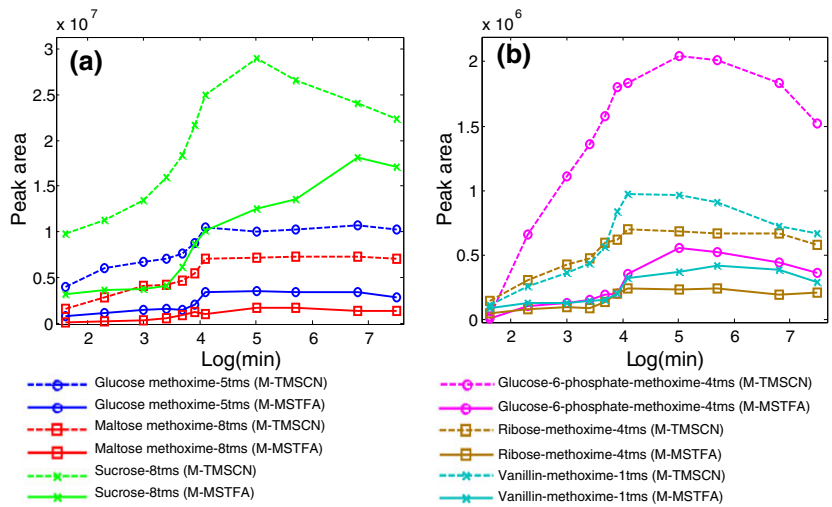
repeatability test of the methods calculated based on the relative standard deviations of metabolites.

The second PCA model compared M-TMSCN method with M-MSTFA (28 samples) using all common metabolites including methoxime-trimethylsilylated (MEOX-TMS) derivatives of reducing sugars (41 variables). The scores plot of this PCA model (see ESM, Fig. S3) showed better separation of the M-TMSCN samples from the M-MSTFA samples when compared with the previous PCA model. The samples corresponding to the M-TMSCN method had significantly higher scores on the PC1 than M-MSTFA samples and showed a general trend in development of derivatization reactions over different silylation times. The loadings plot of the model also showed higher loadings of all variables in PC1 compared with PC2. These results suggest that the M-TMSCN method outperformed M-MSTFA in terms of silylation reaction speed, efficiency, and repeatability.

#### Derivatization of sugars

A comparison of the M-TMSCN and M-MSTFA methods for GC-MS detection of MEOX-TMS derivatives of reducing sugars, vanillin and the TMS-derivative of sucrose at silylation time range of 5 min to 30 h is shown in Fig. 3. In both methods, the relative peak abundances of metabolites increase gradually to reach their maximum between a silylation time of 60 min to 5 h. At the optimal silylation time, abundances of all metabolites were 1.5–6 times higher in the M-TMSCN method compared with M-MSTFA. Accordingly, the M-TMSCN method outperformed M-MSTFA in terms completeness and the silylation reaction rates of mono- and disaccharides, including the sterically hindered C2' hydroxyl

**Fig. 3** Relative peak abundances of methoxime trimethylsilyl derivatives (MEOX-TMS) of glucose, maltose, glucose-6-phosphate, ribose, vanillin, and TMS derivative of sucrose in four different derivatization methods (TMSCN, M-TMSCN, MSTFA, and M-MSTFA) over silylation time range of 5 min to 30 h. Relative peak abundances of TMS derivatives were calculated using PARAFAC2 modeling and ln(min) scale of silylation time was used for better visualization



groups of sucrose and maltose. At the optimal silylation times the peak abundances of *trans*- and *cis*-glucose-MEOX-5TMS derivatives were 3- to 4-fold higher when using the M-TMSCN method as compared with the M-MSTFA method. Similarly, at the optimal derivatization time the peaks of *trans*- and *cis*-glucose-6-phosphate-MEOX-4TMS were 3.6- to 4.5-fold higher in the case of the M-TMSCN method as compared with the M-MSTFA method. Likewise, in the cases of *trans*- and *cis*-maltose-MEOX-8TMS, methoximation followed by TMSCN (M-TMSCN) outperformed M-MSTFA, as the detected derivatives peak abundances were up to 4.6 times higher at the optimal silylation times.

To evaluate the derivatization efficiency of carbohydrates, e.g., a number of derivatives and their ratios, six different sugars were individually derivatized by the four different methods (see ESM, Figs. S11, S12, S13, and S14). Direct silylation using either TMSCN or MSTFA resulted in four derivatives of glucose,  $\alpha$ - and  $\beta$ -glucopyranose-5TMS (represented 90 % in a ratio of 1:1) and  $\alpha$ - and  $\beta$ -glucofuranose-5TMS, whereas only two derivatives, *cis* and *trans* isomers of glucose-methoxime-5TMS were detected using M-TMSCN or M-MSTFA. Three sucrose TMS-derivatives were detected using all derivatization methods. Methoxime sucrose derivatives were not observed since sucrose is a non-reducing sugar and cannot react with the methoximation reagent. Of the three sucrose TMS derivatives, sucrose-8TMS was the most abundant one and eluted last while the two earlier eluting peaks correspond to sucrose TMS-derivatives with a lower silylation level. With increasing sucrose silylation time, the ratios between the TMS-derivatives changed in all derivatization methods, and in the case of TMSCN, sucrose was fully converted to the sucrose-8TMS derivative, at the silylation time of 150 min. By contrast, MSTFA was not able to fully silylate sucrose, and even after 15 h silylation, all three

sucrose derivatives were present. In conclusion, direct silylation of sucrose with TMSCN and MSTFA derivatization was more efficient in terms of reaction rate and detection as compared with the M-TMSCN and M-MSTFA methods.

#### Derivatization of triterpenes

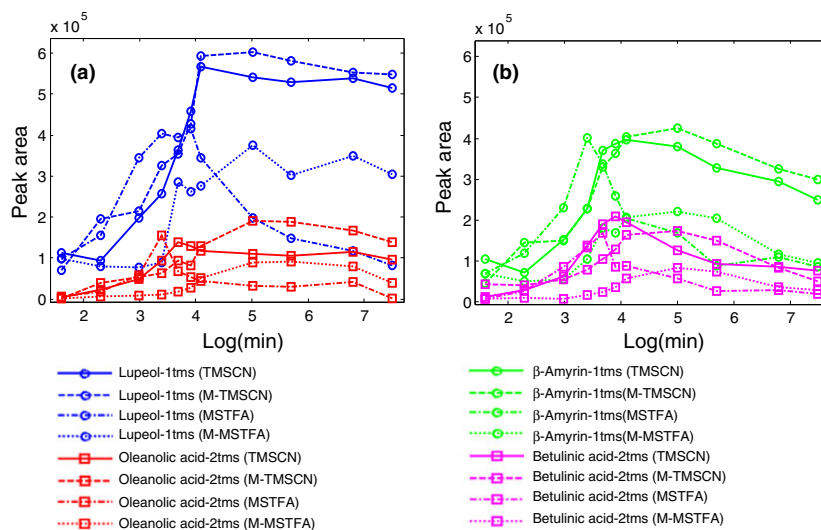
At the optimal silylation times, detection of silylated triterpenes was comparable when using TMSCN and MSTFA methods, while the M-TMSCN method has significantly outperformed the M-MSTFA and abundances of peaks were 1.4–3.4 times higher (Fig. 4; Table 1). TMSCN-based silylation depicted a maximum peak intensities of TMS-derivatives of lupeol,  $\beta$ -amyrin, and oleanolic acid after 50–60 min of derivatization and peak intensities remained stable even after the silylation time of 15 h. Whereas, in the case of MSTFA-based silylation metabolite peaks reached their maximum at 30 min of silylation and further increase of silylation time resulted in a significant decrease of the peak intensities.

#### Derivatization of polyphenols, organic and amino acids

Relative abundances of caffeic, malic, *p*-coumaric, and phenyllactic acids' TMS-derivatives were two to six times higher using M-TMSCN compared with M-MSTFA method in the silylation time range of 10 min to 15 h (see ESM, Fig. S4). These TMS-derivatives reached maximum abundances within the first 2.5 h of silylation and after 15 h of derivatization, peak abundances start to decrease, presumably due to degradation of the TMS derivatives. TMSCN and MSTFA methods performed almost equally well for the detection of caffeic, *p*-coumaric, and phenyllactic acids, and peak intensities increased much faster and reached maximum at 30–50 min of silylation. Although the



**Fig. 4** Relative peak abundances of trimethylsilyl derivatives of triterpenes, such as lupeol, oleanolic acid,  $\beta$ -amyrin, and betulinic acid in four different derivatization methods (TMSCN, M-TMSCN, MSTFA, and M-MSTFA) over silylation time range of 5 min to 30 h. Relative peak abundances of TMS derivatives were calculated using PARAFAC2 modeling and  $\ln(\text{min})$  scale of silylation time was used for better visualization



stability of TMS-derivatives of these metabolites were significantly lower when using the MSTFA method. At the optimal silylation times, abundances of TMS-derivatives were 1.5–4 times higher when using TMSCN and MSTFA compared with M-TMSCN and M-MSTFA. This indicates a higher silylation efficiency of the solvent free reagents compared with the solvent, pyridine, interaction that is used during the methoximation. Higher silylation efficiency of the reagent alone may also be due to the better availability of the silylating reagent to the exchangeable hydrogen atoms.

Higher sensitivity and stability of the TMS-derivative was observed in detection of 2-hydroxy-3-methoxybenzoic acid, naringenin, catechin, succinic acid, and 2-hydroxycinnamic acid by using TMSCN-based methods (TMSCN and M-TMSCN) than MSTFA and M-MSTFA methods (Table 1; and see ESM, Figs. S5, 6, 7, 9, and S9). By contrast, optimal peak intensity of 2-hydroxycinnamic acid was 3.2 times higher using MSTFA than with TMSCN. However, the intensity of this metabolite was reduced up to 37 % when MSTFA silylation time was prolonged, which indicates a possible non-stability of TMS-derivatives in a MSTFA reagent. Direct silylation of vanillin using TMSCN resulted in detection of the vanillin-1TMS derivative, where the hydroxyl functional group was silylated, whereas the aldehyde functional group remained intact. At the optimal silylation time, intensity of the vanillin-1TMS peak was almost twofold higher when MSTFA was applied as compared with TMSCN (Table 1). However, in the case of M-TMSCN method, the intensity of vanillin-MEOX-1TMS was 2.23 times higher than in M-MSTFA derivatization.

At the optimal silylation times of amino acids, valine, glycine, aspartic acid, and phenylalanine, the direct silylation methods performed much better than methoximation followed by silylation. Peak intensities were 1.6–58 times higher when

using MSTFA compared with M-MSTFA, and 1.5–15 times higher when using TMSCN compared with M-TMSCN method. Derivatization of serine and threonine were reagent specific (see ESM, Figs. S10 and S11), in the case of MSTFA, serine was mainly detected in the form of serine-2TMS, whereas with TMSCN, the abundance of serine-3TMS exceeded serine-2TMS several times, as serine-3TMS increased and serine-2TMS declined over silylation time. However, both serine derivatives were detected when using M-TMSCN and M-MSTFA. The silylation reaction rate and peak abundances were higher in M-TMSCN method compared with M-MSTFA. A similar pattern was observed in the derivatization of threonine.

#### Choosing the optimal derivatization time

One of the compromise measures in GC-MS analysis of complex mixtures is the derivatization time. An ideal situation would be that the TMS-derivatives of complex mixture metabolites reach their maximum at the same derivatization time and remain stable during the analysis. Unfortunately, this is not the case as TMS-derivatives of different classes of metabolites reach maximum at the different derivatization times (even in the same reaction conditions) which varies peak abundances as a function of derivatization time. Uneven degradation of the derivatives over time is a further complication. Therefore, it is critical to choose an optimal derivatization time for simultaneous detection of a wide variety of metabolites in complex mixtures and with a reasonable sensitivity. However, until kinetic derivatization sampling becomes feasible, the derivatization time must be kept constant over all samples despite not an optimal derivatization time for all metabolites. Consequently, the stability of the TMS-derivatives depends on the chemistry of the TMS derivative, moisture content, derivatization reagents, time, and

temperature. In this study, the significance of the silylation time in four different derivatization methods were evaluated and variations in detection at the three silylation time points closest to the optimal silylation time were calculated. The observed variations were between 7–21, 6–18, 16–39, and 9–24 % for TMSCN, M-TMSCN, MSTFA, and M-MSTFA, respectively. For MSTFA the high peak variations can be explained by an aggressive nature of the reagent when applied alone. For more than half of the metabolites of the standard mixture, MSTFA and TMSCN were equally rapid. However, a more significant influence of the silylation time and relatively faster product degradation was observed in the MSTFA method. The relative low variations when using M-TMSCN or M-MSTFA may be due to a proton scavenging activity of the solvent pyridine as it prevents TMS derivative hydrolysis. Despite many metabolites of standard mixture were derivatized and detected well using MSTFA, the level of reagent derived unexpected peaks and variations were relatively higher than the other derivatization methods.

#### Derivatization method comparison of blueberry extracts

As direct silylation based on MSTFA exhibit low repeatability, it was omitted from the GC-MS analysis of the two blueberry

extracts. The first extract is the phenolic extract A, which is obtained from basic hydrolysis, while the second one (phenolic extract B) is an acid hydrolyzed extract. Fourteen known and five unknown metabolites were detected from the GC-MS profiles of the phenolic extract A. All these metabolites were quantified from each derivatization method, TMSCN, M-TMSCN and M-MSTFA that were evaluated at the 13 different silylation times varying from 5 min to 60 h. Some of the identified phenolic acids such as vanillic, caffeic, syringic, *p*-coumaric, *m*-coumaric, protocatechuic, and gallic acid have previously been found in a GC-MS study of small polish berries [30]. Moreover, metabolites like, succinic, malic, vanillic, *p*-coumaric, and caffeic acids that are identified from the Pphenolic extract A were also included in the standard mixture. This facilitated a comparison of derivatization method performances from two different complex sample matrices. For all the detected metabolites, TMSCN method proved to be superior to MSTFA in terms of sensitivity, derivatization rate and metabolite stability. At the optimal silylation times, metabolite abundances were 1.5–3.0 times higher with M-TMSCN compared with the M-MSTFA method. However, direct silylation with TMSCN outperformed the methoximation based methods and at optimal silylation times, peak intensities were 1.8–8.8 times higher (Table 2).

**Table 2** Trimethylsilyl (TMS) derivatives identified from blueberry phenolic extract A, sorted according to their retention time

No.	Substance	TMSCN <sup>a</sup>	M-TMSCN <sup>a</sup>	M-MSTFA <sup>a</sup>	R <sub>I</sub> <sup>b</sup>	EI-MS <sup>c</sup>
1	Glycerol-3TMS	4.96 (40)	2.64 (40)	(150)	1,267	94
2	Succinic acid-2TMS	2.72 (40)	2.00 (40)	(300)	1,308	95
3	Maleic acid-2TMS	3.84 (40)	1.73 (50)	(150)	1,351	92
4	Lactic acid dimer-2TMS	4.97 (60)	1.96 (60)	(300)	1,391	90
5	Malic acid-3TMS	2.71 (50)	1.73 (60)	(300)	1,489	91
6	Unknown-1	2.84 (50)	1.60 (40)	(50)	1,760	
7	Vanillic acid-2 TMS	4.65 (50)	2.14 (150)	(150)	1,768	96
8	Unknown-2	2.10 (40)	1.66 (40)	(150)	1,784	
9	<i>m</i> -coumaric acid-2TMS	6.71 (60)	2.72 (40)	(150)	1,807	93
10	Protocatechuic acid-3TMS	7.43 (50)	2.94 (40)	(40)	1,830	96
11	Unknown-3	2.91 (50)	1.23 (60)	(150)	1,850	
12	Unknown-4	8.48 (40)	2.25 (30)	(60)	1,866	
13	Unknown-5	9.6 (50)	1.5 (50)	(60)	1,886	
14	Syringic acid-2TMS	8.81 (50)	2.33 (40)	(150)	1,907	89
15	<i>p</i> -coumaric acid-2TMS	5.15 (40)	1.51 (60)	(150)	1,942	91
16	2-methyl-2-methoxy-mandelate-2TMS	5.30 (50)	1.82 (40)	(60)	1,954	86
17	Gallic acid-4TMS	4.87 (50)	2.09 (40)	(50)	1,975	89
18	( <i>cis</i> )-caffeic acid-3TMS	4.69 (50)	2.13 (60)	(150)	1,992	92
19	( <i>trans</i> )-caffeic acid-3TMS	2.46 (60)	1.81 (150)	(150)	2,151	93

<sup>a</sup> PARAFAC2-based relative abundances of derivatives at their optimal silylation times in minutes (indicated in the brackets), for each derivatization method are illustrated as their ratio to the relative abundance of the corresponding derivative in the method M-MSTFA

<sup>b</sup> Retention indices of metabolite derivatives were calculated based on Van den Dool and Kratz equation by using retention times of C10–C40 alkanes

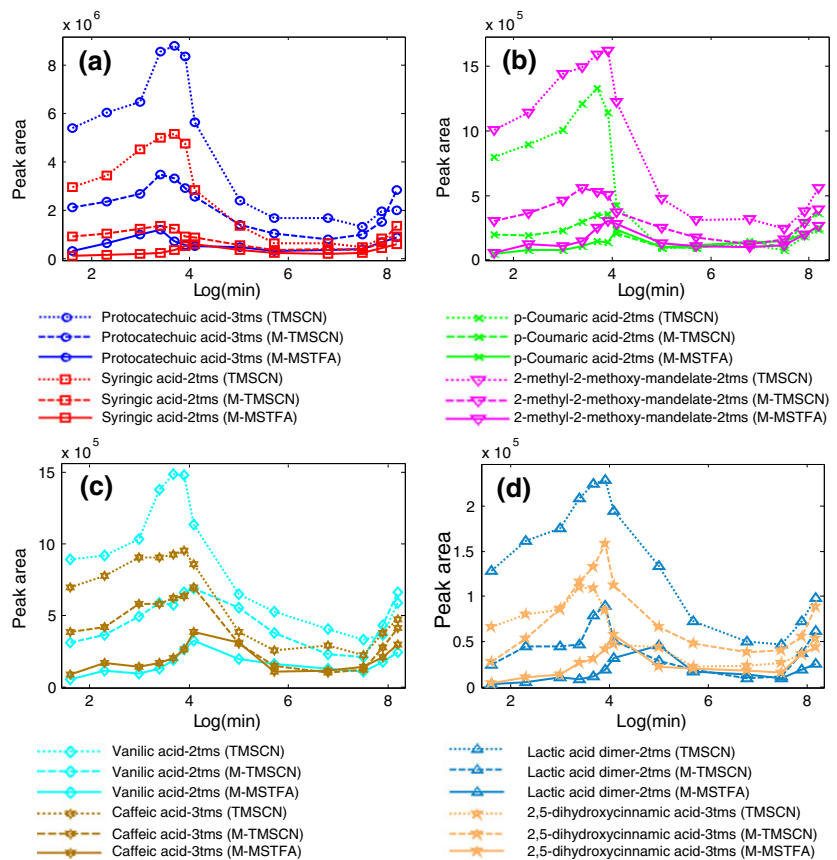
<sup>c</sup> EI-MS-based mass spectra comparison involved Wiley08, NIST05 as well as in-house triterpenes libraries

Figure 5 displays the relative abundances TMS-derivatives of eight most abundant metabolites at different silylation time points. The figure shows that the direct silylation using TMSCN resulted in higher metabolite peak intensities than M-TMSCN and MSTFA methods, at all silylation time points. Likewise, at every silylation time point, abundances of all metabolites were higher when using the derivatization method M-TMSCN compared with M-MSTFA. Figure 5 also shows an increase in metabolite abundances at the later silylation times (45 and 60 h). This might be due to an unintended up-concentration because of the evaporation of the solvent pyridine as a course of long incubation time. Thus, it is important to find the best compromise in setting the derivatization time to obtain an optimal profile with high repeatability.

Repeatability of GC-MS analysis of the blueberry extracts was evaluated by calculating the relative standard deviations of 19 quantified metabolites in four replicates of each derivatization method at the optimal silylation time. The mean errors of the derivatization methods TMSCN, M-TMSCN

and M-MSTFA were 4.6 (varying from 2.1 to 8.6 % for all metabolites), 9.0 (3.5 to 18.1 %), and 15.2 % (3.7 to 21.3 %), respectively. These repeatability tests are in agreement with the robustness of the derivatization methods evaluated in the standard mixture analysis. As the direct silylation method TMSCN performed best in terms of silylation reaction speed, efficiency, and repeatability, this method was used for the GC-MS profiling of the phenolic extract B. This extract contained phenolic and other organic acids that are present in free forms, and conjugated forms with other cell membrane components via ester and glycosidic bonds. Derivatization of these extracts with TMSCN for 40 min enabled identification of 27 metabolites based on RI and EI-MS patterns, including the triterpenoids such as,  $\beta$ -amyrin,  $\alpha$ -amyrin, and oleanolic acid. This result confirmed the presence of triterpenoid saponins derived from  $\beta$ -amyrin,  $\alpha$ -amyrin, and oleanolic acid, as previously reported in a study of anticancer properties of blueberry fruits (*V. myrtilus*) [36, 37]. The developed protocol promises more insight into secondary metabolites of the

**Fig. 5** Relative peak abundances of trimethylsilyl derivatives of some phenolic and organic acids identified from blueberry phenolic extract in four different derivatization methods (TMSCN, M-TMSCN, MSTFA, and M-MSTFA) over silylation time range of 5 min to 30 h. Relative peak abundances of TMS derivatives were calculated using PARAFAC2 modeling and  $\ln(\text{min})$  scale of silylation time was used for better visualization



plant-derived samples and enables detection of low concentration metabolites.

## Conclusions

The GC-MS profiles of complex metabolite mixtures obtained from four derivatization methods were different and significantly influenced by silylation time. The results suggest that for the majority of the investigated metabolites of the complex mixtures, TMSCN-based methods outperformed MSTFA-based methods in terms of silylation reaction speed, sensitivity, and repeatability of the methods. However, direct silylation methods, TMSCN and MSTFA performed equally well for detection of most phenolic and organic acids. Although, the MSTFA method displayed significantly lower TMS-derivative stability over time and lower repeatability of the GC-MS profiles. In general, direct silylation methods provided better sensitivity and more rapid silylation, though methoximation based methods illustrated higher metabolite stability. Thus, based on the results of this study, it is recommended to pay a special attention to the consistency of the sample preparation and derivatization practice prior to comparing GC-MS profiles. In the case of the direct silylation, it is advised to use 50–100  $\mu\text{L}$  pure TMSCN for complete dried extracts of the 100- to 200- $\mu\text{L}$  sample and to incubate at 37  $^{\circ}\text{C}$  for 40 min. If an initial methoximation is required then the trimethylsilylation time must be increased to 120–150 min to allow silylation of all the labile protons in the presence of pyridine. The average silylation variations of methods increased in the following order, M-TMSCN<TMSCN<M-MSTFA<MSTFA, whereas robustness of methods increased almost in opposite direction, MSTFA<M-MSTFA<M-TMSCN<TMSCN. This study showed an unbiased and rapid silylation of various classes of metabolites by using TMSCN either alone or in conjunction with a prior methoximation step. These results illustrate high potential of TMSCN as an alternative silylation reagent for derivatization of complex biological mixtures in GC-MS metabolomics.

**Acknowledgments** The authors thank the Faculty of Science for support to the elite-research area “Metabolomics and bioactive compounds” with a PhD stipendium to B. Khakimov and The Ministry of Science and Technology for a grant to University of Copenhagen (S.B. Engelsen) with the title “Metabolomics infrastructure” under which the GC-MS was acquired.

## References

- Pasikanti KK, Ho P, Chan E (2008) *J Chromatogr B-Anal Technol Biomed Life Sci* 871:202–211
- Gu Q, David F, Lynen F, Rumpel K, Dugardeyn J, Van Der Straeten D, Xu G, Sandra P (2011) *J Chromatogr A* 1218: 3247–3254
- Fiehn O (2008) *Trac-Trends Anal Chem* 27:261–269
- Roessner U, Wagner C, Kopka J, Trethewey RN, Willmitzer L (2000) *Plant J* 23:131–142
- Fiehn O, Kopka J, Trethewey RN, Willmitzer L (2000) *Anal Chem* 72:3573–3580
- Lisec J, Schauer N, Kopka J, Willmitzer L, Fernie AR (2006) *Nat Protoc* 1:387–396
- Poole CF (1978) Recent advances in the silylation of organic compounds for gas chromatography. In: Blau K, King G (eds) *Handbook of derivatives for chromatography*. Heydon & Son Inc, Philadelphia, PA, pp 152–200
- Fales HM and Luukkainen T (1965) *Analytical Chemistry* 37:955
- Horning MG, Moss AM, Horning EC (1968) *Anal Biochem* 22:284–294
- Gehrke CW, Nakamoto H, and Zumwalt RW (1969) *Journal of Chromatography* 45:24–51
- Schweer H (1982) *J Chromatogr* 236:355–360
- Kanani H, Chrysanthopoulos PK, Klapa MI (2008) *J Chromatogr B-Anal Technol Biomed Life Sci* 871:191–201
- Poole CF (2013) *J Chromatogr A* 1296:2–14
- Little JL (1999) *J Chromatogr A* 844:1–22
- Chromatography Catalogue 1998–99*, Regis Technologies, 1998; pp. 86–88
- GC Derivatization, Pierce 2003–2004. Applications handbook and catalog
- Gullberg J, Jonsson P, Nordstrom A, Sjoström M, Moritz T (2004) *Anal Biochem* 331:283–295
- Danielsson APH, Moritz T, Mulder H, Spigel P (2012) *Metabolomics* 8:50–63
- Pierce AE (1968) *Silylation of organic compounds*. Pierce Chemicals Co, Rockford, IL, p 58
- Kashutina MV, S. L. Ioffe, V. A. Tartakovskii (1975) *Russian Chem. Rev.* 44, 733
- Orata F (2012) Derivatization reactions and reagents for gas chromatography analysis, advanced gas chromatography—progress in agricultural, biomedical and industrial applications. In: Mustafa Ali Mohd (ed), *InTech 2012*. ISBN: 978-953-51-0298-4
- Summer LH, Parker GA, Lloyd NC, Frey CL, Michael KW (1967) *J Amer Chem Soc* 89:857
- Prasad H (2002) *Resonance* 7:48
- Pike RM (1961) *J Org Chem* 26:232
- Hulshoff A, Lingeman H (1984) *J Pharm Biomed Anal* 2:337–380
- Birkofer L and Brokmeie D (1968) *Tetrahedron Letters* 9: 1325–1328
- Matsukawa S, Fujikawa S (2012) *Tetrahedron Lett* 53:1075–1077
- Mai K, Patil G (1986) *J Org Chem* 51:3545–3548
- Riggio PP, Karasiewicz RJ, Rosen P, Toome V (1992) *J Chromatogr Sci* 30:29–31
- Zadernowski R, Nacz M, Nesterowicz J (2005) *J Agric Food Chem* 53:2118–2124
- Amigo JM, Popielarz MJ, Callejon RM, Morales ML, Troncoso AM, Petersen MA, Toldam-Andersen TB (2010) *J Chromatogr A* 1217: 4422–4429
- Khakimov B, Amigo JM, Bak S, Engelsen SB (2012) *J chromatogr A* 1266:84–94
- Hotelling H (1933) *J Educ Psychol* 24:417–441
- Vandendool H and Kratz PD (1963) *Journal of Chromatography* 11: 463–471
- Skov T, Bro R (2008) *Anal Bioanal Chem* 390:281–285
- Ono M, Koto M, Komatsu H, Igoshi K, Kobayashi H, Ito Y, Nohara T (2004) *Food Sci Technol Res* 10:56–59
- Szakiel A, Paczkowski C, Koivuniemi H, Huttunen S (2012) *J Agric Food Chem* 60:4994–5002



## **The use of trimethylsilyl cyanide derivatization for robust and broad-spectrum high-throughput gas chromatography-mass spectrometry based metabolomics**

**Bekzod Khakimov**, Mohammed Saddik Motawia, Søren Bak, Søren Balling Engelsen

### **Additional safety consideration**

Most silylation reagents require careful handling and users must follow all safety instructions. According to European Regulation (EC) No 1272/2008, the silylation reagents TMSCN, MSTFA, as well as other reagents such as BSTFA, MTBSTFA, BSA, TMCS are all flammable volatile liquids and harmful in contact with skin, eye and inhalation. Safety data of all these chemicals suggest careful handling (avoid air, moisture, contact with skin, eye, and swallow) and storage (cool, well-ventilated place, under inert gas, protected from direct sunlight and water which causes their decomposition). Stability and reactivity data show that all these reagents require similar conditions to avoid possible danger and degradations: no heat, flame, sparks and water. Avoid strong acids, base, aldehydes, ketones. The difference in boiling points of TMSCN (115°C) and MSTFA (131°C) is not significant, while BSA (72°C) and TMCS (57°C) have relative lower boiling point that increases their volatility.

As part of this study, we evaluated evaporation of silylation reagents TMSCN, MSTFA as well as methoximation reagents and silylation reagent (M-TMSCN, M-MSTFA) both, at room temperature and at the incubation temperature used during derivatization. 100 µl of reagents were sealed in GC-MS vials using the same magnetic-silicon septum caps that are used throughout the analysis and penetrated four times by needles installed in the autosampler. The volume of the reagents, TMSCN and MSTFA did not change after 48h of incubation at both temperatures, while the volume of M-TMSCN and M-MSTFA was reduced by 10-20%. The volume reduction of M-TMSCN and M-MSTFA may rather be due to the readily volatile pyridine used in methoximation step. Our data documents, that when using appropriate needles (OD: 0.5 mm) and GC-MS vial septum caps, the reagents do not evaporate out of the GC-MS encapsulated vials even after two days which ensure both a safeness of derivatization reactions and high reproducibility. Most GC-MS labs that perform high-throughput analysis utilize autosamplers that facilitates further derivatization accuracy and safer handling of reagents.

All silylation reagents produce byproducts during the reaction: MSTFA (byproduct: N-Methyltrifluoroacetamide), BSTFA (byproduct: mono(trimethylsilyl)trifluoroacetamide and trifluoroacetamide), BSA (byproduct: N-trimethylsilyl-pivalimidic acid) and TMCS (byproduct: hydrochloric acid) fall into the similar categories of hazard classifications (according to European Regulation (EC) No 1272/2008) as the byproduct of TMCN, hydrogen cyanide (HCN). Most silylation protocols use 30-100 µl of silylation reagents for derivatization of dried complex extracts of plants and/or animal origin. In this study,

according to the reaction stoichiometry, the amount of TMSCN used for derivatization of standard mixture was 200–400 times more than the amount of the reagent needed for silylation of all available active protons in the standard mixture. Moreover, the byproduct, HCN might be consumed during protonation of TMSCN to form  $\text{TMSCNH}^+$ , since the basicity of TMSCN is much greater than the basicity of HCN. Therefore, the silylation reaction rate may even increase because  $\text{TMSCNH}^+$  is more electrophile than TMSCN and easily attracts nucleophile substrate [1].

In addition, we have estimated the amount of the byproduct, HCN formed during a standard silylation reaction to evaluate the potential toxicity. Potential toxic concentration of HCN in the air is  $300 \text{ mg/m}^3$  [2], while the smallest size of most standard laboratories is  $60 \text{ m}^3$ . One mole of TMSCN produces one mole of HCN, thus  $0.0003197$  mole (an amount that is used in this study) of TMSCN produces  $0.0003197$  mole of HCN, that is equivalent to  $0.0003197 \times 27.03 = 0.0086422 \text{ g} = 8.6422 \text{ mg}$ . For example, if one hundred samples are derivatized simultaneously, a total of  $864.22 \text{ mg}$  of HCN will be formed. In an average laboratory of  $60 \text{ m}^3$ , this amount corresponds to a concentration of  $14.40367 \text{ mg/m}^3$ . This amount is 20 times less than the toxic amount of HCN for human health, which is  $300 \text{ mg/m}^3$ . A HCN concentration of  $14.40367 \text{ mg/m}^3$  may form only if all TMSCN ( $40 \mu\text{l}$ ) is used, in all 100 samples, and evaporated. Liberation of this amount of HCN is very unlikely, mainly due to the small amount of TMSCN that reacts (as mentioned earlier TMSCN is 200–400 in excess for complete silylation, and accordingly  $1/200$  of used TMSCN will form HCN), sample is moisture free and vials are tightly sealed with septum leads. Therefore, application of TMSCN in high-throughput GC-MS analysis is safe and may provide easy and powerful silylation.

1. M.V.Kashutina, S. L. Ioffe, V. A. Tartakovskii, *Russian Chem. Rev.* 44 (1975) 733.
2. Environmental and Health Effects. Cyanidecode.org. Retrieved on 2012-06-02.

### **MultiPurpose Sampler (MPS) and Cooled Injection System (CIS) parameters**

GC-MS injection parameters of the left MPS equipped with  $10 \mu\text{l}$  syringe were as follows: injection was performed in sandwich mode, top air volume  $1.0 \mu\text{l}$ , air volume below  $1.0 \mu\text{l}$ , air volume above  $1.0 \mu\text{l}$ , injection volume  $1.0 \mu\text{l}$ , injection speed  $50.0 \mu\text{l s}^{-1}$ , fill speed  $0.2 \mu\text{l s}^{-1}$ , viscosity delay 4 seconds, pre and post injection delay 2 seconds, vial penetration 30 mm, and injection penetration 40 mm. Parameters of the right MPS, equipped with  $100 \mu\text{l}$  syringe were as follows: add volume 40 and/or  $80 \mu\text{l}$ , add speed  $5 \mu\text{l s}^{-1}$ , viscosity delay 4 seconds, post add delay 2 seconds, eject speed  $50 \mu\text{l s}^{-1}$ , source vial penetration 30 mm and destination vial penetration 28 mm. Both syringes were pre and post washed two times with acetone followed by n-hexane. The agitator parameters for sample incubation were as follows: incubation time varied depending on the chosen silylation time, agitator speed was 750 rpm, agitator on time was 59 seconds, agitator off time was 1

second and agitator temperatures were 30°C and 37°C for methoxiamination and silylation, respectively. The CIS port parameters were as follows: initial temperature 40°C, equilibration time 1.0 min, initial time 0.5 min, heating rate 12.0°C s<sup>-1</sup>, end temperature 320°C and hold time 0.5 min.

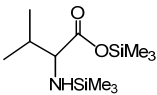
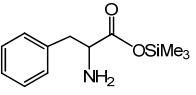
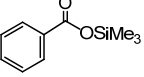
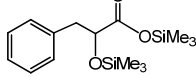
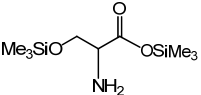
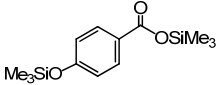
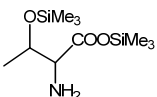
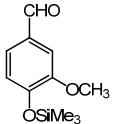
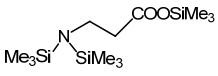
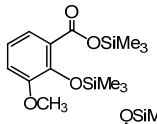
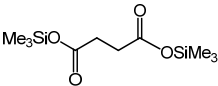
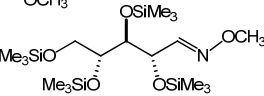
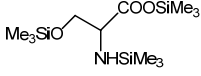
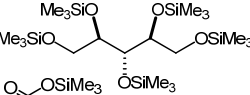
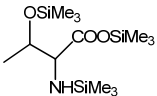
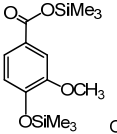
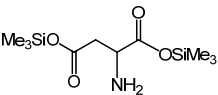
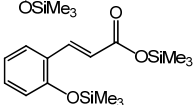
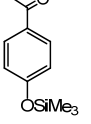
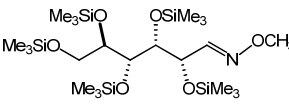
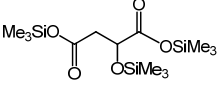
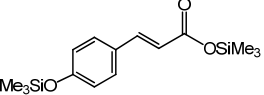
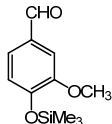
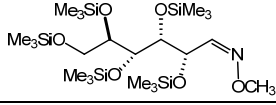
### **Important considerations in automation of derivatization and GC-MS analysis**

High-throughput GC-MS metabolomics requires the use of robots prior to reduce the level of experimental error. However, automated and simultaneous derivatization of samples requires derivatization reagents to be present in the autosampler at all times throughout the analysis. Most of the commercially available derivatization reagents requires specific storing conditions prior to avoid degradation and reactivity lost. It is important to fulfill the storing conditions of the reagents recommended by the manufacturers when keeping them on the autosampler for a long time. Moisture and direct sunlight can easily cause degradation of the reagents, silyl-derivatives and/or alter the original content and in turn introduce a bias in the silylation of samples [13, 14]. It is also important to use an excess amount of reagent for complete silylation of metabolites and to suppress a little amount of water, which might be present in the reaction mixture. Hydrolytic stability of the silyl-derivatives highly depends on the structural and steric features of the molecules. The general hydrolytic stability of silyl derivatives of the different classes of compounds decreases in the following order: alcohols > phenols > carboxylic acid > amines > amides [15]. Moreover, stability of trimethylsilyl derivatives depends on the type of the injection port used in the analysis. Despite most silyl-derivatives of metabolites are thermally stable, they may degrade in contact with stainless steel injection ports at high temperature, and therefore it is recommended to use glass injection ports (e.g. glass liners) for high-temperature GC injections [15].

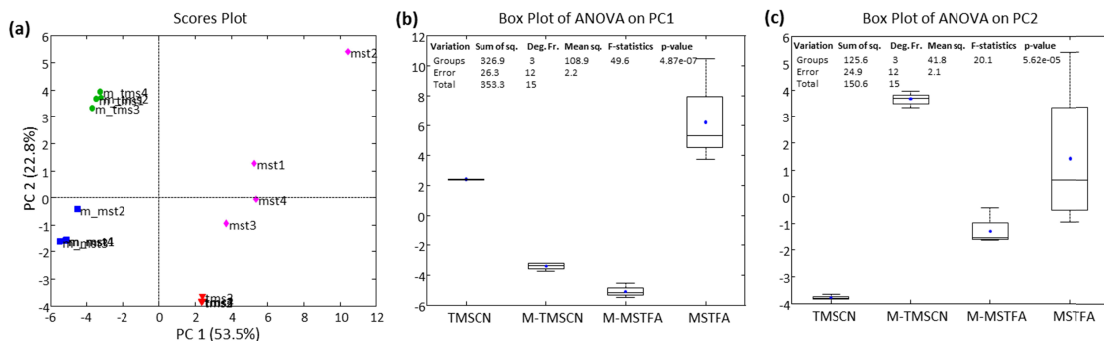
Most high-throughput GC-MS metabolomic studies aimed to obtain a quantitative data. In order to compare GC-MS profiles of these samples, it is crucial to keep the derivatization reaction time (time interval between the addition of derivatization reagent into sample and GC-MS injection) constant over the entire analysis, as derivatization time differences alters the GC-MS profiles of the identical samples significantly, both qualitatively and quantitatively. Moreover, caution must be taken not to use polar solvents, with active protons (e.g. ROH, RCOOH), throughout the derivatization (even for the injection needle wash), since they can easily react with the silylation reagent. Likewise, appropriate GC column with inert stationary phase (e.g. silicon-based columns) must be used for the analysis of trimethylsilylated samples. Injection of silylation reagent into the GC column with a polar stationary phase (e.g. polyethylene glycol based and free fatty acid based columns) will result in unreliable GC profiles with artifacts and column degradation products.

**Table S1.** Derivatization products of the metabolites of standard mixture, in the same order as Table 1.

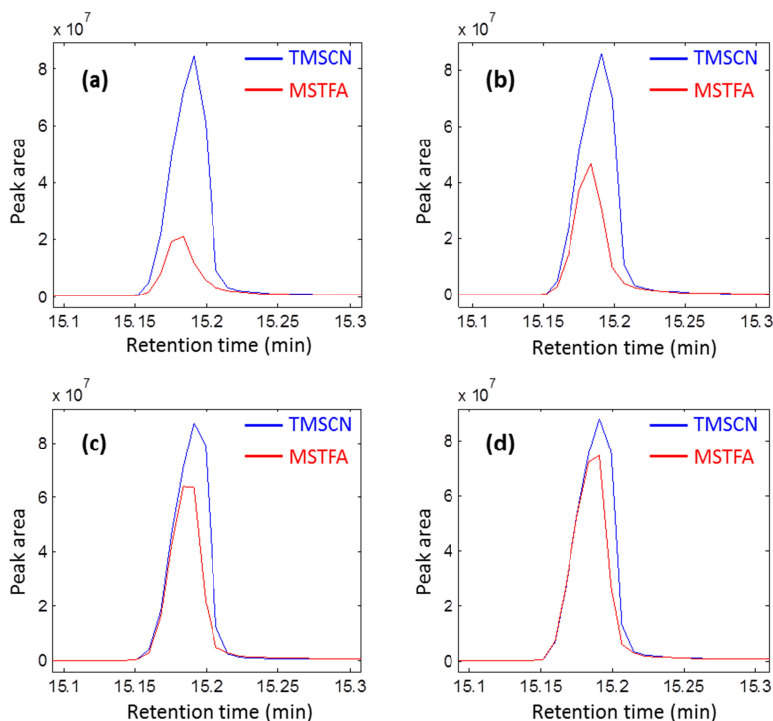


Entry	Substrate	Derivatized product	Entry	Substrate	Derivatized product
1	Valine		13	Phenylalanine	
2	Benzoic acid		14	Phenylacetic acid	
3	Serine		15	4-Hydroxybenzoic acid	
4	Threonine		16	Vanillin-MEOX	
5	Glycine		17	2-Hydroxy-3-methoxybenzoic acid	
6	Succinic acid		18	(trans)-Ribose	
7	Serine		19	Ribitol	
8	Threonine		20	Vanillic acid	
9	Aspartic acid		21	2-Hydroxycinnamic acid	
10	4-Hydroxyacetophenone		22	(trans)-Glucose	
11	Malic acid		23	p-Coumaric acid	
12	Vanillin		24	(cis)-Glucose	

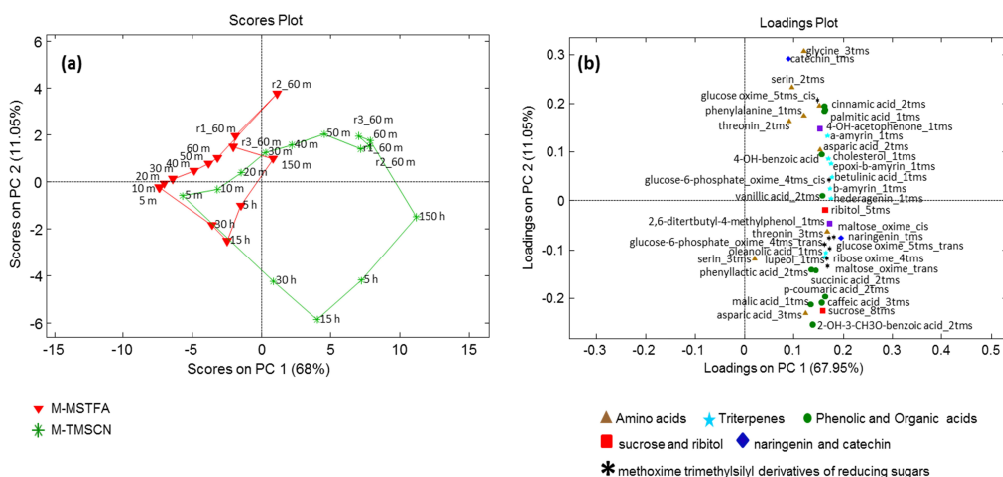
Entry	Substrate	Derivatized product	Entry	Substrate	Derivatized product
25	Palmitic acid		34	Cholesterol	
26	(trans)-Caffeic acid		35	β-Amyrin	
27	(trans)-Glucose-6-phosphate-MEOX		36	α-Amyrin	
28	(cis)-Glucose-6-phosphate-MEOX		37	Lupeol	
29	Sucrose		38	Oleanolic acid	
30	(trans)-Maltose-MEOX		39	Betulinic acid	
31	Naringenin		40	α-Epoy-β-amyrin	
32	Catechin		41	Hederagenin	



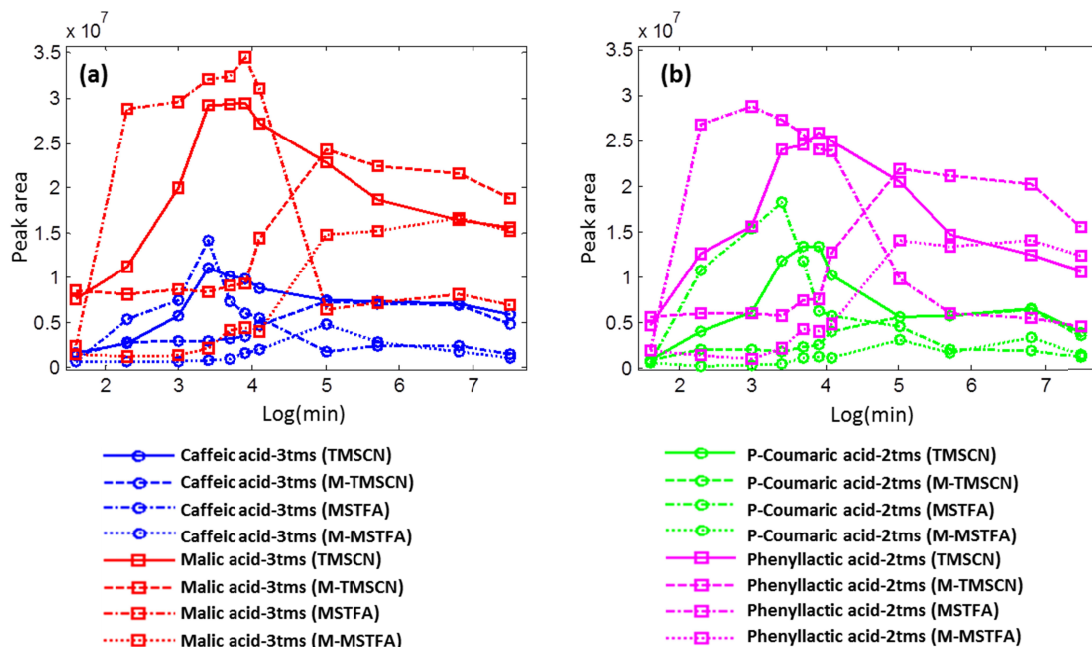
**Figure S1.** (a) PCA scores plot of the replicate data matrix (16 x 41), with 16 samples, 4 replicates per method and abundances of 41 metabolites listed in Table 1. (b) ANOVA analysis of the four derivatization methods (TMSCN, M-TMSCN, MSTFA and M-MSTFA) based on their scores on PC1, (c) ANOVA analysis of the four derivatization methods based on their scores on PC2.



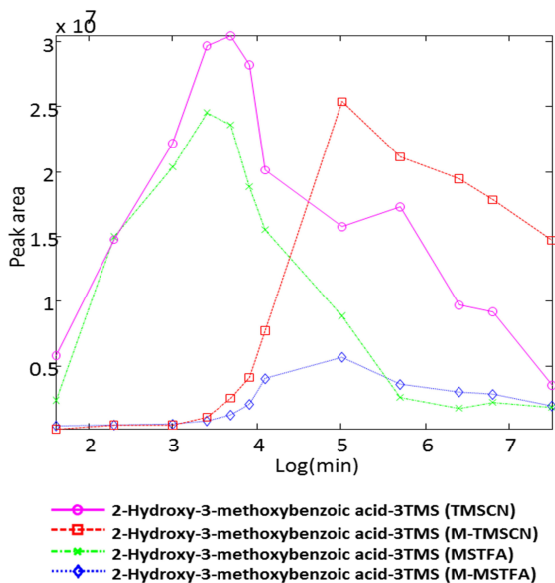
**Figure S2.** Comparison of total ion current chromatograms of trimethylsilyl derivative of 2,6-diphenyl-phenol at four different silylation time points, (a) 5 seconds, (b) 30 seconds, (c) 1 minute and (d) 5 minutes with two silylation reagents, TMSCN and MSTFA. \*Silylation time refers to the incubation time of sample in agitator after addition of the reagent, excluding 35 seconds of robot operation time for washing needle, delays and injection.



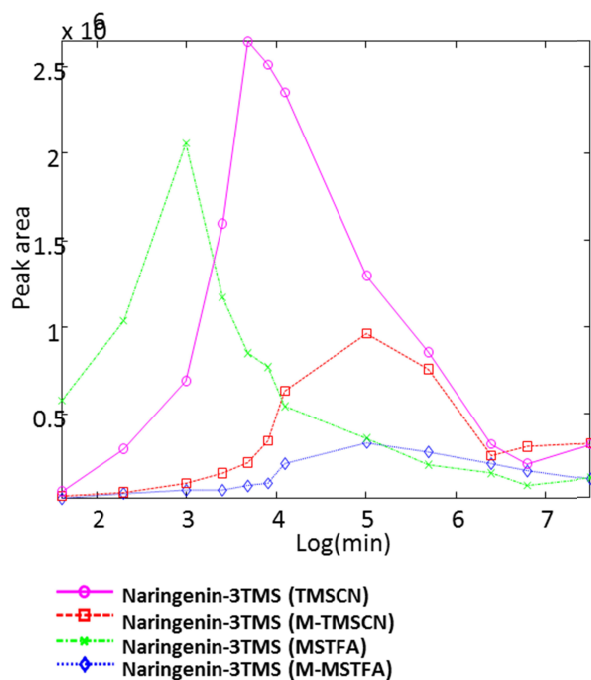
**Figure S3.** Scores (a) and loadings (b) plots of PCA model developed on a matrix containing PARAFAC2 scores (relative abundances of all peaks including MEOX-TMS derivatives of carbohydrates) of TMS-derivatives of standard mixture detected by four derivatization methods, at 12 different silylation time points (including four replicates at the optimal silylation times).



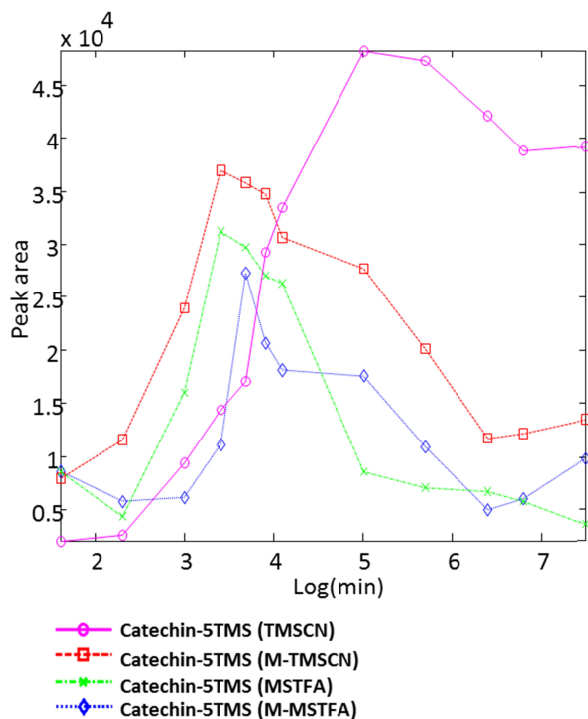
**Figure S4.** Relative peak abundances of trimethylsilyl derivatives of caffeic, malic, p-coumaric and phenyllactic acids in four different derivatization methods (TMSCN, M-TMSCN, MSTFA, M-MSTFA) over silylation time range of 5 minutes to 30 hours. \*Relative peak abundances of TMS-derivatives were calculated using PARAFAC2 modeling and  $\ln(\text{min})$  scale of silylation time was used for better visualization.



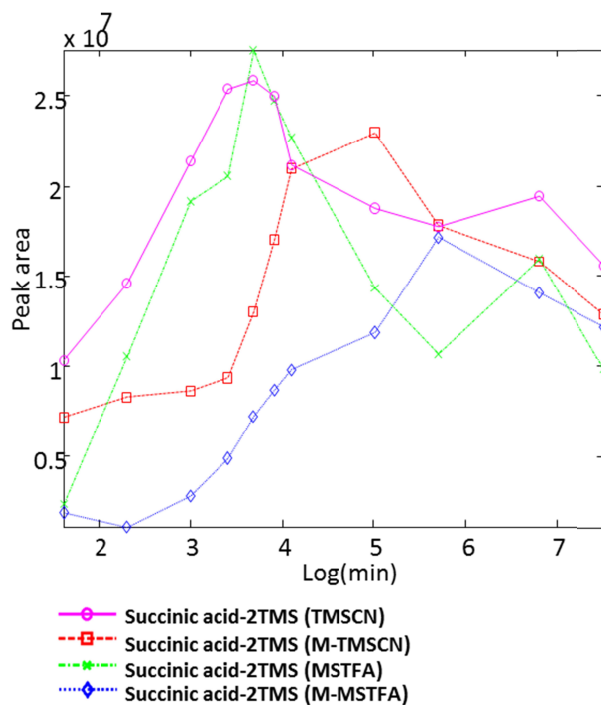
**Figure S5.** Relative peak abundance of trimethylsilyl derivative of 2-Hydroxy-3-methoxybenzoic acid in four different derivatization methods (TMSCN, M-TMSCN, MSTFA, M-MSTFA) over silylation time range of 5 minutes to 30 hours. \*Relative peak abundances of TMS-derivatives were calculated using PARAFAC2 modeling and logarithmic scale of silylation time was used for better visualization.



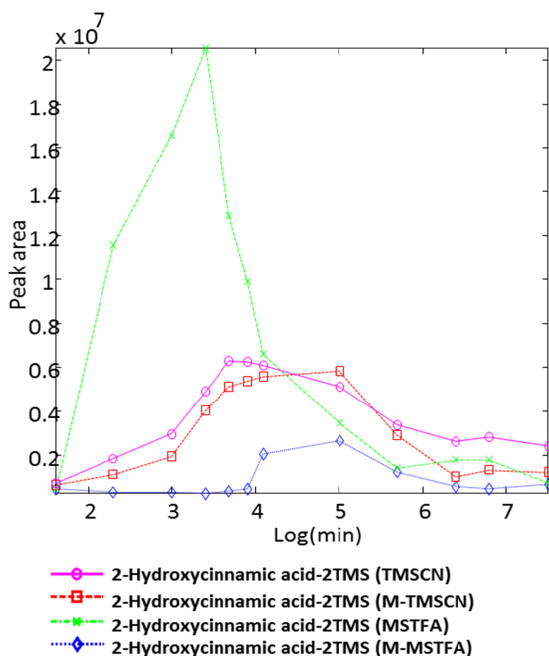
**Figure S6.** Relative peak abundance of trimethylsilyl derivative of Naringenin in four different derivatization methods (TMSCN, M-TMSCN, MSTFA, M-MSTFA) over silylation time range of 5 minutes to 30 hours. \*Relative peak abundances of TMS-derivatives were calculated using PARAFAC2 modeling and logarithmic scale of silylation time was used for better visualization.



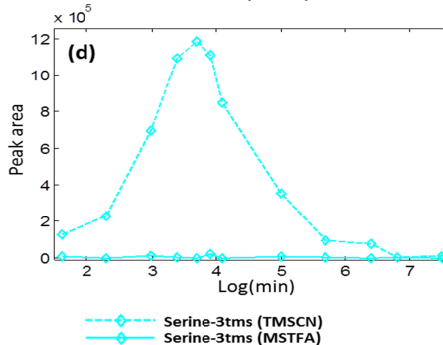
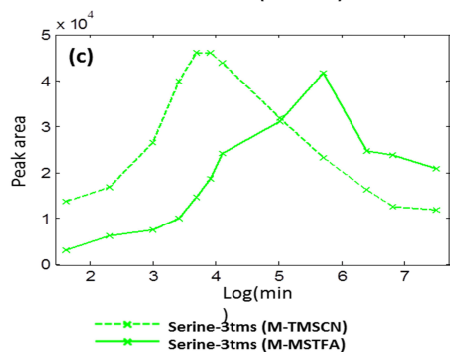
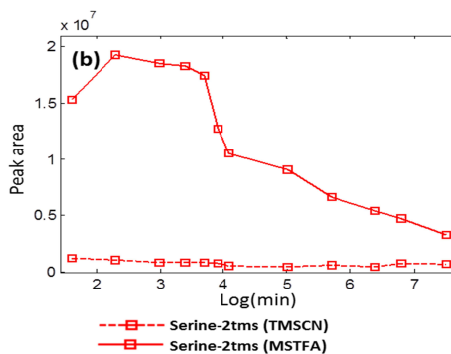
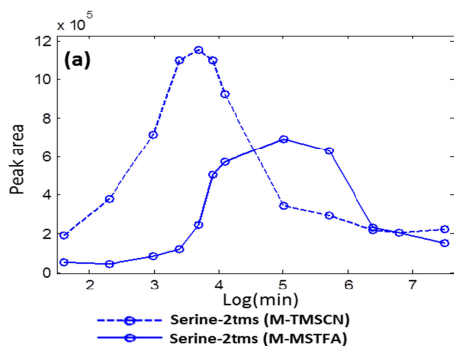
**Figure S7.** Relative peak abundance of trimethylsilyl derivative of Catechin in four different derivatization methods (TMSCN, M-TMSCN, MSTFA, M-MSTFA) over silylation time range of 5 minutes to 30 hours. \*Relative peak abundances of TMS-derivatives were calculated using PARAFAC2 modeling and logarithmic scale of silylation time was used for better visualization.



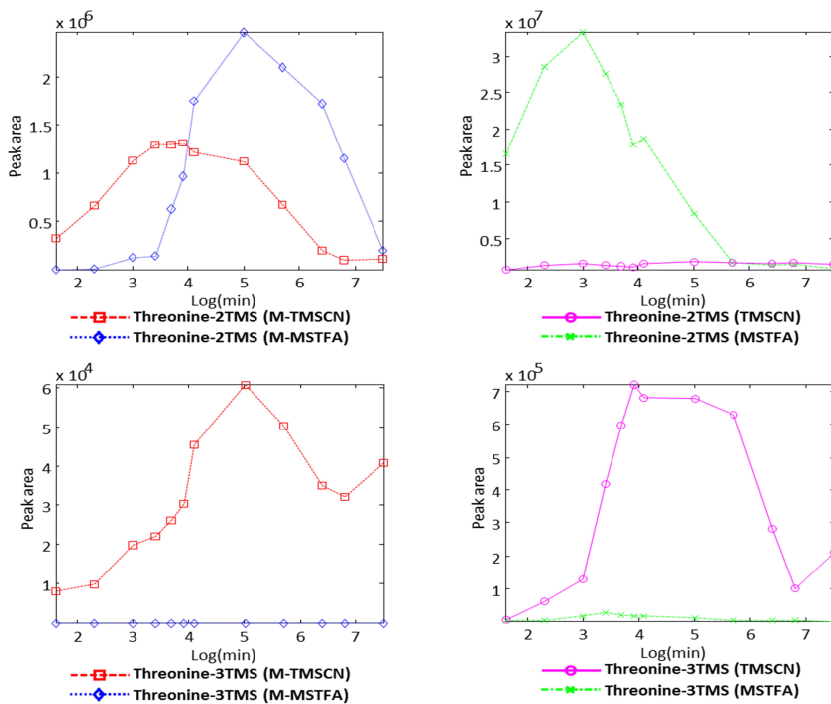
**Figure S8.** Relative peak abundance of trimethylsilyl derivative of Succinic acid in four different derivatization methods (TMSCN, M-TMSCN, MSTFA, M-MSTFA) over silylation time range of 5 minutes to 30 hours. \*Relative peak abundances of TMS-derivatives were calculated using PARAFAC2 modeling and logarithmic scale of silylation time was used for better visualization.



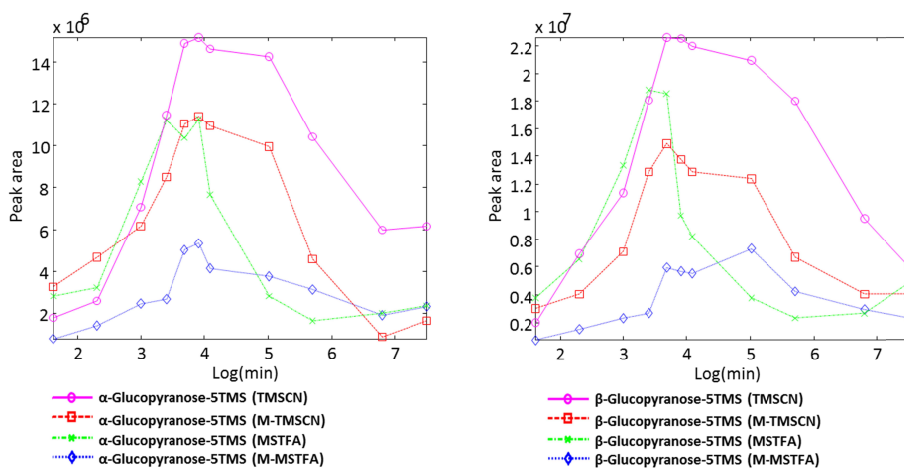
**Figure S9.** Relative peak abundance of trimethylsilyl derivative of 2-Hydroxycinnamic acid in four different derivatization methods (TMSCN, M-TMSCN, MSTFA, M-MSTFA) over silylation time range of 5 minutes to 30 hours. \*Relative peak abundances of TMS-derivatives were calculated using PARAFAC2 modeling and logarithmic scale of silylation time was used for better visualization.



**Figure S10.** Relative peak abundances of trimethylsilyl derivatives of serine in four different derivatization methods (TMSCN, M-TMSCN, MSTFA, M-MSTFA) over silylation time range of 5 minutes to 30 hours. \*Relative peak abundances of TMS-derivatives were calculated using PARAFAC2 modeling and logarithmic scale of silylation time was used for better visualization.

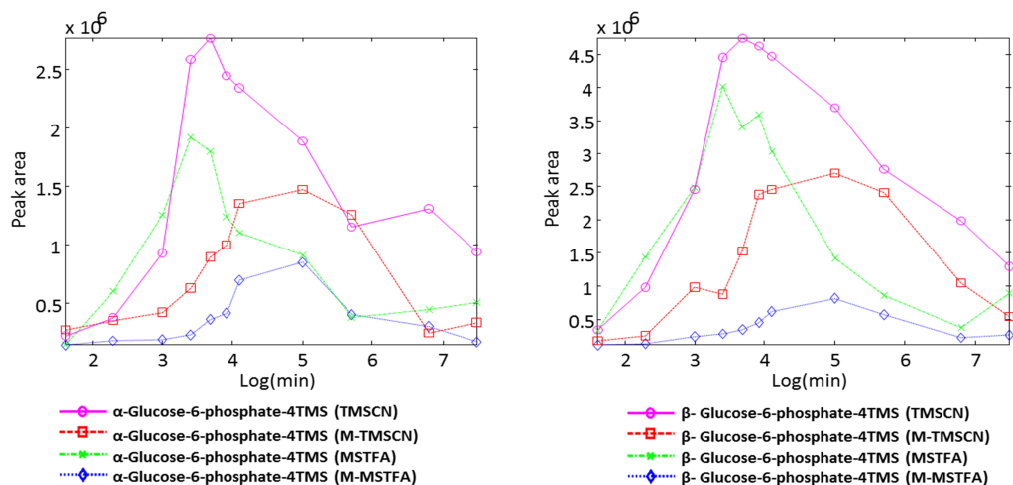


**Figure S11.** Relative peak abundances of trimethylsilyl derivatives of threonine in four different derivatization methods (TMSCN, M-TMSCN, MSTFA, M-MSTFA) over silylation time range of 5 minutes to 30 hours. \*Relative peak abundances of TMS-derivatives were calculated using PARAFAC2 modeling and logarithmic scale of silylation time was used for better visualization.

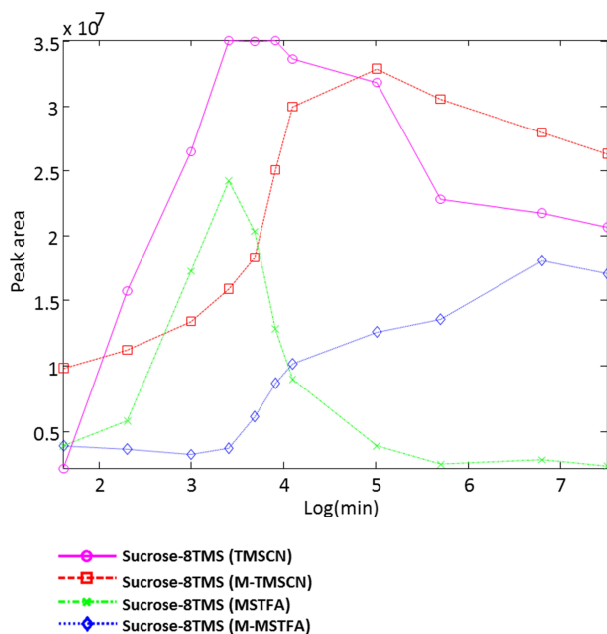


**Figure S12.** Relative peak abundances of trimethylsilyl derivative of glucose in four different derivatization methods (TMSCN, M-TMSCN, MSTFA, M-MSTFA) over silylation time range of 5 minutes to 30 hours. \*Relative peak abundances of TMS-derivatives were calculated using PARAFAC2 modeling and logarithmic scale of silylation time was used for better visualization.





**Figure S13.** Relative peak abundances of trimethylsilyl derivative of glucose-6-phosphate in four different derivatization methods (TMSCN, M-TMSCN, MSTFA, M-MSTFA) over silylation time range of 5 minutes to 30 hours. \*Relative peak abundances of TMS-derivatives were calculated using PARAFAC2 modeling and logarithmic scale of silylation time was used for better visualization.



**Figure S14.** Relative peak abundance of trimethylsilyl derivative of sucrose in four different derivatization methods (TMSCN, M-TMSCN, MSTFA, M-MSTFA) over silylation time range of 5 minutes to 30 hours. \*Relative peak abundances of TMS-derivatives were calculated using PARAFAC2 modeling and logarithmic scale of silylation time was used for better visualization.

# Paper 3

**Bekzod Khakimov**, Søren Bak, Søren Balling Engelsen

High-throughput cereal metabolomics: Current analytical technologies, challenges and perspectives

*Journal of Cereal Science, In press, DOI: 10.1016/j.jcs.2013.10.002*



**Title: High-throughput cereal metabolomics: Current analytical technologies, challenges and perspectives**

Bekzod Khakimov<sup>a,b</sup>, Søren Bak<sup>b</sup>, Søren Balling Engelsen<sup>a,\*</sup>

<sup>a</sup> Spectroscopy & Chemometrics, Department of Food Science, Faculty of Science, University of Copenhagen, Rolighedsvej 30, 1958 Frederiksberg C, Denmark

<sup>b</sup> Plant Biochemistry, Department of Plant and Environmental Sciences, Faculty of Science, University of Copenhagen, Thorvaldsensvej 40, 1871 Frederiksberg C, Denmark

\*Corresponding author: prof. Søren Balling Engelsen

Email: [se@life.ku.dk](mailto:se@life.ku.dk)

Tel.: +45 20 20 00 64, +45 35 33 32 05

## **Abstract**

Metabolomics attempts to answer questions that lie beyond the powers of genomics, transcriptomics and proteomics to facilitate an understanding and assessment of the phenotype based on the metabolome. Metabolomics can serve as (1) a direct tool to explicit secondary metabolites, (2) as an epigenetic gene amplification on the whole phenome level to define the whole genotype by a metabolome marker pattern and (3) as a marker for optimal adaptation of a specific genotype to the environment. Several biologically important questions such as influence of genetic engineering, breeding, climate change, fertilizers, biotic and abiotic stresses in bioactive components and nutritional properties of crop plants have been addressed by using metabolomic approaches. This article focusses on application of high-throughput metabolomics in cereals. Cereal metabolomics is a newly emerged and rapidly developing omics area that assists in the evaluation of cereals and cereal products and plays a key role in the development and improvement of cereal cultivars, by quantitative (and qualitative) global metabolome analysis of phenotypes. In this review, all steps of the metabolomic workflow, from sample harvesting to data analysis are discussed in detail. Main sources of errors that lead to an increase in non-sample-related variations are addressed and current recommended solutions are highlighted. Analytical platforms are discussed and compared in terms of their sensitivity, resolution and applications. Several raw metabolomic data preprocessing and analyses methods are illustrated with examples and their advantages and limitations are addressed. Finally, selected metabolomic studies applied to main cereals are summarized and discussed with emphasis on analytical technologies and protocols focusing on targeted and untargeted metabolomics.

*Keywords:* cereal metabolomics, whole-grains, analytical platforms, chemometrics

# Contents

1. Introduction .....	4
2. Cereal metabolomics .....	6
2.1. Background, definition and motivation .....	6
2.2. Beyond polysaccharides and proteins .....	7
2.3. Cereal phenomics.....	9
3. Experimental design and sampling .....	10
3.1. Experimental design and optimization .....	10
3.2. Sample preparation and metabolite extraction .....	11
4. High-throughput analytical platforms .....	12
4.1. LC-MS .....	13
4.2. CE-MS.....	17
4.3. GC-MS .....	18
4.4. NMR .....	23
4.5. Vibrational spectroscopy .....	24
4.6. Electronic spectroscopy .....	26
5. Turning metabolomics data into information .....	27
5.1. Metabolomic data processing .....	28
5.2. Unsupervised multivariate methods .....	35
5.3. Supervised multivariate methods.....	37
5.4. Exploiting the experimental design: ASCA.....	41
5.5. Network analysis.....	42
6. Application of cereal metabolomics.....	43
6.1. Maize.....	43
6.2. Rice.....	45
6.3. Wheat, barley, oat and rye .....	46
7. Outlook and perspectives.....	51
8. Abbreviations.....	53
9. Acknowledgment .....	54
10. References .....	55

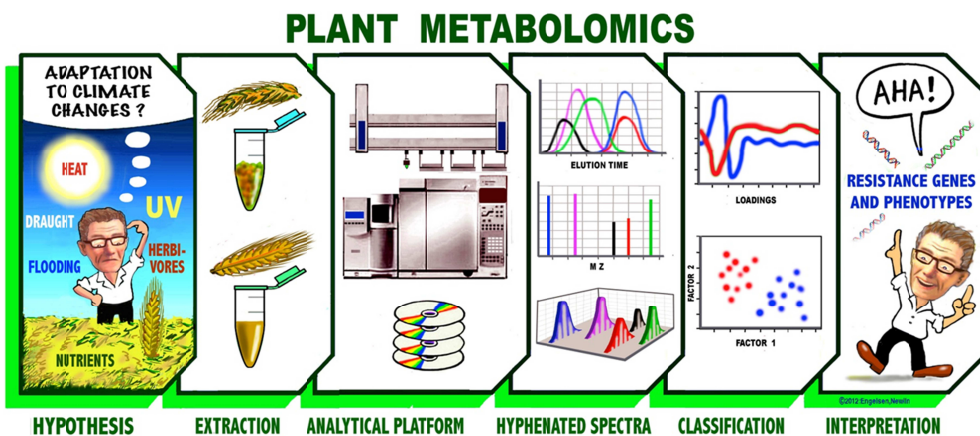
# 1 Introduction

Cereals are ubiquitously grown primarily for food and feed. Cereal plants such as rice, maize, wheat, barley, rye and oat are grown in many different geographical regions of the world. In each region, several different cereal varieties might be grown and these varieties have been bred to grow under the surrounding climate conditions and to give a high yield. The same variety of a cereal plant grown in two different regions will have phenotypic differences, and these differences will become even more pronounced if the growing conditions in two regions differ significantly. Unfortunately, the rapidly increasing climate changes and natural disasters will not always allow cereal plants to recover and instead prevent their growth and diminish their nutritional value. The current trends of climate change and increasing natural disasters such as drought, flooding and extreme temperatures challenge the cereal production and the value of the important agricultural varieties grown today. Therefore, it is of prime importance to develop new varieties that are resistant and/or easily adaptable to such changes and maintain the health benefits of the crops. Moreover, there is a huge demand for increasing the production of cereal crops and at the same time to reduce the use of fertilizers, pesticides and water.

The major effort in adjusting the cereal genotype to changing environments is done with classical plant breeding employing cloning, mutation and selection in the field and in the laboratory. In addition, to solve the global challenges mentioned above, molecular scientists employ state-of-the-art techniques such as genomics, proteomics, transcriptomics and metabolomics. Metabolomics is the newer approach, which found wide application after the development of high-throughput hyphenated analytical techniques. Metabolomics was founded as a powerful screening approach in toxicology (Nicholson et al., 1999), but now it has become a key tool to investigate biological questions that are not easily addressed by applying other 'omic' technologies. State-of-the-art metabolomic techniques allow detection of hundreds of different cell metabolites at the given state of the cell. Since metabolites are synthesized and turned over in cells within a very short time, the metabolomic equilibrium of organisms (the metabolome) is constantly changing. For example, the plant leaf metabolome changes according to the season and to the time and temperature of the day. Metabolomic changes become more pronounced when the plant is challenged with biotic or abiotic stresses. Moreover, the metabolome reflect the changes that occur due to breeding and/or genetic engineering. Therefore, metabolites and/or metabolite patterns are effective biomarkers for evaluating effects of internal and external stresses and therefore widely used in systems biology and biotechnology. Several studies have shown the application of metabolomics in crop breeding, genetic modification and biomarker discovery, for evaluation of intended/unintended changes and for assessing the quality of the final products (Fernie and Schauer, 2009; Kusano and Saito, 2012; Larkin and Harrigan, 2007).

Metabolomic analyses of biological systems consist of several steps that are equally important in order to draw meaningful conclusions. Metabolomic analysis of biological systems usually follow experimental design, sample preparation, metabolite extraction, data acquisition, data pre-treatment,

data analysis and interpretation. In this review, we focus on each step of the metabolomic workflow that is part of most cereal metabolomic studies (Figure 1). Metabolite analysis of cereal plants goes back to the early 20<sup>th</sup> century and the chemical composition of different cereals was always of central interest among plant biologists, plant breeders and farmers. Breeding strategies of cereal cultivars have aimed at improving the yield and/or improving desired quality traits of the plants, and this has boosted the application of metabolomics in cereal science.



**Figure 1.** General Overview of Plant Metabolomics Studies

The chemical composition of a cereal is one of the most important characteristics that define the value of the product. The main part of the global cereal production is utilized as animal feed where amino acid (protein) composition and metabolizable energy are the most important quality traits. Quantitative and qualitative analyses of health beneficial components such as dietary fibres, proteins, vitamins, sterols, polyphenols and other primary and secondary metabolites largely determine the phenotype and thus is a fundamental base in most cereal science studies focussing on human foods. Most metabolomic studies related to cereals are aimed at determining the chemical composition of cereals and/or cereal products and at understanding the plants' response to internal and/or external factors. This review summarises, current cereal metabolomic studies, differentiates the purpose-orientated metabolomics from untargeted approaches and highlights the role of the current analytical platforms including their advantages and limitations. In this review, the main steps in the quantitative metabolomics technology workflow are discussed in the order in which they are performed. Minimization of non-sample-related variation is illustrated and main sources of experimental errors are highlighted. A concise tutorial on the preparation of raw metabolomic data for multivariate data analysis is provided. The main purpose of each step of metabolomic data preprocessing, including



noise reduction, metabolite alignment, peak deconvolution, normalization and scaling are explained with examples and useful tools are described in detail. Advantages and limitations of the state-of-the-art, semi-automated complex chromatographic data processing tools are described as well.

Interpretation of metabolomic data and subsequent biological interpretation require an appropriate statistical treatment. Some of the commonly applied statistical approaches including unsupervised and supervised multivariate methods are discussed and useful recommendations are provided. Sources of common errors that lead to misinterpretations are illustrated. The most frequently used classification methods such as PCA, OPLS-DA, PLS-DA, SIMCA and ECVA are demonstrated with examples and their advantages and drawbacks are compared. In addition, the review compiles recent studies conducted on metabolomic analysis of the main cereals, maize, rice, wheat, barley, oat and rye. The current trends in metabolomics are finally set in perspective of future developments of integrated cereal phenomics laboratories.

## **2 Cereal metabolomics**

### **2.1 Background, definition and motivation**

Cereal metabolomics comprise all kinds of qualitative and/or quantitative measurements of metabolites from cereal plants such as maize, rice, wheat, barley, rye and oat (Figure 1). Historically, analysis of chemical composition of cereals has attracted much attention, mainly because cereals have always been the main food products consumed by humans since development of agriculture. The very first attempts at chemical analysis of cereals were primarily focused on measurement of nitrogen containing compounds (Teller, 1935), phosphorus containing compounds (Anderson, 1912; Rooke et al., 1949), dietary fibre (Vandekamer and Vanginkel, 1952), sugars (Clegg, 1955; Ponte et al., 1969) and protein content (Bietz and Kruger, 1988; Wisner and Jones, 1971). From the middle of the 20<sup>th</sup> century, by development of chromatography, mass spectrometry and various spectroscopic techniques, chemical composition analysis of cereal plants has significantly broadened. In the 1950s, substantial amounts of research were performed on wheat plants to understand the regulation of protein (Bilinski and Mcconnell, 1958b), carbohydrate (Mcconnell et al., 1958) and energy expenditure systems (Bilinski and Mcconnell, 1958a; Mcconnell, 1959) by using <sup>14</sup>C labelling. One of the early phenolic profile screens of various cereal plants was performed in 1962 (Bardinskaya and Shubert, 1962). Until 2000, most cereal metabolomic studies were based on targeted analysis of vitamins (Sampson et al., 1996), sterols (Berry et al., 1968; Kemp and Mercer, 1968), phenolics (Collins et al., 1991; Maier et al., 1995; Sridhar and Ou, 1974), volatile compounds (Withycom et al., 1974) and other metabolites that are known to be related to responses to biotic and/or abiotic stresses (Baker and Smith, 1977; Tsai and Tood, 1972).

The first comprehensive metabolomic analyses applied to cereal science dealt with metabolomic fingerprinting of field-grown transgenic wheat samples by using 1D  $^1\text{H}$  NMR and GC-MS (Baker et al., 2006) and metabolomic profiling of rice plants during plant development by using GC-MS (Tarpley et al., 2005). Continuous development of analytical techniques and advances in the analysis and subsequent interpretation of highly complex metabolomic data sets have significantly broadened the role of metabolomics in cereal sciences. Today, metabolomics based studies assist the elucidation of important biological phenomena that a few years ago could not be effectively resolved.

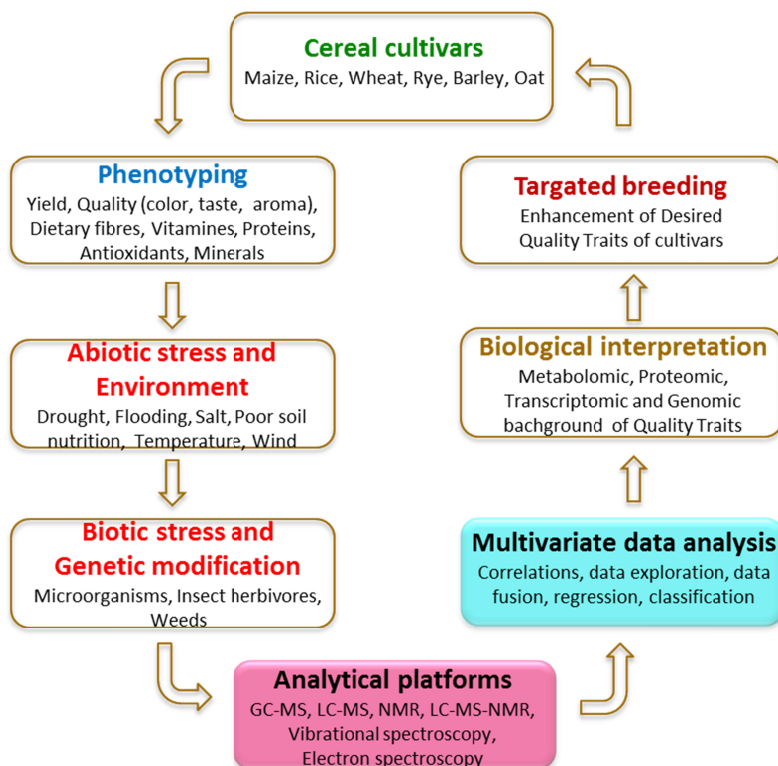
Based on the purpose and the type of information gained, metabolomics has been divided into different approaches e.g., targeted analysis, metabolomic profiling, metabolomic fingerprinting (Dunn, 2008; Fiehn, 2002). However, it is becoming increasingly evident that biological systems are multivariate and very unlikely to have only a few changes related to one specific effect. In most cases, breeding, gene modifications and biotic/abiotic stresses result in multiple changes in the whole phenome of plants (Munck et al., 2001; Munck et al., 2010) also affecting the metabolome in a characteristic way. Therefore, today, most metabolomic studies aim at covering as broad range of metabolites as possible to be able to evaluate both, expected and unexpected metabolomic changes. Up to date, there is no a single method that enables detection of the complete metabolome of a biological system. Therefore, almost all metabolomic approaches require a compromise to be made to obtain qualitative and quantitative metabolomic data. Most cereal metabolomic studies aim to assist in development and/or improvement of cultivars towards increasing their production, resistance against various biotic/abiotic stresses and enhance their health beneficial properties such as e.g., dietary fibre and antioxidant phenolic acids. When conducting such comprehensive studies it is worth covering as wide a range of metabolites as possible. For example, when investigating an effect of gene modifications on the metabolite composition of maize, it is important not only to focus on carbohydrates or proteins, but also evaluate changes in metabolites that occurred in low concentration levels (e.g., sterols, vitamins, and polyphenols). Since the function of one gene may be closely related to the functions of the other gene, and one or more metabolic pathways may be interconnected to some extent. Thus, drawing conclusions on the positive and negative effects of one specific gene modification must consider multiple factors. In addition, one must consider the high-throughput capabilities of the applied analytical techniques (number of samples involved in the study) and ensure that it is suitable for obtaining a required quality metabolomic data.

## **2.2 Beyond polysaccharides and proteins**

Cereal grains mainly consist of starch, dietary fiber, proteins, sugars, lipids and minerals (Kong et al., 1995; Quinde et al., 2004). Despite this, the presence of small metabolites such as phenolic acids, plant sterols and flavonoids contribute significantly to the important values of the grains. Recent studies

have documented the health beneficial effects of these low molecular weight phytochemicals and revealed their antioxidant, radical scavenging and antiproliferative properties (Madhujith and Shahidi, 2007). Among these, polyphenols have recently received much attention due to their preventive properties against various biotic and abiotic stresses (Amarowicz et al., 2007; Manach et al., 2004; Zielinski and Kozłowska, 2000). Cereals and cereal products are one of the richest sources of human polyphenol intake. The main health beneficial polyphenols of whole-grain cereals are phenolic compounds derived from 4-hydroxybenzoic acid (e.g., vanillic, gallic, and protocatechuic acids) and hydroxycinnamic (e.g., ferulic, caffeic, and coumaric) acids. Cereal phenolic compounds are mainly present in free, conjugated and bonded forms with sugars or other cell membrane components that alter their solubility and thus their bioavailability and bioactivity. Phenolic acids are important texturizing agents in cooking-extrusion of cereals (Gibson and Strauss, 1991) and recognized as the main antioxidant constituents of cereal and cereal products (Vinson et al., 2009).

In addition, some argue that the main health beneficial effects of barley, associated with its  $\beta$ -glucan content, in fact might be highly dependent on the content of polyphenols and antioxidants (Thondre et al., 2011). The phytochemical composition of cereals have been studied in a number of projects within the HEALTHGRAIN diversity-screening program (<http://www.healthgrain.org/>) (Andersson et al., 2008; Li et al., 2008; Nyström et al., 2008; Shewry et al., 2008; Ward et al., 2008). These studies showed that phytochemical composition of even a single cereal variety may greatly vary based on the growing conditions and geographical regions and thus differentiate the value of the final grain products. Other studies have opened new perspectives on targeted breeding of cereal cultivars for increased health beneficial phytochemicals. All these issues emphasize the importance of low molecular weight metabolites of cereals in development of future crops, and opens for a new research field related to cereal science through metabolomics.



**Figure 2.** Overview of Integrated PHENOMICS Laboratory

### 2.3 Cereal phenomics

Cereal phenomics is an emerging field of cereal science that comprises several ‘omics’ technologies to measure the physical and biochemical properties of cereal plants. Cereal phenomics play a key role in assessment, development and improvement of cereal cultivars. The importance of plant phenotype analysis, its objectives and main methodologies have previously been addressed (Eberius and Lima-Guerra, 2009; Gerlai, 2002). Today, technological developments allow high-throughput acquisition of qualitative and quantitative phenotype data of plants. However, in cereal science, these technologies are not yet combined in a way that would facilitate comprehensive phenotype analysis. Cereal phenomics require close cooperation between several different disciplines, from farming to bioinformatics and chemistry. Development of a new cereal phenotype usually start from the consumer and the farmer, by determination of the desired quality traits of the cereal cultivars followed by identification of the biological background of that specific quality trait. In fact, this is one of the most challenging steps for cross-disciplinary fields such as genomics, transcriptomics, proteomics and metabolomics. An overview of cereal phenomics is illustrated in Figure 2. The

establishment of all the steps involved in cereal phenomics is challenging and sometimes the application of the latest technologies and the most expensive analytical platforms may not be able to solve the problem. Even today, the most efficient way to solve these issues is to gain more experience, understand the multivariate nature of the biological phenomena and develop comprehensive analytical and statistical methods.

### **3 Experimental design and sampling**

#### **3.1 Experimental design and optimization**

A comprehensive metabolomic study of complex biological samples requires an appropriate experimental design and optimization of protocols. Design of a quantitative metabolomics experiment must enhance the statistical properties of the results obtained based on absolute and/or relative concentrations of the metabolites (Broadhurst and Kell, 2006). For example, if the aim of a metabolomic study is to evaluate the effect of two different treatments on growing barley plants, it is essential to collect enough and representative sample material (mass reduction) from both treatments and to collect control samples grown under ordinary conditions. Representative sampling can be effectively investigated by the theory of sampling, but so far sampling studies have been limited to Near-Infrared Transmission (NIT) spectroscopic sampling of single kernels (Tonning et al., 2006). Then, the collected samples must be analysed in randomized order and each biological replicate should be divided into at least two technical replicates to determine the variation caused by the analytical measurements. It is important to be able to discriminate variation caused by the treatment and measurement errors from the total variation and elucidate only the true biological variation related to the effects of two treatments.

Optimization of metabolomic data acquisition protocols is mainly determined by the purpose of the study. Targeted metabolomics focuses on quantitative detection of one or few metabolites and optimization of protocols for such a study is simpler than in untargeted metabolomics. Development of an optimal measurement protocol usually requires a compromise, since metabolites of biological samples are very diverse and cannot be detected by applying a single protocol. However, when plants are exposed to different treatments or stress, it becomes difficult, if not impossible, to know a priori all of the metabolomic pathways and networks that will be perturbed. Therefore, most metabolomics studies are designed for detection of as wide a range of metabolites as possible. When optimizing a metabolomic protocol, it is very important to determine the correct response variable(s) that will improve data quality. For example, in targeted analysis the method can be optimized to increase the signal to noise (s/n) ratio of the desired metabolite or to reduce the experiment time and cost. In untargeted metabolomics, the prioritized response variable must be the reproducibility of the protocol, since only reproducible metabolomic profiles can provide reliable biological information.

Normally, at the early stages of the metabolomics study, analysts will not be aware of the ratio of the true biological variation to the variation caused by experimental errors. Therefore, irreproducible metabolomic protocols may hide true information and result in misinterpretation. Appropriate optimization and validation of the protocols will ensure that the possible sources of error are minimized and reproducibility optimized. However, protocols should facilitate detection of as many metabolites as possible to increase the chance of finding biomarkers or patterns that may explain biological phenomena.

Metabolomic data acquisition protocols include parameters such as solvent concentration, extraction time and temperature and variation of these parameters will affect the observed metabolomic profile. By varying these parameters in different combinations, it is possible to find an optimal condition. This will require a large number of experiments to be performed which is expensive and time consuming. If the metabolomic analysis involves extra steps such as sample derivatization, several new parameters such as derivatization time, temperature need to be optimized. For instance, the optimization of a GC-MS protocol may include five very important parameters: extraction solvent concentration, time, temperature, derivatization time and temperature. Then, optimisation of the protocol by varying all the parameters in four different levels individually, will require a total of  $4^5 = 1024$  experiments to be performed. By applying, a DoE approach e.g. fractional factorial analysis or D-Optimal design; it is possible to find an optimal metabolomic protocol by performing much less experiments.

Several approaches of design of experiment have been applied to optimize metabolomic protocols. As an example, (Gullberg et al., 2004) applied fractional factorial analysis followed by Multiple Linear Regression (MLR) to optimize the metabolite extraction protocol from *Arabidopsis thaliana* leaves and for optimization of a metabolite derivatization protocol for GC-MS. Prior to the GC-MS analysis, they used a D-optimal design and Partial Least Squares regression (PLS). Recently, (Danielsson et al., 2012) applied statistical design of experiments for optimization of a derivatization method in GC-MS metabolomic analysis of blood plasma samples.

## **3.2 Sample preparation and metabolite extraction**

Harvesting of plant material, sample preparation and metabolite extraction are crucial steps in cereal metabolomics. These are the major steps when experimental errors occur that may significantly deteriorate quantitative metabolomic data. Most cereal metabolomic studies use plant leaves and/or grain samples (hulled or hull less grains). As mentioned earlier, the leaf metabolome is always in an equilibrium state determined by internal and external factors. At the harvest time plants experience stress that may lead to additional alterations of the metabolome. Therefore, metabolomic changes in harvested plant samples are usually halted by snap freezing in liquid nitrogen or by rapid cooling and freeze-drying. Particularly for high-throughput metabolomic studies that deal with hundreds or even

thousands of samples, plant materials are directly sampled into liquid nitrogen followed by freeze-drying. This serves to preserve the plant tissue and inhibit enzymatic activity and can be used for both leave and grain samples. Harvested samples must always be at least at – 20 °C freezer until they are used for extraction.

When quantitative data are required from the metabolomic study, it is essential that samples are analysed in randomized order and that all samples are handled identically. Practical aspects of the metabolomics protocol must be considered before starting the experiments:

1) solvents (preferably, not too volatile, since volatile solvents will cause volume errors due to evaporation and pipetting difficulties), 2) extraction tubes (must be inert to the used solvent, prevent evaporation, and resistant to the applied temperature), 3) extraction temperature (temperature stability of metabolites must be taken into account), 4) samples must be processed in smaller batches (extraction of large number of samples at the same time may increase the error, 5) all the samples that are subject to comparison must be processed by exactly the same protocol preferably by using a robot.

Different metabolomic protocols have been applied in cereal studies. For example, an extraction protocol for wheat phenolic acids using 80% methanol followed by basic hydrolysis was provided by (Li et al., 2008), while comprehensive metabolomic protocol for measuring a polar metabolomic fingerprints of barley plants by using 100% methanol has been reported (Gorzolka et al., 2012). Another study reports a screening protocol for plant sterols in cereals based on GC-MS metabolomics (Piironen et al., 2002). Extraction of lipids and polar metabolites from rice flour samples was demonstrated by Frank et al. by using a single protocol based on 100% methanol extraction (Frank et al., 2007). After complete drying of the extract, the lipophilic metabolites were recovered by dichloromethane, while polar metabolites were extracted using 80% methanol. A simple protocol for extraction of free amino acids present in the wheat flour samples was presented in a study that dealt with four different wheat lines to study the effect of environment and genotype on the metabolite composition of cereals (Curtis et al., 2009). They used 0.01 M hydrochloric acid to extract the amino acids at room temperature and the obtained extracts were analysed by GC-MS analysis.

## **4 High-throughput analytical platforms**

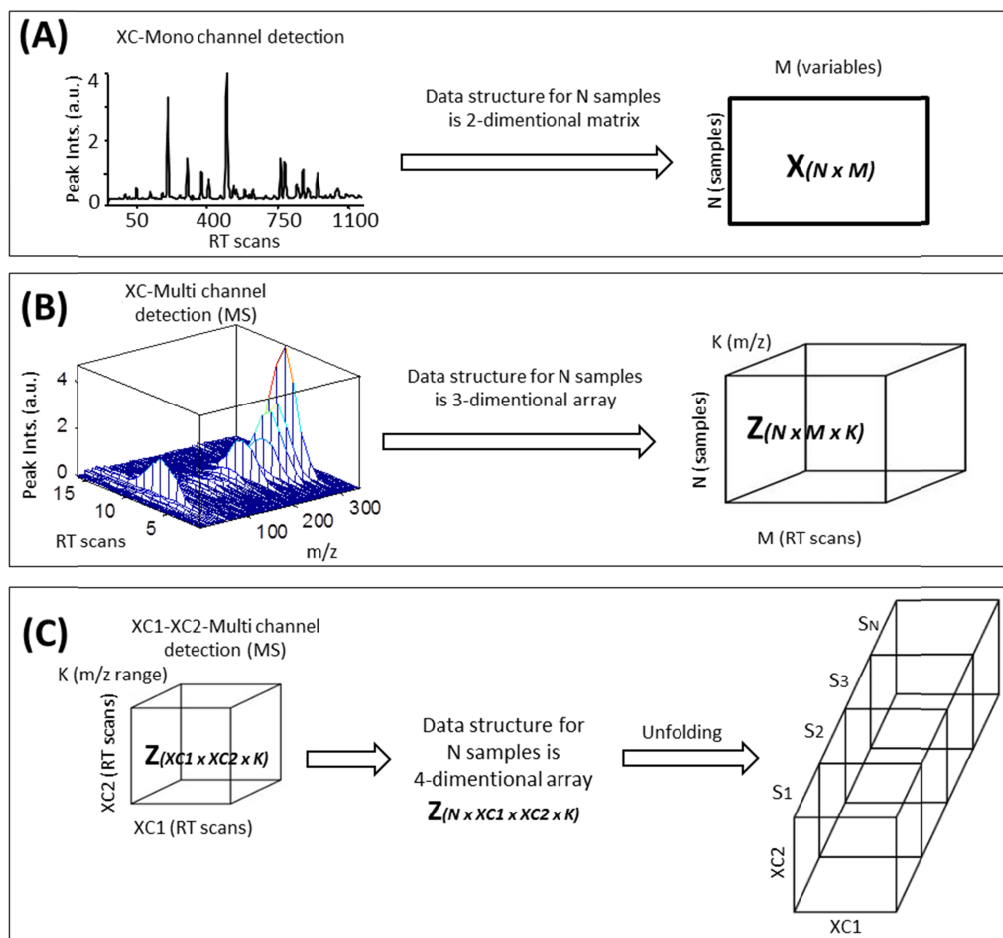
Most of the recent achievements in systems biology, metabolome flux analysis and biomarker discovery are the results of advances in the analytical platforms such as GC and LC coupled to different types of detectors (e.g., MS, UV, DAD, FID), high resolution NMR, hyphenated NMR (LC-NMR, LC-solid phase extraction-NMR) and other non-destructive spectroscopic methods (e.g., IR, NIR). Depending on the detectors' capabilities and analysis mode, these analytical platforms may generate two, three, four

and even higher dimensional metabolomic data (Figure 3A-3C). In table 1 and 2, we provide a general comparison of the most often used analytical platforms in high-throughput cereal metabolomics. These platforms are further discussed briefly in the sub-sections below. Detailed descriptions of the recent developments in GC-MS, LC-MS, NMR and other spectroscopic methods can e.g. be found in (Holcapek et al., 2012; Santos and Galceran, 2003).

#### **4.1 LC-MS**

Liquid chromatography coupled to mass spectrometry (LC-MS) can be defined in its simplest terms as a technique that allows mass spectrometric detection of metabolites that are separated based on their different partitioning coefficients between the mobile phase (solvent) and stationary phase (column). LC-MS has become the favourite choice of separation after implementation of electrospray ionization (ESI). ESI offers a good compromise between ionization of non-volatile/polar metabolites that could not be ionized by the GC-MS conventional ionization techniques (EI and CI), and minimization problems caused by the LC solvent entering into the sample interface. In terms of ionization, ESI can be considered as a method between EI and CI, since it generates less fragmentation ions than EI, but more than CI. Today, ESI is the most commonly used ionization technique in LC-MS analysis. ESI fragmentation pattern of metabolites are highly depend on metabolite structure and may provide an informative fragmentation pattern as well as the molecular mass of the precursor ion. In addition, some LC-MS platforms employ atmospheric pressure chemical ionization (APCI) and atmospheric pressure photoionization (APPI) techniques that are limited to ionization of volatile compounds.





**Figure 3.** Structure of metabolomic data. **(A)** Chromatography coupled to mono channel detectors (one measurement per one scan) generate two dimensional data, for example, GC-FID, GC-SIM(MS), LC-UV, LC-SIM(MS) and CE-UV. Data obtained from 1D 1H NMR experiments also falls into this category. **(B)** Chromatography coupled to multi-channel detectors generates three-dimensional data, for example, GC-MS, LC-MS, CE-MS, LC-DAD. Data obtained from 2D NMR experiments also fall into this category. **(C)** Sequential combination of two chromatographic separation and multi-channel detectors generate four-dimensional data, for example, GC x GC-MS, LC x GC-MS and LC x LC-MS. These type of data usually analyzed either by unfolding or summing one of the dimensions.

While GC-MS is mainly applicable to the analysis of volatile organic compounds and derivatization-based detection of non-volatile/polar metabolites, LC-MS allows detection of wider range of metabolites without a prior derivatization step. As metabolites of complex mixtures separately elute into the LC-MS ionization chamber, they form  $m/z$  ions representative to metabolites, followed by separation of produced ions based on their  $m/z$  values and ions finally reach a MS detector. The resolution and mass accuracy of the detected  $m/z$  peaks depend on the type of the mass analyser

(Table 2). As in all other separation coupled to mass spectrometry instruments, LC-MS records a mass spectrum at each elution scan point and the generated data will have a three-dimensional structure (Figure 3B). The resolution of LC-MS chromatographic signals is highly depend on both the LC separation and on the scan speed of the mass analyser. Recent advances in LC-MS allowed significant improvements in the resolution power and peak capacity of the techniques (Allwood and Goodacre, 2010). Ultra-high performance LC-MS systems allowed detection of up to several hundred metabolites from a single analysis of a complex plant samples (De Vos et al., 2007).

Compared to GC-MS, the application of atmospheric pressure ionization (API) in LC-MS allows detection of ions in positive and negative modes, which consequently improve the sensitivity of LC-MS towards the analysis of trace compounds. However, LC-MS based metabolite databases are not as rich as GC-MS libraries, which make LC-MS peak annotation difficult. Nevertheless, the choice of ionization and ion separation methods of mass spectrometry is much greater in LC-MS than in GC-MS. Several types of mass analysers (Q, TOF, Q-TOF, QQQ and IT) have successively been coupled with LC (Table 2). Quadrupole (Q) mass analysers are one of the most frequently used techniques in LC-MS based plant metabolomics. The scanning mode nature of Q mass analysers through the whole m/z range allows simultaneous performance of full (e.g., 50-1200 m/z) as well as selected ion monitoring (SIM) experiments. SIM increases sensitivity and selectivity of predetermined metabolites and is widely used in targeted metabolomics. On the other hand, IT and TOF based mass analysers offer higher mass accuracy and enable identification of unknown compounds. LC-MS experiments utilizing different mass analysers have been successfully applied in plant and cereal metabolomics (Chang et al., 2012; Grata et al., 2008; Guerard et al., 2011; Kuzina et al., 2009; Qiu et al., 2010). The application of tandem mass spectrometry (MS/MS) has further assisted the identification of unknown metabolites (Xu et al., 2007). In tandem mass spectrometry, two or more mass analysers are coupled to each other by collision-induced ionization chambers (e.g., Q-TOF, Q-IT) where one m/z ion is trapped, further ionized and the generated fragment m/z ions are separated by the second mass analyser. This approach will allow tentative characterization of unknowns such as the number of the sugar moieties in glycosides.

**Table 1.** Advantages and drawbacks of analytical platforms

Analytical platforms	Advantages	Drawbacks
LC-MS	Allow analysis of a wide range of metabolites without prior derivatization (up to 60k Da) No requirements for metabolites to be volatile Great sensitivity towards polar/easily ionized metabolites	Sensitivity of the technique (especially towards polar metabolites) depend on a mobile phase (pH, polarity, gradient program) Ion suppression Difficult to ionize volatile

	<p>High mass accuracy combined with databases allow identification of unknown</p> <p>Tandem MS provide valuable structural information and lays strong foundation for development of LC-MS based libraries</p> <p>Larger volume of sample e.g., 1 to 50 ml (depending on the system) can be injected and allow metabolite purification</p>	<p>metabolites</p> <p>Low chromatographic resolution of structurally similar metabolites</p> <p>Expensive (prices for high mass accuracy platforms exceptionally high)</p>
CE-MS	<p>Provide simpler method of metabolite separation based on their charge, mass, and size</p> <p>Provide higher resolution of metabolites compared to LC (usually width of peaks are few seconds)</p> <p>Allow separation of proteins, nucleic acids, ionic and very polar metabolites that are complicated in LC and GC</p> <p>Provide very consistent migration times of metabolites, provided that experimental temperature and buffer is stable</p> <p>Simpler sample preparation (allow analysis of heterogeneous samples with interfering constituents like lipids and precipitates)</p>	<p>Only few <math>\mu</math>l of sample can be loaded, therefore less sensitive than LC and metabolite purification is not possible</p> <p>Resolution power highly depends on polarity and pH of solvent and require prior optimization</p> <p>Migration times of the same metabolites fluctuate by changing the temperature of environment</p> <p>Limitations in electrolyte selections</p>
GC-MS	<p>High chromatographic resolution allow separation of several hundred metabolites, including structurally similar metabolites like <i>cis</i> and <i>trans</i> stereoisomers of fatty acids</p> <p>Greater sensitivity towards non-polar and volatile metabolites compared to LC-MS</p> <p>Harsh ionization technique, EI provides valuable fingerprint of metabolites</p> <p>Rich EI-MS libraries are available that comprise several hundred thousand metabolites</p> <p>Cheaper than LC-MS and lower running cost (solvent free)</p>	<p>Narrower range of metabolites can be detected compared to LC-MS</p> <p>Requires prior derivatization for detection of polar and non-volatile metabolites</p> <p>EI approach usually gives no information about the mass of molecular ion</p> <p>High temperature may increase level of column and derivatization reagent based artifact peaks</p>
NMR	<p>Allow complete structure elucidation of unknown metabolites</p> <p>Non-destructive (analyzed samples can be reused)</p> <p>Less biased than MS based techniques (all metabolites containing NMR active nucleus e.g., <math>^1\text{H}</math>, <math>^{13}\text{C}</math> can be detected, no matter of their volatility, polarity, molecular weight, size, chemical structure and the sample matrix)</p>	<p>Less sensitive than MS based methods (usually provide detection of most abundant metabolites)</p> <p>NMR signals of different metabolites may be overlapped and hamper quantification</p> <p>Requires expensive, NMR suitable, deuterated solvents and higher running cost than most MS based methods</p>

	Provide higher reproducibility and lower experimental error than MS based methods Minimal sample preparation methods Compound coverage is excellent Good for untargeted profiling of primary metabolism, phenolic acids and oligomerized polyphenols	Expensive (prices for high resolution platforms exceptionally high) Measurement speed is medium Reproducibility is medium
Vibrational spectroscopy	Non-destructive analysis Provide high-throughput analysis of large number of samples with a minimum sample preparation Highly reproducible High measurement speed Applicable in on-line process control Well established and validated methods are available Compound coverage is excellent Good for untargeted profiling of the bulk constituents e.g., carbohydrates, proteins and lipids with high speed and accuracy	Limited structural information can be gained and it highly depend on the metabolites e.g., type of functional groups, polarity Poor resolution

## 4.2 CE-MS

Capillary electrophoresis-mass spectrometry (CE-MS) is one of the most versatile analytical techniques widely used in proteomics, metabolomics and forensic science (Kolch et al., 2005; Mischak et al., 2009). The separation principle of CE is simpler than in GC and LC, and molecules are separated based on their charge and size by using a capillary tube and electric field. One end of the capillary tube is connected to the source vial, which is coupled to the anode and the other end is connected to the destination vial, which is coupled to the cathode. Both vials are filled with electrolytic solution and by applying electric power, cations generated at the source vial start to migrate to the destination vial and form an electroosmotic flow. At the same time, molecules of the sample mixture also migrate from the source vial to the destination vial. The migration time of molecular ions is directly proportional to their electrophoretic mobility ( $\mu_{eo}$ ), which is a function of their electrophoretic velocity ( $veo$ ) and applied electric field ( $E$ ) ( $\mu_{eo} = veo/E$ ). The migration times of cations through the capillary differ depending on their charge, size and the electroosmotic flow of the solution. Cations with the same charges, but different sizes will have different migration times; smaller size cations will

migrate faster than larger cations. In a similar manner, the migration times of cations of equal size that differ in their charges will be different and highly charged cations arrive at the cathode earlier.

In contrast to the cations, anions are attracted to the anode and attempt to remain in the source vial. Since the velocity of the electroosmotic flow exceeds the velocity of the anions that are attracted to the anode eventually all ions will migrate to the destination vial. Therefore, the migration rates of most anions are higher than the migration rates of cations. Highly charged and bigger anions migrate slower compared to the lower charged and smaller anions. An applied electric field does not influence neutral molecules and their migration towards the destination vial depends only on the physical interaction with the electroosmotic flow of the solution and on their size.

The nature of the capillary tube coating largely determines the resolution and sensitivity of CE-MS (Erny et al., 2006). Acidic coatings with pH of 2-2.5 are frequently used for the separation of peptides and proteins which significantly reduce their interactions with the capillary wall (Kasicka, 2012). The most common CE-MS combination involves ESI and MALDI ionization techniques (Hommerson et al., 2011). ESI is a harsher way of ionization and is used in analysis of a wider range of metabolites and cellular macromolecules than MALDI, which is mostly employed in proteomics and peptidomics. Due to its high-resolution power CE-MS has been widely applied in metabolomic analysis of complex samples from e.g., plants, microorganisms, blood and urine (Ramautar et al., 2009). One of the advantages of CE over LC is that it allows analysis of heterogeneous samples e.g. with interfering lipids and precipitates. Stable buffer-gradient (electrolytic solution) minimize the interference that is caused by the continuous gradient change in LC. In addition, the resolution of CE-MS, in many cases, exceeds the resolution of LC-MS. The higher resolution power of CE is mainly due to the capillary tube wall that drives the mobility of the ions and provides an even flow and narrow peaks. In contrast, LC-MS columns resist metabolite mobility and cause differences in mobility between the centre and sides of the column that in turn result in broader peaks. Currently available CE-MS techniques only allow loading of up to 1  $\mu$ l sample, while several ml of sample can be injected into LC-MS. This limits the comprehensive profiling of complex samples where the detection of low concentration metabolites is difficult and limits the performance of CE when it is coupled to tandem mass spectrometry.

### **4.3 GC-MS**

Gas chromatography coupled to mass spectrometry (GC-MS) is the best established hyphenated analytical technique used in metabolomics. The first GC-MS was developed in the 1950s (Gohlke and McLafferty, 1993) and today, modern GC-MS equipment allow simultaneous separation and detection of several hundred metabolites in complex biological mixtures from a single analysis (Fiehn, 2008; Lisec et al., 2006; Roessner et al., 2000). Its wide applicability in a broad range of metabolomic analyses has resulted in detailed studies of all the steps involved in comprehensive GC-MS analysis (e.g. sample

preparation, derivatization, optimization, GC separation, MS settings) (Danielsson et al., 2012; Kanani et al., 2008; Khakimov et al., 2013; Koek et al., 2011; Lisec et al., 2006; Pasikanti et al., 2008; Xu et al., 2010). GC-MS applies isothermal or gradient temperature programs that allow vaporization of metabolites and a constant flow of a carrier gas (e.g., helium or hydrogen) enables separation of the metabolites inside the GC column. The eluted gas phase metabolites are then detected using mass spectrometry. At each elution time point a full mass spectrum will be recorded at the specified  $m/z$  range (e.g., 50-500  $m/z$ ) and the GC-MS data obtained for one sample will be represented by the intensities of  $m/z$  ions recorded at each elution time points (Figure 3B).

For GC-MS analysis, the metabolites must be volatile and thermally stable under the given conditions of the instrument. Recent advances in chemical derivatization of polar and non-volatile compounds have considerably widened the coverage of metabolites that can be analysed by GC-MS (Khakimov et al., 2013). Currently, GC-MS allows the detection of various primary and secondary metabolites with molecular mass of up to 1200 Da. The high chromatographic resolution and high reproducibility of GC-MS has resulted in wide use of the platform in high-throughput metabolomic profiling of complex biological samples (Fiehn, 2008). Several equipment manufacturers have developed autosamplers that can be integrated with the GC-MS and which enable automatic sample preparation, derivatization and injection. This, in turn, has provided even greater reproducibility, reduced experimental time, reduced human interference and enhanced high-throughput analysis. Advanced injection systems developed for GC-MS allow split or splitless injections of small or large volume, cold or hot injections as well as temperature programmable sample injection modes (e.g., the cooled injection system (CIS) from GERSTEL or programmed temperature vaporization (PTV) from Agilent). Depending on the purpose of the analysis, the different injection modes of these systems can be used interchangeably. Moreover, headspace GC-MS systems allow the detection of already volatile metabolites (e.g., aroma and small molecular organic compounds) and provide quantitative volatile profiles. Increased chromatographic resolution is achievable by applying GC x GC-MS where two GC columns are coupled to allow increased separation of closely eluted metabolites from the first column onto the second column. This platform may provide rich metabolomic fingerprints and the data obtained for each sample will have three-dimensional structure (GC1 x GC2-MS) (Figure 3C).

Ionization techniques in GC-MS are limited to those that ionize metabolites in a gas phase e.g. electron ionization (EI), chemical ionization (CI), atmospheric pressure chemical ionization (APCI), field ionization (FI) and electron-capture negative ionization (ECNI). Historically, electron ionization method (EI), which is considered as a harsh ionization, has successfully been used in conjunction with GC-MS. In EI, gas phase compounds are ionized by a beam of high-energy (70 eV) electrons and provide representative and reproducible mass spectra of each compound with a unique fragmentation pattern. Therefore, the EI-MS spectra of compounds are independent of the sample matrix and comparable between labs. These fragmentation patterns ( $m/z$  ions and their ratios to each other) are stored in libraries as a fingerprint of the compound. Another type of ionization method, which is less frequently

used in GC-MS, is chemical ionization (CI). CI is considered to be a soft form of ionization and may provide information about the mass of the molecular ion but provides less fragmentation ions.

The mass spectra of compounds recorded by GC-MS also depend on the type of mass analyser. In modern GC-MS analysis, three different kinds of mass analysers: quadrupole (Q), ion trap (IT) and time-of-flight (TOF) are mainly employed and differ in their mass resolution power, mass accuracy, mass range, sensitivity, linear dynamic range and scan speed. GC coupled to quadrupole (Q) MS has found widest due to its simplicity, robustness, great dynamic range, sensitivity, low cost and small physical size. However, most commercial GC-MS instruments that are based on quadrupole MS have less mass resolution power (100-1000 FWHM) and mass accuracy (100 ppm) compared to TOF-MS (resolution power: 1000-40 000 FWHM, accuracy: 5-50 ppm) and less sensitivity than IT based instruments (Hart-Smith and Blanksby, 2011). Since metabolites are continuously eluted from the GC column into the ionization chamber, the MS analyser must be able to scan the required  $m/z$  range with a reasonable scan speed in order to obtain high quality mass spectra, symmetric peaks and to be able to resolve closely eluted compounds. A high scan speed of the MS analyser becomes even more important when quadrupole MS is used, since the mass spectra are recorded in the whole  $m/z$  range within a small time interval (ions with different  $m/z$  values pass through quadrupole one-after-another, from low  $m/z$  to high  $m/z$ ). In addition to the overlapping and non-resolved peaks, a low scan rate in quadrupole MS may cause a distortion of the original relative mass spectral peak intensities of pure metabolites (mass spectral skewing). This is due to the unstable partial pressure of the metabolite inside the ionization chamber, since metabolites are continuously eluted into the ionization chamber (Watson and Sparkman, 2007c). It is a commonly accepted rule that the acquisition time of complete mass spectra must not exceed one-fifth of the duration of the chromatographic peak. More importantly, there should at least be nine scan points per chromatographic peak for precise quantification. In fact, most quadrupole GC-MS techniques readily provides a high quality mass spectra with 2 to 3 scans per second when recording spectra in the range of 50-500  $m/z$  at unit resolution (Watson and Sparkman, 2007c). Such configuration of quadrupole GC-MS allow resolution of few hundred peaks from complex mixture samples and high quality mass spectra of resolved peaks. Data sets obtained by such a quadrupole GC-MS configuration are favourably accepted by the scientific community and serve as a powerful tool in high-throughput metabolomic profiling of complex biological samples in last two decades.

In contrast to Q and IT, TOF based GC-MS instruments provide better sensitivity, more rapid spectral acquisition rate and higher quality mass spectra without any spectral skewing. This is due to their pulsed mode  $m/z$  separation nature, when all the ions of all  $m/z$  values are detected at each scan point. In quadrupole MS,  $m/z$  ions are separated by scanning through the  $m/z$  range but only a single  $m/z$  ion passes through the quadrupole rods and reaches the MS detector at a time. Thus, this results in poorer sensitivity due to loss of some ions during spectral acquisition time. The rapid and pulsed spectral acquisition mode of TOF mass analyser provides not only greater sensitivity, but also better

chromatographic peak resolution and shape. At a mass range of 70-600 m/z, GC-TOF-MS allow spectral acquisition rate of 20 scans per second with an ion abundance sensitivity of  $10^3$  and great mass resolution power and accuracy (Hart-Smith and Blanksby, 2011; Lisec et al., 2006). This facilitates identification of unknown metabolites, not only based on EI fragmentation pattern, but also on elemental composition. For more information about the principles, advantages and disadvantages of different ionization and mass separation techniques that are used in GC-MS, readers are advised to see references (Hart-Smith and Blanksby, 2011; Watson and Sparkman, 2007a; Watson and Sparkman, 2007b; Watson and Sparkman, 2007c). GC-MS data processing that involve retention time shift correction, deconvolution, baseline elimination and data analysis will be discussed in detail in section 5.1.



**Table 2.** Overview of performance parameters of most common mass spectrometers combined with LC, CE and GC. <sup>a</sup> Upper mass range limit of the specified mass analyzer, though it is not often possible to work in this range due to the limitation of the chromatographic separation. <sup>b</sup> Full width at half maximum height (FWHM) is often used measure of resolution performance and it is defined as  $RP = M/\Delta M$ , where M is m/z and  $\Delta M$  is the width of this m/z at 50% of the maximum height. <sup>c</sup> Upper limit of acquisition speed of GC and LC coupled mass spectrometers, however, this value highly depend of the investigated m/z range and in some cases it can even be higher.

Analytical platforms	LC-MS	CE-MS	GC-MS
Ionization Technique	Electrospray ionization (ESI) Atmospheric Pressure CI (APCI) Photoionization (APPI), Matrix-Assisted Laser Desorption Ionization (MALDI), Fast Atom Bombardment (FAB)	ESI, APCI, APPI, MALDI, FAB, Sonic Spray Ionization (SSI), Thermospray ionization (TSI)	Electron Ionization (EI) Chemical Ionization (CI)
Mass Separation Technique	Q, IT, TOF, quadrupole-TOF (Q-TOF), IT-TOF, triple-quadrupole (QQQ), Q-IT, IT-Orbitrap, Q-Orbitrap, IT-Fourier Transform Ion Cyclotron Resonance (FTICR), Q-FTICR	Q, IT, TOF, Q-TOF, QQQ	Quadrupole (Q) Ion trap (IT) Time-of-flight (TOF)
Mass range (m/z) <sup>a</sup>	Q-TOF: 100,000 Q-IT: 4,000 QQQ: 4,000 LIT-Orbitrap: 6,000 LIT-FTICR: 10,000	Q: 4,000 IT: 4,000 TOF: 100,000	Q: 4,000 IT: 4,000 TOF: 100,000
Mass resolving power (FWHM) <sup>b</sup>	Q-TOF: 10,000 Q-IT: 2,000 QQQ: 2,000 IT-Orbitrap: 100,000 IT-FTICR: 500,000	Q: 100-1000 IT: 1000-10,000 TOF: 1000-40,000 Q-TOF: 10,000 QQQ: 2,000	Q: 100-1000 IT: 1000-10,000 TOF: 1000-40,000
Mass accuracy (ppm)	Q-TOF: 2-5 Q-IT: 100 QQQ: 100 IT-Orbitrap: 2 IT-FTICR: < 2	Q: 100 IT: 50-100 TOF: 5-50 Q-TOF: 2-5 QQQ: 100	Q: 100 IT: 50-100 TOF: 5-50
Acquisition speed (scan/sec) <sup>c</sup>	Q-TOF: 50 Q-IT: 20 QQQ: 20 IT-Orbitrap: 10 IT-FTICR: 2	Q: 20 IT: 20 TOF: up to 500 Q-TOF: 50 QQQ: 20	Q: 20 IT: 20 TOF: up to 500
Linear Dynamic Range	Q-TOF: 10 <sup>5</sup> Q-IT: 10 <sup>5</sup> QQQ: 10 <sup>5</sup> IT-Orbitrap: 10 <sup>4</sup> IT-FTICR: 10 <sup>4</sup>	Q: 10 <sup>5</sup> -10 <sup>6</sup> IT: 10 <sup>4</sup> -10 <sup>5</sup> TOF: 10 <sup>4</sup> -10 <sup>5</sup> Q-TOF: 10 <sup>5</sup> QQQ: 10 <sup>5</sup>	Q: 10 <sup>5</sup> -10 <sup>6</sup> IT: 10 <sup>4</sup> -10 <sup>5</sup> TOF: 10 <sup>4</sup> -10 <sup>5</sup>

## 4.4 NMR

Nuclear magnetic resonance (NMR) spectroscopy is probably the most commonly used analytical technique in metabolomics and it allows structure elucidation of unknown compounds, evaluation of metabolomic changes associated with biotic and/or abiotic perturbations. In comparison with GC-MS, LC-MS and CE-MS, NMR is more unbiased, less destructive to the sample matrix and provides a simpler way of measuring the metabolome of the biological systems. The main advantage of NMR over MS based techniques is that it can quantitatively detect all metabolites present in the complex mixtures, no matter of their volatility, polarity, molecular weight, size, chemical structure and the sample matrix, provided that they possess chemical elements with non-zero spin quantum number, such as proton ( $^1\text{H}$ ), carbon ( $^{13}\text{C}$ ), phosphorous ( $^{31}\text{P}$ ) and nitrogen ( $^{15}\text{N}$ ). Since proton ( $^1\text{H}$ ) is the most suitable (highly abundant, most sensitive, produce sharp and informative NMR signals and allow rapid data acquisition) nuclei for NMR and it is the part of most metabolites, NMR analyses of metabolites are mainly based on measuring protons. Other important advantages of NMR is that it is a non-destructive method (analysed samples can be reused), relatively faster and requires much less labour for sample preparation compared to MS based methods.

In its simpler term, NMR spectroscopic analysis can be described as recording the energy released from the nucleus of the NMR active atoms in a molecule when they are returned to the original low energy spin state after being excited by external magnetic field ( $B_0$ ). This energy is defined by  $\Delta E = h\nu_0$  and  $\nu_0 = \gamma B_0 / (2\pi)$ , where  $B_0$  is the magnetic field,  $\gamma$  is the gyromagnetic ratio which is an atom specific parameter and  $\nu_0$  is the radiofrequency required for excitation of the low energy spin state (+1/2) nucleus to high energy spin state (-1/2). Depending on the chemistry of the molecule, nuclei of the same atoms may give NMR signals at the different frequencies ( $\nu_0$ ) of the applied magnetic field, which is generally referred to as the chemical shift (ppm). Chemical shift is the main qualitative characteristics of the nucleus gained from NMR analysis, and it describes the investigated atom's (e.g.,  $^1\text{H}$ ) chemical environment. In  $^1\text{H}$  NMR spectroscopy, the other two very important qualitative characteristics are the splitting pattern of NMR signals of proton (e.g., singlet, doublet, triplet, quartet or multiplet) and spin-spin coupling constant (measured in Hz). The spin-spin coupling constant refers to the distance between the splitted NMR signals of one type of proton due to the magnetic field effects of the neighbour protons. Thus, splitting patterns spin-spin coupling constants and chemical shifts carry important information on the functionality of and neighbouring protons of the investigated NMR signal. The coupling constant play a key role in structure elucidation of unknown metabolites and it becomes even more important when NMR signals are not well resolved and/or represent more than one metabolite. Resolution of overlapped NMR signals derived from different functional groups and/or metabolites can be resolved by application of two-dimensional NMR experiments. Two-dimensional NMR experiments such as correlation spectroscopy (COSY), total correlation spectroscopy (TOCSY), nuclear overhauser effect spectroscopy (NOESY), heteronuclear single quantum coherence

(HSQC) and heteronuclear multiple-bond correlation spectroscopy (HMBC) allow structure elucidation of unknown metabolites with complex chemical structures.

The inherent quantitative nature of NMR spectroscopy rely on the fact that the NMR signal intensity of the metabolites are directly related to their concentrations (Winning et al., 2008). This quantitative information is essential for both structure elucidation and quantitative metabolomics. These two important aspects of NMR have been widely applied in cereal metabolomics (Barding et al., 2013; Gavaghan et al., 2011). Apart from being a powerful tool in structure elucidation based targeted metabolomics, NMR plays a key role in untargeted metabolomic analysis of various complex biological samples (Gavaghan et al., 2011; Lopez-Rituerto et al., 2012). NMR spectra of complex cereal extracts may be referred as a fingerprint and represent the global metabolome at the given state. These complex NMR spectra may represent the most abundant metabolites of the sample (e.g. 30-50 metabolites), while low concentration metabolites will not be detected and/or their insignificant signals will be hidden by the signals of much more abundant metabolites. The main drawback of NMR is its lower sensitivity compared to MS based detection techniques as it requires a few  $\mu\text{m}$  of metabolites for high quality data. However, in many cases NMR based quantification is more precise and unbiased compared to MS based techniques.

Recent advancements in hyphenation of NMR with separation techniques (Jaroszewski, 2005a; Jaroszewski, 2005b) development of microflow NMR (Olson et al., 2004), high capacity autosamplers and high field magnets lead to minimization of sensitivity related issues, enhanced resolution and shortened analysis time. However, today, prices for these techniques are high and complications in routine operational procedures greatly hamper their utilization. The majority of high-throughput NMR metabolomics performed on cereal samples utilize only 1D  $^1\text{H}$  NMR experiments. However, 1D  $^1\text{H}$  NMR based cereal metabolomics illustrated a high potential to uncover elusive biological variations related to diseased and/or healthy states (Defeo et al., 2011; Lodi et al., 2013), effect of gene modifications (Barros et al., 2010), and influence of the environment on chemical composition of cereals (Graham et al., 2009). Further examples on application of NMR in cereal metabolomics and importance of preprocessing of raw NMR metabolomic data (baseline correction, alignment, normalization) for multivariate data analysis will be discussed in the following sections.

## 4.5 Vibrational spectroscopy

*IR.* Infrared spectroscopy is one of the mostly used unbiased fingerprinting techniques and is based on measuring the physico-chemical properties of metabolites. IR measures the energy absorption that occurs when the frequency of the applied electromagnetic radiation matches with the transitional energy of one of the vibrational modes of molecules. The IR region of the electromagnetic spectrum covers broad wavelength (0.8-1000  $\mu\text{m}$ ) and it usually divided into three smaller regions (near IR 0.8-

2.5, mid IR 2.5-25 and far IR 25-1000  $\mu\text{m}$ ). The IR spectrum of a metabolite can be considered as a fingerprint, since it reflects energy absorptions specific to the molecular structure and the absorption peak abundance is directly proportional to concentration. Depending on the complexity of molecular structure (e.g., chemical bonds), different vibration modes (e.g., stretching, scissoring, wagging, rocking and twisting) can be present in each molecular and they give a resonance at different frequencies of the electromagnetic radiation. An IR spectrum of an unknown metabolite is normally not sufficient qualitative information to obtain complete structure elucidation, but it may assist to identify the types and gross amounts of chemical bonds and functional groups.

Most modern IR applications apply Fourier transform IR (FTIR) spectrometers. This technique is based on the interferometer principle and uses light output at all frequencies of the applied IR radiation simultaneously. In plant science IR spectroscopy has mainly been used to study well-defined components such as plant cell wall polysaccharides (Kacurakova and Wilson, 2001; Mccann et al., 1992; Robert et al., 2005), but it has also been used to study the accumulation of mixed linkage beta-glucan during grain filling in barley (Seefeldt et al., 2009) and to study the structure of zein, the main maize seed storage protein (Mejia et al., 2012)

*NIR*. Near infrared spectroscopy is probably the most commonly used spectroscopic technique and has found wide applications in different research fields (e.g., plant, food, agriculture, medicine and pharmaceutical science). NIR is based on measuring the energy absorbance due to molecular overtone and combination vibrations. As mentioned earlier, NIR spectra are usually complex and carry valuable information about the overall physico-chemical state of the sample. In other words, NIR is able to give a representative snap-shot of the cereal phenome (Munck et al., 2004; Munck et al., 2010). Since NIR spectroscopy is a notoriously highly reproducible, rapid and non-invasive technique and requires minimum sample preparation it is a perfect proxy method for contrasting complex biological samples prior to more destructive analytical metabolomics platforms. NIR has for long been undervalued due to its complex unresolvable spectra, but there is now a slowly growing awareness that the cell and self-organizing biological systems are much too complex to be understood based on destructive analysis (Lander, 2011). Therefore, integrated analysis of phenotypes to capture an overall phytochemical signature require unbiased and high-throughput analytical techniques such as by NIR spectroscopy (Munck et al., 2010).

The combination of NIR spectroscopy with multivariate calibration techniques has found a very broad use in many different areas of cereal science. NIR technology has been proven to be able to provide precise and accurate measurements of the bulk constituent in cereal flours (Williams and Norris , 1988) and high energetic shortwave NIR is able to penetrate single seeds to provide very accurate information on the protein content of individual seeds (Delwiche, 1995; Pedersen et al., 2002; Tønning et al., 2006). This property have great potential for cereal breeding and cereal sorting such as utilized in the TriQ SKNIR sorter (BoMill AB, Lund, Sweden), which use the second and third overtone NIR

spectra to sort cereal grains according to a preselected quality traits. Additional examples of NIR applications in cereal science are presented in the following sections and references are compiled in table 3.

## 4.6 Electronic spectroscopy

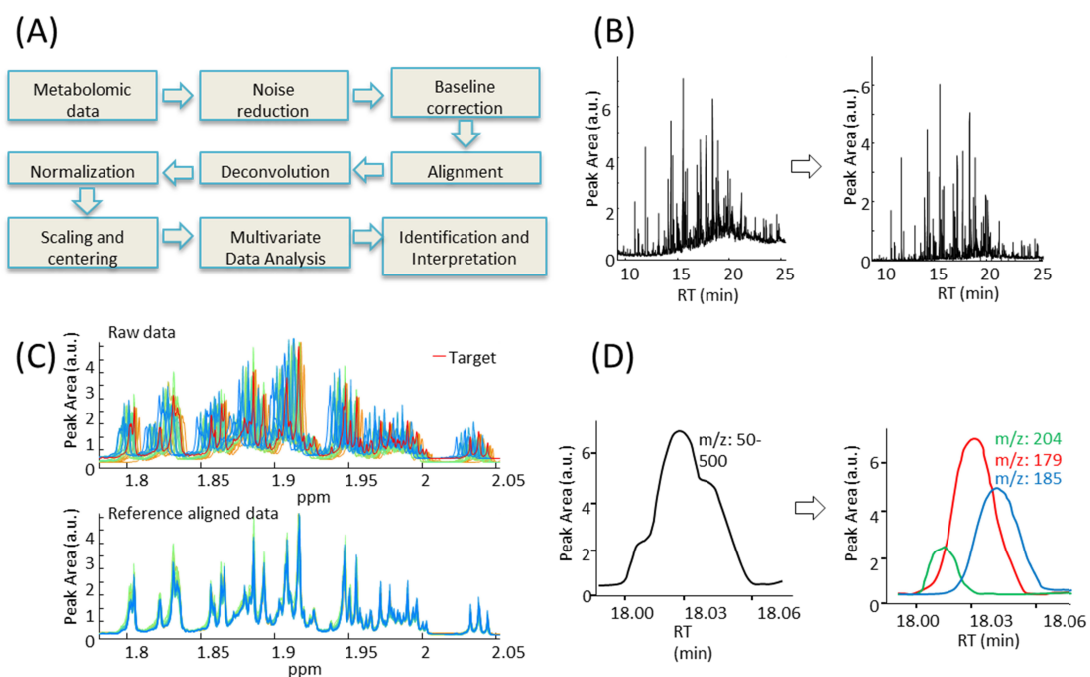
*Fluorescence.* Electrons of molecules are normally in a low energy state, which ensures their stability. However, if the electrons are exposed to external energy they will move to a high energy state which is referred as excitation. Fluorescence spectroscopy is based on electron excitation by using light (visible or ultraviolet), and records the emitted light formed when the excited molecules return to a lower energy state. In the process of energy loss, the excited molecules may drop down to different vibrational low energy states and emit photons at different wavelengths. In order to enhance the information gained from fluorescence spectroscopy, the samples are usually excited at different light wavelengths and the emitted light is recorded at a broad range of emission wavelengths (typically from 200-800 nm). This will form a fluorescence landscape (data recorded for each sample will be a cub, excitation x emission x light intensities) of the investigated sample and can be referred as a fingerprint. As emitted light intensity is directly proportional to the metabolite concentration fluorescence spectroscopy is a powerful quantification technique. However, fluorescence does not occur in all molecules, as it requires the presence of fluorophores. Fluorophores are the rigid parts of molecules such as an aromatic ring and double or triple bond that prevent molecular relaxation by rotational energy. For example, the amino acid tryptophan is the main fluorophores of the most proteins and it is widely used in fluorescence based proteomics. As in many other quantitative metabolomics methods, fluorescence spectroscopy requires a robust and optimized sampling protocol. The development of robust sampling protocols is probably more important in fluorescence since the fluorescence landscape of a sample is highly depended on a number of matrix effects such as pH, solvent polarity, concentration quenching, inner filter effects and scatter effects (Christensen et al., 2006). Nevertheless fluorescence has found a wide application in cereal research primarily due to its capability to probe the aromatic amino acids and riboflavin (Christensen et al., 2006). For example, fluorescence in combination with chemometric methods has been used for prediction and classification of botanical tissue components of complex wheat flour (Jensen et al., 1982) and rye flour (Kissmeyernielsen et al., 1985; Zandomeneghi et al., 2003). Fluorescence imaging methodology based on ferulic acid and riboflavin has also been applied to monitor wheat flour refinement and milling efficiency (Symons and Dexter, 1996). More recently, the method has been developed for classification of intact flour samples from different cereals (Zekovic et al., 2012).

*UV-VIS.* In contrast to fluorescence spectroscopy, ultraviolet-visible spectroscopy is based on measuring the energy absorbed by molecules during the transition from a low energy state to a high

energy state. UV-VIS uses ultraviolet (10-400 nm) and visible (400-800 nm) regions of the electromagnetic spectrum. In UV-VIS,  $\pi$ -electrons and other non-bonding electrons absorb energy in the form of ultraviolet or visible light that causes their excitation. As in many other spectroscopic methods, absorption of UV-VIS light (at the specific frequency/frequencies) is unique to the structure of the metabolite. Although, the method is not sufficient for complete structure assignment, it provides valuable qualitative and quantitative measure of the sample. In other words, UV-VIS can be described as an analysis method based on colours that are also observable by human eye. Application of UV-VIS spectroscopy plays an important role in quantitative analysis of colour components of plant, food and food raw materials. For example, it has been used in quantitative analysis of silica trace compounds in different rice varieties (Samadi-Maybodi and Atashbozorg, 2006), non-destructive analysis of plant leaf chlorophyll content (Li et al., 2009) and prediction of chlorophyll content in Anthracnose infected leaves of the oil camellia plants (Wu et al., 2012).

## 5 Turning metabolomics data into information

Comprehensive metabolomic studies involve several steps performed in a sequential order before the pre-defined biological question can be answered (Figure 1). For example, in order to investigate an effect of growing conditions and/or abiotic stresses on the final chemical composition of whole-grain barley samples, one must first decide to what extent the metabolome must be measured (e.g., specific pathway or all primary and/or secondary metabolites. It is not always possible or straightforward to screen the whole metabolome using a single protocol. In many cases, extraction of metabolites that greatly differ by their physico-chemical properties requires different protocols, while precision of the quantitative detection largely depend on the applied analytical platform. Once the metabolomic protocol is defined and optimized to enhance its efficiency (e.g., reproducibility,  $s/n$  ratio, time and cost), the investigated samples must be analysed in a randomized order combining both control and blank samples. In chromatography, control samples must constitute at least 30% and blank samples at least 5%, of the total number of samples and must be treated in the same way as the real biological samples. After obtaining the metabolomic data, it is crucial to inspect and preprocess the raw data prior to analysis. Figure 4A shows the most commonly applied raw metabolomic data preprocessing steps. Most of the data preprocessing steps and useful preprocessing tools as well as metabolomics data analysis, including explorative analysis, classification and regression will be discussed in detail in the following sub-sections. Applications of various multivariate data analysis techniques applied on main cereals are compiled in table 3.



**Figure 4.** Turning metabolomic data into information. (A) Overview of the route from metabolomic data to the biological interpretation. (B) Baseline corrected GC-MS total ion current chromatogram of the barley flour phenolics. (C) *iCoshift* based alignment of the NMR spectral interval of complex biological mixture. (D) Deconvolution of three closely eluted metabolites from the GC-MS chromatographic interval.

## 5.1 Metabolomic data processing

Multivariate analysis of quantitative metabolomic data usually requires preprocessing of raw data. Preprocessing of chromatography-mass spectrometry and NMR data usually involve data cleaning, noise reduction, baseline correction, alignment, peak deconvolution, normalization and scaling.

**Baseline correction.** The baseline of chromatographic and spectroscopic data is a uniformly or randomly introduced variation caused by artefacts that occur during the analysis. In chromatography, main sources of these artefact effects might be a gradient program of the mobile phase, column bleed and temperature/pH fluctuations. In general, sample specific baseline variations that occur in chromatography-mass spectrometry platforms have more severe consequences compared to spectroscopic methods. Depending on the analysed sample matrices complexity, number of samples and repeatability of the separation techniques, GC-MS and LC-MS metabolomic data may possess a significant baseline level that may shield important biological variations. In contrast, some

spectroscopic methods such as NIR generate a complex background (no baseline) for all samples due to the dominating substances present in the sample matrix.

In quantitative metabolomics, baseline drifts must be corrected prior to data analysis. Non-sample-related variations introduced by baseline variations may exceed the true biological variation several times, and hamper information level. Most of the chromatographic and NMR data visualization and processing software (DataAnalysis from Bruker, ChemStation from Agilent and Topspin from Bruker) do provide build-in baseline correction functions. Most of these functions allow baseline correction of a single sample at a time. In metabolomics, all samples used in quantitative comparison, must be treated in the same manner throughout data preprocessing and analysis. Therefore, one must be aware of that no bias is introduced by baseline correction of each sample separately when using commercial software. An example of baseline correction of a whole-grain barley phenolic extract GC-MS metabolomic profile is shown in Figure 4B. Several methods have been developed for baseline correction of LC-DAD, LC-Raman data (Boelens et al., 2004) and GC-MS data (Xu et al., 2011). FastChrom is a recently developed Matlab based software, (available via [www.models.life.ku.dk](http://www.models.life.ku.dk)), which enabled rapid baseline correction, peak detection and grouping of similar peaks across many samples simultaneously (Johnsen et al., 2013).

*Alignment.* Alignment of peaks, which belong to the same metabolite, across all samples, is one of the most crucial steps of the metabolomic data preprocessing prior to multivariate data analysis. Multivariate data analysis is able to provide a reliable solution when certain conditions in the data are met: the intensity axis must be the same for all samples since the basic assumption is that the signals obey the quantitative Lambert-Beer law.

In chromatography, peak retention times are not always consistent throughout the samples. Retention time shifts depend on the chemistry of the metabolites, chromatographic system (column, mobile phase, pressure) and number of samples involved in the study. If injected samples have a high affinity to react with the stationary phase of the column, retention time shifts, column and metabolite degradation will occur, which in turn, may deteriorate the analysis results. Retention time inconsistencies are mainly caused by slight alterations of gradient solvent (LC) or temperature (GC) programs, inconsistent temperature and/or pressure of column, contaminated injection port or when number of samples analysed in one sequence exceeds the limit of the one or more parts of the instrument that require either cleaning or replacement.

In NMR spectroscopy, chemical shift require alignment prior to data analysis. Chemical shifts of molecules depend on their surrounding chemical environment characterized by unique electron density patterns of the nucleus. Chemical shifts of the same metabolite observed in two different sample matrices might slightly differ if the pH of two sample matrices differ, NMR data is recorded at different temperatures and/or inter- and intra-molecular interactions of metabolites are different in two sample matrices. In comprehensive metabolomic analysis of complex biological mixtures, almost



all the above-mentioned sources of misalignments may be present. This requires robust, rapid and user-friendly metabolomic data alignment tools.

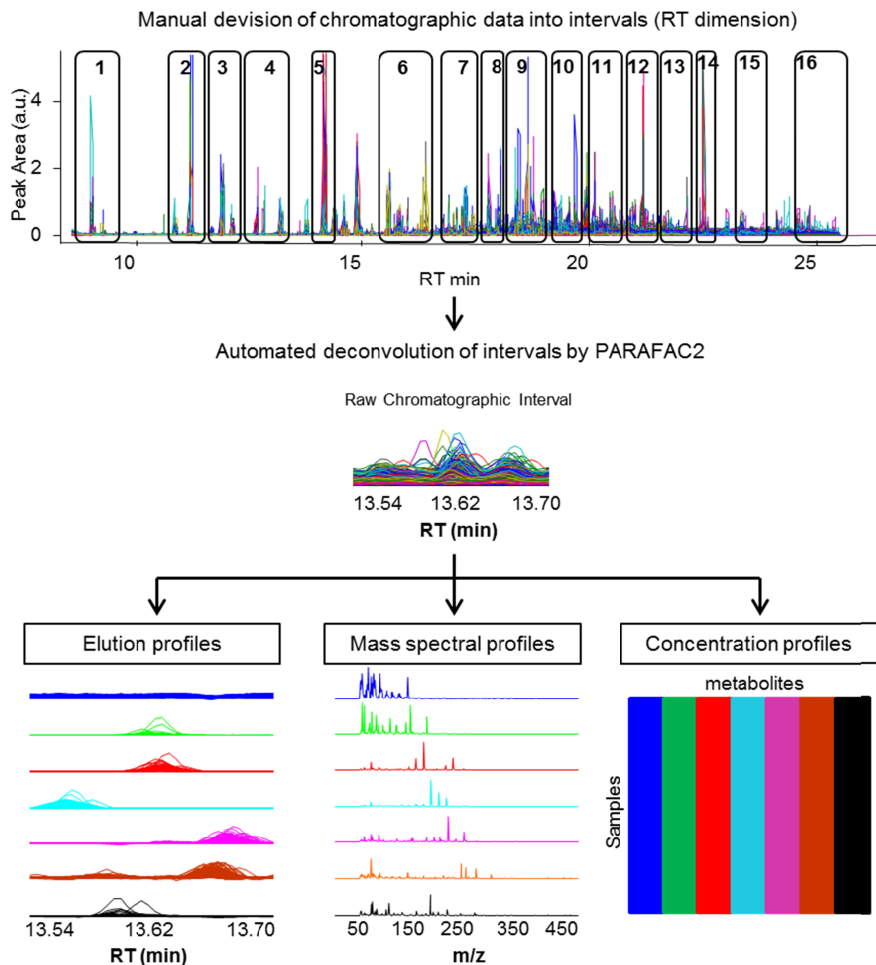
There are several ways to correct for unwanted shifts, but two models are commonly used for the alignment: compression/expansion (C/E) or insertion/deletion (I/D). The former implicitly assumes that peak widths can be correlated to the spectral axis. The latter, on the contrary, assumes that the peak width be invariant within limited ranges of the spectral axis and remains unchanged in case of a shift. Therefore, the C/E model is most commonly used in the alignment of chromatographic signals with methods such as Correlation Optimised Warping (COW) (Nielsen et al., 1998; Tomasi et al., 2004) whereas the I/D model which conserve the peak shapes is used for example in the interval Correlation Optimised Shifting (*icoshift*) (Savorani et al., 2010; Tomasi et al., 2011). *icoshift* aligns two dimensional chromatographic and NMR data based on correlation shifting of spectral intervals and/or whole spectra for many samples simultaneously. One of the main advantages of *icoshift* over several other alternative methods is that it is very flexible, easy to use, rapid, no artifact will be introduced on the shape and abundances of the peaks and requires minimum user interfere. However, it requires that the user is familiar with metabolomic data (e.g., artifact peaks, noise regions, outlier samples) and has a basic knowledge of Matlab. Illustrative example of *icoshift* based alignment of complex 1D  $^1\text{H}$  NMR data is presented in Figure 4C. The *icoshift* Matlab code is available via [www.models.life.ku.dk](http://www.models.life.ku.dk).

*Deconvolution.* In chromatography, deconvolution refers to the reconstruction of the true profiles of overlapped and/or closely eluted metabolite peaks. Depending on the dimensions of the chromatographic data, deconvolution methods greatly differ. In this review, we describe the most commonly utilized deconvolution techniques applied to one and two-dimensional metabolomic data. A general overview of deconvolution of two-dimensional chromatographic data is illustrated in Figure 4D. Simplifying a complex sample matrix by dividing it into different fractions (e.g., polar and non-polar fractions) is laborious and it assists to solve overlapping issues to a rather limited extend. Chromatographic separation system optimization (development of ultra-high performance liquid chromatography (UPLC) and two-dimensional chromatography (GC-GC, LC-LC)) can partly solve some overlapping issues, though comprehensive metabolomic analysis of complex biological samples such as plant tissue extracts and bio-fluid analysis remain challenging. Deconvolution becomes crucial in the situations when the resolution power of chromatographic system is not sufficient to resolve peaks of structurally similar metabolites and/or steric isomers of the complex mixtures.

One of the mostly used deconvolution method is the Automated Mass Spectral Deconvolution and Identification System (AMDIS) (Stein, 1999) which is developed for GC-MS data analysis. AMDIS allows automatic extraction of pure metabolite mass spectra from complex GC-MS data and compares the obtained pure spectra against reference library to identify the metabolite. The software is relatively easy to use and requires few parameters to set. However, it allows analysis of one sample at a time and requires validation of deconvolution and search results. Another deconvolution approach is Multivariate Curve Resolution (MCR) (Lawton and Sylvestr, 1971) . In contract to AMDIS, MCR is a

versatile technique and can be applied to various kinds of data sets, including LC-UV, GC-FID, GC-MS, LC-MS and LC-DAD. However, both methods (AMDIS and MCR) can handle only one two-dimensional GC-MS type of data structure at a time. Deconvolution of chromatographic peaks by MCR requires chemometric knowledge to choose an optimal number of components that will represent the data best. MCR based deconvolution of chromatographic data becomes biased if the data is complex, due to the underestimation of low *s/n* peaks. Therefore, the best MCR solution can be gained in interval based chromatographic analysis. MCR has successively been applied to deconvolution of overlapped and embedded peaks of LC-DAD, LC-MS and GC-MS data (Pere-Trepat et al., 2005; Rodriguez-Cuesta et al., 2005; Salau et al., 1998). In addition, a graphical user interface has been developed for MCR as a freely available Matlab toolbox. (Jaumot et al., 2005).

*PARAFAC2*. PARAllel FACtor Analysis 2 (PARAFAC2) (Bro et al., 1999; Kiers et al., 1999) has several advantages over AMDIS and MCR. PARAFAC2 is a multi-way decomposition method that is increasingly being used in chemometrics and metabolomics. In contrast to AMDIS and MCR, PARAFAC2 can be applied to three-dimensional GC-MS or LC-MS data (Figure 3C) which include many samples and in turn greatly enhances the methods capability and robustness in high-throughput data processing. PARAFAC2 is not only a powerful chromatographic deconvolution technique, but it is also an efficient and rapid method for solving several metabolomic data distortions such as baseline correction, alignment of retention time shifts and quantification of low *s/n* peaks, simultaneously. The method has successfully been applied in resolution of elusive LC-MS peaks (retention time shifted, overlapped and low *s/n*) from complex plant extracts (Khakimov et al., 2012) and for exploring GC-MS data (Amigo et al., 2008). Application of PARAFAC2 in complex GC-MS profiles of barley seed extracts are shown in Figure 5. The PARAFAC2 model of the three-dimensional chromatographic data (retention time x mass spectra x samples) provide three very important outputs that are (1) elution profiles, (2) mass spectral profiles and (3) concentration profiles. The PARAFAC2 elution profiles represent the true elution profiles of the modelled chromatographic peaks, facilitate easy and rapid visualization of peak deconvolution and model validation. The PARAFAC2 mass spectral profiles represent the actual mass spectra of resolved peaks and enable metabolite identification. The PARAFAC2 concentration profiles are the area of the each resolved peak that is observed from elution profiles to be used for quantification purposes. This approach allows deconvolution of severely overlapped peaks profiting from the fact that the individual peaks have a combined unique mass spectra x elution time profile (second order advantage).



**Figure 5.** Processing of raw GC-MS metabolomic data using Parallel Factor Analysis 2 (PARAFAC2).

Today, PARAFAC2 provides one of the most powerful semi-automated deconvolution of overlapped, closely eluted and completely embedded peaks from the complex sample mixtures, in a high-throughput manner. PARAFAC2 is able to model the peaks of the same metabolites as one component even if they do not represent the same elution profiles (due to the retention time shifts) and peak shape, provided that they have identical mass spectra. This feature is considered as the automatic correction of retention time shifts. Moreover, PARAFAC2 eliminates the baseline that hampers analysis of low s/n ratio peaks by separately modeling a baseline. This in turn leads to more pure mass spectra that represent metabolites only and improves identification. Thus, PARAFAC2 is a powerful semi-automated tool for extracting both qualitative and quantitative information without any pre-

processing of raw data. However, the users must define the number of components based on the data and developed model statistics. In order to simplify PARAFAC2 modelling and to reduce the complexity, GC-MS and LC-MS data are divided into smaller intervals in retention time dimension followed by separate modeling of each interval. Future advances in automation of PARAFAC2 modelling will facilitate to use the method by non-experts in a daily routine analysis and obtain high quality quantitative and qualitative data directly from raw data.

*Normalization.* Normalization is an important step in quantitative metabolomics and attempts to correct data for any non-sample-related variations caused during sample preparation and/or data acquisition. For example, variations caused by inconsistent sample weight or volume are corrected by normalizing measured variables of each sample by its weight or volume. Normalization attempts to minimize non-sample-related variations by identifying some measure of characteristics of the samples that must be identical throughout all samples and correct the scale of each measured variable using the sample characteristics. Normalization is usually performed after baseline elimination and before autoscaling (centering followed by scaling) (van den Berg et al., 2006). Prior elimination of inconsistent baseline is crucial when the normalization factor does not consider these baseline variations and in turn, it may increase non-sample-related variations. Normalization is the sample-vice (horizontal) data correction and it assists to give an equal importance to all the samples of the data set to influence on a global model.

Addition of the internal standard (IS) to each sample is a commonly applied approach to correct for variations due to experimental errors e.g., pipetting errors, derivatization, sample injection, detector sensitivity lose during analysis. Main prerequisite for ISs is that they must be well analysed by a chosen metabolomic protocol and their signals must have a one-to-one relation with their concentrations. Moreover, the IS must be added to samples in an earlier stage of analysis and the IS must not interact with the sample components. Variations observed in the response to the internal standard throughout the samples will indicate the level of the experimental error and it can be corrected by dividing each variable response to the internal standard response of the corresponding sample. Unfortunately, many other sources of non-sample-related variations cannot be eliminated by this approach. This includes, instrumental errors (e.g., detector variations, scattering effects) and physico-chemical effects (e.g., solvent, reagent and stationary phase) that influence the scale of the variables. In order to correct for such variations, an application of appropriate normalization methods is required that can eliminate and/or minimize the scaling effects. To date, several normalization methods are used in metabolomics data obtained from different analytical platforms. This include 1-Norm, 2-Norm, Inf-Norm, standard normal variate (SNV) (Barnes et al., 1989), multiplicative signal correction (MSC) (Geladi et al., 1985) and probabilistic quotient normalization (PQN) (De Jong, 1990). These methods apply different transformation algorithms. For example, 1-Norm divides each variable by the sum of absolute value of all variables for the given sample, while SNV divides each variable by the standard deviation of all the pooled variables for the given sample. Normalization methods are not necessarily analytic platform

and/or sample specific. For instance, one kind of normalization may correct NMR metabolomic data well and when the same sample set is analysed by GC-MS this method may fail to remove interfering variations and vice versa. As a rule of thumb, spectral data are usually more robust and generally do not require strong normalization while chromatographic data often requires normalization to total area (1-Norm). However, normalization should only be part of preprocessing if there is a need to correct the data from variability that disturbs quantitative analysis. It is always recommended to compare the final analysis results obtained from normalized and raw data. Normally, if normalization reduces non-sample-related variations, then analysis results will improve (e.g., reduced cross validation error, better prediction models).

*Centering.* Centering is column-wise pre-processing, where each measured variables is centered relative to a reference point (the column average). It is normally performed at the end of the preprocessing steps, just before data analysis. Centering is the default preprocessing step for many multivariate data analysis methods. For example, in order to perform PCA, data must often be mean centered. Mean centering subtracts the mean value of the column from each element of the corresponding column. This ensures zero centre of the data and adjusts offset differences between high and low magnitude variables, and thus provide easily interpretable models. Mean centered data exposes important relative variations present between the samples and significantly improves data analysis (van den Berg et al., 2006).

*Scaling.* Metabolomic data must be scaled if the applied data analysis method assumes that the relative magnitude of the variable is unrelated to its importance and if measured variables have significantly different scales. Scaling provides an equal importance to all the variables to influence to the final model and eliminates scale differences. Scaling is usually performed after centering and it is normally the last preprocessing step. During scaling, each variable will be divided by a scaling factor that is different for each variable (column) (van den Berg et al., 2006). As with other preprocessing steps, scaling may significantly change the outcome from the data analysis, and therefore it must be performed carefully. Different scaling methods are currently being used in multivariate data analysis. The most commonly used method is autoscaling which covers centering followed by scaling by dividing each variable by the standard deviation of the corresponding column. However, if variables contain significant amount of noise, autoscaled data will be highly influenced by the noise which may obscure the data analysis. Several data analysis software provides various kinds of scaling methods as optional preprocessing techniques. Such scaling methods e.g., pareto, range, vast, level, group and log decay emphasize different characteristics of the data.

*Software for metabolomic data processing.* The rapidly developing field of metabolomics and the generation of huge data sets require user friendly methods that facilitate extraction of relevant information. Several different metabolomic data processing software have been developed to perform automated and/or semi-automated peak detection. For example, there are number of commercial

(e.g., LineUp, MarkerLynx, MarkerView, , MS Resolver, Metabolic Profiler, Profile and Sieve) and freely available (XCMS, MZmine, metAlign, MSFACTs, MathDAMP, Compare, COMSPARI, HiRes and MET-IDEA) metabolomic data processing software that deal with LC-MS, GC-MS, CE-MS and NMR data (Katajamaa and Oresic, 2007).

Most of these software packages perform noise reduction, baseline correction, alignment, deconvolution and provide a table, where each row corresponds to a samples and each columns corresponds to a detected peak. These methods require several threshold parameters to be set by the user and these parameters will have a high influence on the obtained results which has been underlined in a comparative LC-MS metabolomics study where the same data was processed by three kinds of freely available software, MarkerLynx, XCMS, and MZmine (Gürdeniz et al., 2012). Most of the algorithms implemented inside these software are not accessible which hampers the understanding of how these methods work. In order to obtain quantitative metabolomic data, users of such all-in-one metabolomic data processing software must understand and carefully observe the processed data after each step of the data processing. When using such methods, it is also important to know the nature of the data (e.g., baseline, resolution, mass accuracy, adduct ions and liner range) and correctly set the required parameters of the methods.

When data preprocessing has been carefully performed, the metabolomics data is ready for multivariate data exploration. This can either be performed as unsupervised data analysis where no priori information is used in the data modelling or by supervised data analysis where design parameters and/or other response variables will guide the data modelling.

## 5.2 Unsupervised multivariate methods

Unsupervised multivariate methods comprise data analysis techniques that are based on comparison of objects using variables measured on those objects and do not receive any priory information on design of experiment or object groups. In plant metabolomics, mostly utilized unsupervised methods comprising principal component analysis (PCA) (Hotelling, 1933) multivariate curve resolution (MCR) (Lawton and Sylvestr, 1971) and hierarchical cluster analysis (HCA) are used.

*PCA.* Principal component analysis (PCA) is the most commonly used multivariate method. PCA displays the intrinsic data structure in a simple, low-dimensional orthogonal projection and highlights similarities and differences among groups as well as the variables involved. (Hotelling, 1933; Pearson, 1901). In its simplest term, PCA can be explained as the method that decomposes multivariate data (e.g.,  $\mathbf{X}(n,m)$  matrix,  $n$  samples and  $m$  variables) into a smaller number of principal components (PC) that is good approximation of the original data matrix. If one could imagine the plot of  $\mathbf{X}(n,m)$  matrix where each sample  $n$  is plotted in  $m$  dimensional space, the first PC is the plane that crosses all these

data points in  $m$  dimensional space in such a way that the sum of distances between each data point to the PC is at a minimum. Likewise, the second PC also crosses all the data points in  $m$  dimensional space with a minimum distance to  $n$  samples in such a way that it will be orthogonal to the first PC. Accordingly, if one would include the third PC, then it would be orthogonal to the first and second PC's and so on. Each PC possesses scores and loadings that represent samples and variables, respectively. In PCA analysis, the most crucial step is to define the number of PCs that would describe the data best. The number of PCs included in PCA of  $\mathbf{X}(n,m)$  matrix cannot exceed  $n$  and  $m$ . PCA that uses two principal components can be written as  $X = t_1 \times p_1' + t_2 \times p_2' + E$ , where  $t_1$  is the score vector and  $p_1$  is the loading vector of the first PC and  $E$  is the unexplained part of the data.

Prior to PCA analysis, data must be mean centered and preferably scaled to give equal importance to all variables. In order to choose an optimal number of PCs, each PC is evaluated by the variance it can explain. Moreover, outliers that shield the true variation present in the data must be identified by examining the scores and residual plots and must be removed prior to model validation. PCA is especially useful for explorative purposes and may assist to evaluate general trends such as how samples differ and/or correlate to each other, which variables co-vary and identify variables that cause groupings and/or differentiations. Thus, PCA results are normally evaluated by visual inspection of scores and loadings plots.

*MCR.* While a discriminative PCA is an important indicator of whether the study design has been effective in revealing effects, it has a weakness in relation to result interpretation. Variable (fingerprinting) or metabolite (profiling) interpretation has to be performed through the loading plot, which is the backbone of a PCA model. However, due to the orthogonality constraint in the PCA, spectral loadings are typically not easy to interpret. However, an alternative to unsupervised classification that does not include the orthogonality constraint is multivariate curve resolution (MCR), which is similar to PCA, but without the constraint that the PC must be orthogonal. This gives MCR the appealing property that it can provide resolution of complex profiles into the "true" underlying components. However, MCR solutions are generally not unique, hence the solution can be assumed to be just one arbitrary solution out of an infinity of equally well-fitting possible solutions. For this reason MCR is often applied with non-negativity constraints and in smaller regions of interest. It is interesting to note that the ambiguity of MCR can be "overruled" if higher order data is recorded such as GC-MS data. Then the multi-way MCR relatives PARAFAC (Bro, 1997; Harshman, 1970) and PARAFAC2 (Bro et al., 1999) can resolve uniquely the underlying profiles. This is now being exploited in diverse metabolomics applications (Khakimov et al., 2012).

*HCA.* Hierarchical cluster analysis (HCA) is one of the mostly used clustering methods in plant metabolomics. HCA is mainly based on two principles. The first principle is to consider each sample as a separate cluster and then gradually merge it with other similar samples to form clusters. This approach is termed agglomerative. In contrast, the second approach, called divisive, assume that all

the samples constitute a single cluster and recursively splits the samples moving down the hierarchy. In order to form a cluster, HCA uses a metric that estimates similarities between samples and linkage criteria. It is worth to mention that, even for the same data set, the results of a HCA analysis will differ depending on the applied metric methods (e.g., Euclidean distance, Manhattan distance or Mahalanobis distance) and linkage criteria (complete linkage clustering, single linkage clustering or minimum energy clustering). However, all of these distance measures between samples are valid for any type of data set. Results of the HCA are normally presented in dendrograms. The method has been successively applied in LC-MS based plant metabolomic study (Kuzina et al., 2009), human blood plasma lipidomic study (Draisma et al., 2013), NMR metabolomics based classification (Kim et al., 2010) and plant phenotype differentiations using untargeted GC-MS and LC-MS metabolomics (Arbona et al., 2009).

In unsupervised models, such as the PCA model, a priori knowledge can be used to color objects in the score plot and thereby emphasizing potential groupings and/or quantitative gradients found in data. However, in metabolomics studies it is common to have a priori knowledge about the data, typically from a controlled experimental design e.g. cereals grown under condition A and B, that can be modeled by supervised methods.

### 5.3 Supervised multivariate methods

In supervised multivariate methods, some of the known data facts (e.g., classes or groups of the data) can be used to guide the multivariate data analysis. Development and implementation of supervised methods resulted in a giant leap for classification and regression analyses. Today these methods play a key role in metabolomic studies dealing with biomarker discovery, quality control, biosynthetic pathway elucidation and to understand the influence of external effects in living systems. Most commonly utilized supervised methods in metabolomics include partial least squares analysis (PLS) (Wold et al., 1983), interval based PLS (iPLS) (Nørgaard et al., 2000), linear discriminant analysis (LDA), partial least squares discriminant analysis (PLS-DA) (Stahle and Wold, 1987), orthogonal partial least squares discriminant analysis (OPLS-DA) (Bylesjo et al., 2006), multiple linear regression (MLR), canonical variate analysis (CVA), extended canonical variate analysis (ECVA) (Nørgaard et al., 2006), soft independent modelling of class analogy (SIMCA) (Wold, 1976), support vector machine (SVM) and artificial neural networks (ANN).

*PLS.* PLS is the most commonly applied multivariate regression analysis that aims to build a linear regression model, which enables prediction of a desired characteristic from a measured multivariate variable (e.g., metabolites). PLS is similarly to PCA, but its scope is to regress (or force) the result in a given direction (reference method), and it is thus called a supervised method, while PCA can be compared to shopping (in the data) without a shopping list (e.g., the data analysis is performed



without the use of a priori knowledge), PLS regression is similar to shopping with a specific shopping list. The method was developed since early 1970s by the Swedish statistician Herman Wold and in 1977 the PLS algorithm found its final and present form (Wold, 1979; Wold, 1975). PLS tries to correlate **X** data matrix with another **Y** data matrix and enables estimation of the correlation level. If **X** data matrix is relevant to the information present in the **Y** matrix, PLS will be able to find a common variance that will facilitate explanation of **Y** based on **X** matrix (Wold et al., 2001). The simplest example can be prediction of dietary fibre concentrations in whole-grain cereals (Seefeldt et al., 2009). In this study, spectroscopic data from 50 whole-grain samples (**X** matrix) were obtained and at the same time, their dietary fibre content was measured using a laborious chemical method (**Y** matrix). Since the spectroscopic fingerprint of the grain samples reflects the fibre content, information present in the **X** data matrix is very likely to have strong correlation with the corresponding **Y** data matrix. In this case, PLS regression modelling can estimate **Y** (fibre content) using information present in **X** (spectroscopic data) with high accuracy and it will be possible to predict the fibre content of new grain samples. Thus, the main application of PLS is focused on predicting some important sample features that are expensive and laborious to measure from cheaper, easier and more accurately measured variables.

The structure of PCA and PLS models are similar (e.g., both possess scores and loadings), however, the criteria for finding PCs are different. PCA finds the best approximation of the **X** data matrix, while PLS maximizes the covariance between matrices **X** and **Y**. Principal components of the PLS model explain the maximum amount of the variation present in **X** and **Y** that are correlated to each other. Therefore, it is possible to predict the scores of **Y** from the scores of **X**. Scores and loadings plots of PLS models can be explained in the same way as PCA models. Evaluation of the PLS models are mainly based on cross-validation and test set validation results. In supervised data analysis it is important to perform a validation in order to obtain a reliable and robust model. Depending on the sample set design, appropriate validation must be performed. For example, if one thousand samples are randomly collected and have no design, full cross-validation or random sub-set cross-validation can be used. If samples are divided in groups and/or subgroups, have some design and replicates, the data must be divided into test and calibration sets, respectively, in such a way that both data sets will have representative samples from each group. In the Matlab based PLS-toolbox (Eigenvector Research, Inc.) these options are already implemented in a graphical user interface that non-specialist users can easily understand and use it.

One of the most important outputs of the PLS modeling are predicted *versus* measured plot that indicates how well the model can predict **Y** values of unknown samples from their **X** data matrix and the Variable Influence on Projection (VIP) plot, which highlights the most important variables that played a key role in the prediction.

In classical empirical research, a model requires that the number of variables must be less than the number of observations, but developments in modern analytical platforms have pushed scientists far beyond the classical model. Nowadays it is common that more than 10.000 variables are recorded for each sample, which pushes the chemometric tools to the limit, as they will also increase the extent of spurious correlations and interferences. In regression analysis, variable selection can often improve the developed regression models. Several techniques and strategies are available, not only for reducing the number of variables before the actual model, but also to reduce while modeling. In typical foodomics studies it is normal to include 100-1000 samples. Subdividing metabolite profiles down into smaller regions can also be useful to break down GIGA-variate structures into smaller (and logical) subunits or regions of interest (ROI). An obvious example is the different regions in an NMR spectrum: aromatic, carbohydrate and aliphatic, but the principle is generic (Savorani et al., 2013a). Interval PLS (iPLS) (Nørgaard et al., 2000) has proven efficient in improving and simplifying classification models by breaking the model up into smaller intervals (either consisting of many metabolites or one metabolite per interval) of data (Di Anibal et al., 2011; Ferrari et al., 2011). The use of interval models is generally a healthy principle when analyzing raw metabolomics data sets and not “just” metabolite tables. Interval models use fewer variables which contain fewer interferences which in turn will lead to more parsimonious models and lead to enhanced model performances and interpretability (Savorani et al., 2013b). When few intervals and/or fewer metabolites are found to be optimal for the best classification, this also makes the subsequent biological interpretation simpler. The method found a wide application in conjunction with metabolomic data obtained from spectroscopic (Borin and Poppi, 2005; Kristensen et al., 2010; Paschoal et al., 2003) and spectrometric (Marhuenda-Egea et al., 2013) methods.

*PLS-DA.* Partial least squares discriminant analysis (PLS-DA) (Stahle and Wold, 1987) is one of the favourite classification techniques applied in plant metabolomics. A PLS-DA problem is based on finding variables and directions in multidimensional space that can distinguish samples that belong to the different classes. As in PLS, PLS-DA uses **X** and **Y** matrices and predicts the response variable **Y** that is not a measured feature of the samples, as in the case of PLS, but class/or group categories of the samples (dummy matrix). The **Y** “dummy” matrix used in PLS-DA consist of zeros and ones and contains as many columns as there are classes (each column defines one class and ones mean that these samples belong to the class and the rest are zeros). PLS-DA decomposes X and Y matrices into two matrices of scores and loadings in a dependent manner. Thus, the scores of Y block determine the loadings of X block, while scores of X block will determine the loadings of Y block. Therefore, PLS-DA is classified as supervised method. Detailed techniques for using of PLS-DA classification method is given in (Barker and Rayens, 2003).

Where a normal PLS model is optimized according to the prediction error (e.g. RMSECV), the PLS-DA should be optimized based on classification parameters (e.g. rate or percentage of misclassified samples). Due to strong classification performance, the PLS-DA based classification is so widespread in

plant metabolomics (and more frequently used than classical PLS modeling) and most probably it will be dominating classification tool in future metabolomics studies. However, in order to interpret the biological information gained from the PLS-DA modelling, model must be validated. Most commonly used validation methods in PLS-DA are cross-validation (type of CV highly depends on the design of the data) and test set validation (data must be splitted into calibration and test sets, where test set samples must comprise at least 30% of total samples). Validation results are crucial when selecting a number of latent variables (LVs). Optimal numbers of LVs are determined by considering the variance captured by each LV, RMSEC and RMSECV values obtained from each modelled classes, and by plotting the measured versus predicted class categories for each modelled classes. The most important PLS-DA model parameter is the number of misclassifications (NMC) which is the sum of the false positive (FP) and false negative (FN) samples ( $NMC = FP + FN$ ). In addition, PLS-DA models can be checked for sensitivity and specificity, for all modelled classes (both for calibration and cross-validation results), by plotting the Receiver Operating Characteristics Curve (ROC curve) (Zweig and Campbell, 1993). Sensitivity illustrates the models' ability to correctly classify the samples to the class that they belong to and it can be written as  $Se = \text{true positive (TP)} / \text{TP} + \text{FN}$ . Specificity is another measure of the model that shows how well it can predict the class of the control samples and it will be described as  $Sp = \text{true negative (TN)} / \text{TN} + \text{FP}$  (Szymanska et al., 2012; Westerhuis et al., 2008). When these two parameters of the model are close to one, the model is usually considered as a valid model.

As in PLS modeling, PLS-DA allows to perform variable selection to improve the model quality. It is known that the variable selection feature of the PLS-DA is best suited when two class problems are analysed. However, improvement of the PLS-DA model by variable selection is also possible during model development by excluding those variables that have insignificant classification power in all classes. In addition to PLS-DA, orthogonal PLS-DA (OPLS-DA) is also a commonly used classification method in plant metabolomics. The main difference between these methods is that OPLS-DA imposes orthogonality that allows extraction of two kinds of variations present in the **X** data. The first variation explains **Y** matrix and is used for developing the model, and the second variation is orthogonal to the information present in the **Y** matrix (Trygg et al., 2007). Some have argued that this feature of the OPLS-DA makes it more powerful for interpretation of classification models compared to PLS-DA and SIMCA (Bylesjo et al., 2006). However, one must ensure to use an appropriate validation method when using such a method to avoid model overfitting.

*SIMCA*. Soft independent modeling of class analogy (SIMCA) (Wold, 1976) is another powerful classification technique used in metabolomics. SIMCA is based on developing separate PCA models for each class independently and compare samples of different classes in principal components' space that is less complex than the original variable space (Wold and Sjöström, 1977). In SIMCA, each class is represented by PCs and the number of PCs may differ between classes. As in PCA, the optimal number of PCs that provides the best approximation of the class is the most important parameter that needs to be validated. Since under- (lower number of PCs) or over- (higher number of PCs) estimated models

may lead to false positive as well as false negative results. In order to evaluate SIMCA model performance, it needs to be validated by using cross-validation or test set validation. One of the main advantages of SIMCA over other competitive classification methods is that it uses PCA approximation of the data to capture the variance of each class separately and allows comparison of classes using PCs that are free of any noise that may be present in the data. Moreover, SIMCA assigns the class of an unknown sample only if it falls to some category of samples with high probability, otherwise the unknown sample will not be assigned to any of the classes. SIMCA is also very flexible with respect to data size (smaller sample sets with only 10 samples and bigger data sets with several thousand variables can also be modelled), and this makes it a very useful tool for metabolomics.

ECVA. Sometimes it is useful to evaluate data using an alternative classification method such as Canonical Variates Analysis (CVA) and in particular its extension that is able to handle datasets with more variables than objects: Extended Canonical Variates Analysis (ECVA) (Nørgaard et al., 2006). In analogy to CVA, ECVA optimizes the within class variation divided by the between class variation criterion by finding new multivariate directions. ECVA has a great potential within metabolomics classification analysis, but it should be noticed that there will not necessarily be a large difference in the misclassification rate between different methods such as SIMCA, PLS-DA, OPLS-DA, and ECVA as they all have their advantages and disadvantages. However, there will be special cases such as for example that ECVA in contrast to PLS-DA is able to handle situations where three groups are separated along one direction. So far ECVA has only been applied to a limited number of metabolomics application but displayed an effective classification potential (Lopez-Rituerto et al., 2012). ECVA and iECVA methods are very flexible, but requires that the user can work in the Matlab environment and it allows the user to decide the interval division boundaries to model separately. Complete ECVA algorithm is freely available from [www.models.life.ku.dk](http://www.models.life.ku.dk).

## 5.4 Exploiting the experimental design: ASCA

Biological systems exhibit sources of variation due to a large number of factors, e.g., varieties, days after flowering, field, fertilizer, abiotic stress, biotic stress etc. Realizing this led R.A. Fisher (Fisher, 1918) (the father of modern statistics) to develop experimental designs suited for estimation and handling of variation due to such factors. The paired t-test and analysis of variance (ANOVA) are examples of models used to analyse univariate data from designed experiments. The backbone of these methods is to estimate variance related to a nuisance (orthogonal) factor (e.g., subject) and remove it. In this way, the variation of interest, e.g., treatment, is emphasized, which in turn increases the chance of finding something interesting (often referred to as statistical power) (Engelsen et al., 2013). The multivariate equivalents of these methods include analysis of variance simultaneous component analysis (ASCA) (Smilde et al., 2005). The principle of ASCA is to split up the original data

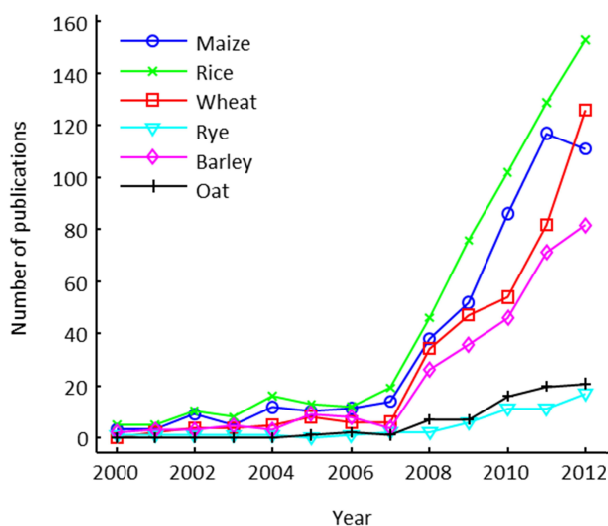
matrix into contributions from the different design factors (e.g. abiotic stress and days after flowering). This splitting of information into two (or possibly even more) orthogonal subspaces, one part that is unrelated and hence irrelevant for the study design factor of interest and one part that contains the relevant information, is the crux of ASCA. Whereas the variable selection can be considered as a horizontal elimination of interferences, orthogonalization can be considered as a vertical elimination of interferences in the data matrix. If the study design is balanced it will enable estimation of the different design factors' contributions and in contrast to OPLS-DA (Bylesjo et al., 2006) that does not utilize any information other than the factor of interest (e.g., treatment), ASCA utilize the different design factors. We foresee a rapidly increasing use of ASCA in balanced study designs in cereal metabolomics and a feasibility study has already been conducted in plants metabolomics in which effects of growth conditions and processing on *Rehmannia glutinosa* (traditional Chinese Medicine) was studied using a GC-MS fingerprint strategy and ASCA analysis (Chang et al., 2006).

## 5.5 Network analysis

Network analysis plays one of the important roles in evaluation of plant-environment, gene-metabolome interactions, and effects of various biotic/abiotic stresses in plants, and thus assist in understanding of systems biology, metabolite biosynthesis and functions of genes (Barabasi and Oltvai, 2004; Bassel et al., 2012; Yamada and Bork, 2009). Network analysis covers various statistical methods of measure e.g. correlation-based networks (CN), similarity measure (SM), euclidean distance (ED), knowledge-based approaches, to evaluate (1) relationships between different components of biological systems e.g., genes, proteins, and metabolites, (2) covariance of these components as a response to biotic/abiotic factors and genetic modifications (Toubiana et al., 2013). Network analysis assist in combined analysis of complex data sets obtained from different 'omics' platforms such as transcriptomics, proteomics and metabolomics. In other words it can be describes as the method that allows identification of origins of one or more quality traits of phenotypes e.g., resistance to salt and drought or pathogenesis, higher yield, in a genetic level. Thus, networks generated from metabolomics and other data sets may provide initial knowledge in complex biosynthetic pathways, crosstalk between distinct pathways and responses of different components of biological systems to environmental and endogenous changes. Recently, several web-based tools have been developed for gene co-expression network analysis for gene discovery in specialized metabolism (Higashi and Saito, 2013). Most recent reviews in network analysis (Higashi and Saito, 2013; Toubiana et al., 2013) comprise several examples of a use of the method in plant metabolomics, improvement and development of crop plant species such as maize and rice.

## 6 Application of cereal metabolomics

In the last two decades, metabolomics has found a wide application in cereal science. Recently, Balmer et al., 2013, published a review paper on metabolomics of cereals under biotic stress (Balmer et al., 2013). That review highlights some applications of metabolomic analyses applied for uncovering biotic effects and illustrates the current knowledge and techniques. Figure 6 shows the current trend of cereal metabolomic studies registered in PubMed Central between 2000 and 2012. In this section, we review metabolomic studies applied to main cereals, maize, rice, wheat, barley, oat and rye. In addition, table 3 comprises more than one hundred cereal metabolomics studies performed on different analytical platforms.



**Figure 6.** Number of cereal metabolomic studies published in PubMed Central between 2000 – 2013. Number of publications were counted by using key words “cereal name” and “metabolomics”.

### 6.1 Maize

Most of the comprehensive maize metabolomic studies focus on evaluation of transgenic lines, comparison of transgenic lines versus non-transgenic lines and influence of abiotic stresses. A study compared two GMO maize lines (Bt-maize) and their corresponding non-GMO parental lines grown under the same conditions utilized CE-TOF-MS based metabolomics (Levandi et al., 2008). They showed that some metabolites as possible biomarkers for transgenic maize varieties and illustrated the potentials of CE-MS based approach in the evaluation of GMOs. A comparison of phenotypic changes influenced by genetic modifications and by the environment was studied by a combined

transcriptomics, proteomic and metabolomic analysis of two GMO (Bt-maize and herbicide-tolerant Roundup Ready maize) and corresponding non-GMO parental maize lines (Barros et al., 2010). This study involved GC-MS based metabolomics and showed that environmental factors caused more variation in the different metabolites profiles than the different genotypes. The other study also dealt with the impact of genetics and environment on the maize grain metabolome based on inbred lines (Skogerson et al., 2010). In that study, metabolomic comparison of GMO maize lines was based on 119 identified metabolites by using GC-TOF-MS. In contrast to the previous study, they have reported that more metabolomic variations were caused by genotypes than by different geographical locations. This study also depicted a high genotypic dependence of small molecular metabolite pool where one genotype showed association with increased concentration of fatty acids and organic acids, while another line illustrated increased levels of amino acids and carbohydrates.

A NMR-based metabolomic profiling approach showed a strong potential to evaluate the influence of salt stress on the maize metabolome (Gavaghan et al., 2011). The study showed a clear difference in amino acids, small molecular organic acids and sugars concentrations detected from the shoots and roots of control and treated maize lines. They concluded that there is a higher salinity effect in shoots than in roots. Frank et al., 2012, performed a comparative study to evaluate genetic modifications versus environmental influence using GMO (Bt-maize) and non-GMO maize varieties grown under different conditions, including several growing locations and seasons. The study was based on GC-MS metabolomics and maize metabolite extracts were analysed in separate fractions containing lipids and polar metabolites. They showed a total of 3% of metabolome difference related to genotype and up to 42% difference caused by different growing locations and seasons. Another maize metabolomics study focused on herbivore-induced metabolites in leaves and roots of resistant and susceptible maize (Marti et al., 2013). This study mainly involved UPLC-TOF-MS analysis (for untargeted analysis of leaves and root samples), CapNMR (for structure elucidation of 32 differentially regulated compounds) and direct infusion tandem MS/MS. The paper shows that by infestation of maize leaves, concentrations of 1,3-benzoxazin-4-ones, phospholipids, *N*-hydroxycinnamoyltyramines, azelaic acid and tryptophan were significantly increased, while only minor changes occurred in the metabolome of the roots.

GC-MS metabolomic analysis of 289 diverse inbred maize lines showed the power of metabolomics for linking genotype and phenotype (Riedelsheimer et al., 2012b). This study dealt with genome-wide association mapping (GWAS) of maize leaf metabolome including 118 distinct metabolites, 56,110 single nucleotide polymorphisms (SNPs) and several agronomic traits of mature maize plants. The GWAS approach demonstrated that 26 distinct metabolites were highly associated with 26 SNPs and allowed identification of lignin precursors, *p*-coumaric and caffeic acids to be strongly associated with a region of chromosome 9 harboring a key enzyme in monolignol synthesis.

## 6.2 Rice

GC-MS based metabolomic profiling of two low phytic acid rice mutants generated by  $\gamma$ -irradiation and the parental wild-type variety exhibited, on average of, 34% and 42% of differences in the total metabolome (Frank et al., 2007). The rice varieties grown in four different field trials did not display consistent metabolomic changes due to the environmental effect. However, some metabolites were up- or down-regulated in the same manner in low phytic acid rice mutants grown in different fields. A rice metabolomics study evaluated metabolomic changes in rice seeds during germination by using GC-MS (Shu et al., 2008). The rice metabolite extracts were divided into lipophilic and hydrophilic fractions covering a broad range of fatty acids, hydrocarbons, sterols, sugars and amino acids. A total of 615 GC-MS peaks were semi-quantified, out of which 174 were identified based on MS. The first principal component of a PCA model developed by using this metabolomic data showed a germination time trend that was identical for the three rice materials investigated. It was shown that mainly the polar metabolites contributed to the time dependent separation of rice samples and illustrated a dynamic pattern. Chang et al., 2012, evaluated unintended effects of GMO rice varieties by using LC-MS and developed a protocol that allowed detection of the metabolomic differences caused by the environmental and genetic effects (Chang et al., 2012). The obtained results illustrate a greater role of the environment than on gene modification for most of the metabolites, including amino acids, fatty acids, and small molecular organic compounds.

Antioxidant properties of commercial wild rice were evaluated by LC-DAD and LC-TOF-MS analysis of soluble and insoluble phenolic acids from methanol extract of flour samples (Qiu et al., 2010). The study showed that the radical scavenging activity of the wild rice exceeded that of the white rice (control) up to ten times. Quantitative metabolomic analysis of wild rice extracts proved that the main phenolic acid constituents were ferulic and sinapic acids that were mainly present in insoluble fractions. Another study compared GC-MS and NMR metabolomic data for understanding the rice metabolome response to submergence stress (as e.g. under flooding) (Barding et al., 2013). This study revealed advantages of multi-platform metabolomics and showed low molecular mass response metabolites that could be detected either by NMR or by GC-MS. Sana et al., 2010, studied a bacterial pathogen (*Xanthomonas oryzae*) effects on two different rice varieties that are resistant and susceptible to the pathogen, respectively. LC-TOF-MS and GC-TOF-MS based metabolomic approaches revealed detection of almost 800 metabolites, out of which 154 were identified based on their retention indices and MS. Multivariate analysis of the metabolomic data assisted to uncover several biosynthetic pathways (acetophenone, xanthophyll, fatty acids, alkaloids, glutathione, carbohydrate and lipid) that were affected by the bacterial pathogen.



### 6.3 Wheat, barley, oat and rye

#### *Wheat*

Proton NMR metabolomics study performed on GMO and their corresponding parental lines of wheat varieties showed greater metabolome variations related to the growing year and locations than genotype (Baker et al., 2006). Multivariate analysis of NMR data obtained from the wheat lines grown in the same year revealed some separation between GMO and parental lines that were mainly due to maltose and/or sucrose concentrations. GC-MS analysis of the same wheat lines grown in the same year illustrated that plants differ in the content of glutamic and aspartic acids and their amide derivatives. Another study evaluated the effects of genotype and environment on free amino acids of two commercial and four double haploid wheat lines (Curtis et al., 2009). In that study total amino acid profile of the wheat lines were screened using GC-MS and environmental and genetic effects on asparagine and other amino acid content was evaluated using canonical variate analysis. The study showed that desirable wheat lines could be identified by this selection method. NMR based metabolomic profiling of four different European wheat varieties revealed differences in concentrations of glucose, fructose, betaine, aspartate, choline and some other small molecular metabolites (Graham et al., 2009). That study illustrated a rapid biochemical mapping of wheat varieties using NMR spectroscopy. A recent study performed comprehensive GC-MS metabolomic profiling of four durum wheat lines grown under conventional and organic farming systems over three years (Beleggia et al., 2013). The study focused on evaluation of genotype, environment and genotype-by-environment on metabolome and quality and once again showed a dominated influence of environment over genotype. A number of studies performed within the HEALTHGRAIN diversity screen program also demonstrated the effect of environment and genotype on the metabolome of wheat cultivars (Andersson et al., 2010; Fernandez-Orozco et al., 2010; Lampi et al., 2010; Li et al., 2008; Shewry et al., 2010). In these studies essential bioactive compounds e.g., phenolic acids, tocopherols, sterols, folates and dietary fibers of 26 wheat cultivars were screened using different metabolomic approaches. A targeted metabolomic approach that utilized GC-MS revealed an increased level of amino acids such as proline, tryptophan, valine, and leucine in one drought intolerant and two tolerant lines, while the drought tolerant wheat lines also exhibited a decreased level of organic acids under drought stress conditions (Bowne et al., 2012). Moreover, they showed that drought stress has a similar effect on the metabolomic alterations of the two tolerant wheat lines, though the level of the effect was different.

### *Barley*

Within the HEALTHGRAIN diversity screen program, Andersson et al., 2008, performed a wide phytochemical analysis of ten different barley lines grown in one location (Andersson et al., 2008). They studied phytosterols, phenolic acids, tocopherols, folate, alkylresorcinol as well as dietary fibers using LC-DAD, GC-MS and other chemical methods. Phenolic acids of barley lines were quantified from three different fractions where phenolics were found in free forms, conjugated and bounded. In the other study, GC-MS based metabolomics assisted to uncover the salt stress responses of two barley cultivars that differ in salinity tolerance (Widodo et al., 2009). The study showed that the more salt sensitive barley line has a tendency to accumulate more amino acids, including proline, and the polyamine putrescine. While, the salt tolerant line showed an induced concentration of hexose phosphates and tricarboxylic acid cycle intermediates. Others have employed LC-MS metabolomics to identify metabolites of barley plants related to the resistance mechanism against the fungi *Fusarium graminearum* (Bollina et al., 2011). They studied metabolomic profiles of five *F. graminearum* resistant genotypes and one susceptible genotype and out of 1430 peaks, 115 metabolites were putatively identified as being related to the resistance. These include different groups of metabolites such as flavonoids, fatty acids, phenylpropanoids, linolenic acid, *p*-coumaric and sinapic acids. In another study, metabolite fingerprints of barley whole seeds, endosperms and embryos was analyzed by using GC-MS, during industrial malting, and the obtained data were analyzed by PCA (Gorzolka et al., 2012). This study revealed potential metabolite markers for a specific developmental stage that could be used in industrial process control. Another barley metabolomic study performed untargeted LC-MS analysis that was aimed at elucidating effects of pathogenic fungi *Gibberella zeae* (Kumaraswamy et al., 2011). In that study, LC-MS metabolomic data obtained from aqueous methanol extracts of spikelets from different barley genotypes were processed by the XCMS metabolomics software. The study included barley genotypes that are resistant to the fusarium head blight (FHB) pathogen (*Gibberella zeae*) and susceptible lines that were infected by this pathogen. Statistical analysis of the XCMS output revealed 161 metabolites as resistance related (RR) and/or pathogenicity related (PR) metabolites. 53 out of 161 metabolites were tentatively identified and they were mainly derived from fatty acids (jasmonic acid and methyl jasmonate), phenolic acids (*p*-coumaric acid, caffeoyl alcohol, dimethoxy-4-phenylcoumarin and rosmarinic acid) and flavonoids (naringenin, catechin, quercetin, and alpinumisoflavone).

### *Oat*

Phytochemical composition of five oat varieties were screened using LC-DAD and GC-MS within the HEALTHGRAIN diversity screen program (Shewry et al., 2008). The study provides metabolomic protocols for detection of sterols, tocopherols, folates and phenolic acids. Oat is considered as one of the most antioxidant rich cereal plants and its triterpene saponin content is much higher than in many other cereals (Osbourn, 2003). The triterpenoid saponin content of 16 different oat varieties were evaluated using HPLC-UV detection and the study showed that the main portion of oat saponins are

present in the endosperm of oat kernels (Onning et al., 1993). They also illustrated that the total saponin content of oat varieties differ in a range of 0.02-0.05 % (dry matter basis) and was not correlated to the lipid content. In another study, 21 different oat varieties were screened for their dietary phenolics, ferulic and coumaric acids by HPLC (Kovacova and Malinova, 2007). They illustrated a high correlation of ferulic ( $r = 0.92$ ) and coumaric acids ( $r = 0.91$ ) to the total phenolic content of the samples as determined by the Folin-Ciocalteu reagent. Ferulic acid concentration varied between 0.16-1.5 mg/g of dry grain sample, while coumaric acid concentrations were between 0.08-2.1 mg/g. Several studies were conducted on decomposition of antioxidant constituents of oat by using metabolomics approaches (Dimberg et al., 2005; Peterson, 2001). GC-MS based high-throughput metabolomic analysis of wild type and mutant oat plants depicted that the major change occurred in the metabolome flux, due to the genetic modifications (Qin et al., 2010). This study proves that the constructive changes made in in the biosynthetic pathway of triterpenoids resulted in a significant decrease of  $\beta$ -Amyrin content, while the concentration of the primary sterols such as Delta-7-campesterol and Delta-7-avenasterol increased several folds, compared to wild type oat.

### *Rye*

Metabolomic analysis of ten different old and modern rye varieties, that originated from five different European countries demonstrated that French rye varieties have the highest concentrations of most phytochemicals, including phenolic acids, sterols, tocopherols, folates and alkylresorcinols, while rye varieties that originated from Poland had the lowest level (Nyström et al., 2008). The study also illustrated a general metabolomic pattern and showed that the increased concentration of folates were related to the low level of alkylresorcinols, whereas elevated amount of arabinoxylans showed a correlation with an increased level of sterols. Free amino acids and sugars are the main precursors of acrylamide formed during the cooking process, and were analyzed from the various rye flour samples originated from different locations within Europe (Curtis et al., 2010). The concentration of free amino acids showed a high dependency on both the environment and the genotype, while the level of sucrose was largely determined by the genotype. In another study, GC-MS and HPLC-UV approaches were utilized for monitoring process-induced changes on bioactive compounds of whole-grain rye (Liukkonen et al., 2003). The study shows stabilities and fluctuations of bioactive compounds such as sterols, folates, tocopherols and phenolic acids during germination and sourdough baking processes. The study illustrates that the concentrations of folate and phenolics were increased during both processing time, while the amount of tocopherols were reduced due to the sourdough fermentation.

**Table 3.** Selected examples of cereal metabolomics by different analytical platforms

Cereals	LC-MS	CE-MS	GC-MS	NMR	Vibrational spectroscopy	Electronic spectroscopy
<b>Maize</b>	Amino acids [24] Sugars [24] Phenolic acids [21],[22],[23],[24],[29] Organic acids [24],[27] Flavonoids [22],[24],[25],[26] Benzoxazinones [7],[9] Mycotoxins [57],[58],[62]	Amino acids [1],[10] Organic acids [10] Nucleotides [1]	Amino acids [2],[4],[6],[24],[109] Sugars [2],[6],[24],[109] Sugar alcohols [6] Fatty acids [4],[6],[109] Phenolic acids [2],[24],[109] Organic acids [2],[6],[109] Sterols [4],[6] Phytic acid [2] Nucleotides [2] Amines [6]	Amino acids [3],[4],[5] Sugars [3],[4],[5] Phenolic acids [5],[28],[29],[30] Flavonoids [30] Organic acids [3],[4],[5],[30]	NIR [11],[12],[13] , [103]	UV-VIS [40],[41],[42] Fluorescence [105]
<b>Rice</b>	Amino acids [16],[106] Phenolic acids [18] Fatty acids [16] Organic acids [16],[106] Mycotoxins [62] Triacylglycerols [106]	Amino acids [43],[44],[46],[106] Amines [44] Sugars [43],[44],[46] Sugar phosphates [44],[106] Organic acids [43],[44],[46],[106] Nucleotides [43],[44],[106]	Amino acids [17],[19],[20] Sugars [17],[20],[106] Sugar alcohols [17] Fatty acids [17] Organic acids [17],[19],[20] Phenolic acids [106] Sterols [106]	Amino acids [20],[46],[47] Sugars [20],[46],[47] Sugar phosphates [46] Fatty acids [47] Organic acids [20],[46],[47] Sterols [47] Nucleotides [47]	NIR [48],[49] Raman [49] IR [51]	Fluorescence [50],[105] UV-VIS [51] UV-VIS [52]
	Amino acids [108] Phenolic acids [21],[108] Flavonoids [59],[108] Mycotoxins	Lignans [67] Sugars [68] Glycophosphates [69]	Amino acids [53],[54],[56],[108] Sugars [53],[56],[108] Fatty acids [56],[108] Phenolic acids	Amino acids [53],[55],[65],[66] Sugars [53],[55],[66] Phenolic acids [65] Organic acids	NIR [69],[70],[103] Raman [85]	UV-VIS [86],[87] Fluorescence [105]

<b>Wheat</b>	[57],[58],[60], [62] Pyrazole fungicides [61] Sterols [63] Benzoxazino nes [64]		[56],[108] Organic acids [53],[56],[108] Sterols [56],[63],[108] Volatile compounds [107]	[53],[55]		
<b>Barley</b>	Amino acids [79] Sugars [79],[80] Sugar phosphates [79],[80] Fatty acids [73],[76] Organic acids [73],[76],[79], [80] Flavonoids [73],[76] Mycotoxins [78]		Amino acids [8],[72],[74],[79] Sugars [8],[72],[74],[79] Sugars alcohols [8],[74] Sugar phosphates [74],[79] Fatty acids [8],[74] Phenolic acids [8],[72] Organic acids [8],[74],[79] Sterols [8],[75] Odorants [77] Tocopherols [8]	Amino acids [84] Sugars [84] Organic acids [84] Lipids [81],[84] Bulk carbohydrate s [83] Flavones [82]	NIR [14],[15], [103] Ramana [88]	UV-VIS [89] Fluorescence [90],[105]
<b>Oat and Rye</b>	Mycotoxins [95],[96] Proteins [97] Avenanthram ides [98] Proteins [97]		Aroma compounds [91] Fatty acids [93] Terpenes [93],[94] Sterols [92] Alkenes [93] Fatty acids [93] Terpenes [93] Alkenes [93]	Fatty acids [99] beta-Glucan [100],[101] Cellulose [102]	FT-IR [99] FT-Raman [100] NIR [103]	UV-VIS [104] Fluorescence [105]
<b>Refs.</b>	[1].(Levandi et al., 2008),[2].(Hazebroek et al., 2007),[3].(Gavaghan et al., 2011),[4].(Barros et al., 2010),[5].(Manetti et al., 2006), [6].(Skogerson et al., 2010),[7].(Walker et al., 2011),[8].(Frank et al., 2011),[9].(Hanhineva et al., 2011),[10].(Leon et al., 2009),[11].(Zhang et al., 2012),[12].(Williams et al., 2012),[13].(Zimmer et al., 1990),[14].(Rudi et al., 2006),[15].(Seefeldt et al., 2009),[16].(Chang et al., 2012),[17].(Frank et al., 2007),[18].(Qiu et al., 2010),[19].(Long et al., 2013),[20].(Barding et al., 2013),[21].(Chiremba et al., 2012),[22].(LeClere et al., 2007),[23].(Culhaoglu et al., 2011),[24].(Lozovaya et al., 2006),[25].(Biesaga, 2011),[26].(Li et al., 2007),[27].(Erro et al., 2009),[28].(Bunzel et al., 2005),[29].(Rouau et al., 2003),[39].(Kuhnen et al., 2010),[40].(CHEN et al., 2013),[41].(Elsark et al., 1993),[42].(Dowell et al., 2002),[43].(Ishikawa et al., 2010),[44].(Sato et al., 2008),[45].(Takahashi et al., 2006),[46].(Barding et al., 2012),[47].(Jones et al., 2011),[48].(Chen et al., 2010),[49].(Sohn et al., 2004),[50].(Shrestha et al., 2012),[51].(Samadi-Maybodi and Atashbozorg, 2006),[52].(Tangkhavanich et al., 2012),[53].(Baker et al., 2006),[54].(Curtis et al.,					

<p>2009),[55].(Graham et al., 2009),[56].(Beleggia et al., 2013), [57].(Tang et al., 2013),[58].(Skrbic et al., 2013),[59].(Wojakowska et al., 2013),[60].(Nakagawa et al., 2013),[61].(Dong et al., 2012),[62].(Warth et al., 2012),[63].(Nurmi et al., 2012),[64].(Farres et al., 2012),[65].(Lamanna et al., 2011),[66].(Browne and Brindle, 2007),[67].(Dinelli et al., 2007),[68].(Kabel et al., 2006),[69].(Goodwin et al., 2003),[70].(Liu et al., 2013),[71].(Salgo and Gergely, 2012),[72].(Widodo et al., 2009),[73].(Bollina et al., 2011),[74].(Gorzolka et al., 2012),[75].(Andersson et al., 2008),[76].(Kumaraswamy et al., 2011),[77].(Fickert and Schieberle, 1998),[78].(Solfrizzo et al., 2013),[79].(Huang et al., 2008),[80].(Rolletschek et al., 2004),[81].(Seefeldt et al., 2011),[82].(Norbaek et al., 2000),[83].(Seefeldt et al., 2008),[84].(Wu et al., 2013),[85].(Barron and Rouau, 2008),[86].(Balcerowska et al., 2009),[87].(Siuda et al., 2006),[88].(Greene and Bain, 2005),[89].(Berghold et al., 2004),[90].(Shalygo et al., 1998),[91].(Ren and Tian, 2012),[92].(Shewry et al., 2008),[93].(Perkowski et al., 2012),[94].(Qin et al., 2010),[95].(Liao et al., 2011),[96].(Gottschalk et al., 2007),[97].(Sorensen et al., 2010),[98].(Jastrebova et al., 2006),[99].(Manolache et al., 2013),[100].(Mikkelsen et al., 2013),[101].(Cui and Wang, 2009),[102].(Cyran and Saulnier, 2007),[103].(Kays et al., 2000),[104].(Feucht et al., 2007),[105].(Zekovic et al., 2012),[106].(Matsuda et al., 2012),[107].(Beleggia et al., 2009),[108].(Beleggia et al., 2011),[109].(Riedelsheimer et al., 2012a)</p>
---

## 7 Outlook and perspectives

This review focuses on current analytical technologies, including metabolomic profiling platforms and chemometric methods commonly used in cereal metabolomic studies. A general cereal metabolomic workflow is discussed with special attention on data acquisition and analysis steps. Metabolomics is usually driven by the purpose of the study and the analysts must decide on metabolomic protocols and applied platforms. The qualitative and quantitative nature of the obtained metabolomic data is highly dependent on these factors. It is important to mention that all the metabolomic workflow steps will have a significant influence on the final data. State-of-the-art metabolomics start from the design of experiment followed by optimization and validation of sample preparation, metabolite extraction and data acquisition protocols. Only reproducible protocols can generate reliable metabolomic data, since fluctuations on measurement of metabolites will lead to increased non-sample-related variations of the data. In section 3, we describe the main sources of errors made during data acquisition, which in fact constitutes the major part of the non-sample-related variations. Current approaches to overcome these issues and useful tools can partly solve the problem, but requires further research and technological innovations.

The amount of the information gained from cereal metabolomics are continuously increasing due to the developments of analytical technologies and the recent discoveries made in the field of plant biology and biotechnology. The metabolome of cereals are complex and covers a broad range of

metabolites. Despite recent technological advances, it is still not possible to detect the whole metabolome of cereals by using a single method which is why truly untargeted metabolomics approaches are not performed today. The current state-of-the-art approach is rather to use a biology driven selection approach in which a priori knowledge about the compound classes of interest and the matrices that they are imbedded in will drive the selection of the sampling protocol and the analytical method. Nevertheless, studies that deal with complex problems usually have as target to detect as many metabolites as possible and apply comprehensive analytical platforms such as NMR, LC-MS and GC-MS. These analytical platforms differ by their detection limit, sensitivity, reproducibility, chromatographic and mass resolution power and accuracy. Spectroscopic techniques such as NMR concede to MS based methods by their sensitivity and selectivity, while chromatography and MS based techniques mostly suffer from low reproducibility.

Today, up to several hundred metabolites of complex biological samples can be detected in a quantitative manner. However, complete phenotyping of biological systems requires even deeper studies of the metabolome. Therefore, screening of biological samples by metabolomic approaches are driven by phenomics, and modern cereal phenomics is driven by the global challenges of the world such as continuously increasing world population, rapidly changing and increasingly unpredictable climate, pollution and natural disasters. The main objectives of cereal metabolomics is to develop and improve the cereal varieties with desired quality traits that will be resistant and/or easily adaptable to the environmental changes, biotic stresses and provide high yield, nutritional value and food security (Figure 2). This in fact brings various different research fields together, from farming up to chemistry and statistics. It is worth to mention the role of the cereal metabolomics, since it provides the majority of quantitative phenomics data. Therefore, cereal plant metabolomic profiling protocols must be further improved and more sensitive and selective analytical platforms must be developed.

## 8 Abbreviations

AMDIS - automated mass spectral deconvolution and identification system

ANN - artificial neural networks

APCI – atmospheric pressure chemical ionization

APPI - atmospheric pressure chemical ionization

ASCA - analysis of variance simultaneous component analysis

CE – capillary electrophoresis

CI – chemical ionization

CIS - cooled injection system

COSY - correlation spectroscopy

COW - correlation optimised warping

CV - cross-validation

CVA - canonical variate analysis

DAD – diode array detector

DoE – design of experiment

ECNI – electron-capture negative ionization

ECVA – extended canonical variate analysis

EI – electron ionization

ESI – electrospray ionization

FI – field ionization

FID - flame ionization detector

FTIR – fourier transform infrared

FWHM - full width at half maximum

GC – gas chromatography

GMO – genetically modified organism

HCA - hierarchical cluster analysis

HMBC - heteronuclear multiple-bond correlation spectroscopy

HSQC - heteronuclear single quantum coherence

icoshift - interval correlation optimised shifting

IR - infrared

IT – ion trap

LC – liquid chromatography

LDA - linear discriminant analysis

LV - latent variables

MALDI - matrix assisted laser desorption ionization

MCA - multivariate curve resolution



MLR - multiple linear regression  
MS – mass spectrometry  
MSC - multiplicative signal correction  
NIR – near infrared  
NMR – nuclear magnetic resonance  
NOESY - nuclear overhauser effect spectroscopy  
OPLS-DA - orthogonal partial least squares discriminant analysis  
PARAFAC - Parallel Factor Analysis  
PARAFAC2 - Parallel Factor Analysis2  
PC – principal component  
PCA – principal component analysis  
PLS – partial least squares analysis  
PLS-DA - partial least squares discriminant analysis  
PQN - probabilistic quotation normalisation  
PTV - programmed temperature vaporization  
Q – quadrupole  
QQQ – triple quadrupole  
RMSEC - root mean square error of calibration  
RMSECV - root mean square error of cross validation  
SIMCA - soft independent modeling of class analogy  
SNV - standard normal variate  
SVM - support vector machine  
TOCSY - total correlation spectroscopy  
TOF – time-of-flight  
UPLC - ultra-high performance liquid chromatography  
UV - ultraviolet  
VIP - variable influence on projection  
VIS – visible spectroscopy

## 9 Acknowledgment

Faculty of Science is acknowledged for support to the elite-research area “Metabolomics and bioactive compounds” with a PhD stipendium to Bekzod Khakimov. We thank Professor Lars Munck for his valuable comments on the review.

## 10 References

- Allwood, J. W. and Goodacre, R. (2010). An Introduction to Liquid Chromatography-Mass Spectrometry Instrumentation Applied in Plant Metabolomic Analyses. *Phytochemical Analysis* 21, 33-47.
- Amarowicz, R., Zegarska, Z., Pegg, R. B., Karamac, M. and Kosinska, A. (2007). Antioxidant and radical scavenging activities of a barley crude extract and its fractions. *Czech Journal of Food Sciences* 25, 73-80.
- Amigo, J. M., Skov, T., Coello, J., MasPOCH, S. and Bro, R. (2008). Solving GC-MS problems with PARAFAC2. *Trac-Trends in Analytical Chemistry* 27, 714-725.
- Anderson, R. J. (1912). Concerning the organic phosphoric acid compound of wheat bran. *Journal of Biological Chemistry* 12, 447-464.
- Andersson, A. A. M., Lampi, A. M., Nyström, L., Piironen, V., Li, L., Ward, J. L., Gebruers, K., Courtin, C. M., Delcour, J. A., Boros, D. et al. (2008). Phytochemical and Dietary Fiber Components in Barley Varieties in the HEALTHGRAIN Diversity Screen. *J. Agric. Food Chem.* 56, 9767-9776.
- Andersson, A. A., Kamal-Eldin, A. and Aman, P. (2010). Effects of Environment and Variety on Alkylresorcinols in Wheat in the HEALTHGRAIN Diversity Screen. *J. Agric. Food Chem.* 58, 9299-9305.
- Arbona, V., Iglesias, D. J., Talon, M. and Gomez-Cadenas, A. (2009). Plant Phenotype Demarcation Using Nontargeted LC-MS and GC-MS Metabolite Profiling. *J. Agric. Food Chem.* 57, 7338-7347.
- Baker, E. A. and Smith, I. M. (1977). Antifungal Compounds in Winter-Wheat Resistant and Susceptible to *Septoria-Nodorum*. *Annals of Applied Biology* 87, 67-73.
- Baker, J. M., Hawkins, N. D., Ward, J. L., Lovegrove, A., Napier, J. A., Shewry, P. R. and Beale, M. H. (2006). A metabolomic study of substantial equivalence of field-grown genetically modified wheat. *Plant Biotechnology Journal* 4, 381-392.
- Balcerowska, G., Siuda, R., Skrzypczak, J., Lukanowski, A. and Sadowski, C. (2009). Effect of particle size and spectral sub-range within the UV-VIS-NIR range using diffuse reflectance spectra on multivariate models in evaluating the severity of fusariosis in ground wheat. *Food Additives and Contaminants Part A-Chemistry Analysis Control Exposure & Risk Assessment* 26, 726-732.
- Balmer, D., Flors, V., Glauser, G. and Mauch-Mani, B. (2013). Metabolomics of cereals under biotic stress: current knowledge and techniques. *Frontiers in plant science* 4, 82.
- Barabasi, A. L. and Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics* 5, 101-U15.
- Barding, G. A., Beni, S., Fukao, T., Bailey-Serres, J. and Larive, C. K. (2013). Comparison of GC-MS and NMR for Metabolite Profiling of Rice Subjected to Submergence Stress. *Journal of Proteome Research* 12, 898-909.

- Barding, G. A., Fukao, T., Beni, S., Bailey-Serres, J. and Larive, C. K. (2012). Differential Metabolic Regulation Governed by the Rice SUB1A Gene during Submergence Stress and Identification of Alanyl-glycine by H-1 NMR Spectroscopy. *Journal of Proteome Research* 11, 320-330.
- Bardinskaya, M. S. and Shubert, T. A. (1962). Phenolic Compounds of Cereals. *Biochemistry-Moscow* 27, 46.
- Barker, M. and Rayens, W. (2003). Partial least squares for discrimination. *J. Chemometrics* 17, 166-173.
- Barnes, R. J., Dhanoa, M. S. and Lister, S. J. (1989). Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra. *Applied Spectroscopy* 43, 772-777.
- Barron, C. and Rouau, X. (2008). FTIR and Raman signatures of wheat grain peripheral tissues. *Cereal Chemistry* 85, 619-625.
- Barros, E., Lezar, S., Anttonen, M. J., van Dijk, J. P., Rohlig, R. M., Kok, E. J. and Engel, K. H. (2010). Comparison of two GM maize varieties with a near-isogenic non-GM variety using transcriptomics, proteomics and metabolomics. *Plant Biotechnology Journal* 8, 436-451.
- Bassel, G. W., Gaudinier, A., Brady, S. M., Hennig, L., Rhee, S. Y. and De Smet, I. (2012). Systems Analysis of Plant Functional, Transcriptional, Physical Interaction, and Metabolic Networks. *Plant Cell* 24, 3859-3875.
- Beleggia, R., Platani, C., Nigro, F., De Vita, P., Cattivelli, L. and Papa, R. (2013). Effect of genotype, environment and genotype-by-environment interaction on metabolite profiling in durum wheat (*Triticum durum* Desf.) grain. *Journal of Cereal Science* 57, 183-192.
- Beleggia, R., Platani, C., Papa, R., Di Chio, A., Barros, E., Mashaba, C., Wirth, J., Fammartino, A., Sautter, C., Conner, S. et al. (2011). Metabolomics and Food Processing: From Semolina to Pasta. *J. Agric. Food Chem.* 59, 9366-9377.
- Beleggia, R., Platani, C., Spano, G., Monteleone, M. and Cattivelli, L. (2009). Metabolic profiling and analysis of volatile composition of durum wheat semolina and pasta. *Journal of Cereal Science* 49, 301-309.
- Berghold, J., Eichmuller, C., Hortensteiner, S. and Krautler, B. (2004). Chlorophyll breakdown in tobacco: On the structure of two nonfluorescent chlorophyll catabolites. *Chemistry & Biodiversity* 1, 657-668.
- Berry, C. P., Youngs, V. L. and Gilles, K. A. (1968). Analysis of Free and Esterified Sterols in Wheat Flour and Semolina. *Cereal Chemistry* 45, 616.
- Biesaga, M. (2011). Influence of extraction methods on stability of flavonoids. *Journal of Chromatography A* 1218, 2505-2512.
- Bietz, J. A. and Kruger, J. E. (1988). The Evolution of Cereal Protein-Analysis by Hplc. *Cereal Foods World* 33, 682-683.
- Bilinski, E. and Mcconnell, W. B. (1958a). Studies on Wheat Plants Using C-14 Compounds .7. Utilization of Pyruvate-2-C-14. *Canadian Journal of Biochemistry and Physiology* 36, 381-388.
- Bilinski, E. and Mcconnell, W. B. (1958b). Studies on Wheat Plants Using Carbon-14 Compounds .6. Some Observations on Protein Biosynthesis. *Cereal Chemistry* 35, 66-81.

- Boelens, H. F. M., Dijkstra, R. J., Eilers, P. H. C., Fitzpatrick, F. and Westerhuis, J. A. (2004). New background correction method for liquid chromatography with diode array detection, infrared spectroscopic detection and Raman spectroscopic detection. *Journal of Chromatography A* 1057, 21-30.
- Bollina, V., Kushalappa, A. C., Choo, T. M., Dion, Y. and Rioux, S. (2011). Identification of metabolites related to mechanisms of resistance in barley against *Fusarium graminearum*, based on mass spectrometry. *Plant Molecular Biology* 77, 355-370.
- Borin, A. and Poppi, R. J. (2005). Application of mid infrared spectroscopy and iPLS for the quantification of contaminants in lubricating oil. *Vibrational Spectroscopy* 37, 27-32.
- Bowne, J. B., Erwin, T. A., Juttner, J., Schnurbusch, T., Langridge, P., Bacic, A. and Roessner, U. (2012). Drought Responses of Leaf Tissues from Wheat Cultivars of Differing Drought Tolerance at the Metabolite Level. *Molecular Plant* 5, 418-429.
- Bro, R. (1997). PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems* 38, 149-171.
- Bro, R., Andersson, C. A. and Kiers, H. A. L. (1999). PARAFAC2 - Part II. Modeling chromatographic data with retention time shifts. *J. Chemometrics* 13, 295-309.
- Broadhurst, D. I. and Kell, D. B. (2006). Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* 2, 171-196.
- Browne, R. A. and Brindle, K. M. (2007). H-1 NMR-based metabolite profiling as a potential selection tool for breeding passive resistance against *Fusarium* head blight (FHB) in wheat. *Molecular Plant Pathology* 8, 401-410.
- Bunzel, M., Ralph, J., Funk, C. and Steinhart, H. (2005). Structural elucidation of new ferulic acid-containing phenolic dimers and trimers isolated from maize bran. *Tetrahedron Letters* 46, 5845-5850.
- Bylesjo, M., Rantalainen, M., Cloarec, O., Nicholson, J. K., Holmes, E. and Trygg, J. (2006). OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *J. Chemometrics* 20, 341-351.
- Chang, W. T., Thissen, U., Ehlert, K. A., Koek, M. M., Jellema, R. H., Hankemeier, T., van der Greef, J. and Wang, M. (2006). Effects of growth conditions and processing on *Rehmannia glutinosa* using fingerprint strategy. *Planta Medica* 72, 458-467.
- Chang, Y. W., Zhao, C. X., Zhu, Z., Wu, Z. M., Zhou, J., Zhao, Y. N., Lu, X. and Xu, G. W. (2012). Metabolic profiling based on LC/MS to evaluate unintended effects of transgenic rice with cry1Ac and sck genes. *Plant Molecular Biology* 78, 477-487.
- Chen, K. J., Huang, M. and Sun, X. (2010). Potential Use of Nir Spectroscopy for the Estimation of Milled Rice Yield. *Transactions of the Asabe* 53, 497-501.
- CHEN, Z. q., WANG, L., BAI, Y. I., YANG, L. p., LU, Y. I., WANG, H. and WANG, Z. y. (2013). Spectral Response of Maize Leaves and Prediction of Their Nitrogen Content. pp. 1066-1070.
- Chiremba, C., Taylor, J. R. N., Rooney, L. W. and Beta, T. (2012). Phenolic acid content of sorghum and maize cultivars varying in hardness. *Food Chemistry* 134, 81-88.

- Christensen, J., Norgaard, L., Bro, R. and Engelsen, S. B. (2006). Multivariate autofluorescence of intact food systems. *Chemical Reviews* 106, 1979-1994.
- Clegg, K. M. (1955). Method for the Estimation of Starch and Sugars in Cereals Using the Anthrone Reagent. *Biochemical Journal* 61, R17.
- Collins, F. W., Mclachlan, D. C. and Blackwell, B. A. (1991). Oat Phenolics - Avenaluminic Acids, A New Group of Bound Phenolic-Acids from Oat Groats and Hulls. *Cereal Chemistry* 68, 184-189.
- Cui, S. W. and Wang, Q. (2009). Cell wall polysaccharides in cereals: chemical structures and functional properties. *Structural Chemistry* 20, 291-297.
- Culhaoglu, T., Zheng, D., Mechin, V. and Baumberger, S. (2011). Adaptation of the Carrez procedure for the purification of ferulic and p-coumaric acids released from lignocellulosic biomass prior to LC/MS analysis. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences* 879, 3017-3022.
- Curtis, T. Y., Muttucumar, N., Shewry, P. R., Parry, M. A. J., Powers, S. J., Elmore, J. S., Mottram, D. S., Hook, S. and Halford, N. G. (2009). Effects of Genotype and Environment on Free Amino Acid Levels in Wheat Grain: Implications for Acrylamide Formation during Processing. *J. Agric. Food Chem.* 57, 1013-1021.
- Curtis, T. Y., Powers, S. J., Balagiannis, D., Elmore, J. S., Mottram, D. S., Parry, M. A. J., Rakszegi, M., Bedo, Z., Shewry, P. R. and Halford, N. G. (2010). Free Amino Acids and Sugars in Rye Grain: Implications for Acrylamide Formation. *J. Agric. Food Chem.* 58, 1959-1969.
- Cyran, M. R. and Saulnier, L. (2007). Association and structural diversity of hemicelluloses in the cell walls of rye outer layers: Comparison between two ryes with opposite breadmaking quality. *J. Agric. Food Chem.* 55, 2329-2341.
- Danielsson, A. P. H., Moritz, T., Mulder, H. and Spiegel, P. (2012). Development of a gas chromatography/mass spectrometry based metabolomics protocol by means of statistical experimental design. *Metabolomics* 8, 50-63.
- De Jong, S. (1990). Multivariate calibration, H. Martens and T. Naes, Wiley, New York, 1989. ISBN 0 471 90979 3. Price: -ú75.00, US\$138.00. No. of pages: 504. *J. Chemometrics* 4, 441.
- De Vos, R. C., Moco, S., Lommen, A., Keurentjes, J. J., Bino, R. J. and Hall, R. D. (2007). Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. *Nature Protocols* 2, 778-791.
- Defeo, E. M., Wu, C. L., McDougal, W. S. and Cheng, L. L. (2011). A decade in prostate cancer: from NMR to metabolomics. *Nature Reviews Urology* 8, 301-311.
- Delwiche, S. R. (1995). Single Wheat Kernel Analysis by Near-Infrared Transmittance - Protein-Content. *Cereal Chemistry* 72, 11-16.
- Di Anibal, C. V., Callao, M. P. and Ruisanchez, I. (2011). H-1 NMR variable selection approaches for classification. A case study: The determination of adulterated foodstuffs. *Talanta* 86, 316-323.

- Dimberg, L. H., Gissen, C. and Nilsson, J. (2005). Phenolic compounds in oat grains (*Avena sativa* L.) grown in conventional and organic systems. *Ambio* 34, 331-337.
- Dinelli, G., Marotti, I., Bosi, S., Benedettelli, S., Ghiselli, L., Cortacero-Ramirez, S., Carrasco-Pancorbo, A., Segura-Carretero, A. and Fernandez-Gutierrez, A. (2007). Lignan profile in seeds of modern and old Italian soft wheat (*Triticum aestivum* L.) cultivars as revealed by CE-MS analyses. *Electrophoresis* 28, 4212-4219.
- Dong, F. S., Chen, X., Liu, X. G., Xu, J., Li, Y. B., Shan, W. L. and Zheng, Y. Q. (2012). Simultaneous determination of five pyrazole fungicides in cereals, vegetables and fruits using liquid chromatography/tandem mass spectrometry. *Journal of Chromatography A* 1262, 98-106.
- Dowell, F. E., Pearson, T. C., Maghirang, E. B., Xie, F. and Wicklow, D. T. (2002). Reflectance and Transmittance Spectroscopy Applied to Detecting Fumonisin in Single Corn Kernels Infected with *Fusarium verticillioides*. *Cereal Chemistry Journal* 79, 222-226.
- Draisma, H. H. M., Reijmers, T. H., Meulman, J. J., van der Greef, J., Hankemeier, T. and Boomsma, D. I. (2013). Hierarchical clustering analysis of blood plasma lipidomics profiles from mono- and dizygotic twin families. *European Journal of Human Genetics* 21, 95-101.
- Dunn, W. B. (2008). Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. *Physical Biology* 5.
- Eberius, M. and Lima-Guerra, J. (2009). High-Throughput Plant Phenotyping G $\hat{C}$  Data Acquisition, Transformation, and Analysis. In *Bioinformatics* (eds. D. Edwards, J. Stajich and D. Hansen), pp. 259-278: Springer New York.
- Elsark, N. S., Rizk, L. F. and Doss, H. R. (1993). Uv Spectra Parameters to Investigate the Influences of Intensifications of Soybean with Sorghum and Maize on the Physical-Properties of Soybean Seed Oils. *Grasas y Aceites* 44, 243-248.
- Engelsen, S. B., Savorani, F. and Rasmussen, M. A. (2013). Chemometric Exploration of Quantitative NMR Data. *eMagRes*, 267-278.
- Erny, G. L., Elvira, C., San Roman, J. and Cifuentes, A. (2006). Capillary electrophoresis using copolymers of different composition as physical coatings: A comparative study. *Electrophoresis* 27, 1041-1049.
- Erro, J., Zamarreno, A. M., Yvin, J. C. and Garcia-Mina, J. M. (2009). Determination of Organic Acids in Tissues and Exudates of Maize, Lupin, and Chickpea by High-Performance Liquid Chromatography-Tandem Mass Spectrometry. *J. Agric. Food Chem.* 57, 4004-4010.
- Farres, M., Villagrasa, M., Eljarrat, E., Barcelo, D. and Tauler, R. (2012). Chemometric evaluation of different experimental conditions on wheat (*Triticum aestivum* L.) development using liquid chromatography mass spectrometry (LC-MS) profiles of benzoxazinone derivatives. *Analytica Chimica Acta* 731, 24.
- Fernandez-Orozco, R., Li, L., Harflett, C., Shewry, P. R. and Ward, J. L. (2010). Effects of Environment and Genotype on Phenolic Acids in Wheat in the HEALTHGRAIN Diversity Screen. *J. Agric. Food Chem.* 58, 9341-9352.
- Fernie, A. R. and Schauer, N. (2009). Metabolomics-assisted breeding: a viable option for crop improvement? *Trends in Genetics* 25, 39-48.

- Ferrari, E., Foca, G., Vignali, M., Tassi, L. and Ulrici, A. (2011). Adulteration of the anthocyanin content of red wines: Perspectives for authentication by Fourier Transform-Near InfraRed and H-1 NMR spectroscopies. *Analytica Chimica Acta* 701, 139-151.
- Feucht, W., Dithmar, H. and Polster, J. (2007). Variation of the nuclear, subnuclear and chromosomal flavanol deposition in hemlock and rye. *International Journal of Molecular Sciences* 8, 635-650.
- Fickert, B. and Schieberle, P. (1998). Identification of the key odorants in barley malt (caramalt) using GC/MS techniques and odour dilution analyses. *Nahrung-Food* 42, 371-375.
- Fiehn, O. (2002). Metabolomics - the link between genotypes and phenotypes. *Plant Molecular Biology* 48, 155-171.
- Fiehn, O. (2008). Extending the breadth of metabolite profiling by gas chromatography coupled to mass spectrometry. *Trac-Trends in Analytical Chemistry* 27, 261-269.
- Fisher, R. A. (1918). The Correlation between Relatives on the Supposition of Mendelian Inheritance. pp. 399-433: *Philosophical Transactions of the Royal Society of Edinburgh*.
- Frank, T., Meuleye, B. S., Miller, A., Shu, Q. Y. and Engel, K. H. (2007). Metabolite profiling of two low phytic acid (lpa) rice mutants. *J. Agric. Food Chem.* 55, 11011-11019.
- Frank, T., Scholz, B., Peter, S. and Engel, K. H. (2011). Metabolite profiling of barley: Influence of the malting process. *Food Chemistry* 124, 948-957.
- Gavaghan, C. L., Li, J. V., Hadfield, S. T., Hole, S., Nicholson, J. K., Wilson, I. D., Howe, P. W., Stanley, P. D. and Holmes, E. (2011). Application of NMR-based Metabolomics to the Investigation of Salt Stress in Maize (*Zea mays*). *Phytochemical Analysis* 22, 214-224.
- Geladi, P., Macdougall, D. and Martens, H. (1985). Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat. *Applied Spectroscopy* 39, 491-500.
- Gerlai, R. (2002). Phenomics: fiction or the future? *Trends in Neurosciences* 25, 506-509.
- Gibson, S. M. and Strauss, G. (1991). Implication of Phenolic-Acids As Texturizing Agents During Cooking-Extrusion Cereals. *Abstracts of Papers of the American Chemical Society* 202, 150.
- Gohlke, R. S. and McLafferty, F. W. (1993). Early Gas-Chromatography Mass-Spectrometry. *Journal of the American Society for Mass Spectrometry* 4, 367-371.
- Goodwin, L., Startin, J. R., Keely, B. J. and Goodall, D. M. (2003). Analysis of glyphosate and glufosinate by capillary electrophoresis - mass spectrometry utilising a sheathless microelectrospray interface. *Journal of Chromatography A* 1004, 107-119.
- Gorzolka, K., Lissel, M., Kessler, N., Loch-Ahring, S. and Niehaus, K. (2012). Metabolite fingerprinting of barley whole seeds, endosperms, and embryos during industrial malting. *Journal of Biotechnology* 159, 177-187.
- Gottschalk, C., Barthel, J., Engelhardt, G., Bauer, J. and Meyer, K. (2007). Occurrence of type A trichothecenes in conventionally and organically produced oats and oat products. *Molecular Nutrition & Food Research* 51, 1547-1553.

- Graham, S., Amigues, E., Migaud, M. and Browne, R. (2009). Application of NMR based metabolomics for mapping metabolite variation in European wheat. *Metabolomics* 5, 302-306.
- Grata, E., Boccard, J., Guillarme, D., Glauser, G., Carrupt, P. A., Farmer, E. E., Wolfender, J. L. and Rudaz, S. (2008). UPLC-TOF-MS for plant metabolomics: A sequential approach for wound marker analysis in *Arabidopsis thaliana*. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences* 871, 261-270.
- Greene, P. R. and Bain, C. D. (2005). Total internal reflection Raman spectroscopy of barley leaf epicuticular waxes in vivo. *Colloids and Surfaces B-Biointerfaces* 45, 174-180.
- Guerard, F., Petriacq, P., Gakiere, B. and Tcherkez, G. (2011). Liquid chromatography/time-of-flight mass spectrometry for the analysis of plant samples: A method for simultaneous screening of common cofactors or nucleotides and application to an engineered plant line. *Plant Physiology and Biochemistry* 49, 1117-1125.
- Gullberg, J., Jonsson, P., Nordstrom, A., Sjostrom, M. and 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, 283-295.
- Gürdeniz, G., Kristensen, M., Skov, R. and Dragsted, L. O. (2012). The Effect of LC-MS Data Preprocessing Methods on the Selection of Plasma Biomarkers in Fed vs.Fasted Rats. *Metabolites*, 77-99.
- Hanhineva, K., Rogachev, I., Aura, A. M., Aharoni, A., Poutanen, K. and Mykkanen, H. (2011). Qualitative Characterization of Benzoxazinoid Derivatives in Whole Grain Rye and Wheat by LC-MS Metabolite Profiling. *J. Agric. Food Chem.* 59, 921-927.
- Harshman, R. A. (1970). Foundations of the PARAFAC procedure: Model and conditions for an GÇÿexplanatoryGÇÖ multi-mode factor analysis. *UCLA Working Papers Phonetics* 16.
- Hart-Smith, G. and Blanksby, S. J. (2011). Mass Analysis. In *Mass Spectrometry in Polymer Chemistry*, pp. 5-32: Wiley-VCH Verlag GmbH & Co. KGaA.
- Hazebroek, J., Harp, T., Shi, J. and Wang, H. (2007). Metabolomic Analysis of Low Phytic Acid Maize Kernels. In *Concepts in Plant Metabolomics* (eds. B. Nikolau and E. Wurtele), pp. 221-238: Springer Netherlands.
- Higashi, Y. and Saito, K. (2013). Network analysis for gene discovery in plant-specialized metabolism. *Plant Cell and Environment* 36, 1597-1606.
- Holcapek, M., Jirasko, R. and Lisa, M. (2012). Recent developments in liquid chromatography-mass spectrometry and related techniques. *Journal of Chromatography A* 1259, 3-15.
- Hommerson, P., Khan, A. M., de Jong, G. J. and Somsen, G. W. (2011). Ionization Techniques in Capillary Electrophoresis-Mass Spectrometry: Principles, Design, and Application. *Mass Spectrometry Reviews* 30, 1096-1120.
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24, 417-441.



- Huang, C. Y., Roessner, U., Eickmeier, I., Genc, Y., Callahan, D. L., Shirley, N., Langridge, P. and Bacic, A. (2008). Metabolite profiling reveals distinct changes in carbon and nitrogen metabolism in phosphate-deficient barley plants (*Hordeum vulgare* L.). *Plant and Cell Physiology* 49, 691-703.
- Ishikawa, T., Takahara, K., Hirabayashi, T., Matsumura, H., Fujisawa, S., Terauchi, R., Uchimiya, H. and Kawai-Yamada, M. (2010). Metabolome Analysis of Response to Oxidative Stress in Rice Suspension Cells Overexpressing Cell Death Suppressor Bax Inhibitor-1. *Plant and Cell Physiology* 51, 9-20.
- Jaroszewski, J. W. (2005a). Hyphenated NMR methods in natural products research, Part 1: Direct hyphenation. *Planta Medica* 71, 691-700.
- Jaroszewski, J. W. (2005b). Hyphenated NMR methods in natural products research, Part 2: HPLC-SPE-NMR and other new trends in NMR hyphenation. *Planta Medica* 71, 795-802.
- Jastrebova, J., Skoglund, M., Nilsson, J. and Dimberg, L. H. (2006). Selective and sensitive LC-MS determination of avenanthramides in oats. *Chromatographia* 63, 419-423.
- Jaumot, J., Gargallo, R., de Juan, A. and Tauler, R. (2005). A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB. *Chemometrics and Intelligent Laboratory Systems* 76, 101-110.
- Jensen, S. A., Munck, L. and Martens, H. (1982). The Botanical Constituents of Wheat and Wheat Milling Fractions .1. Quantification by Autofluorescence. *Cereal Chemistry* 59, 477-484.
- Johnsen, L. G., Skov, T., Houlberg, U. and Bro, R. (2013). An automated method for baseline correction, peak finding and peak grouping in chromatographic data. *Analyst* 138, 3502-3511.
- Jones, O. A. H., Maguire, M. L., Griffin, J. L., Jung, Y. H., Shibato, J., Rakwal, R., Agrawal, G. K. and Jwa, N. S. (2011). Using metabolic profiling to assess plant-pathogen interactions: an example using rice (*Oryza sativa*) and the blast pathogen *Magnaporthe grisea*. *European Journal of Plant Pathology* 129, 539-554.
- Kabel, M. A., Heijnis, W. H., Bakx, E. J., Kuijpers, R., Voragen, A. G. J. and Schols, H. A. (2006). Capillary electrophoresis fingerprinting, quantification and mass-identification of various 9-aminopyrene-1,4,6-trisulfonate-derivatized oligomers derived from plant polysaccharides. *Journal of Chromatography A* 1137, 119-126.
- Kacurakova, M. and Wilson, R. H. (2001). Developments in mid-infrared FT-IR spectroscopy of selected carbohydrates. *Carbohydrate Polymers* 44, 291-303.
- Kanani, H., Chrysanthopoulos, P. K. and Klapa, M. I. (2008). Standardizing GC-MS metabolomics. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences* 871, 191-201.
- Kasicka, V. (2012). Recent developments in CE and CEC of peptides (2009-2011). *Electrophoresis* 33, 48-73.
- Katajamaa, M. and Oresic, M. (2007). Data processing for mass spectrometry-based metabolomics. *Journal of Chromatography A* 1158, 318-328.
- Kays, S. E., Barton, F. E. and Windham, W. R. (2000). Predicting protein content by near infrared reflectance spectroscopy in diverse cereal food products. *Journal of Near Infrared Spectroscopy* 8, 35-43.

- Kemp, R. J. and Mercer, E. I. (1968). Sterol Esters of Maize Seedlings. *Biochemical Journal* 110, 111.
- Khakimov, B., Amigo, J. M., Bak, S. and Engelsens, S. B. (2012). Plant metabolomics: Resolution and quantification of elusive peaks in liquid chromatography-mass spectrometry profiles of complex plant extracts using multi-way decomposition methods. *Journal of Chromatography. A* 1266, 84-94.
- Khakimov, B., Motawia, M. S., Bak, S. and Engelsens, S. B. (2013). The use of trimethylsilyl cyanide derivatization for robust and broad spectrum high-throughput gas-chromatography-mass spectrometry based metabolomics. *Analytical and Bioanalytical Chemistry*. DOI: 10.1007/s00216-013-7341-z
- Kiers, H. A. L., Ten Berge, J. M. F. and Bro, R. (1999). PARAFAC2 - Part I. A direct fitting algorithm for the PARAFAC2 model. *Journal of Chemometrics* 13, 275-294.
- Kim, H. K., Saifullah, Khan, S., Wilson, E. G., Kricun, S. D. P., Meissner, A., Goraler, S., Deelder, A. M., Choi, Y. H. and Verpoorte, R. (2010). Metabolic classification of South American Ilex species by NMR-based metabolomics. *Phytochemistry* 71, 773-784.
- Kissmeyernielsen, A., Jensen, S. A. and Munck, L. (1985). The Botanical Composition of Rye and Rye Milling Fractions Determined by Fluorescence Spectrometry and Amino-Acid Composition. *Journal of Cereal Science* 3, 181-192.
- Koek, M. M., Jellema, R. H., van der Greef, J., Tas, A. C. and Hankemeier, T. (2011). Quantitative metabolomics based on gas chromatography mass spectrometry: status and perspectives. *Metabolomics* 7, 307-328.
- Kolch, W., Neussus, C., Peizing, M. and Mischak, H. (2005). Capillary electrophoresis - Mass spectrometry as a powerful tool in clinical diagnosis and biomarker discovery. *Mass Spectrometry Reviews* 24, 959-977.
- Kong, D., Choo, T. M., Jui, P., Ferguson, T., Therrien, M. C., Ho, K. M., May, K. W. and Narasimhalu, P. (1995). Variation in starch, protein, and fibre of Canadian barley cultivars. *Canadian Journal of Plant Science* 75, 865-870.
- Kovacova, M. and Malinova, E. (2007). Ferulic and coumaric acids, total phenolic compounds and their correlation in selected oat genotypes. *Czech Journal of Food Sciences* 25, 325-332.
- Kristensen, M., Savorani, F., Ravn-Haren, G., Poulsen, M., Markowski, J., Larsen, F. H., Dragsted, L. O. and Engelsens, S. B. (2010). NMR and interval PLS as reliable methods for determination of cholesterol in rodent lipoprotein fractions. *Metabolomics* 6, 129-136.
- Kuhnen, S., Oglari, J. B., Dias, P. F., Santos, M. D., Ferreira, A. G., Bonham, C. C., Wood, K. V. and Maraschin, M. (2010). Metabolic Fingerprint of Brazilian Maize Landraces Silk (Stigma/Styles) Using NMR Spectroscopy and Chemometric Methods. *J. Agric. Food Chem.* 58, 2194-2200.
- Kumaraswamy, G. K., Bollina, V., Kushalappa, A. C., Choo, T. M., Dion, Y., Rioux, S., Mamer, O. and Faubert, D. (2011). Metabolomics technology to phenotype resistance in barley against *Gibberella zeae*. *European Journal of Plant Pathology* 130, 29-43.
- Kusano, M. and Saito, K. (2012). Role of Metabolomics in Crop Improvement. *Journal of Plant Biochemistry and Biotechnology* 21, S24-S31.

- Kuzina, V., Ekstrom, C. T., Andersen, S. B., Nielsen, J. K., Olsen, C. E. and Bak, S. (2009). Identification of Defense Compounds in *Barbarea vulgaris* against the Herbivore *Phyllotreta nemorum* by an Ecometabolomic Approach. *Plant Physiology* 151, 1977-1990.
- Lamanna, R., Cattivelli, L., Miglietta, M. L. and Troccoli, A. (2011). Geographical origin of durum wheat studied by H-1-NMR profiling. *Magnetic Resonance in Chemistry* 49, 1-5.
- Lampi, A. M., Nurmi, T. and Piironen, V. (2010). Effects of the Environment and Genotype on Tocopherols and Tocotrienols in Wheat in the HEALTHGRAIN Diversity Screen. *J. Agric. Food Chem.* 58, 9306.
- Lander, E. S. (2011). Initial impact of the sequencing of the human genome. *Nature* 470, 187-197.
- Larkin, P. and Harrigan, G. G. (2007). Opportunities and surprises in crops modified by transgenic technology: metabolic engineering of benzylisoquinoline alkaloid, gossypol and lysine biosynthetic pathways. *Metabolomics* 3, 371-382.
- Lawton, W. H. and Sylvestr, E. A. (1971). Self Modeling Curve Resolution. *Technometrics* 13, 617.
- LeClere, S., Schmelz, E. A. and Chourey, P. S. (2007). Phenolic compounds accumulate specifically in maternally-derived tissues of developing maize kernels. *Cereal Chemistry* 84, 350-356.
- Leon, C., Rodriguez-Meizoso, I., Lucio, M., Garcia-Canas, V., Ibanez, E., Schmitt-Kopplin, P. and Cifuentes, A. (2009). Metabolomics of transgenic maize combining Fourier transform-ion cyclotron resonance-mass spectrometry, capillary electrophoresis-mass spectrometry and pressurized liquid extraction. *Journal of Chromatography A* 1216, 7314-7323.
- Levandi, T., Leon, C., Kaljurand, M., Garcia-Canas, V. and Cifuentes, A. (2008). Capillary electrophoresis time-of-flight mass spectrometry for comparative metabolomics of transgenic versus conventional maize. *Anal. Chem.* 80, 6329-6335.
- Li, H. H., Flachowsky, H., Fischer, T. C., Hanke, M. V., Forkmann, G., Treutter, D., Schwab, W., Hoffmann, T. and Szankowski, I. (2007). Maize Lc transcription factor enhances biosynthesis of anthocyanins, distinct proanthocyanidins and phenylpropanoids in apple (*Malus domestica* Borkh.). *Planta* 226, 1243-1254.
- Li, L., Shewry, P. R. and Ward, J. L. (2008). Phenolic Acids in Wheat Varieties in the HEALTHGRAIN Diversity Screen. *J. Agric. Food Chem.* 56, 9732-9739.
- Li, Q. B., Huang, Y. W., Zhang, G. J., Zhang, Q. X., Li, X. and Wu, J. G. (2009). Chlorophyll Content Nondestructive Measurement Method Based on Vis/NIR Spectroscopy. *Spectroscopy and Spectral Analysis* 29, 3275-3278.
- Liao, C. D., Lin, H. Y., Chiueh, L. C. and Shih, D. Y. C. (2011). Simultaneous Quantification of Aflatoxins, Ochratoxin A and Zearalenone in Cereals by LC-MS/MS. *Journal of Food and Drug Analysis* 19, 259.
- Lisec, J., Schauer, N., Kopka, J., Willmitzer, L. and Fernie, A. R. (2006). Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nature Protocols* 1, 387-396.
- Liu, L. L., Zhao, B., Zhang, Y. Q. and Zhang, X. C. (2013). Research on Development and Experiment of NIR Wheat Quality Quick Detection System. *Spectroscopy and Spectral Analysis* 33, 92-97.

- Liukkonen, K. H., Katina, K., Wilhelmsson, A., Myllymaki, O., Lampi, A. M., Kariluoto, S., Piironen, V., Heinonen, S. M., Nurmi, T., Adlercreutz, H. et al. (2003). Process-induced changes on bioactive compounds in whole grain rye. *Proceedings of the Nutrition Society* 62, 117-122.
- Lodi, A., Tiziani, S., Khanim, F. L., Gunther, U. L., Viant, M. R., Morgan, G. J., Bunce, C. M. and Drayson, M. T. (2013). Proton NMR-Based Metabolite Analyses of Archived Serial Paired Serum and Urine Samples from Myeloma Patients at Different Stages of Disease Activity Identifies Acetylcarnitine as a Novel Marker of Active Disease. *Plos One* 8.
- Long, X. H., Liu, Q. Q., Chan, M. L., Wang, Q. and Sun, S. S. M. (2013). Metabolic engineering and profiling of rice with increased lysine. *Plant Biotechnology Journal* 11, 490-501.
- Lopez-Rituerto, E., Savorani, F., Avenozza, A., Busto, J. H., Peregrina, J. M. and Engelsen, S. B. (2012). Investigations of La Rioja Terroir for Wine Production Using H-1 NMR Metabolomics. *J. Agric. Food Chem.* 60, 3452-3461.
- Lozovaya, V., Ulanov, A., Lygin, A., Duncan, D. and Widholm, J. (2006). Biochemical features of maize tissues with different capacities to regenerate plants. *Planta* 224, 1385-1399.
- Madhujith, T. and Shahidi, F. (2007). Antioxidative and antiproliferative properties of selected barley (*Hordeum vulgare* L.) cultivars and their potential for inhibition of low-density lipoprotein (LDL) cholesterol oxidation. *J. Agric. Food Chem.* 55, 5018-5024.
- Maier, W., Peipp, H., Schmidt, J., Wray, V. and Strack, D. (1995). Levels of A Terpenoid Glycoside (Blumenin) and Cell Wall-Bound Phenolics in Some Cereal Mycorrhizas. *Plant Physiology* 109, 465-470.
- Manach, C., Scalbert, A., Morand, C., Remesy, C. and Jimenez, L. (2004). Polyphenols: food sources and bioavailability. *American Journal of Clinical Nutrition* 79, 727-747.
- Manetti, C., Bianchetti, C., Casciani, L., Castro, C., Di Cocco, M. E., Miccheli, A., Motto, M. and Conti, F. (2006). A metabonomic study of transgenic maize (*Zea mays*) seeds revealed variations in osmolytes and branched amino acids. *Journal of Experimental Botany* 57, 2613-2625.
- Manolache, F. A., Hanganu, A., Duta, D. E., Belc, N. and Marin, D. I. (2013). The Physico-chemical and Spectroscopic Composition Characterization of Oat Grains and Oat Oil Samples. *Revista de Chimie* 64, 45-48.
- Marhuenda-Egea, F. C., Gonsalvez-Alvarez, R. D., Lledo-Bosch, B., Ten, J. and Bernabeu, R. (2013). New Approach for Chemometric Analysis of Mass Spectrometry Data. *Anal. Chem.* 85, 3053-3058.
- Marti, G., Erb, M., Boccard, J., Glauser, G., Doyen, G. R., Villard, N., Robert, C. A. M., Turlings, T. C. J., Rudaz, S. and Wolfender, J. L. (2013). Metabolomics reveals herbivore-induced metabolites of resistance and susceptibility in maize leaves and roots. *Plant Cell and Environment* 36, 621-639.
- Matsuda, F., Okazaki, Y., Oikawa, A., Kusano, M., Nakabayashi, R., Kikuchi, J., Yonemaru, J. I., Ebana, K., Yano, M. and Saito, K. (2012). Dissection of genotype-phenotype associations in rice grains using metabolome quantitative trait loci analysis. *Plant Journal* 70, 624-636.
- Mccann, M. C., Hammouri, M., Wilson, R., Belton, P. and Roberts, K. (1992). Fourier-Transform Infrared Microspectroscopy Is A New Way to Look at Plant-Cell Walls. *Plant Physiology* 100, 1940-1947.

- McConnell, W. B. (1959). Studies on Wheat Plants Using Carbon-14 Labelled Compounds .10. the Incorporation of Glutamic Acid-1-C-14. *Canadian Journal of Biochemistry and Physiology* 37, 933.
- McConnell, W. B., Mitra, A. K. and Perlin, A. S. (1958). Studies on Wheat Plants Using C-14 Compounds .8. Formation of Amylose and Amylopectin in the Wheat Kernel. *Canadian Journal of Biochemistry and Physiology* 36, 985-991.
- Mejia, C. D., Gonzalez, D. C., Mauer, L. J., Campanella, O. H. and Hamaker, B. R. (2012). Increasing and Stabilizing beta-Sheet Structure of Maize Zein Causes Improvement in Its Rheological Properties. *J. Agric. Food Chem.* 60, 2316-2321.
- Mikkelsen, M. S., Jespersen, B. M., Larsen, F. H., Blennow, A. and Engelsen, S. B. (2013). Molecular structure of large-scale extracted beta-glucan from barley and oat: Identification of a significantly changed block structure in a high beta-glucan barley mutant. *Food Chemistry* 136, 130-138.
- Mischak, H., Coon, J. J., Novak, J., Weissinger, E. M., Schanstra, J. P. and Dominiczak, A. F. (2009). Capillary Electrophoresis-Mass Spectrometry As A Powerful Tool in Biomarker Discovery and Clinical Diagnosis: An Update of Recent Developments. *Mass Spectrometry Reviews* 28, 703-724.
- Munck, L., Jespersen, B. M., Rinnan, A., Seefeldt, H. F., Engelsen, M. M., Nørgaard, L. and Engelsen, S. B. (2010). A physiochemical theory on the applicability of soft mathematical models-experimentally interpreted. *J. Chemometrics* 24, 481-495.
- Munck, L., Moller, B., Jacobsen, S. and Søndergaard, I. (2004). Near infrared spectra indicate specific mutant endosperm genes and reveal a new mechanism for substituting starch with (1 -> 3,1 -> 4)-beta-glucan in barley. *Journal of Cereal Science* 40, 213-222.
- Munck, L., Nielsen, J. P., Moller, B., Jacobsen, S., Søndergaard, I., Engelsen, S. B., Nørgaard, L. and Bro, R. (2001). Exploring the phenotypic expression of a regulatory proteome-altering gene by spectroscopy and chemometrics. *Analytica Chimica Acta* 446, 171-186.
- Nakagawa, H., Sakamoto, S., Sago, Y. and Nagashima, H. (2013). Detection of Type A Trichothecene Di-Glucosides Produced in Corn by High-Resolution Liquid Chromatography-Orbitrap Mass Spectrometry. *Toxins* 5, 590-604.
- Nicholson, J. K., Lindon, J. C. and Holmes, E. (1999). 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* 29, 1181-1189.
- Nielsen, N. P. V., Carstensen, J. M. and Smedsgaard, J. (1998). Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A* 805, 17-35.
- Norbaek, R., Brandt, K. and Kondo, T. (2000). Identification of flavone C-glycosides including a new flavonoid chromophore from barley leaves (*Hordeum vulgare* L.) by improved NMR techniques. *J. Agric. Food Chem.* 48, 1703-1707.
- Nørgaard, L., Bro, R., Westad, F. and Engelsen, S. B. (2006). A modification of canonical variates analysis to handle highly collinear multivariate data. *J. Chemometrics* 20, 425-435.

- Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J. P., Munck, L. and Engelsen, S. B. (2000). Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. *Applied Spectroscopy* 54, 413-419.
- Nurmi, T., Lampi, A. M., Nystrøm, L., Hemery, Y., Rouau, X. and Piironen, V. (2012). Distribution and composition of phytosterols and steryl ferulates in wheat grain and bran fractions. *Journal of Cereal Science* 56, 379-388.
- Nystrøm, L., Lampi, A. M., Andersson, A. A. M., Kamal-Eldin, A., Gebruers, K., Courtin, C. M., Delcour, J. A., Li, L., Ward, J. L., Fras, A. et al. (2008). Phytochemicals and Dietary Fiber Components in Rye Varieties in the HEALTHGRAIN Diversity Screen. *J. Agric. Food Chem.* 56, 9758-9766.
- Olson, D. L., Norcross, J. A., O'Neil-Johnson, M., Molitor, P. F., Detlefsen, D. J., Wilson, A. G. and Peck, T. L. (2004). Microflow NMR:GC Concepts and Capabilities. *Anal. Chem.* 76, 2966-2974.
- Onning, G., Asp, N. G. and Sivik, B. (1993). Saponin Content in Different Oat Varieties and in Different Fractions of Oat Grain. *Food Chemistry* 48, 251-254.
- Osborn, A. E. (2003). Saponins in cereals. *Phytochemistry* 62, 1-4.
- Paschoal, J., Barboza, F. D. and Poppi, R. J. (2003). Analysis of contaminants in lubricant oil by near infrared spectroscopy and interval partial least-squares. *Journal of Near Infrared Spectroscopy* 11, 211-218.
- Pasikanti, K. K., Ho, P. and Chan, E. (2008). Gas chromatography/mass spectrometry in metabolic profiling of biological fluids. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences* 871, 202-211.
- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 2, 559-572.
- Pedersen, D. K., Martens, H., Nielsen, J. P. and Engelsen, S. B. (2002). Near-infrared absorption and scattering separated by extended inverted signal correction (EISC): Analysis of near-infrared transmittance spectra of single wheat seeds. *Applied Spectroscopy* 56, 1206-1214.
- Pere-Trepat, E., Lacorte, S. and Tauler, R. (2005). Solving liquid chromatography mass spectrometry coelution problems in the analysis of environmental samples by multivariate curve resolution. *Journal of Chromatography A* 1096, 111-122.
- Perkowski, J., Stuper, K., Busko, M., Goral, T., Kaczmarek, A. and Jelen, H. (2012). Differences in metabolomic profiles of the naturally contaminated grain of barley, oats and rye. *Journal of Cereal Science* 56, 544-551.
- Peterson, D. M. (2001). Oat antioxidants. *Journal of Cereal Science* 33, 115-129.
- Piironen, V., Toivo, J. and Lampi, A. M. (2002). Plant sterols in cereals and cereal products. *Cereal Chemistry* 79, 148-154.
- Ponte, J. G., Destefan, V. A. and Titcomb, S. T. (1969). Application of Thin-Layer Chromatography to Sugar Analysis in Cereal-Based Products. *Cereal Science Today* 14, 101.

- Qin, B., Eagles, J., Mellon, F. A., Mylona, P., Pena-Rodriguez, L. and Osbourn, A. E. (2010). High throughput screening of mutants of oat that are defective in triterpene synthesis. *Phytochemistry* 71, 1245.
- Qiu, Y., Liu, Q. and Beta, T. (2010). Antioxidant properties of commercial wild rice and analysis of soluble and insoluble phenolic acids. *Food Chemistry* 121, 140-147.
- Quinde, Z., Ullrich, S. E. and Baik, B. K. (2004). Genotypic variation in color and discoloration potential of barley-based food products. *Cereal Chemistry* 81, 752-758.
- Ramautar, R., Somsen, G. W. and de Jong, G. J. (2009). CE-MS in metabolomics. *Electrophoresis* 30, 276-291.
- Ren, Q. and Tian, Y. L. (2012). Studies of aroma active components in naked oat by GC-MS. *Journal of Food Agriculture & Environment* 10, 67-71.
- Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C., Lisec, J., Technow, F., Sulpice, R., Altmann, T., Stitt, M., Willmitzer, L. and Melchinger, A. E. (2012a). Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nature Genetics* 44, 217-220.
- Riedelsheimer, C., Lisec, J., Czedik-Eysenberg, A., Sulpice, R., Flis, A., Grieder, C., Altmann, T., Stitt, M., Willmitzer, L. and Melchinger, A. E. (2012b). Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. *Proceedings of the National Academy of Sciences of the United States of America* 109, 8872-8877.
- Robert, P., Marquis, M., Barron, C., Guillon, F. and Saulnier, L. (2005). FT-IR investigation of cell wall polysaccharides from cereal grains. Arabinoxylan infrared assignment. *J. Agric. Food Chem.* 53, 7014-7018.
- Rodriguez-Cuesta, M. J., Boqué, R., Rius, F. X., Martinez Vidal, J. L. and Garrido Frenich, A. (2005). Development and validation of a method for determining pesticides in groundwater from complex overlapped HPLC signals and multivariate curve resolution. *Chemometrics and Intelligent Laboratory Systems* 77, 251-260.
- Roessner, U., Wagner, C., Kopka, J., Trethewey, R. N. and Willmitzer, L. (2000). Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant Journal* 23, 131-142.
- Rolletschek, H., Weschke, W., Weber, H., Wobus, U. and Borisjuk, L. (2004). Energy state and its control on seed development: starch accumulation is associated with high ATP and steep oxygen gradients within barley grains. *Journal of Experimental Botany* 55, 1351-1359.
- Rooke, H. S., Lampitt, L. H. and Jackson, E. M. (1949). The Phosphorus Compounds of Wheat Starch. *Biochemical Journal* 45, 231-236.
- Rouau, X., Cheynier, V., Surget, A., Gloux, D., Barron, C., Meudec, E., Louis-Montero, J. and Criton, M. (2003). A dehydrotrimer of ferulic acid from maize bran. *Phytochemistry* 63, 899-903.
- Rudi, H., Uhlen, A. K., Harstad, O. M. and Munck, L. (2006). Genetic variability in cereal carbohydrate compositions and potentials for improving nutritional value. *Animal Feed Science and Technology* 130, 55-65.

- Salau, J. S., Honing, M., Tauler, R. and Barcelo, D. (1998). Resolution and quantitative determination of coeluted pesticide mixtures in liquid chromatography thermospray mass spectrometry by multivariate curve resolution. *Journal of Chromatography A* 795, 3-12.
- Salgo, A. and Gergely, S. (2012). Analysis of wheat grain development using NIR spectroscopy. *Journal of Cereal Science* 56, 31-38.
- Samadi-Maybodi, A. and Atashbozorg, E. (2006). Quantitative and qualitative studies of silica in different rice samples grown in north of Iran using UV-vis, XRD and IR spectroscopy techniques. *Talanta* 70, 756-760.
- Sampson, D. A., Wen, Q. B. and Lorenz, K. (1996). Vitamin B6 and pyridoxine glucoside content of wheat and wheat flours. *Cereal Chemistry* 73, 770-774.
- Santos, F. J. and Galceran, M. T. (2003). Modern developments in gas chromatography-mass spectrometry-based environmental analysis. *Journal of Chromatography A* 1000, 125-151.
- Sato, S., Arita, M., Soga, T., Nishioka, T. and Tomita, M. (2008). Time-resolved metabolomics reveals metabolic modulation in rice foliage. *BMC Systems Biology* 2.
- Savorani, F., Rasmussen, M. A., Mikkelsen, M. S. and Engelsen, S. B. (2013a). A primer to nutritional metabolomics by NMR spectroscopy and chemometrics. *Food Research International*, In press, DOI: 10.1016/j.foodres.2012.12.025.
- Savorani, F., Rasmussen, M. A., Rinna, Å. and Engelsen, S. B. (2013b). Interval based chemometric methods in NMR-Foodomics. In *Chemometrics in Food Chemistry* (ed. F. Marini), Elsevier, UK, p. 449.
- Savorani, F., Tomasi, G. and Engelsen, S. B. (2010). icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal of Magnetic Resonance* 202, 190-202.
- Seefeldt, H. F., Larsen, F. H., Viereck, N., Petersen, M. A. and Engelsen, S. B. (2011). Lipid composition and deposition during grain filling in intact barley (*Hordeum vulgare*) mutant grains as studied by H-1 HR MAS NMR. *Journal of Cereal Science* 54, 442-449.
- Seefeldt, H. F., Blennow, A., Jespersen, B. M., Wollenweber, B. and Engelsen, S. B. (2009). Accumulation of mixed linkage (1 → 3) (1 → 4)-beta-D-glucan during grain filling in barley: A vibrational spectroscopy study. *Journal of Cereal Science* 49, 24-31.
- Seefeldt, H. F., Larsen, F. H., Viereck, N., Wollenweber, B. and Engelsen, S. B. (2008). Bulk carbohydrate grain filling of barley beta-glucan mutants studied by H-1 HR MAS NMR. *Cereal Chemistry* 85, 571.
- Shalygo, N. V., Mock, H. P., Averina, N. G. and Grimm, B. (1998). Photodynamic action of uroporphyrin and protochlorophyllide in greening barley leaves treated with cesium chloride. *Journal of Photochemistry and Photobiology B-Biology* 42, 151-158.
- Shewry, P. R., Piironen, V., Lampi, A. M., Edelmann, M., Kariluoto, S., Nurmi, T., Fernandez-Orozco, R., Ravel, C., Charmet, G., Andersson, A. A. M. et al. (2010). The HEALTHGRAIN Wheat Diversity Screen: Effects of Genotype and Environment on Phytochemicals and Dietary Fiber Components. *J. Agric. Food Chem.* 58, 9291-9298.



- Shewry, P. R., Piironen, V., Lampi, A. M., Nyström, L., Li, L., Rakszegi, M., Fras, A., Boros, D., Gebruers, K., Courtin, C. M. et al. (2008). Phytochemical and Fiber Components in Oat Varieties in the HEALTHGRAIN Diversity Screen. *J. Agric. Food Chem.* 56, 9777-9784.
- Shrestha, S., Brueck, H. and Asch, F. (2012). Chlorophyll index, photochemical reflectance index and chlorophyll fluorescence measurements of rice leaves supplied with different N levels. *Journal of Photochemistry and Photobiology B-Biology* 113, 7-13.
- Shu, X. L., Frank, T., Shu, Q. Y. and Engel, K. R. (2008). Metabolite Profiling of Germinating Rice Seeds. *J. Agric. Food Chem.* 56, 11612-11620.
- Siuda, R., Balcerowska, G. and Sadowski, C. (2006). Comparison of the usability of different spectral ranges within the near ultraviolet, visible and near infrared ranges (UV-VIS-NIR) region for the determination of the content of scab-damaged component in blended samples of ground wheat. *Food Additives and Contaminants* 23, 1201-1207.
- Skogerson, K., Harrigan, G. G., Reynolds, T. L., Halls, S. C., Ruebelt, M., Iandolo, A., Pandravada, A., Glenn, K. C. and Fiehn, O. (2010). Impact of Genetics and Environment on the Metabolite Composition of Maize Grain. *J. Agric. Food Chem.* 58, 3600-3610.
- Skrbic, B., Koprivica, S. and Godula, M. (2013). Validation of a method for determination of mycotoxins subjected to the EU regulations in spices: The UHPLC-HESI-MS/MS analysis of the crude extracts. *Food Control* 31, 461-466.
- Smilde, A. K., Jansen, J. J., Hoefsloot, H. C. J., Lamers, R. J. A. N., van der Greef, J. and Timmerman, M. E. (2005). ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics* 21, 3043-3048.
- Sohn, M., Himmelsbach, D. S. and Barton, F. E. (2004). A comparative study of Fourier transform Raman and NIR spectroscopic methods for assessment of protein and apparent amylose in rice. *Cereal Chemistry* 81, 429-433.
- Solfrizzo, M., De Girolamo, A., Lattanzio, V. M. T., Visconti, A., Stroka, J., Alldrick, A. and van Egmond, H. P. (2013). Results of a proficiency test for multi-mycotoxin determination in maize by using methods based on LC-MS/(MS). *Quality Assurance and Safety of Crops & Foods* 5, 15-48.
- Sorensen, H. P., Madsen, L. S., Petersen, J., Andersen, J. T., Hansen, A. M. and Beck, H. C. (2010). Oat (*Avena sativa*) Seed Extract as an Antifungal Food Preservative Through the Catalytic Activity of a Highly Abundant Class I Chitinase. *Applied Biochemistry and Biotechnology* 160, 1573-1584.
- Sridhar, R. and Ou, S. H. (1974). Phenolic Compounds Detected in Rice Blast Lesions. *Biologia Plantarum* 16, 67-70.
- Stahle, L. and Wold, S. (1987). Partial Least Squares Analysis with Cross-Validation for the Two-Class Problem A Monte Carlo Study. *J. Chemometrics* 1, 185-196.
- Stein, S. E. (1999). An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *Journal of the American Society for Mass Spectrometry* 10, 770-781.

- Symons, S. J. and Dexter, J. E. (1996). Aleurone and pericarp fluorescence as estimators of mill stream refinement for various Canadian wheat classes. *Journal of Cereal Science* 23, 73-83.
- Szymanska, E., Saccenti, E., Smilde, A. K. and Westerhuis, J. A. (2012). Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics* 8, S3-S16.
- Takahashi, H., Hayashi, M., Goto, F., Sato, S., Soga, T., Nishioka, T., Tomita, M., Kawai-Yamada, M. and Uchimiya, H. (2006). Evaluation of metabolic alteration in transgenic rice overexpressing dihydroflavonol-4-reductase. *Annals of Botany* 98, 819-825.
- Tang, Y. Y., Lin, H. Y., Chen, Y. C., Su, W. T., Wang, S. C., Chiueh, L. C. and Shin, Y. C. (2013). Development of a Quantitative Multi-Mycotoxin Method in Rice, Maize, Wheat and Peanut Using UPLC-MS/MS. *Food Analytical Methods* 6, 727-736.
- Tangkhavanich, B., Kobayashi, T. and Adachi, S. (2012). Properties of Rice Straw Extract after Subcritical Water Treatment. *Bioscience Biotechnology and Biochemistry* 76, 1146-1149.
- Tarpley, L., Duran, A. L., Kebrom, T. H. and Sumner, L. W. (2005). Biomarker metabolites capturing the metabolite variance present in a rice plant developmental period. *Bmc Plant Biology* 5.
- Teller, G. L. (1935). Changes in nitrogen compounds in the wheat grain at different stages of development. *Plant Physiology* 10, 499-509.
- Thondre, P. S., Ryan, L. and Henry, C. J. K. (2011). Barley beta-glucan extracts as rich sources of polyphenols and antioxidants. *Food Chemistry* 126, 72-77.
- Tomasi, G., van den Berg, F. and Andersson, C. (2004). Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *J. Chemometrics* 18, 231-241.
- Tomasi, G., Savorani, F. and Engelsen, S. B. (2011). icoshift: An effective tool for the alignment of chromatographic data. *Journal of Chromatography A* 1218, 7832-7840.
- Tonning, E., Norgaard, L., Engelsen, S. B., Pedersen, L. and Esbensen, K. H. (2006). Protein heterogeneity in wheat lots using single-seed NIT - A Theory of Sampling (TOS) breakdown of all sampling and analytical errors. *Chemometrics and Intelligent Laboratory Systems* 84, 142-152.
- Toubiana, D., Fernie, A. R., Nikoloski, Z. and Fait, A. (2013). Network analysis: tackling complex data to study plant metabolism. *Trends in Biotechnology* 31, 29-36.
- Trygg, J., Holmes, E. and Lundstedt, T. (2007). Chemometrics in metabonomics. *Journal of Proteome Research* 6, 469-479.
- Tsai, S. D. and Tood, G. W. (1972). Phenolic Compounds of Wheat Leaves Under Drought Stress. *Phyton-International Journal of Experimental Botany* 30, 67-75.
- van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K. and van der Werf, M. J. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. *Bmc Genomics* 7.
- Vandekamer, J. H. and Vanginkel, L. (1952). Rapid Determination of Crude Fiber in Cereals. *Cereal Chemistry* 29, 239-251.

- Vinson, J. A., Erk, K. M., Wang, S. Y., Marchegiani, J. Z. and Rose, M. F. (2009). Total polyphenol antioxidants in whole grain cereals and snacks: Surprising sources of antioxidants in the US diet. *Abstracts of Papers of the American Chemical Society* 238, 246.
- Walker, V., Bertrand, C., Bellvert, F., Moenne-Loccoz, Y., Bally, R. and Comte, G. (2011). Host plant secondary metabolite profiling shows a complex, strain-dependent response of maize to plant growth-promoting rhizobacteria of the genus *Azospirillum*. *New Phytologist* 189, 494-506.
- Ward, J. L., Poutanen, K., Gebruers, K., Piironen, V., Lampi, A. M., Nyström, L., Andersson, A. A. M., Aman, P., Boros, D., Rakszegi, M. et al. (2008). The HEALTHGRAIN Cereal Diversity Screen: Concept, Results, and Prospects. *J. Agric. Food Chem.* 56, 9699-9709.
- Warth, B., Parich, A., Atehnkeng, J., Bandyopadhyay, R., Schuhmacher, R., Sulyok, M. and Krska, R. (2012). Quantitation of Mycotoxins in Food and Feed from Burkina Faso and Mozambique Using a Modern LC-MS/MS Multitoxin Method. *J. Agric. Food Chem.* 60, 9352-9363.
- Watson, J. T. and Sparkman, O. D. (2007a). Chemical Ionization. In *Introduction to Mass Spectrometry*, pp. 449-484: John Wiley & Sons, Ltd.
- Watson, J. T. and Sparkman, O. D. (2007b). Electron Ionization. In *Introduction to Mass Spectrometry*, John Wiley & Sons, p. 315-448.
- Watson, J. T. and Sparkman, O. D. (2007c). Gas Chromatography/Mass Spectrometry. In *Introduction to Mass Spectrometry*, John Wiley & Sons, p. 571-638.
- Westerhuis, J. A., Hoefsloot, H. C. J., Smit, S., Vis, D. J., Smilde, A. K., van Velzen, E. J. J., van Duijnhoven, J. P. M. and van Dorsten, F. A. (2008). Assessment of PLSDA cross validation. *Metabolomics* 4, 81-89.
- Widodo, Patterson, J. H., Newbigin, E., Tester, M., Bacic, A. and Roessner, U. (2009). Metabolic responses to salt stress of barley (*Hordeum vulgare* L.) cultivars, Sahara and Clipper, which differ in salinity tolerance. *Journal of Experimental Botany* 60, 4089-4103.
- Williams, P. and Norris, K. (1988). Near-Infrared Technology in the Agriculture and Food Industries. 330 Seiten, zahlr. Abb. und Tab. American Association of Cereal Chemists, Inc., St. Paul, Minnesota, USA. *Nahrung* 32.
- Williams, P. J., Geladi, P., Britz, T. J. and Manley, M. (2012). Investigation of fungal development in maize kernels using NIR hyperspectral imaging and multivariate data analysis. *Journal of Cereal Science* 55, 272-278.
- Winning, H., Larsen, F. H., Bro, R. and Engelsen, S. B. (2008). Quantitative analysis of NMR spectra with chemometrics. *Journal of Magnetic Resonance* 190, 26-32.
- Wiser, W. J. and Jones, G. E. (1971). Rapid Predigestion Technique for Automatic Analysis of Protein in Rice and Other Cereal Grains. *Cereal Science Today* 16, 305.
- Withycom, D. A., Stuiber, D. A. and Lindsay, R. C. (1974). Isolation and Identification of Volatile Compounds from Wild Rice (*Zizania-Aquatica*). *Abstracts of Papers of the American Chemical Society*, 36.

- Wojakowska, A., Perkowski, J., Goral, T. and Stobiecki, M. (2013). Structural characterization of flavonoid glycosides from leaves of wheat (*Triticum aestivum* L.) using LC/MS/MS profiling of the target compounds. *Journal of Mass Spectrometry* 48, 329-339.
- Wold, H. (1979). Model Construction and Evaluation when Theoretical Knowledge is Scarce: An Example of the Use of Partial Least Squares. Université de Genève, Faculté des Sciences Économiques et Sociales, Geneva.
- Wold, H. (1975). Path models with latent variables: the NIPALS approach. *Quantitative Sociology: International perspectives on mathematical and statistical modeling*. Academic Press, New York, p. 307-357.
- Wold, S. (1976). Pattern-Recognition by Means of Disjoint Principal Components Models. *Pattern Recognition* 8, 127-139.
- Wold, S., Martens, H. and Wold, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method. In *Matrix Pencils* (eds. B. Kågström and A. Ruhe), Springer, Berlin, Heidelberg, pp. 286.
- Wold, S., Sjostrom, M. and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* 58, 109-130.
- Wold, S. and Sjöström, M. (1977). SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy. In *Chemometrics: Theory and Application*, American Chemical Society, pp. 243-282.
- Wu, D. Z., Cai, S. G., Chen, M. X., Ye, L. Z., Chen, Z. H., Zhang, H. T., Dai, F., Wu, F. B. and Zhang, G. P. (2013). Tissue Metabolic Responses to Salt Stress in Wild and Cultivated Barley. *Plos One* 8.
- Wu, N., Liu, J. A., Zhou, G. Y., Yan, R. K. and Zhang, L. (2012). Prediction of Chlorophyll Content of Leaves of Oil Camelliae after Being Infected with Anthracnose Based on Vis/NIR Spectroscopy. *Spectroscopy and Spectral Analysis* 32, 1221-1224.
- Xu, F. G., Zou, L. and Ong, C. N. (2010). Experiment-originated variations, and multi-peak and multi-origination phenomena in derivatization-based GC-MS metabolomics. *Trac-Trends in Analytical Chemistry* 29, 269-280.
- Xu, R. N. X., Fan, L. M., Rieser, M. J. and El-Shourbagy, T. A. (2007). Recent advances in high-throughput quantitative bioanalysis by LC-MS/MS. *Journal of Pharmaceutical and Biomedical Analysis* 44, 342.
- Xu, Z. F., Sun, X. B. and Harrington, P. D. (2011). Baseline Correction Method Using an Orthogonal Basis for Gas Chromatography/Mass Spectrometry Data. *Anal. Chem.* 83, 7464-7471.
- Yamada, T. and Bork, P. (2009). Evolution of biomolecular networks - lessons from metabolic and protein interactions. *Nature Reviews Molecular Cell Biology* 10, 791-803.
- Zandomenighi, M., Carbonaro, L., Calucci, L., Pinzino, C., Galleschi, L. and Ghiringhelli, S. (2003). Direct fluorometric determination of fluorescent substances in powders: The case of riboflavin in cereal flours. *J. Agric. Food Chem.* 51, 2888-2895.
- Zekovic, I., Lenhardt, L., Dramicanin, T. and Dramicanin, M. D. (2012). Classification of Intact Cereal Flours by Front-Face Synchronous Fluorescence Spectroscopy. *Food Analytical Methods* 5, 1205-1213.

- Zhang, X., Liu, F., He, Y. and Li, X. (2012). Application of Hyperspectral Imaging and Chemometric Calibrations for Variety Discrimination of Maize Seeds. *Sensors* 12, 17234-17246.
- Zielinski, H. and Kozłowska, H. (2000). Antioxidant activity and total phenolics in selected cereal grains and their different morphological fractions. *J. Agric. Food Chem.* 48, 2008-2016.
- Zimmer, E., Gurrath, P. A., Paul, C., Dhillon, B. S., Pollmer, W. G. and Klein, D. (1990). Near infrared reflectance spectroscopy analysis of digestibility traits of maize stover. *Euphytica* 48, 73-81.
- Zweig, M. H. and Campbell, G. (1993). Receiver-Operating Characteristic (Roc) Plots - A Fundamental Evaluation Tool in Clinical Medicine. *Clinical Chemistry* 39, 561-577.

# Paper 4

**Bekzod Khakimov**, Morten Arendt Rasmussen, Birthe Møller Jespersen,  
Lars Munck, Søren Balling Engelsen

The emerging barley seed metabolome studied by mutant analysis and advanced GC-MS: evaluation of the effects of development stage, genotype and growth temperature by ASCA

*Journal of Experimental Botany, Submitted*



**Title:** The emerging barley seed metabolome studied by mutant analysis and advanced GC-MS: evaluation of the effects of development stage, genotype and growth temperature by ASCA

**Authors:** Bekzod Khakimov<sup>\*</sup>, Morten Arendt Rasmussen, Birthe Møller Jespersen, Lars Munck, Søren Balling Engelsen

**Institution:** Spectroscopy & Chemometrics, Department of Food Science, Faculty of Science, University of Copenhagen, Rolighedsvej 30, 1958 Frederiksberg C, Denmark

**\*Corresponding author:** Bekzod Khakimov  
E-mail: bzo@food.ku.dk, Tel.: +45- 35 33 29 74

Morten Arendt Rasmussen: mortenr@food.ku.dk  
Birthe Møller Jespersen: bm@food.ku.dk  
Lars Munck: lmu@food.ku.dk  
Søren Balling Engelsen: se@food.ku.dk

**Submission date:** 28 August 2013

**Number of tables:** 2

**Number of color figures:** 7

**Number of words:** 6900

**Number of supplementary figures:** 6

**Number of supplementary tables:** 1

**Short running title:** The development of barley seed metabolome



## Abstract

The immensely complex plant metabolome is dynamically emerging through the developmental stages during epigenesis of the seed. This study demonstrates a gene specific metabolomic analysis of barley endosperm seed model including two mutant genotypes, low-starch-high- $\beta$ -glucan (*lys5.f*), and high lysine (*lys3.a*) mutants isogenic to the mother line (*Bomi*). The three barley genotypes were grown at 15 and 25 °C and analyzed in duplicates by GC-MS metabolomic profiling at eight developmental stages during the grain filling period. The study facilitated the detection of 247 metabolites that mainly included phenolic acids, aldehydes, esters, organic acids, alcohols and fatty acids. Metabolic changes related to the design parameters: development stage, barley genotype and growth temperature were separated by ANOVA-simultaneous component analysis (ASCA). The study revealed three dominating metabolomic patterns during the seed development, common throughout the all genotypes. In addition, some organic acids exhibited genotype specific dynamics and increased in one genotype and decreased or remained status quo in the two other genotypes. The study further revealed the presence of “signature” metabolites for the three barley genotypes and effects of growth temperature on metabolome. The high lysine mutant contained higher amount of most phenolic acids, whereas malic, citric, gallic, p-coumaric acids were more abundant in the mother line. The most affected metabolites with respect to growth temperature were 4-hydroxyphenylethanol (more abundant at high temperature), p-coumaric and mandelic acids (more abundant at low temperature). Global correlation tables between the metabolites at different developmental stages revealed a significant deregulation of the metabolome for the two mutants compared to *Bomi* and a significant deregulations due to high temperature that were more pronounced in the mutants.

**Keywords:** ANOVA-simultaneous component analysis (ASCA); barley grain filling; gas chromatography-mass spectrometry (GC-MS); metabolomics; PARAllel FACtor Analysis2 (PARAFAC2); pleiotropy

## Introduction

Current trends of cereal sciences are mainly focused on development and improvement of cereal cultivars by increasing their health promoting properties, yield, resistance to various abiotic stresses, including temperature (Frederiks *et al.* 2012; Soltesz *et al.* 2013), salt (Widodo *et al.* 2009), drought (Manavalan *et al.* 2012; Winning *et al.* 2009) and various abiotic stresses (Balmer *et al.* 2013). Metabolomics has become one of the well-established and powerful approaches in cereal sciences enabling the understanding of biochemical and genetic backgrounds of plants' quality traits (Fernie *et al.* 2009; Balmer *et al.* 2013; Bino *et al.* 2004).

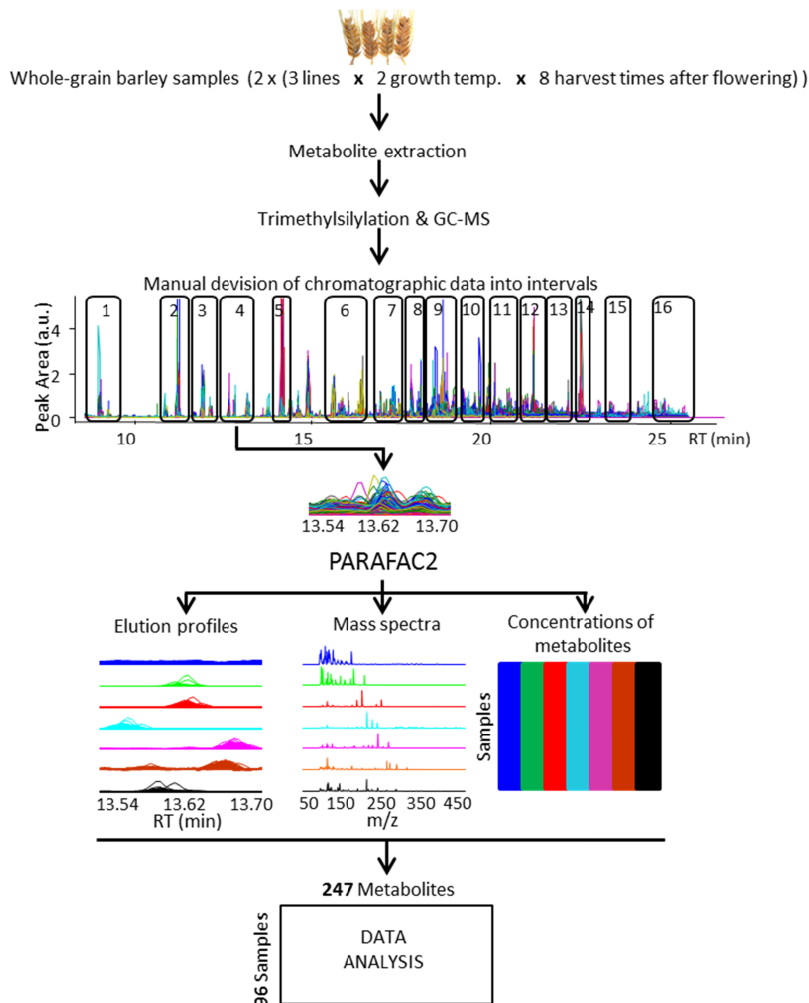
The health beneficial effects associated with the consumption of barley have been attributed to high content of the dietary fiber,  $\beta$ -glucan (Wood, 2007; Mcintosh *et al.* 1991; AbuMweis *et al.* 2010), antioxidants, radical scavenging and antiproliferative phytochemicals (Madhujith *et al.* 2007; Amarowicz *et al.* 2007). In addition to the health beneficial value, polyphenols of barley are the main fraction of micronutrients that are highly related to the preventive properties against various biotic and abiotic stresses (Amarowicz *et al.* 2007; Zielinski *et al.* 2000; Madhujith *et al.* 2007; Taketa *et al.* 2010). Possible bioactive properties of polyphenols that reduce the risk of cancer, cardiovascular diseases, improve immune system and general well-being, have prompted a great number of studies on polyphenol composition of food and food raw materials and their effects on human health (Korkina *et al.* 2013; Quinde-Axtell *et al.* 2006; Manach *et al.* 2004). Moreover, phenolic acids were found to be important texturizing agents in cooking-extrusion of cereals (Gibson *et al.* 1991) and recognized as the main antioxidant constituents of cereals (Vinson *et al.* 2009). The phytochemical composition of barley and other cereals have been studied in a number of projects within the HEALTHGRAIN diversity-screening program (Shewry *et al.* 2008; Nystrom *et al.* 2008; Li *et al.* 2008; Ward *et al.* 2008). Ten barley genotypes from different geographical locations have been screened for dietary fibers and phytochemicals, including ten phenolic acids, sterols and folates (Andersson *et al.* 2008). The study showed a significant influence of the barley genotypes in the content of all phytochemicals.

This study investigates the expression of two specific genes on the metabolomics level by monitoring mutants and comparing them to their near isogenic mother line in a barley seed (endosperm) mutant model. The study involved, two genetically modified barley genotypes, low starch – high- $\beta$ -glucan mutant, *lys5.f* (Munck *et al.* 2004) and high lysine mutant, *lys3.a* (Munck *et al.* 2001) and their mother genotype *Bomi* (normal wild type) that were grown at two different temperatures and harvested at eight different time points after flowering. The barley isogenic mutant model with two major gene mutants employed in this study is especially favorable to show the potential of the new analysis tools because these major mutants generates significant secondary (pleiotropic) changes in the metabolome.

In this study GC-MS metabolomic profiling was mainly focused on polyphenols and organic acids of barley seeds. Polyphenols of barley whole-grains primarily present in conjugated forms with carbohydrates, lipids and other cell membrane components which alter their solubility and thus bioavailability (Andersson *et al.* 2008). A protocol for extraction of conjugated phenolic acids from whole-grain wheat flour samples was presented by Li *et al.* (Li *et al.* 2008). They obtained phenolic extracts using 80% ethanol followed by alkaline hydrolysis using 2 M sodium hydroxide solution to cleave ester bonds through which phenolics are bonded. Many other studies also applied alkaline hydrolysis using bases to enhance phenolic acid extraction (Max *et al.* 2010). It is worth to mention that the basic hydrolysis can only cleave the ester bonds and stabilize de-esterification reactions. However, phenolics and other organic acids of cereals are bonded not

only via ester bonds, but also through glycosidic bonds to the carbohydrates. In contrast to basic hydrolysis, acidic hydrolysis cleaves both, ester and glycosidic bonds and in aqueous media, the reaction is favorable to de-esterification. Advantages of this approach have been shown in polyphenol analysis of the wheat and rice grains (Arranz *et al.* 2010; Sani *et al.* 2012). Therefore, in this study, a simple and robust protocol was developed which comprised hydrochloric acid based hydrolysis of complete dried methanol extracts of the barley flour samples. This study represent the first application of a novel derivatization method developed for unbiased GC-MS analysis of complex biological samples (Khakimov *et al.* 2013). Prior to the GC-MS analysis barley seed extracts were derivatized using a novel trimethylsilylation method that exhibit several advantages over existing methods in terms of reaction speed, yield and reproducibility.

This study aims at demonstrating the tools of cutting-edge chemometric methodologies where a multi-way chemometric method PARAllel FACTor Analysis2 (PARAFAC2) (Bro *et al.* 1999) is applied for processing the complex metabolomics data obtained from the new GC-MS protocol. The effects of development stage, growth temperature and genotype was analyzed by ANOVA - Simultaneous Component Analysis (ASCA) (Smilde *et al.* 2005) followed by PCA (Hotelling, 1933), PLS (Wold, 1979) and PLS-DA (Stahle *et al.* 1987). The obtained raw GC-MS data were processed by semi-automated multi-way decomposition method,PARAFAC2. The PARAFAC2 approach leads to improved comprehensive analysis of the three dimensional GC-MS metabolomics data when compared to the other alternative methods (Amigo *et al.* 2008; Khakimov *et al.* 2012). The PARAFAC2 processing of raw GC-MS data obtained from the barley extracts lead to the precise quantification of all the metabolites and enabled resolution of elusive peaks such as, overlapped, retention time shifted, low s/n peaks and peaks that were below the noise level. Figure 1 illustrates an overview of the GC-MS metabolomic workflow used in this study.



**Figure 1.** Overview of the barley grain metabolomics by using GC-MS metabolomic profiling and interval based semi-automated PARALLEL FACTOR Analysis 2 (PARAFAC2).

## Materials and methods

Three barley genotypes with contrasting chemical composition the low-starch-high  $\beta$ -glucan Risø mutant (Doll, 1983), *lys5.f* in chromosome 6 in *Bomi* background), the high lysine Risø mutant, *lys3.a* (*lys3* allele in chromosome 5 in *Bomi* background (Munck, 1992), and the near isogenic mother genotype *Bomi* were included in this study. The mutants were selected as “high lysine mutants” (Munck *et al.* 2010) by the dye-binding method (acilane orange) at Risø, Denmark in the 1970’s (Doll, 1983). Lysine is increased from 3.5% of protein in *Bomi* to 5.5% in *lys3.a* with minor changes in protein and starch. *lys3.a* is a regulatory gene that inhibit the synthesis of the hordein proteins low in lysine by inducing a lack of demethylation of the promoter DNA for these genes (Von Wettstein, 1992), whereas the other gene, *lys5.f* is a

structural starch gene that lacks an isoenzyme for transport of ADP-glucose into the plastid. The low starch content was compensated by increasing the content of  $\beta$ -glucan (Munck *et al.* 2004). The surprising  $\beta$ -glucan overproduction in the starch mutant *lys5.f* was found by near infrared spectroscopic analysis of seeds. In the *lys5.f* starch content was reduced from 55% in *Bomi* to 30% in the mutant while  $\beta$ -glucan increased from 5.5% to 20%. Protein and lysine content was slightly increased in the mutant compared to *Bomi*. All barley genotype were grown under the same conditions using a semifield pot experiment at two different growing temperatures, high (15°C) and low (25°C). The spikes on the main and the first side tillers were harvested at eight different time points during development stage: 9, 13, 16, 20, 23, 30, 39 and 47 days after flowering (DAF). Two biological replicates were sampled at each time point. One measurement was lost during the sample preparation (low temperature grown *Bomi* genotype at 23 DAF) leading to a total of 95 samples. More detailed protocols of plant growing conditions, whole-grain seed harvesting and chemical analysis are described by Seefeldt *et al.* (Seefeldt *et al.* 2009). It should be noticed that many of the metabolites focused on in this study are produced in the first part of the seed synthesis localized in the outer layers of the barley seed – in the awns, the pericarp, the testa and in the endosperm aleurone also including the germ.

#### *Metabolite extraction and sample derivatization*

Metabolites of the milled seeds were extracted using methanol and injected into GC-MS after trimethylsilylation. Phenolic and organic acids of all barley samples were extracted using 50 mg flour samples obtained from whole-grain seeds including awns (palea and lemma). The flour sample were soaked into 600  $\mu$ L 85% methanol and vortexed for 20 sec at 3000 rpm followed by 20 min incubation at 30°C using a Thermomixer at 1400 rpm. After 3 min of centrifugation at 16k g, the supernatant was transferred to a fresh 2 ml Eppendorf tubes and the remaining flour sample was extracted second time by using the same extraction procedure. Then combined supernatants were completely dried under nitrogen gas flow at 40°C and hydrolyzed by using 240  $\mu$ L of 6M HCl at 96°C for 1 h by stirring at 1400 rpm. The hydrolyzed extracts were transferred into fresh 2 ml glass vials and phenolics and organic acids were extracted into diethyl ether. Ether-based extraction of phenolics and organic acids was performed twice, by addition of 800  $\mu$ L diethyl ether and vortexing for 25 sec. The obtained ether fractions were completely dried using nitrogen gas flow and re-solubilized in 200  $\mu$ L 100% methanol. 90  $\mu$ L aliquots out of the final extracts were transferred into 200  $\mu$ L glass inserts and completely dried under nitrogen gas flow, sealed and stored at -20°C until GC-MS analysis (1-3 days). Each sample was spiked with an internal standard (IS) (5  $\mu$ L of 0.2 mg ml<sup>-1</sup> solution of ribitol). Prior to GC-MS the analysis samples were derivatized. In order to avoid moisture, samples stored in the freezer were dried under reduced pressure before they were tightly sealed in GC-MS vials using silicon septum magnetic lids. Sample derivatization and injection were fully automated by using a Multi-Purpose Sampler (MPS, GERSTEL, Mülheim, Germany) with DualRait WorkStation integrated to a GC-MS system from Agilent. Each sample was individually derivatized by addition of 40  $\mu$ L trimethylsilyl cyanide (TMSCN) and incubated for 40 min at 40°C. All samples were randomized and analyzed in one GC-MS sequence and the MPS autosampler allowed a sequential derivatization of all samples in the same manner by keeping the derivatization time constant, throughout the analysis.

#### *GC-MS data acquisition*

The GC-MS consisted of an Agilent 7890A GC and an Agilent 5975C series MSD. GC separation was performed on a Phenomenex ZB 5MSi column (30 m x 250  $\mu$ m x 0.25  $\mu$ m). A derivatized

sample volume of 1  $\mu\text{L}$  was injected into a cooled injection system (CIS port) using Solvent Vent mode at the vent pressure of 7 kPa until 0.3 min after injection at the vent flow of  $100\text{ ml min}^{-1}$ . Detailed information on CIS and MPS parameters are described by Khakimov *et al.* (Khakimov *et al.* 2013). Hydrogen was used as carrier gas, at a constant flow rate of  $1.2\text{ ml min}^{-1}$ , and initial temperature of CIS was set to  $120^\circ\text{C}$  until 0.3 min followed by heating at  $5^\circ\text{C s}^{-1}$  until  $320^\circ\text{C}$  and hold for 10 min. The GC oven program was as follows: initial temperature  $40^\circ\text{C}$ , equilibration time 3.0 min, heating rate  $12.0^\circ\text{C min}^{-1}$ , end temperature  $300^\circ\text{C}$ , hold time 8.0 min and post run time 5 min at  $40^\circ\text{C}$ . Mass spectra were recorded in the range of 50-500  $m/z$  with a scanning frequency of 3.2 scans  $s^{-1}$ , and the MS detector was switched off during the 8.5 min of solvent delay time and after 25.5 min of the run time. The transfer line, ion source and quadrupole temperatures were set to 290, 230 and  $150^\circ\text{C}$ , respectively. The mass spectrometer was tuned according to manufacturer's recommendation by using perfluorotributylamine (PFTBA).

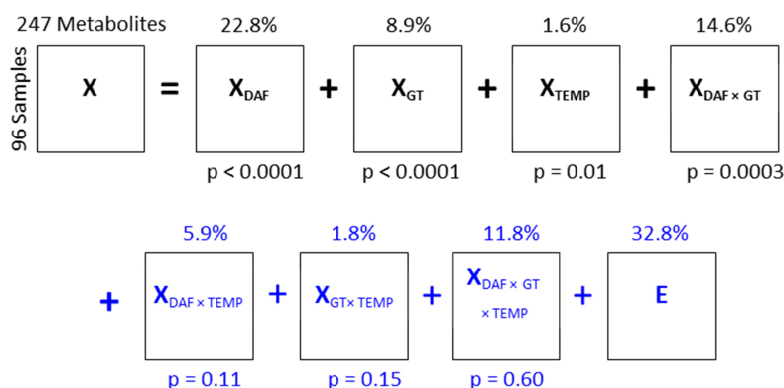
### Data analysis

The raw GC-MS chromatographic data was processed by PARAFAC2 as previously described (Khakimov *et al.* 2012). In the PARAFAC2 model the raw data is keeping its original three-way structure (elution time x mass spectra x samples) and do not require pre-processing of the data e.g. baseline correction or alignment. The only pre-processing required is to divide the data into smaller intervals for reducing complexity of the data and obtain reliable models. Three-way GC-MS data is usually divided in retention time dimension, then each interval can be modeled separately. In this study, the raw GC-MS data of 96 samples was manually divided into 121 smaller intervals where baseline was present, followed by PARAFAC2 modeling of each interval. Prior to metabolite identification the mass spectrum of each metabolite, resolved by PARAFAC2 modeling, was extracted and imported into NIST05 stand-alone software and spectra were compared against the library (NIST, USA). Moreover, metabolite identification involved Wiley08 database integrated into the data analysis software, GCMS Solutions (Shimadzu, Japan) and the in-house EI-MS databases integrated into ChemStation software (Agilent, Germany). Retention indices of detected metabolites were calculated using the Van den Dool and Kratz equation and from retention times of C10-C40 alkanes that were analyzed using the same GC-MS method (Vandendool *et al.* 1963). Finally, the obtained metabolomic matrix **X** containing relative abundances of metabolites (247 columns) in all barley samples (96 rows) was analyzed by multivariate data analysis methods including principal component analysis (PCA) (Hotelling, 1933), as well as Analysis of variance-Simultaneous Component Analysis (ASCA) (Smilde *et al.* 2005), partial least squares (PLS) regression analysis (Wold *et al.* 1983), and partial least squares-discriminant analysis PLS-DA (Stahle *et al.* 1987). All the chemometric analysis were performed in MATLAB® ver. R2012b (8.0.0.783) using the PLS-toolbox ver. 6.0.1 (Eigenvector Inc, Manson, Washington, USA.) and in-house algorithms ([www.models.life.ku.dk](http://www.models.life.ku.dk)).

### ANOVA-simultaneous component analysis (ASCA)

Anova Simultaneous Component Analysis (ASCA) is a method for extracting information from multivariate data derived from a designed experiment. ASCA analysis includes two steps; First a separation of the variance according to the design factors, i.e. main effects, two factor interactions effects etc. in a similar fashion as ANOVA for univariate data. Secondly, the individual effects are explored by e.g. PCA. In the present work, the variation of the data matrix **X** can be partitioned into days after flowering (DAF), barley genotypes (GT), growing temperature (TEMP) and their interaction effects (Figure 2). This resulted in a total of seven systematic terms (three main effects, three two factor interaction effects and a single three factor interaction effect) and a single random effect (E). In order only to interpret effects that

significantly perturbate the system, the systematic effects are tested one by one by random permutation testing (Zwanenburg *et al.* 2011). The individual significant contributions, e.g.  $X_{\text{DAF}}$ , are interpreted by PCA, where a set of principal components ( $P$ ) are estimated directly on the systematic effect matrix. In the construction of a score plot, the residual matrix is added to the effect matrix, e.g.  $X_{\text{DAF}} + E$ , and projected onto the loadings ( $T = (X_{\text{DAF}} + E)P$ ), in order to visualize the spread. In order to extract the most informative metabolites, partitioning of the data (according to Figure 2) are interpreted by PLS and PLS-DA. This is done by combining the systematic matrix of interest, e.g.  $X_{\text{GT}}$ , with the residuals ( $E$ ) and exposing this matrix to a targeted methods (PLS and PLS-DA) including cross validation.



**Figure 2.** ANOVA based overview of the overall variance distribution across 247 metabolites. <sup>5</sup> The null hypothesis ( $H_0$ ) test for effect (DAF: days after flowering, GT: three barley genotypes, TEMP; low and high growing temperatures and their second and third order interactions) of parameter in a model with all nested levels included. Assessed by permutation testing (with 10,000 random permutations). Further analysis included DAF, GT, TEMP and DAF  $\times$  GT effects separated metabolomic data matrices with  $p < 0.05$ .

## Results and Discussions

### GC-MS metabolomics

Validated PARAFAC2 models of 121 intervals of the raw GC-MS data revealed 389 components (resolved peaks). Then, each model was individually evaluated and the components that represent baseline, artifact peaks, column bleed peaks and/or shoulder of the neighbor peaks were eliminated, which resulted in 247 metabolites with unique retention indices and mass spectra (Figure 1 and Supplementary Table S1). Total ion current chromatogram of the raw GC-MS of one example of the mother genotype *Bomi* at the harvest time is illustrated in Supplementary Figure S1. 89 out of 247 metabolites were identified based on their retention indices (RI) and electron ionization-mass spectral (EI-MS) data comparison using Wiley08 and NIST05 metabolite libraries. 33 out of all identified metabolites were trimethylsilyl (TMS) derivatives of phenolic acids, their esters and aldehydes. In addition to the previously found phenolic acids from different barley genotypes (Andersson *et al.* 2008), several other phenolics such as p-salicylic, gallic, gentisic, homovanilic and  $\alpha$ -resorcylic acids and methyl esters of ferulic, caffeic, protocatechuic and sinapinic acids were identified. Small molecular organic acids, alcohols and their methyl esters constituted 27 out of 89 identified metabolites. These included

several metabolites (e.g., succinic, glyceric, maleic, fumaric, malic, pyroglutamic, azelaic acids and methyl esters of aconitic and citric acids) that are part of the same or different metabolic pathways. In addition, TMS-derivatives of seven fatty acids and their esters, four sterols and a flavonoid, catechin-nTMS were identified.

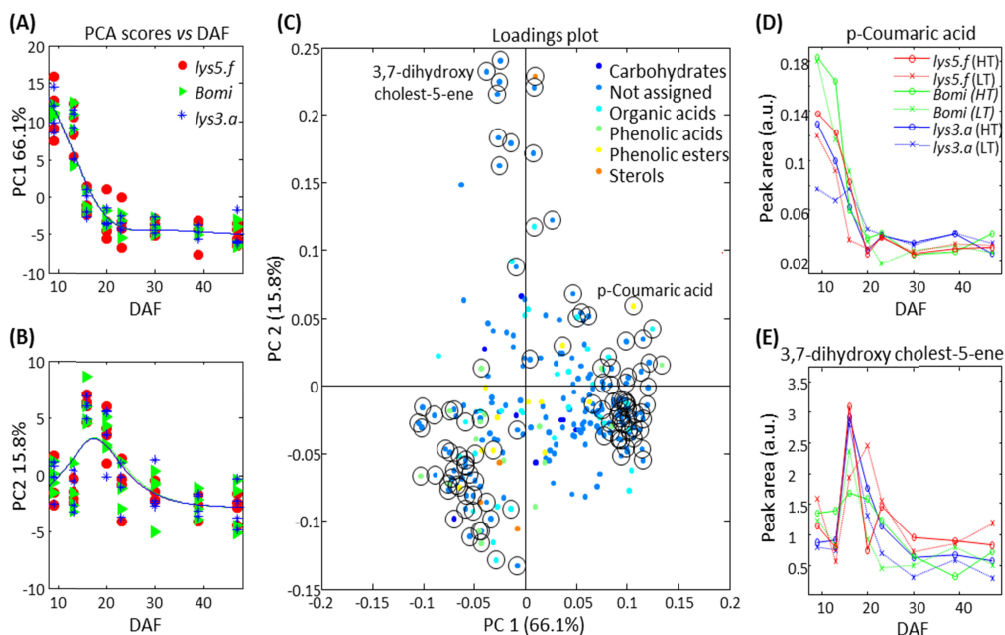
#### *The developing metabolome during grain filling: the common DAF effect*

ASCA analysis revealed that the dynamic changes of the barley metabolome (DAF effect) constituted to the major systematic variation in the data namely 22.8%, whereas metabolic variations associated with growth temperature and genotype represented 1.8% and 8.9%, respectively (Figure 2). Moreover, the barley genotype dependent DAF effect ( $X_{\text{DAF} \times \text{GT}}$ ) also represented significant variance (14.6%).

A PCA of the initial metabolomic data  $X$  reveal a significant difference between the metabolome of the barley seeds at the earlier and later development stages (Supplementary Figure S2 (A)). Principal component 1 (PC1) differentiate 9 and 13 DAF samples from the later DAF samples, while PC5 partially differentiate the 16 and 20 DAF samples from the more mature barley seeds. The loadings plot of the PCA model suggest that several metabolites are equally responsible for separating the 9 and 13 DAF samples. Thus, metabolomic changes were induced during the seed development stages and the PCA reveal groups of metabolites that were distinctly altered at the earlier DAF points. These metabolites mainly included, small molecular organic acids such as laevulic, sorbic, glyceric, maleic, fumaric acids and phenolics such as, 2,5-dimethoxymandelic acid, 4-hydroxycinnamic acid, methyl ester, gentisic acid, p-coumaric acid and methyl vanillactate. However, while PCA could model the main systematic variations of the data such as a significant change of the entire metabolome during the first two weeks after flowering, metabolic alterations at the later stages of the development were poorly described.

In order to investigate the development stage effect the ASCA separated metabolome-DAF effect,  $X_{\text{DAF}}$ , that contained 22.8% of the variation was scrutinized by PCA. Most metabolites detected in this study were expectedly influenced by the developing stage of the seeds. This illustrated common metabolomic alterations influenced by the DAF effect that are captured by PC1 and PC2 (Figure 3). Some metabolites decreased during the development stage (PC1), while others dramatically increased after two weeks of flowering and then gradually decrease after the three weeks of the anthesis time (PC2). This trend of a drastic early change in the developing metabolome was explained by a similar proteome effect during the early development stage in previous studies on barley endosperm mutants (Jacobsen *et al.* 2005). These two trends of the metabolomic development during the development stages were the major DAF effects and constituted more than 80% of the variation, which indicates that most metabolites followed these trends. The loadings plot of this PCA model show which metabolites that are mostly influenced by these two DAF effects. Metabolites with higher loadings on PC1 e.g. p-coumaric acid decreased during the development stage, while those with a negative loading on PC1 increased during the development stage. Likewise, the metabolites with higher loadings on PC2 follow the trend shown in figure 3 (B). Metabolites with black circles are the mostly influenced metabolites by the DAF effect as identified from the PLS regression modeling, which will be discussed later. Figures 3 (D) and (E) show the data of one example metabolite for the common DAF trends captured by PC1 and PC2, respectively (Figures 3 (A) and (B)).





**Figure 3.** Development of barley metabolome during the grain filling period. Barley genotype common metabolomic alterations during seed developmental period by PCA analysis of the days after flowering (DAF) effect separated metabolomics data matrix ( $X_{\text{DAF}}$ ). (A) and (B) demonstrate PC1 and PC2 scores versus DAF that show common patterns for all three barley genotypes. (C) depicts loadings plot of the corresponding PCA model (black circled variables are the most informative metabolites (with variable importance projection (VIP) scores > 1) for predicting DAF, identified from PLS modeling (Supplementary Figure S3). (D) and (E) are the raw data of the two identified metabolites which represent two common patterns detected from the PCA model.

In order to further investigate the metabolomic development during the development stage, a PLS model was developed on the DAF effect separated metabolomic data  $X_{\text{DAF}}$  and the response vector,  $Y$ , describing days after flowering (DAF). The PLS model was able to predict DAF with correlation coefficient ( $r^2$ ) of 0.86 (Supplementary Figure S3 (A)). The importance of the metabolites on predicting DAF was evaluated by inspection of variable importance projection (VIP) scores and loadings of the PLS model. Metabolites that exhibit high prediction power (VIP > 1) are highlighted with black circles in the loadings plot of the PCA model of DAF effect separated data (Figure 3 (C)) and with red color in the loadings plot of the PLS model (Supplementary Figure S3 (B)). Syringaldehyde, which increases during development stage, was found as the most affected metabolite by the DAF effect. Other influential metabolites were found to be organic acids, 3-hydroxyoctanoic, sorbic, malonic, glyceric and maleic acids, and phenolics, p-hydroxybenzaldehyde, vanillin, salicylic acid, protocatechuic acid, vanillic acid and syringic acid. Highly altered metabolites during the development stage period for all three barley genotypes are listed in Table 1.

**Table 1.** Common and barley genotype specific metabolites highly influenced by the DAF and growing temperature effects.

	<b>Metabolites increased by DAF</b>	<b>Metabolites decreased by DAF</b>	<b>Metabolites increased by HT</b>	<b>Metabolites increased by HT</b>
<b>Common effects</b>	Octanol-1 ( <b>4</b> ), p-hydroxybenzaldehyde ( <b>28</b> ), 3-hydroxybutanoic acid ( <b>32</b> ), 3-hydroxyoctanoic acid (51), 4-hydroxybenzoic acid, methyl ester (57), vanillin (58), 4-Hydroxyphenylethanol (67), 2-Ketoglutaric acid (72), 3-methyl-3-hydroxypentanedioic acid (73), 6-hydroxydodecane (75), p-Salicylic acid (76), Suberic acid (89), Syringaldehyde (91), Protocatechuic acid, methyl ester (96), Vanillic acid (100), Azelaic acid (106), Protocatechuic acid (113), $\alpha$ -Resorcylic acid (114), Syringic acid (132), Gallic acid (145), Caffeic acid methyl ester (152), 2-hydroxysebacic acid (160), Caffeic acid (172), 3-hydroxyandrostane-17-one (200), 2,3-dihydroxypalmitic acid, propyl ester (232)	Laevulinic acid (1), Sorbic acid (2), Hepta-2,4-dienoic acid, methyl ester (3), Malonic acid (9), 1,3-dihydroxypropanone-2(18), Glyceric acid (23), Maleic acid (24), Fumaric acid (25), Trimethylaconitate (36), 3-hydroxyanthranilic acid, methyl ester (43), Pyroglutamic acid (53), Erythritol (54), 2,5-dimethoxymandelic acid (98), 4-hydroxycinnamic acid, methyl ester (105), Gentisic acid (127), p-Coumaric acid (140), Methylvanillactate (171)	(5), (12), (15), (18), (38), Pyroglutamic acid (53), 4-Hydroxyphenylethanol (67), (68), 3-methyl-3-hydroxypentanedioic acid (73), (74), Dodecane-6-hydroxy (75), p-Salicylic acid (76), (80), (82), Suberic acid (89), $\beta$ -D-Xylopyranose (93), Vanillic acid (100), Azelaic acid (106), (109), Methyl 2-(oxy)-2-(4-(oxy)phenyl)propanoate (110), Protocatechuic acid (113), Homovanillic acid (118), Syringic acid (132), $\beta$ -D-Glucopyranose (153), 2-hydroxysebacic acid (160), Ferulic acid (165), Sinapinic acid methyl ester (169), (177), Sinapinic acid (187), (194), (195), 19-Norandrosterone (209), 3,7-dihydroxyandrostane-17-one (220), (225), (243)	Hepta-2,4-dienoic acid, methyl ester (3), Octanol (4), (11), Benzoic acid (13), (16), 1,3-dihydroxypropanone-2 (18), (19), (21), Maleic acid (24), (35), (39), (41), 3-hydroxyanthranilic acid, methyl ester (43), (49), 3-hydroxyoctanoic acid (51), (61), (66), (69), (71), (77), 2,5-dimethoxymandelic acid (98), (124), (130), p-Coumaric acid (140), (144), 2-hydroxymandelic acid, ethyl ester (147), (151), (158), (159), (163), Methylvanillactate (171), (175), (180), (185), (188), (196), (202), (206), (211), (212), (219), (221), (226), (231), (239)

<b>lys5.f</b>	2-hydroxyheptanoic acid (29), Maseptol-1 (46), 2-hydroxycyclohexane-1-carboxylic acid (50), (68), (82), (84), (85), 2,3-dihydroxyphosphoric acid, propyl ester (107), (121), (133), (217), (227)	(8), 3-methylfuran-2-carboxylic acid (14), (16), (19), (35), (55), (66), Isocitric acid(116), (143), (149), (173), (198), (211), (212), (226), (231), (246), (247)	(6), (27), Resorcinol (33), Maseptol (46), 4-hydroxybenzeneacetic acid, methyl ester (57), (65), Anozol (70), Methyl Isovanillate (78), (81), (84), (90), Syringaldehyde (91), Protocatechuic acid, methyl ester (96), (108), $\alpha$ -Resorcylic acid (114), (121), (122), Ferulic acid, methyl ester (141), Gallic acid (145), (174), 4,8-dihydroxy-2-quinolinecarboxylic acid (186), 3-hydroxyandrostane-17-one (200), 2,3-dihydroxypalmitic acid, propyl ester (232), (234), (237)	(26), (30), Erythritol (54), (94), (119), D-Galactose (137), (149), (161), (164), (168), (184), (197), (203), 9,10-dihydroxystearic acid (218), (222), (224), (242), 3,7-dihydroxycholest-5-ene (245), (247)
<b>lys3.a</b>	(79), (108), (120), Palmitic acid, methyl ester (136), (139), (150), 2-hydroxysebacic acid (160), (167), (177), (179), 4,8-dihydroxy-2-quinolinecarboxylic acid (186), (194), (195), 2,3-dihydroxypalmitic acid, propyl ester (232), (234), (243)	3-methylfuran-2-carboxylic acid (14), (16), (21), (26), (30), (39), (41), (65), (71), (124), (130), (155), (157), (184), (196), (198)	(20), 2-hydroxyheptanoic acid (29), 2,4-dihydroxy-5-methylpyrimidine (44), 4-hydroxybenzeneacetic acid, methyl ester (57), 4-Hydroxyphenylethanol (67), 2-Ketoglutaric acid (72), Protocatechuic acid, methyl ester (96), (104), 2,3-dihydroxyphosphoric acid, propyl ester (107), $\beta$ -D-Glucopyranose (134), $\alpha$ -D-Glucopyranose (135), Gallic acid (145), (146), Caffeic acid, (172), (179), Linoleic acid (181),	acid (2), (31), Trimethyl aconitate (36), Citric acid, trimethyl ester (42), (55), Dimethyl azelate (56), Vanillin (58), (79), (95), (102), (155), (162), (170), (174), (191)

			(184), 2-hydroxytetracosanoic acid, methyl ester (244)	
<b>Bomi</b>	(27), 2-hydroxycyclohexane-1-carboxylic acid (50), (83), (85), (88), (101), (214)	(30), (37), 2,4-dihydroxy-5-methylpyrimidine (44), (122), (124), (130), (155), (161), (180), 9,10-dihydroxystearic acid, dimethyl ester (230), (235), (239), (241)	(31), Malic acid (48), (55), Anozol (70), (87), (102), (119), D-Galactose (137), (139), (154), (170), (214)	(3,3-Dimethyl-1-cyclohexen-1-yl)oxy (10), Glycerol (17), (40), Pyroglutamic acid (53), Erythritol (54), Dimethyl azelate (56), 4-hydroxycinnamic acid, methyl ester (105), Catechin (117), $\beta$ -D-Galactopyranose (123), Caffeic acid, methyl ester (152)

Further barley genotype-dependent metabolomic variations during the development stage was investigated separately by using a data matrix  $\mathbf{X}_{\text{DAF} \times \text{GT}}$  and it will be discussed in following section. Common DAF effects observed from the PLS modeling were in agreement with the DAF effects observed from the PCA analysis of the  $\mathbf{X}_{\text{DAF}}$  data. The majority of the metabolites that showed high loadings for common DAF effects observed in the PCA, also had high VIP scores for the corresponding effects in the PLS models. However, the PLS model revealed more insight into the mostly affected metabolites. In Supplementary Table S1, all the metabolites that decrease, increase or increase in the two weeks of flowering followed by a decrease after three weeks are highlighted.

#### *Metabolic difference of barley genotypes: the genotype specific effect*

Metabolomic variations associated with DAF and temperature effects mask the variation related to the three genetically different barley genotypes for which reason the barley genotype dependent metabolomic variations were not observed by the PCA analysis of the initial metabolomic data  $\mathbf{X}$  (Supplementary Figure S2 (B)). The ASCA separation of the barley genotype related metabolomic variation (8.9%) revealed the genotype specific metabolomic alterations. PCA analysis of the genotype effect separated data  $\mathbf{X}_{\text{GT}}$  (Figure 2), allowed differentiation of the low starch-high- $\beta$ -glucan

mutant (*lys5.f*) from the other two barley genotypes by PC1, and PC2 was able to discriminate the high lysine mutant (*lys3.a*) from the mother genotype (*Bomi*) (Figure 4 (A)). The loadings plot of the PCA model of  $\mathbf{X}_{\text{GT}}$  show a clustering of the metabolites that were responsible for the separation of the barley genotypes (Figure 4 (B)). Metabolites highlighted with red, green and blue circles represent the main classifiers of the *lys5.f*, *Bomi* and *lys3.a* barley genotypes, respectively. Barley genotype dependent trends of the metabolites observed from the loadings plot of the PCA analysis were confirmed by plotting the raw data of selected metabolites (Figure 4 (C-E)). As it was suggested by the PCA and PLS-DA analysis gentisic, protocatechuic and p-

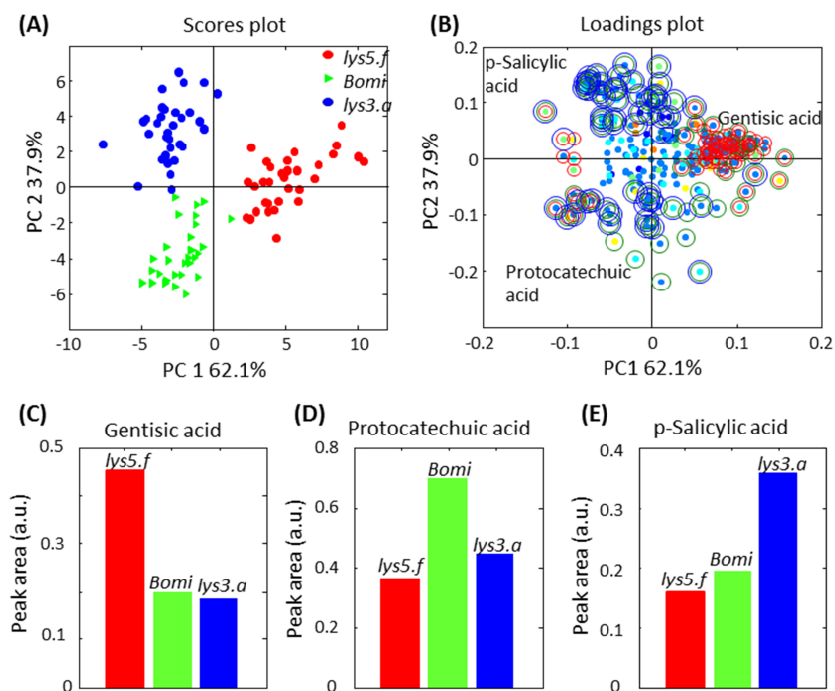
salicylic acids are the main classifiers of the *lys5.f*, *Bomi* and *lys3.a* barley genotypes, respectively. The most important metabolites that separate the high  $\beta$ -glucan barley genotype from the other two genotypes are laevulinic acid, sorbic acid, gentisic acid, 3-methylfuran-2-carboxylic acid, 1,3-dihydroxypropanone, methyl ester of 3-hydroxyanthranilic acid and 4'-cyclohexylacetophenone. The concentrations of these metabolites were significantly higher in the high  $\beta$ -glucan mutant samples compared to the two other barley genotypes. Moreover, the PCA analysis of the three barley genotypes by using only identified phenolic acids and esters (data not shown) as well as organic acids (Supplementary Figure S4 (A)) depicted a similar separation of the barley genotypes as compared to the figure 4 (A). This allowed evaluation of the relative distributions of the metabolites between the different barley genotypes. However, PCA analysis of the barley genotypes by using only identified fatty acids showed a partial separation of the high lysine mutant from the other two genotypes due to the higher content of the fatty acids (Supplementary Figure S4 (B)). This finding was in agreement with the results (increased fat content in flour of *lys3.a* and *lys5.f*) of the previous studies performed on these barley genotypes (Jacobsen *et al.* 2005). Earlier findings have showed that the high  $\beta$ -glucan content (at the expense of starch) in developing mutant seeds in *lys5.f* result in an increase of water content by up to 10% compared to normal barley (Munck *et al.* 2010). The differences in water binding and water activity between the mutant and *Bomi* should influence the metabolome pattern of *lys5.f* as found in this study. Similarly the *lys3.a* mutant contains much more hydrophilic high lysine proteins on the expense of hydrophobic storage proteins (Jacobsen *et al.* 2005) that increase the water content and water activity during epigenesis. This mechanism can partially explain the characteristic *lys3.a* and *lys5.f* metabolome patterns revealed here by PCA and PLS-DA.

In order to gain more insight into the metabolomic differences of the barley genotypes and to search for genotype specific metabolites (metabolites with significantly higher or lower concentration in one barley genotype compared to the other genotypes) a PLS-DA based classification was performed on the genotype effect separated metabolomic data  $X_{GT}$ . PLS-DA enable a more clear separation of the barley genotypes compared to the PCA and the importance of metabolites for the separation can be evaluated in more details (data not shown). The importance of the metabolites for the classification was assessed based on the VIP scores and loadings of the PLS-DA model. Majority of the classifier metabolites of the barley genotypes were those previously identified from the PCA model of the  $X_{GT}$  (Figure 2). These results, analyzing the composition of dried milled flour that is the target for food quality, confirm that the metabolome of the low starch-high- $\beta$ -glucan mutant was significantly different from the metabolome of the mother genotype and high lysine mutant genotypes. It should be emphasized that this study is made on the barley flour and not on the barley seed which is the basic biological unit. During seed development the secondary metabolites are significantly diluted by the effective production of starch but also beta-glucan and cellulose at decreasing water content during the seed ripening process. This effect will be differently regulated in the three genotypes and will cause an apparent decrease of many metabolites on the basis of percent seed flour. The dilution factor would explain a major part of the decreasing content of metabolites over the DAF as well as the deviation of *lys5.f* compared to *lys3.a* and *Bomi* in this respect.

**Table 2.** Main classifier metabolites of the three genetically different barley lines identified by PLS-DA based classification.

	<b>High beta-glucan line (<i>lys5.f</i>)</b>	<b>High lysine line (<i>lys3.a</i>)</b>	<b>Mother line (<i>Bomi</i>)</b>
<b>Main classifier metabolites of barley lines</b>	Laevulinic acid (1), Sorbic acid (2), (8), (3,3-Dimethyl-1-cyclohexen-1-yl)oxy (10), 3-methylfuran-2-carboxylic acid (14), (16), 1,3-dihydroxypropanone-2(18), (19), (21), Maleic acid (24), Trimethyl aconitate (36), 3-hydroxyanthranilic acid, methyl ester (43), 2,4-dihydroxy-5-methylpyrimidine (44), Gentisic acid (127), (142), 4'-Cyclohexylacetophenone (148), (157), (159), (173), (190), (198), (205), (211), (226), (231)	(12), 2-hydroxyheptanoic acid (29), Resorcinol (33), Pyroglutamic acid (53), Dimethyl azelate (56), 4-Hydroxyphenylethanol (67), (68), (74), Dodecane-6-hydroxy (75), p-Salicylic acid (76), (82), Suberic acid (89), $\beta$ -D-Xylopyranose (93), Azelaic acid (106), Homovanilic acid (118), (120), 1-methyl- $\alpha$ -D-Glucopyranose (131), Syringic acid (132), 1-methyl- $\beta$ -D-Glucopyranose (134), $\alpha$ -D-Glucopyranose (135), D-Galactose (137), (139), $\beta$ -D-Glucopyranose (153), 2-hydroxysebacic acid (160), Ferulic acid (165), (167), Sinapinic acid methyl ester (169), (177), (179), Sinapinic acid (187), (194), (195), 9,10-dihydroxystearic acid (218), 9,10-dihydroxystearic acid, dimethyl ester (230), (243), 3,7-dihydroxycholest-5-ene (245)	p-hydroxybenzaldehyde (28), (31), 3-hydroxybutanoic acid (32), Malic acid (48), 2-hydroxycyclohexane-1-carboxylic acid (50), 3-hydroxyoctanoic acid (51), 4-hydroxybenzeneacetic acid, methyl ester (57), Citric acid, trimethyl ester (62), Anozol (70), 2-Ketoglutaric acid (72), 3-methyl-3-hydroxypentanedioic acid (73), (85), (87), (88), Protocatechuic acid, methyl ester (96), 4-hydroxycinnamic acid, methyl ester (105), 1-methyl- $\alpha$ -D-Galactofuranose (111), Protocatechuic acid (113), Catechin (117), (122), p-Coumaric acid (140), Gallic acid (145), 2-hydroxymandelic acid, ethyl ester (147), (178)

Barley genotype classifier metabolites identified from the PLS-DA are also highlighted in the loadings plot of the PCA model (Figure 4 (B)). The high  $\beta$ -glucan genotype had relatively higher concentrations of small molecular organic acids such as, laevulinic, sorbic, maleic acids and phenolic compounds, methyl ester of 3-hydroxyanthranilic acid and gentisic acid. In contrast, the high lysine mutant illustrated greater level of organic acids like, 2-hydroxyheptanoic acid, and suberic acid and phenolic acids such as, ferulic, syringic and sinapinic acids. Table 2 lists the main classifier metabolites that builds the unique pattern of each of the three barley genotypes.



**Figure 4.** Scores (A) and loadings (B) plots of the PCA model developed on the barley genotype effect separated metabolomic data matrix ( $X_{GT}$ ). Red, green and blue color circles of the loadings plot represent VIP variables of the *lys5.f*, *Bomi* and *lys3.a* genotypes that were identified from the PLS-DA model of the  $X_{GT}$  data. Color coding of metabolite classes are the same as in Figure 3 (c). Bar plots (C), (D) and (E) illustrate relative abundances of three examples of genotype specific metabolites, averaged for both, high and low temperature samples and over all DAF points.

#### *The combined genotype and DAF interaction effects: Genotype dependent DAF effect*

It has become clear from NIR spectroscopic studies of the barley mutant seeds and chemical analysis on several expression levels (Munck *et al.* 2004; Jacobsen *et al.* 2005; Munck *et al.* 2010) that the *lys3.a* and *lys5.f* mutants display mutant specific chemical patterns on barley flour basis that are highly reproducible when grown in a controlled environment. Thus, mutation permutation experiments, knocking out one specific gene, are powerful tools to study gene-interaction effects of one gene on all other active genes in the seed (endosperm). We will here demonstrate that genotype specific patterns of these mutants are reflected in raw metabolic data in the secondary metabolism.

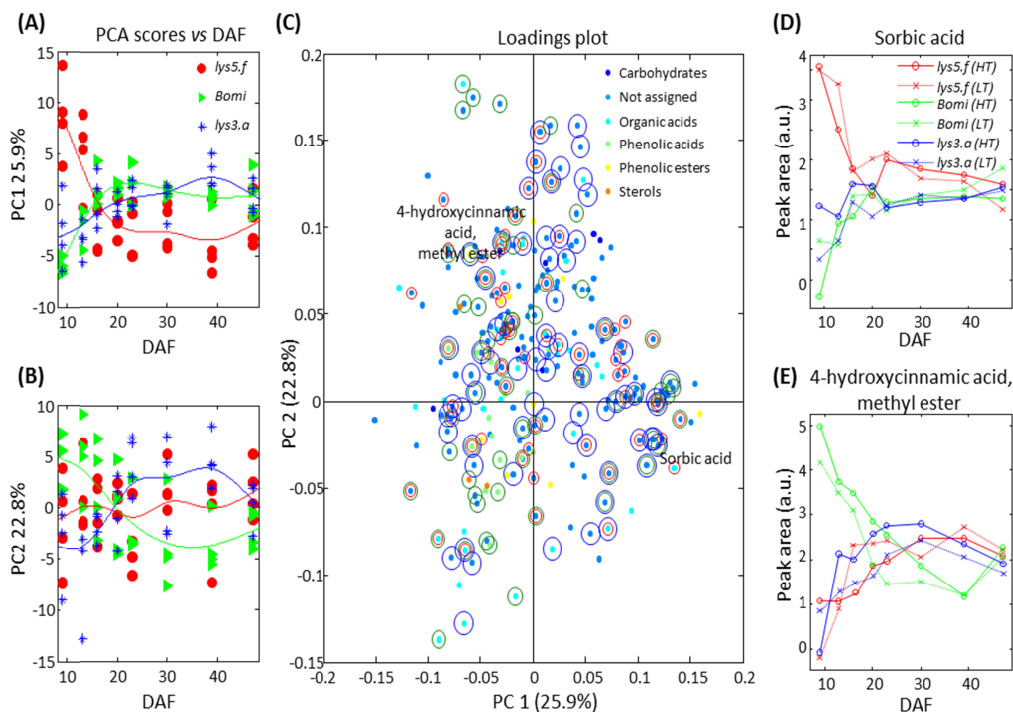
A substantial degree of the DAF variation was associated with barley genotype (14.5%) (Figure 2) and these trends are described by PC1 and PC2 of the PCA model developed on the DAF-genotype interaction effect separated data matrix  $X_{DAF \times GT}$  (Figure 5). PC1 shows genotype dependent DAF effect where scores of low starch-high- $\beta$ -glucan genotype *lys5.f* samples decrease during the development stage, while the scores of the other two barley genotypes increase (Figure 5(A)). These differences are more pronounced at the early development stages and attenuate at later stages. PC2 of this PCA model also capture barley genotype dependent

metabolomic alterations occurring during the seed developmental stages (Figure 5 (B)). The scores of the high lysine mutant samples in PC2 gradually increase over DAF whereas the scores of the parental *Bomi* genotype decrease. The low-starch-high- $\beta$ -glucan mutant remained almost unchanged. These genotype specific DAF effects captured from DAF-genotype interactions were highly pronounced in small molecular organic and phenolic acids. The loadings plot of this PCA model show the metabolites highlighted with red, green and blue circles, which correspond to the main classifiers of the *lys5.f*, *Bomi* and *lys3.a* genotypes, identified from the PLS-DA classification analysis.

The general patterns found by the PCA is illustrated by two examples of metabolites, sorbic acid and 4-hydroxycinnamic acid's methyl ester, in figures 5 (D) and (C), respectively. These patterns are associated with the genetic modification, as demonstrated by the significant deviation of the high  $\beta$ -glucan mutant from its parental genotype *Bomi* and high lysine mutant *lys3.a* where differences in protein synthesis were observed between another high  $\beta$ -glucan mutant *lys5.g*, *lys3.a* and *Bomi*. The initial protein synthesis in *lys5.g* were fastest, followed by *Bomi* and *lys3.a*, whereas from day 23 after anthesis the rate of protein synthesis in *lys3.a* accelerated considerably leading to very high final levels at harvest time (Jacobsen *et al.* 2005). The differential behavior in protein synthesis kinetics at different kernel formation stages is expected to influence the seed metabolome as well.

In order to further investigate genotype specific DAF effects, individual PLS models were developed on the DAF effect separated data matrices of each barley genotype separately. PLS models of the separate barley genotypes had an equal (in case of *lys3.a*) or improved prediction power (in case of *lys5.f* and *Bomi*). Most of the detected predictor metabolites were those previously found in the PLS model of the global data  $\mathbf{X}_{\text{DAF}}$  (Supplementary Figures S5 A-C). However, some important predictor metabolites were different among the barley genotypes, while others showed similar trends, but with different rate of increase or decrease in the flour over the DAF. Some of the genotype specific metabolites highly influenced by the developmental stage (DAF effect) are listed in Table 1. These genotype specific DAF effects observed from the individual PLS models were in agreement with the genotype specific DAF effects detected from the PCA analysis of the DAF-genotype interaction effect separated data  $\mathbf{X}_{\text{DAF} \times \text{GT}}$ . The loadings plot of this PCA model highlights genotype specific DAF predictors that showed high influence on the individual PLS models (VIP>1).

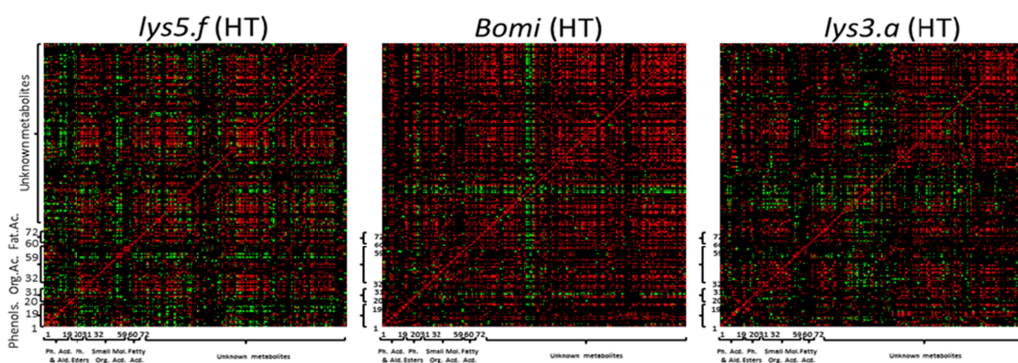




**Figure 5.** Barley genotypes specific metabolomic alterations during seed development by PCA analysis of the DAF and GT interaction effects separated metabolomics data matrix ( $X_{\text{DAF} \times \text{GT}}$ ). **(A)** and **(B)** demonstrate PC1 and PC2 scores versus DAF that show genotypes specific patterns. **(C)** depicts loadings plot of the corresponding PCA model. Most important genotypes specific metabolites for predicting DAF (variable importance projection (VIP) scores  $> 1$ ) are illustrated with red, green and blue circles for *lys5.f*, *Bomi* and *lys3.a* lines, respectively. These genotypes specific DAF effects were identified from the individual PLS models of each barley line (Supplementary Figure S5). **(D)** and **(E)** illustrate barley genotype specific fluctuations of the two metabolites over different DAF points. These reflect the genotypes specific DAF effects captured by PC1 and PC2, **(A)** and **(B)**, respectively.

The covariance of the barley seed metabolites during the development stages shows inter-relationships of several biosynthetic pathways through which metabolites are synthesized. Increase or decrease of metabolite level during the seed development occurs in a barley genotype specific manner when the seed metabolome gradually change during the maturation. In order to compare the covariance of the metabolites during the seed development, the DAF effect separated metabolomic data was used for a simple correlation analysis between the metabolites. The DAF effect separated metabolomic data  $X_{\text{DAF}}$  ( $8 \times 247$ ) for each barley genotypes (*lys5.f*, *lys3.a* and *Bomi*) grown under high or low temperature was analyzed separately. Metabolites for each genotype were correlated across the eight harvest points during the seed development and a correlation matrix ( $247 \times 247$ ) was established for each barley genotype (at a given growth temperature). The heat maps from the high temperature (HT) treatment colored according to the correlation level (Figure 6) shows that the mother genotype *Bomi* possess a higher number of metabolites that are positively correlated during the seed development, both among identified and unidentified metabolites compared to the other two barley genotypes. It is evident that all three genotypes show higher numbers of positive

correlations at the low temperature LT level and that the two mutants especially *lys3a* suffer more from a high temperature environment than *Bomi*. This indicates that the biosynthesis of these metabolites have been altered and deregulated in the mutants. In contrast to this, the number of negatively correlated metabolites during the seed development was higher in the mutant barley genotypes compared to their mother genotype *Bomi*. This show that in the case of *Bomi*, more metabolites were increased or decreased in a similar manner, while the mutation of one specific gene in barley genotypes, *lys5.f* and *lys3.a*, resulted in a deregulation of this metabolomic equilibrium. Therefore, in the cases of the *lys5.f* and *lys3.a* barley mutants the number of positively correlated metabolites have significantly decreased and the number of negatively correlated metabolites have increased, which confirm the more global alteration of the metabolome in the mutants. This phenomenon can be explained by the pleiotropy and the alteration of the several (if not all) pathways simultaneously as a whole pattern, which resulted in a significant perturbation of the barley seed metabolic equilibrium that was especially evident at a high temperature.



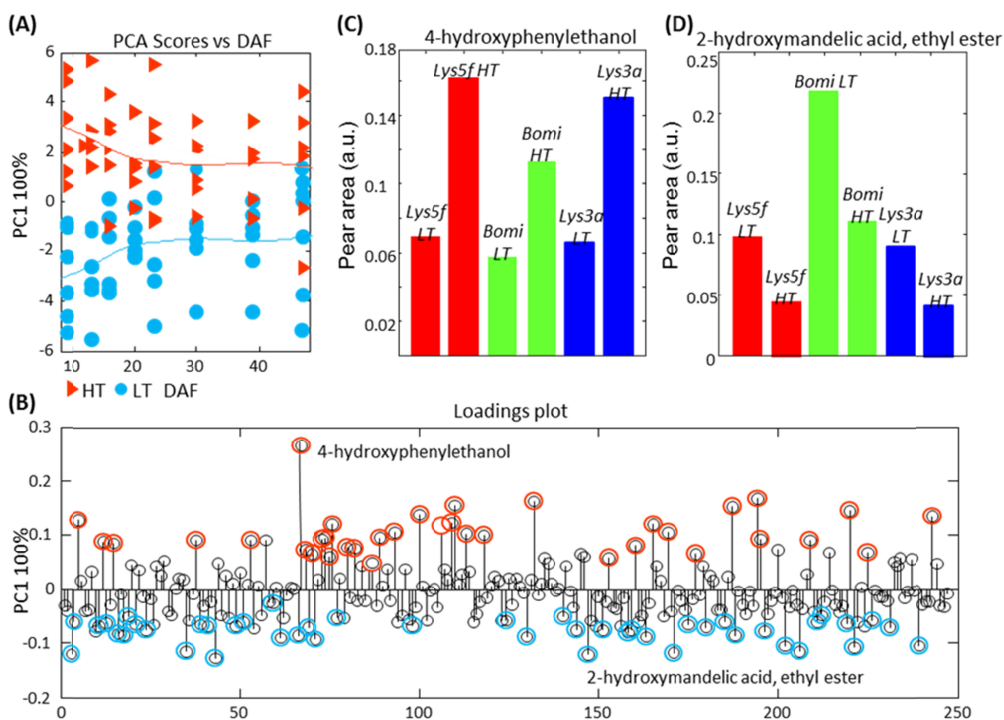
**Figure 6.** The heat maps of the correlation matrix obtained from the DAF effect separated metabolomic data,  $X_{DAF}(8 \times 247)$ , of high temperature (HT) grown barley genotypes harvested at 8 different time points during the development. **Red** squares correspond to the correlation coefficients ( $r$ )  $> 0.7$ , while **green** squares correspond to the ( $r$ )  $< -0.7$  and the **black** squares correspond to the  $r < 0.7$  &  $r > -0.7$ , between the 247 detected metabolites. These findings illustrate the levels of disturbances of whole-grain metabolomic equilibrium between the mother line (*Bomi*) and the mutants (*lys5.f* and *lys3.a*) derived from it. The balances in increase and/or decrease of whole-grain barley metabolite concentrations have been disturbed from the equilibrium present in the mother line. The same trends were observed in the low temperature (LT) grown barley plants. The number of positive correlations between the metabolites, during seed development, of barley genotypes *lys5.f*(LT), *Bomi*(LT), and *lys3.a*(LT) were 8019, 9957 and 7151 (or in the ratio of 1.24 : 1.40 : 1), respectively. While, the number of negative correlations between the 247 metabolites of barley genotypes *lys5.f*(LT), *Bomi*(LT), and *lys3.a*(LT) were 4028, 1706 and 1800 (or in the ratio of 2.36 : 1 : 1.05), respectively. It is worth to mention that the whole-grain metabolomic deviations were slightly more pronounced in the mutant *lys5.f* than in *lys3.a*.

### Influence of the growing temperature

Analysis of variance of the initial metabolomic data  $X$  showed that the variations related to the low and high growing temperature are significantly lower than the variation due to the DAF effect and genotype differences (Figure 2). PCA of the initial metabolomic data revealed no differences between the barley genotypes grown under the two temperatures (Supplementary

Figure S2 (C)). However, a PCA of the temperature effect separated metabolomic data  $X_{TEMP}$  revealed a partial separation of barley plants grown under high and low temperature conditions (Figure 7 (A)). Loadings plot of this PCA model illustrated important metabolites for the separation and provided a good initial overview of the metabolomic changes caused by the growing temperature. Organic acids such as malonic, succinic, suberic, 2-hydroxyheptanoic, 3-methyl-3-hydroxypentanedioic acids, and phenolic acids, ferulic, syringic, p-salicylic, vanillic, sinapinic acids, 4-hydroxyphenylethanol and some fatty acids and sterols were more abundant in the high temperature grown plants (HT) compared to the low temperature grown plants (LT). While, concentrations of organic acids, citric acid trimethyl ester, benzoic acid and phenolics such as, p-hydroxybenzaldehyde, 2,5-dimethoxymandelic acid, methyl ester of 4-hydroxycinnamic acid, ethyl ester of 2-hydroxymandelic acid, and methyl ester of ferulic acid were relatively higher in LT barley samples than in HT samples. Metabolites illustrated with red and blue circles represent the main classifiers of the HT and LT samples identified by PLS-DA classification analysis, which will be discussed later (Figure 6 (B)). The mostly influenced metabolite by higher growing temperature was 4-hydroxyphenylethanol, which significantly increased in HT samples in all three barley genotypes (Figure 6 (C)). In contrast, 2-hydroxymandelic acid's ethyl ester significantly increased when the barley plants were grown under the lower temperature (Figure 6 (D)). The identity of the metabolites that had relatively higher influence in PCA based separation are provided in Supplementary Table S1. This table also lists the relative ratios of the all metabolites (at the harvest time, 47 days after flowering) detected in the barley genotypes grown under two different temperatures. Although, interactions between growing temperature and DAF effects were insignificant ( $p = 0.1$ ) (Figure 2), temperature related variations of some metabolites were more pronounced at the earlier stages of DAF.

While PCA provide a good overview of the differences between HT and LT barley genotypes, it was not effective in evaluating the most influential metabolites. Therefore, the temperature effect separated metabolomic data  $X_{TEMP}$  was further investigated by PLS-DA classification. PLS-DA attempted to classify HT and LT barley samples as two different groups and provided valuable information on the main classifier metabolites. This approach lead to a separation of the HT barley plants from the LT barley plants and assisted in evaluation of common growing temperature dependent metabolomic alterations (Supplementary Figure S6). The majority of the important classifiers identified from the PLS-DA model were those metabolites detected in the PCA modeling that were responsible for partial separation of HT and LT barley samples. In contrast to PCA, VIP scores of the PLS-DA model allowed more holistic evaluation of the importance of individual metabolites for the separation. Among these metabolites, 4-hydroxyphenylethanol, ethyl ester of 2-hydroxymandelic acid and p-salicylic acid displayed the highest influence for the separation, followed by vanillic acid, syringic acid, sinapinic acid, azelaic acid and ferulic acids which were also the main classifiers of the HT barley samples. Variables with high VIP scores ( $>1$ ) are also highlighted in the loadings plot of the PCA model of  $X_{GT}$  (Figure 6 (B)) and it illustrates that these variables also had high loadings for PCA based separation of HT plants from LT plants and confirm the agreement between the PCA and PLS-DA results. Table 1 lists the main classifier metabolites of the HT and LT barley plants that were found based on the VIP scores of the PLS-DA model.



**Figure 7.** Effects of high and low growing temperature on barley metabolome. **(A)** and **(B)** illustrate scores and loadings of the PCA model developed on the temperature effect separated metabolomic data ( $\mathbf{X}_{\text{TEMP}}$ ). The scores plot shows partial separation of the high (HT) and low (LT) temperature grown barley plants, while loadings plot depict metabolites responsible for this separation. Metabolites highlighted with red and light blue circles correspond to the main classifier metabolites (variable importance projection (VIP) scores > 1) of the HT and LT plants identified from the PLS-DA model of the  $\mathbf{X}_{\text{TEMP}}$  data matrix (Supplementary Figure S6). Bar plots **(C)** and **(D)** show examples of two mostly influenced metabolites by growing temperatures (averaged over all DAF points).

Although the interaction effect observed between the genotype and temperature was not significant with p-value of 0.15 ( $\mathbf{X}_{\text{GT}} \times \mathbf{X}_{\text{TEMP}}$ ) (Figure 2), initial PCA and a global PLS-DA model suggested the presence of genotype specific temperature effects. The genotype specific metabolites that were highly influenced by the growing temperatures might be underestimated in the global PLS-DA models, due to their low abundance. Therefore, individual PLS-DA models were developed on the genotype effect separated data,  $\mathbf{X}_{\text{GT}}$ , for each barley genotype separately and revealed additional temperature-altered metabolites. In addition to the most important classifiers found from the global PLS-DA model, individual PLS-DA model of the high  $\beta$ -glucan mutant, revealed 3,7-dihydroxycholest-5-ene, methyl vanillactate, resorcinol, methyl ester of sinapinic acid,  $\alpha$ -resorcylic acid, homovanilic acid and protocatechuic acid as the mostly influential metabolites. This PLS-DA model revealed that 4-hydroxyphenylethanol was the best discriminating metabolite of HT and LT high  $\beta$ -glucan mutant, *lys5.f* genotype and this was also the case in the global PLS-DA model. Classification of the HT and LT barley samples of the high lysine mutant, *lys3.a*, revealed caffeic acid, protocatechuic acid, suberic acid, 2-ketoglutaric acid

and 2,5-dimethoxymandelic acid as the most important classifiers. Importance of these metabolites (VIP scores) for classification of HT and LT plants were significantly higher in genotype separated models than in the global model. This indicates that these metabolites had greater temperature effect in high lysine barley genotype than in the other two genotypes. While organic acids such as maleic, malic and suberic acids, anozol and 4-hydroxycinnamic acid methyl ester had much greater temperature effect in the mother genotype *Bomi* than in mutant genotypes. In addition to the common temperature effects, Table 1 lists the mostly pronounced genotype specific metabolites that were influenced by low and high growing temperatures.

It is worth to mention that more than 50% of the metabolites identified as the main classifiers of HT barley plants were also increased during the development, while none of these HT induced metabolites decreased (Supplementary Table S1). This basically show that the accumulation of metabolites during the seed development is more significant and faster in high temperature grown barley plants than in low temperature grown barley plants. Likewise, almost 60% of the metabolites identified as the main classifiers of low temperature grown barley plants matched with the metabolites that decrease during the development stage, while only 10% of the LT classifier metabolites increased during the maturation. It must be pointed out that there is a considerable dilution effect when e.g. starch and beta-glucan are developed during the maturation of the seed. Here the LT material and *Bomi* has larger seeds with more starch compared to the HT material and the mutants. Larger seeds with higher level of starch could explain a major part of the apparent decrease in the metabolites as percent of the seed flour during maturation while those who increase should increase even more.

## Conclusions

The development of the barley seed (endosperm) metabolome has been studied by a new high-throughput GC-MS method as a function of developmental stage, genotype and growing temperature. The metabolite concentrations were extracted using a new tool, PARAFAC2, that allow detection and quantification of even strongly overlapped and low s/n metabolites. The PARAFAC2 model generated metabolite tables were analyzed by the new ASCA approach which allow the separation and analysis of individual and combined effects in a balanced experimental design. Combination of these two new methods assisted in obtaining an overview over the complex multidimensional metabolomics data set and facilitate in finding the hot spots in the raw data which are of importance for the biological understanding. Our focused barley mutant permutation experiment has shown that it is necessary with a dialogue between results from the data compressive chemometric modeling on one side and selected plots of raw data on the other side. Using this strategy three dominating grain filling metabolomic patterns was found: two in which the metabolites decrease or increase during the development and one which show a significant increase after two weeks of anthesis time and a gradual decrease after three weeks. The ASCA modeling of design effect showed that some organic acids exhibits genotype specific dynamics over the development stage and increased in one genotype and decreased or remained stable in the two other genotypes. Further analysis revealed the presence of classifier metabolites for the three barley genotypes and metabolite effects of different growth temperatures. The high lysine mutant contained greater amount of most antioxidant phenolic acids. The most affected metabolite by the growth temperature was found to be 4-hydroxyphenylethanol that was significantly more abundant in high temperature grown barley plants, while p-coumaric acid and mandelic acid derivatives were much more abundant in the low temperature grown barley plants. Moreover, the barley seed metabolome development

patterns on percent barley flour basis for the different genotypes were partially dependent on the growth temperature revealing metabolites that were markers for high temperature. At the higher growing temperature, metabolites that were increasing during the seed maturation was more significant and rapid than in the case of low temperature growing. Likewise, the metabolites that decrease during the barley seed development, degraded much faster and significantly, when the plant were grown under the low temperature than in high temperature. Finally, correlation maps of the metabolites during seed development illustrated the deregulation on barley seed metabolomic equilibriums influenced by single gene mutations that were significantly amplified at higher growing temperatures.

## Supplementary data

Supplementary data are available at JXB online

Table S1. Common and barley genotype specific metabolites highly influenced by the DAF and growing temperature effects.

Figure S1. TIC chromatogram obtained from GC-MS analysis of the whole-grain flour of barley genotype *Bomi*, grown under high temperature and harvested at the last DAF point.

Figure S2. Scores plots of the PCA model developed on the initial metabolomics data matrix.

Figure S3. PLS model developed on the days after flowering (DAF) effect separated metabolomics data matrix ( $\mathbf{X}_{\text{DAF}}$ ).

Figure S4. PLS models developed on the DAF effect separated metabolomics data matrix ( $\mathbf{X}_{\text{DAF}}$ ), of the individual barley genotypes.

Figure S5. PLS-DA model developed on the temperature effect separated metabolomics data matrix ( $\mathbf{X}_{\text{TEMP}}$ ).

Figure S6. PCA of genotype effect separated data,  $\mathbf{X}_{\text{GT}}$ , including barley samples from 23 to 47 DAF and small organic and fatty acids.

## Acknowledgements

Faculty of Science is acknowledged for support to the elite-research area “Metabolomics and bioactive compounds” with a PhD stipendium to Bekzod Khakimov and The Ministry of Science and Technology is acknowledged for a grant to University of Copenhagen (S.B. Engelsen) with the title “Metabolomics infrastructure” under which the GC-MS was acquired.

## References

- AbuMweis SS, Jew S, Ames NP.** 2010. beta-glucan from barley and its lipid-lowering capacity: a meta-analysis of randomized, controlled trials. *European Journal of Clinical Nutrition* **64**, 1472-1480.
- Amarowicz R, Zegarska Z, Pegg RB, Karamac M, Kosinska A.** 2007. Antioxidant and radical scavenging activities of a barley crude extract and its fractions. *Czech Journal of Food Sciences* **25**, 73-80.
- Amigo JM, Skov T, Coello J, Maspoch S, Bro R.** 2008. Solving GC-MS problems with PARAFAC2. *Trac-Trends in Analytical Chemistry* **27**, 714-725.
- Andersson AAM, Lampi AM, Nystrom L, Piironen V, Li L, Ward JL, Gebruers K, Courtin CM, Delcour JA, Boros D, Fras A, Dynkowska W, Rakszegi M, Bedo Z, Shewry PR, Aman P.** 2008. Phytochemical and Dietary Fiber Components in Barley Varieties in the HEALTHGRAIN Diversity Screen. *Journal of Agricultural and Food Chemistry* **56**, 9767-9776.
- Arranz S, Calixto FS.** 2010. Analysis of polyphenols in cereals may be improved performing acidic hydrolysis: A study in wheat flour and wheat bran and cereals of the diet. *Journal of Cereal Science* **51**, 313-318.
- Balmer D, Flors V, Glauser G, Mauch-Mani B.** 2013. Metabolomics of cereals under biotic stress: current knowledge and techniques. *Frontiers in Plant Science* **4**, 82.
- Bino RJ, Hall RD, Fiehn O, Kopka J, Saito K, Draper J, Nikolau BJ, Mendes P, Roessner-Tunali U, Beale MH, Trethewey RN, Lange BM, Wurtele ES, Sumner LW.** 2004. Potential of metabolomics as a functional genomics tool. *Trends in Plant Science* **9**, 418-425.
- Bro R, Andersson CA, Kiers HAL.** 1999. PARAFAC2 - Part II. Modeling chromatographic data with retention time shifts. *Journal of Chemometrics* **13**, 295-309.
- Doll H.** 1983. Barley Seed Proteins and Possibilities for their Improvement. In: Gottschalk W, M++ller H, eds. *Seed Proteins*, Springer Netherlands, pp. 207-223.
- Fernie AR, Schauer N.** 2009. Metabolomics-assisted breeding: a viable option for crop improvement? *Trends in Genetics* **25**, 39-48.
- Frederiks TM, Christopher JT, Harvey GL, Sutherland MW, Borrell AK.** 2012. Current and emerging screening methods to identify post-head-emergence frost adaptation in wheat and barley. *Journal of Experimental Botany* **63**, 5405-5416.
- Gibson SM, Strauss G.** 1991. Implication of Phenolic-Acids As Texturizing Agents During Cooking-Extrusion Cereals. *Abstracts of Papers of the American Chemical Society* **202**, 150-AGFD.
- Hotelling H.** 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **24**, 417-441.
- Jacobsen S, Sondergaard I, Moller B, Desler T, Munck L.** 2005. A chemometric evaluation of the underlying physical and chemical patterns that support near infrared spectroscopy of barley seeds as a tool for explorative classification of endosperm, genes and gene combinations. *Journal of Cereal Science* **42**, 281-299.

- Khakimov B, Mohammed SM, Bak S, Engelsen SB.** 2013. The use of trimethylsilyl cyanide derivatization for robust and broad spectrum high-throughput gas-chromatography-mass spectrometry based metabolomics. *Analytical and Bioanalytical Chemistry* **In press**. DOI: 10.1007/s00216-013-7341-z
- Khakimov B, Amigo JM, Bak S, Engelsen SB.** 2012. Plant metabolomics: Resolution and quantification of elusive peaks in liquid chromatography-mass spectrometry profiles of complex plant extracts using multi-way decomposition methods. *Journal of Chromatography.A* **1266**, 84-94.
- Korkina LG, Pastore S, Dellambra E, De Luca C.** 2013. New Molecular and Cellular Targets for Chemoprevention and Treatment of Skin Tumors by Plant Polyphenols: A Critical Review. *Current Medicinal Chemistry* **20**, 852-868.
- Li L, Shewry PR, Ward JL.** 2008. Phenolic Acids in Wheat Varieties in the HEALTHGRAIN Diversity Screen. *Journal of Agricultural and Food Chemistry* **56**, 9732-9739.
- Madhujith T, Shahidi F.** 2007. Antioxidative and antiproliferative properties of selected barley (*Hordeum vulgare* L.) cultivars and their potential for inhibition of low-density lipoprotein (LDL) cholesterol oxidation. *Journal of Agricultural and Food Chemistry* **55**, 5018-5024.
- Manach C, Scalbert A, Morand C, Remesy C, Jimenez L.** 2004. Polyphenols: food sources and bioavailability. *American Journal of Clinical Nutrition* **79**, 727-747.
- Manavalan LP, Chen X, Clarke J, Salmeron J, Nguyen HT.** 2012. RNAi-mediated disruption of squalene synthase improves drought tolerance and yield in rice. *Journal of Experimental Botany* **63**, 163-175.
- Max B, Salgado JM, Cortes S, Dominguez JM.** 2010. Extraction of Phenolic Acids by Alkaline Hydrolysis from the Solid Residue Obtained after Prehydrolysis of Trimming Vine Shoots. *Journal of Agricultural and Food Chemistry* **58**, 1909-1917.
- Mcintosh GH, Whyte J, McArthur R, Nestel PJ.** 1991. Barley and Wheat Foods - Influence on Plasma-Cholesterol Concentrations in Hypercholesterolemic Men. *American Journal of Clinical Nutrition* **53**, 1205-1209.
- Munck, L.** The case of high lysine barley breeding, *Barley: Genetics, Biochemistry, Molecular Biology and Biotechnology*. 573-603. 1992. Wallingford, Oxon, UK, C.A.B: International. (GENERIC)  
Ref Type: Edited Book
- Munck L, Jespersen BM, Rinnan A, Seefeldt HF, Engelsen MM, Norgaard L, Engelsen SB.** 2010. A physicochemical theory on the applicability of soft mathematical models-experimentally interpreted. *Journal of Chemometrics* **24**, 481-495.
- Munck L, Moller B, Jacobsen S, Sondergaard I.** 2004. Near infrared spectra indicate specific mutant endosperm genes and reveal a new mechanism for substituting starch with (1 → 3,1 → 4)-beta-glucan in barley. *Journal of Cereal Science* **40**, 213-222.
- Munck L, Nielsen JP, Moller B, Jacobsen S, Sondergaard I, Engelsen SB, Norgaard L, Bro R.** 2001. Exploring the phenotypic expression of a regulatory proteome-altering gene by spectroscopy and chemometrics. *Analytica Chimica Acta* **446**, 171-186.
- Nystrom L, Lampi AM, Andersson AAM, Kamal-Eldin A, Gebruers K, Courtin CM, Delcour JA, Li L, Ward JL, Fras A, Boros D, Rakszegi M, Bedo Z, Shewry PR, Piironen V.** 2008. Phytochemicals and



Dietary Fiber Components in Rye Varieties in the HEALTHGRAIN Diversity Screen. *Journal of Agricultural and Food Chemistry* **56**, 9758-9766.

**Quinde-Axtell Z, Baik BK.** 2006. Phenolic compounds of barley grain and their implication in food product discoloration. *Journal of Agricultural and Food Chemistry* **54**, 9978-9984.

**Sani IM, Iqbal S, Chan KW, Ismail M.** 2012. Effect of Acid and Base Catalyzed Hydrolysis on the Yield of Phenolics and Antioxidant Activity of Extracts from Germinated Brown Rice (GBR). *Molecules* **17**, 7584-7594.

**Seefeldt HF, Blennow A, Jespersen BM, Wollenweber B, Engelsen SB.** 2009. Accumulation of mixed linkage (1 → 3) (1 → 4)-beta-D-glucan during grain filling in barley: A vibrational spectroscopy study. *Journal of Cereal Science* **49**, 24-31.

**Shewry PR, Piironen V, Lampi AM, Nystrom L, Li L, Rakszegi M, Fras A, Boros D, Gebruers K, Courtin CM, Delcour JA, Andersson AAM, Dimberg L, Bedo Z, Ward JL.** 2008. Phytochemical and Fiber Components in Oat Varieties in the HEALTHGRAIN Diversity Screen. *Journal of Agricultural and Food Chemistry* **56**, 9777-9784.

**Smilde AK, Jansen JJ, Hoefsloot HJ, Lamers RJAN, van der Greef J, Timmerman ME.** 2005. ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics* **21**, 3043-3048.

**Soltész A, Smedley M, Vashegyi I, Galiba G, Harwood W, Vagujfalvi A.** 2013. Transgenic barley lines prove the involvement of TaCBF14 and TaCBF15 in the cold acclimation process and in frost tolerance. *Journal of Experimental Botany* **64**, 1849-1862.

**Stahle L, Wold S.** 1987. Partial Least Squares Analysis with Cross-Validation for the Two-Class Problem A Monte Carlo Study. *Journal of Chemometrics* **1**, 185-196.

**Taketa S, Matsuki K, Amano S, Saisho D, Himi E, Shitsukawa N, Yuo T, Noda K, Takeda K.** 2010. Duplicate polyphenol oxidase genes on barley chromosome 2H and their functional differentiation in the phenol reaction of spikes and grains. *Journal of Experimental Botany* **61**, 3983-3993.

**Vandendool H, Kratz PD.** 1963. A Generalization of Retention Index System Including Linear Temperature Programmed Gas-Liquid Partition Chromatography. *Journal of Chromatography* **11**, 463.

**Vinson JA, Erk KM, Wang SY, Marchegiani JZ, Rose MF.** 2009. Total polyphenol antioxidants in whole grain cereals and snacks: Surprising sources of antioxidants in the US diet. *Abstracts of Papers of the American Chemical Society* **238**, 246.

**Von Wettstein, D.** The future of barley as an experimental organism, Barley Genetics IGBS VI. 2, 1087-1098. 1992. Helsingborg, Sweden, 1991, (Munck,L., Ed.), Munksgaard International Publ., Copenhagen. (GENERIC)  
Ref Type: Edited Book

**Ward JL, Poutanen K, Gebruers K, Piironen V, Lampi AM, Nystrom L, Andersson AAM, Aman P, Boros D, Rakszegi M, Bedo Z, Shewry PR.** 2008. The HEALTHGRAIN Cereal Diversity Screen: Concept, Results, and Prospects. *Journal of Agricultural and Food Chemistry* **56**, 9699-9709.

**Widodo, Patterson JH, Newbiggin E, Tester M, Bacic A, Roessner U.** 2009. Metabolic responses to salt stress of barley (*Hordeum vulgare* L.) cultivars, Sahara and Clipper, which differ in salinity tolerance. *Journal of Experimental Botany* **60**, 4089-4103.

**Winning H, Viereck N, Wollenweber B, Larsen FH, Jacobsen S, Sondergaard I, Engelsen SB.** 2009. Exploring abiotic stress on asynchronous protein metabolism in single kernels of wheat studied by NMR spectroscopy and chemometrics. *Journal of Experimental Botany* **60**, 291-300.

**Wold H.** 1979 Model Construction and Evaluation when Theoretical Knowledge is Scarce: An Example of the Use of Partial Least Squares. Université de Genève, Faculté des Sciences Économiques et Sociales.

**Wold S, Martens H, Wold H.** 1983. The multivariate calibration problem in chemistry solved by the PLS method. In: Kågström B, Ruhe A, eds. *Matrix Pencils*, Springer Berlin Heidelberg, pp. 286-293.

**Wood PJ.** 2007. Cereal beta-glucans in diet and health. *Journal of Cereal Science* **46**, 230-238.

**Zielinski H, Kozłowska H.** 2000. Antioxidant activity and total phenolics in selected cereal grains and their different morphological fractions. *Journal of Agricultural and Food Chemistry* **48**, 2008-2016.

**Zwanenburg G, Hoefsloot HCJ, Westerhuis JA, Jansen JJ, Smilde AK.** 2011. ANOVA-principal component analysis and ANOVA-simultaneous component analysis: a comparison. *Journal of Chemometrics* **25**, 561-567.



## Supplementary Data Files

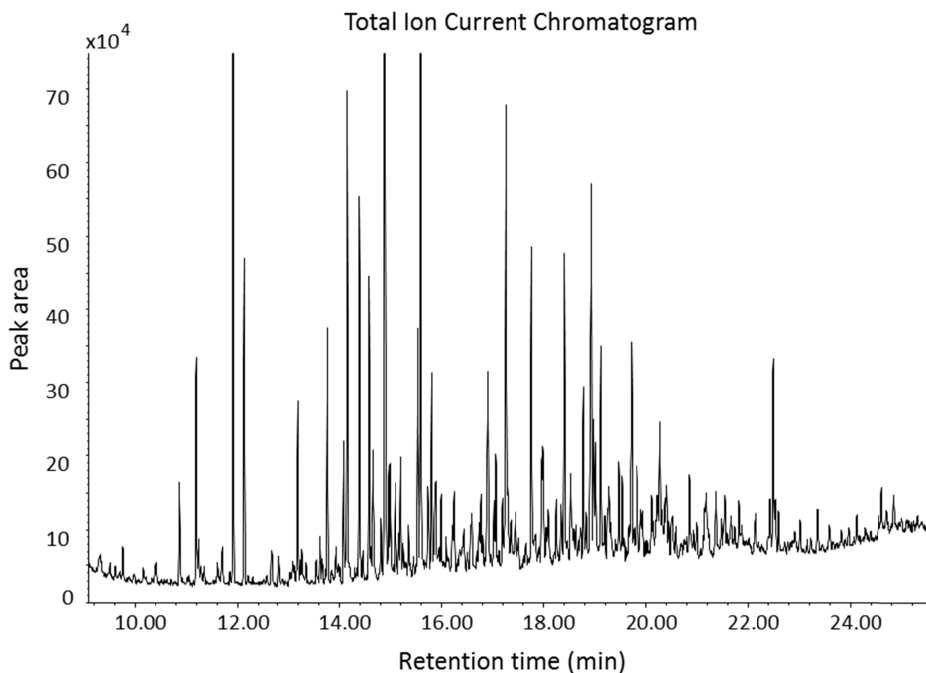
**Title:** The emerging barley seed metabolome studied by mutant analysis and advanced GC-MS: evaluation of the effects of development stage, genotype and growth temperature by ASCA

**Authors:** Bekzod Khakimov\*, Morten Arendt Rasmussen, Birthe Møller Jespersen, Lars Munck, Søren Balling Engelsen

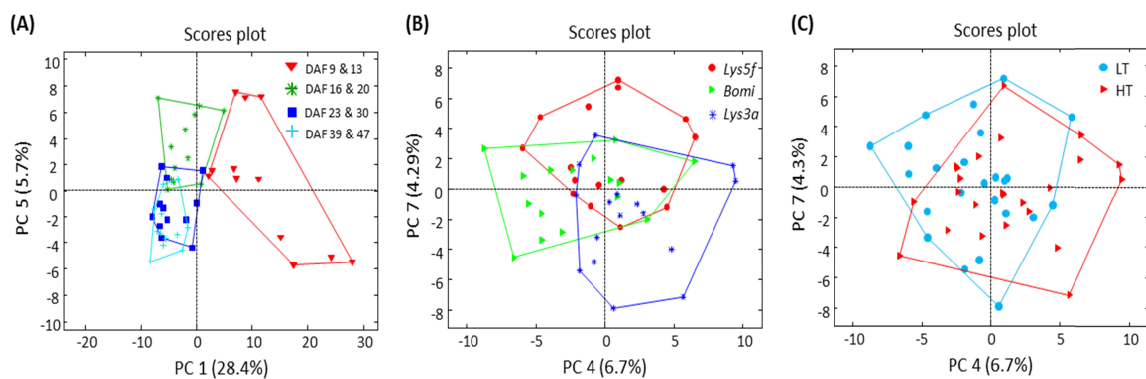
**Institution:** Spectroscopy & Chemometrics, Department of Food Science, Faculty of Science, University of Copenhagen, Rolighedsvej 30, 1958 Frederiksberg C, Denmark

**\*Corresponding author:** Bekzod Khakimov

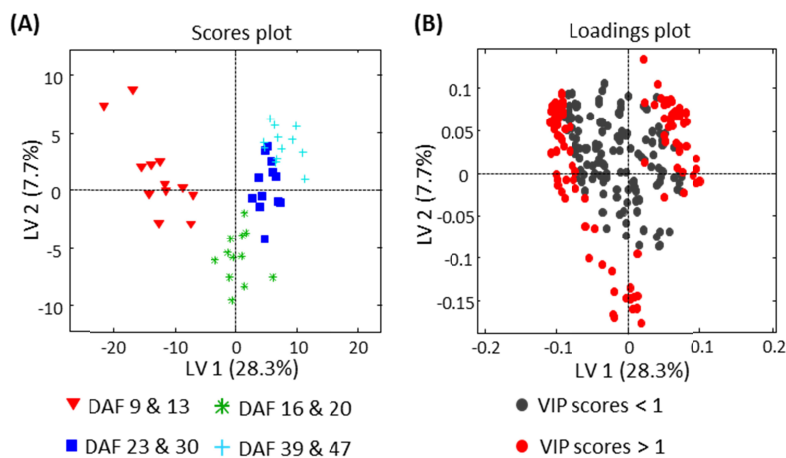
E-mail: [bzo@food.ku.dk](mailto:bzo@food.ku.dk), Tel.: +45- 35 33 29 74



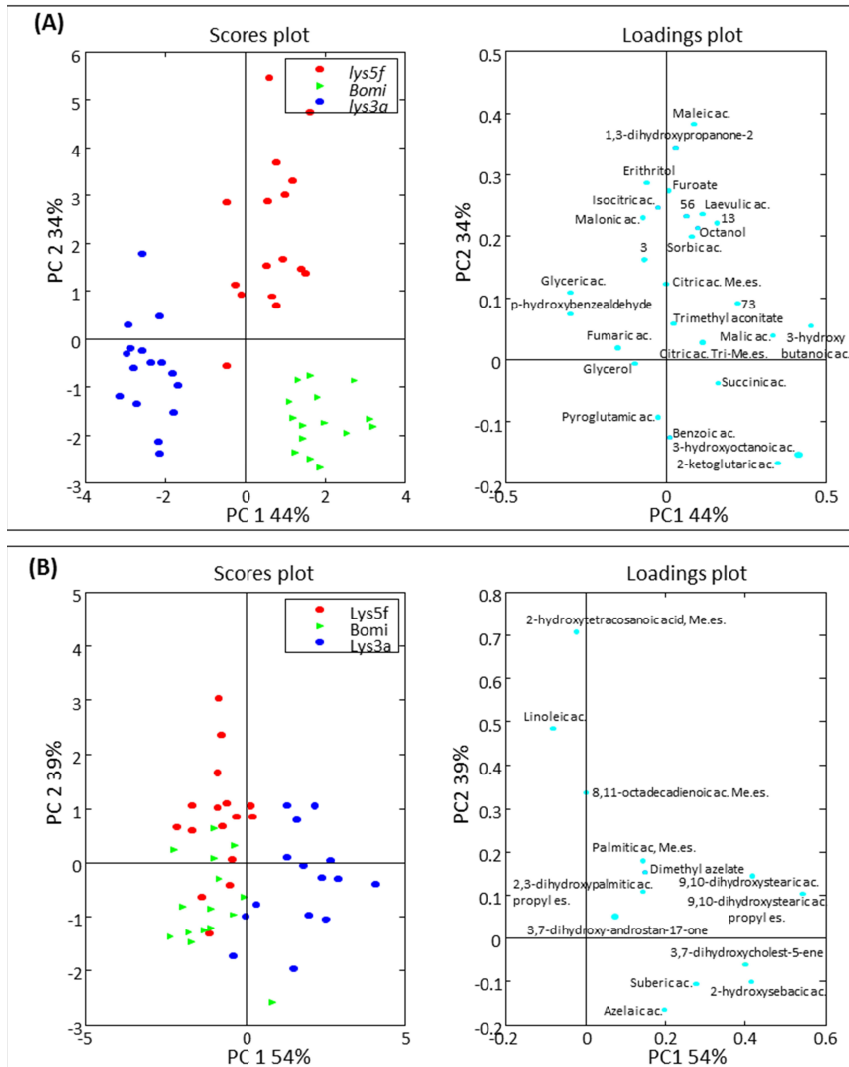
**Figure S1.** TIC chromatogram obtained from GC-MS analysis of the whole-grain flour of the barley genotype *Bomi*, grown under high temperature and harvested at the last DAF point. \* The sample, Bomi\_HT\_DAF47, used for calculation of metabolites relative ratios at the harvest time (Supplementary Table S1).



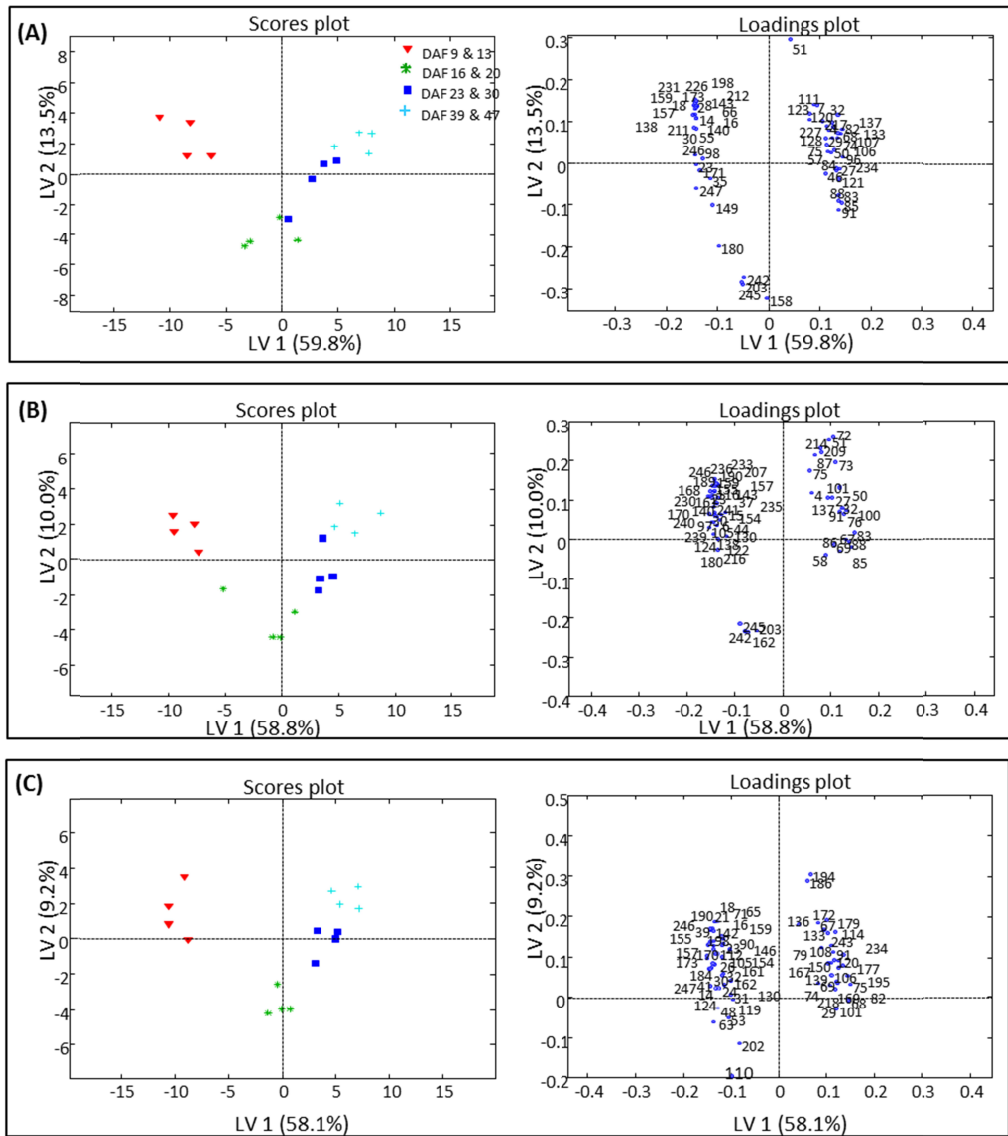
**Figure S2.** Scores plots of the PCA model developed on the initial metabolomics data matrix **X**. **(A)** Samples are colored according to their DAF, **(B)** Samples are colored according to barley genotypes and **(C)** samples are colored by two different growth temperatures, low (LT) and high (HT).



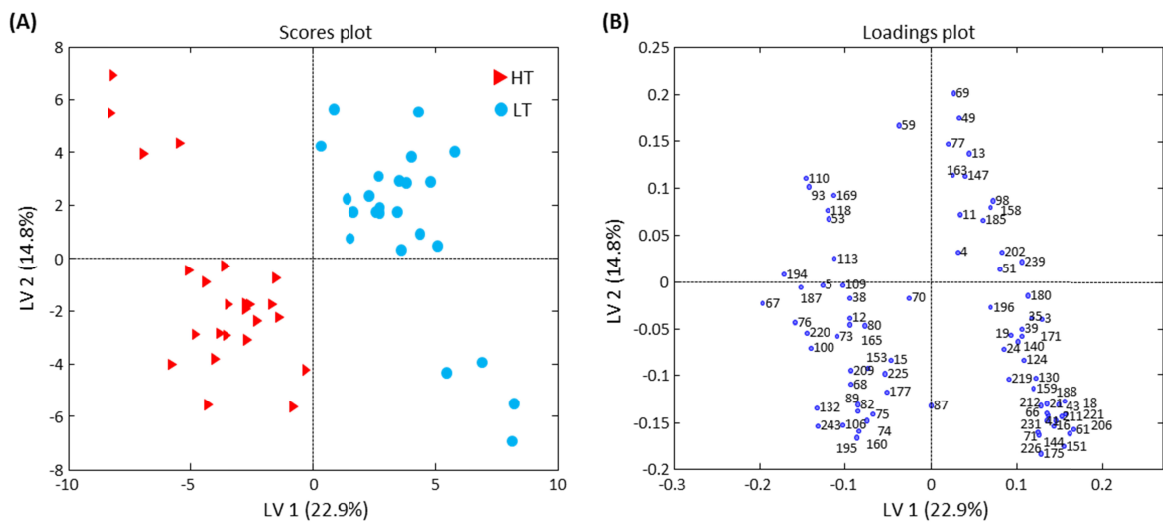
**Figure S3.** Scores **(A)** and loadings **(B)** plots of the PLS model developed on the days after flowering (DAF) effect separated metabolomics data matrix ( $\mathbf{X}_{\text{DAF}}$ ).



**Figure S4.** PCA scores and loadings plot of genotype effect separated data,  $X_{GT}$ , including barley samples from 23 to 47DAF only. **(A)** PCA model developed on 28 small organic acids and alcohols and **(B)** 13 most abundant fatty acid and their esters.



**Figure S5.** Scores and loadings plots of the PLS models developed on the DAF effect separated metabolomics data matrix ( $X_{DAF}$ ), of the individual barley genotypes, lys5.f (A), Bomi (B), and lys3.a (C). \* PLS models include only most important metabolites for predicting DAF that showed VIP scores < 1. Metabolite numbers on the loadings plot correspond to their number in Supplementary Table S1.



**Figure S6.** (A) Scores and (B) loadings plots of the PLS-DA model developed on the temperature effect separated metabolomics data matrix ( $X_{TEMP}$ ). \* The PLS-DA model include only most important metabolites that showed VIP scores < 1, for classifying HT plants from LT barley plants. Metabolite numbers on the loadings plot correspond to their number in Supplementary Table S1.



**Supplementary Table S1.** Relative ratios of the metabolites at the harvest time (47 days after flowering) to the level of metabolites detected from the mother line barley grown in high temperature (*Bomi\_HT*). The table also comprises retention times (RT), reported retention indices (RI(r)), calculated retention indices (RI(c)), EI-MS based library search (LS) by using NIST05 and Wiley08. Mass spectra of the unknown metabolites can be obtained by contacting the corresponding author (Bekzod Khakimov (bzo@life.ku.dk)).

\* Common metabolomic alterations, related to barley seed developmental stages and growth temperature observed for all three barley genotypes are highlighted by **a-e** indices.

a. Metabolites increased during the grain-filling period

b. Metabolites decreased during the grain-filling period

c. Metabolites significantly increased after two weeks of grain-filling period and gradually decreased after the third week of the grain-filing.

d. Metabolites with higher concentrations in barley lines grown under high temperature (HT) conditions compared to the low temperature (LT) grown barley lines.

e. Metabolites with higher concentrations in barley lines grown in under low temperature (LT) conditions compared to the high temperature (HT) grown barley lines.

f. Tentatively identified

No.	Metabolites	<i>lys5.f_LT</i>	<i>lys5.f_HT</i>	<i>lys3.a_LT</i>	<i>lys3.a_HT</i>	<i>Bomi_LT</i>	RT(min)	RI (r)	RI (c)	LS
1.	Laevulinic acid-1TMS (b)	0.43	2.05	0.6	0.6	0.43	9.04	1030	1070	96
2.	Sorbic acid-1TMS (b)	0.55	0.62	0.52	0.8	0.55	9.06	1009	1071	84
3.	Hepta-2,4-dienoic acid, methyl ester (b), (e)	4.34	1.42	1.46	0.94	4.34	9.28	1000	1080	82
4.	Octanol-1-1TMS (a), (e)	1.06	1.69	1.22	0.53	1.06	9.51	1101	1090	79
5.	Unk5 (d)	0.87	1.3	1.06	1.65	0.87	9.48			
6.	Unk6	0.79	0.38	1.19	0	0.79	9.69			
7.	Unk7	1.05	1.48	1.9	0.86	1.05	9.76		1097	
8.	Unk8 (b)	0.82	1.71	0.85	1.11	0.82	9.74			
9.	Malonic acid-2TMS	1.12	2.01	1.19	1.49	1.12	9.99	1205	1207	84
10.	(3,3-Dimethyl-1-cyclohexen-1-yl)oxy]-1TMS (c)	2.54	0.86	0.77	1.46	2.54	9.97	1110	1206	73
11.	Unk11 (e)	1.29	1.43	1.42	1.21	1.29	10.15			
12.	Unk12 (d)	0.71	2.34	3.64	2.89	0.71	10.18			
13.	Benzoic acid-1TMS (e)	1.19	0.85	1.28	0.71	1.19	10.42	1228	1226	91
14.	3-methylfuran-2-carboxylic acid, -1TMS (b)	0.55	0.57	0.54	0.66	0.55	10.38	1107	1224	91
15.	Unk15 (d)	1.25	1.58	1.16	1.47	1.25	10.63			
16.	Unk16 (b), (e)	0.78	1.01	0.92	0.81	0.78	10.60			
17.	Glycerol-3TMS (c)	1.1	1.53	0.86	1.5	1.1	10.88	1282	1246	84
18.	1,3-dihydroxypropanone-2-2TMS (b), (e)	1.61	1.09	0.83	1.03	1.61	11.03		1249	81
19.	Unk19 (b), (e)	1.47	1.9	1.39	1.23	1.47	11.02			
20.	Unk20 (c)	0.94	1.04	0.76	1.28	0.94	11.05			
21.	Unk21 (b), (e)	2.57	3.43	0.96	0.89	2.57	11.19		1259	
22.	Succinic acid-2TMS	0.56	1.14	0.86	0.91	0.56	11.24	1292	1262	82
23.	Glyceric acid-3TMS (b)	3.04	1.45	1.2	3.51	3.04	11.51	1199	1274	93
24.	Maleic acid-2TMS (b), (e)	1.25	3.51	0.54	1	1.25	11.55	1286	1275	75
25.	Fumaric acid-2TMS (b)	0.69	1.27	2.53	1.61	0.69	11.60	1178	1278	93
26.	Unk26	1	1.52	0.58	0.66	1	11.67		1280	
27.	Unk27 (a)	0.87	1.75	1	0.52	0.87	11.70		1282	
28.	p-hydroxybenzaldehyde -1TMS (a)	0.82	0.72	1.15	0.34	0.82	11.85	1280	1289	93
29.	2-hydroxyheptanoic acid-2TMS (a)	2.53	5.82	2.42	6.59	2.53	11.83	1312	1288	91
30.	Unk30 (b)	1.76	0.3	0.46	1.17	1.76	11.82		1287	
31.	Unk31	1.73	0.62	0.56	0.32	1.73	11.91		1291	
32.	3-hydroxybutanoic acid-2TMS (a)	0.78	0.88	0.37	0.19	0.78	12.12	1403	1401	79
33.	Resorcinol-2TMS	0.97	1.06	1.32	1.05	0.97	12.2	1378	1404	79
34.	Unk34 (c)	0.89	1.79	1.04	2.35	0.89	12.28		1409	
35.	Unk35 (e)	1.79	0.19	1.2	0.45	1.79	12.48		1418	
36.	Trimethyl aconitate (c)	1.44	0.25	1.18	0.15	1.44	12.50	1428	1419	90
37.	Unk37	2.54	1.62	1.66	2.03	2.54	12.46		1417	

## Supplementary Data of Khakimov et al., 2013 (submitted to JXB)

38.	Unk38 (d)	1.26	1.53	1.02	1.22	1.26	12.55		1422	
39.	Unk39 (b), (e)	2.34	1.71	0.57	0.7	2.34	12.57		1423	
40.	Unk40 (b)	0.66	0.91	0.95	0.84	0.66	12.6		1424	
41.	Unk41 (b), (e)	0.56	1.16	0.26	0.44	0.56	12.64		1426	
42.	Citric acid, trimethyl ester	2.45	0.96	1.32	0.22	2.45	12.82	1442	1435	99
43.	3-hydroxyanthranilic acid, methyl ester-1TMS (b), (e)	3.38	1.13	0.5	1.16	3.38	12.8		1434	85
44.	2,4-dihydroxy-5-methylpyrimidine, -2TMS	1.16	1.38	0.93	1.16	1.16	12.89	1403	1439	77
45.	5-hydroxy-2-(hydroxymethyl)-4H-pyran-4-one-2TMS (c)	2.01	1.42	1.07	1.29	2.01	13.08	1492	1448	72
46.	Maseptol-1TMS (a)	1	1.35	1.53	0.77	1	13.12	1358	1450	71
47.	Unk47	1.32	0.85	1.15	0.99	1.32	13.1		1449	
48.	Malic acid-2TMS	0.94	0.52	0.44	0.4	0.94	13.19	1494	1453	94
49.	Unk49 (a), (e)	1.33	0.97	1.1	0.69	1.33	13.27		1458	
50.	2-hydroxycyclohexane-1-carboxylic acid-2TMS (a)	0.92	2.22	0.75	0.58	0.92	13.23	1402	1456	71
51.	3-hydroxyoctanoic acid-2TMS (e)	0.43	0.55	0.12	0.1	0.43	13.35	1452	1462	78
52.	Unk52	3.91	2.08	0.31	2.49	3.91	13.33		1461	
53.	Pyroglutamic acid-2TMS (d)	0.78	0.37	1.2	1.79	0.78	13.46	1466	1467	96
54.	Erythritol-4TMS	2.53	1.6	1.16	1.75	2.53	13.47		1467	83
55.	Unk55 (b)	1.72	1.64	0.85	0.49	1.72	13.50		1469	
56.	Dimethyl azelate	3.13	3.52	1.68	0.77	3.13	13.61	1485	1474	90
57.	4-hydroxybenzeneacetic acid, methyl ester-1TMS (a)	0.74	1.27	0.67	0.39	0.74	13.62	1458	1475	92
58.	Vanillin -1TMS (a)	0.9	0.98	1.03	0.32	0.9	13.55	1469	1471	98
59.	Unk59	2.17	2	1.86	1.13	2.17	13.67		1477	
60.	Unk60 (a)	0.88	0.7	0.82	0.86	0.88	13.66		1477	
61.	Unk61 (b), (e)	2.7	3.18	2.29	1.48	2.7	13.58		1473	
62.	Citric acid, trimethyl ester-1TMS	0.96	0.72	0.96	0.15	0.96	13.76		1482	96
63.	2-Furancarboxylic acid, 5-[(oxy)methyl]-1TMS (b)	1.58	0.99	0.84	1.44	1.58	13.72	1540	1480	97
64.	Unk64	1.46	1.03	0.91	0.89	1.46	13.84		1486	
65.	Unk65 (b)	1.14	0.68	1.06	0.75	1.14	13.87		1487	
66.	Unk66 (b), (e)	1.75	1.84	1.29	1.12	1.75	13.82		1485	
67.	4-Hydroxyphenylethanol-2TMS (a), (d)	0.75	1.35	1.05	1.2	0.75	13.92	1475	1490	97
68.	Unk68 (a), (d)	1.49	2.11	1.5	2.12	1.49	13.93		1490	
69.	Unk69 (a), (e)	1.12	0.76	0.78	0.14	1.12	14.08		1498	
70.	Anozol	0.59	1.14	0.25	0.31	0.59	14.15	1603	1601	93
71.	Unk71 (b), (e)	2.64	2.29	0.42	0.57	2.64	14.16		1602	
72.	2-Ketoglutaric acid-3TMS	0.21	0.78	0.13	0.28	0.21	14.34	1622	1612	98
73.	3-methyl-3-hydroxypentanedioic acid-3TMS (a), (d)	0.43	1.54	0.79	1	0.43	14.3	1610	1609	95
74.	Unk74 (a), (d)	1.69	1.92	2	7.97	1.69	14.32		1610	
75.	Dodecane-6-hydroxy-1TMS (a), (d)	1.61	2.07	1.68	2.67	1.61	14.40	1631	1615	88
76.	p-Salicylic acid -2TMS (a), (d)	0.59	0.81	1.41	1.26	0.59	14.45	1618	1618	96
77.	Unk77 (e)	1.23	0.93	1.26	0.97	1.23	14.57		1624	
78.	Methyl Isovanillate-1TMS	0.7	1.21	1.2	0.33	0.7	14.66	1547	1629	90
79.	Unk79 (a)	1.98	0.38	1.41	0.37	1.98	14.80		1637	
80.	Unk80 (c), (d)	0.84	0.6	1.47	0.82	0.84	14.77		1635	
81.	Unk81	1.08	0.99	1.62	0.88	1.08	14.88		1641	
82.	Unk82 (a), (d)	1.64	2.51	1.7	2.87	1.64	14.86		1640	
83.	Unk83 (a)	1.06	1.13	0.93	0.78	1.06	14.89		1642	
84.	Unk84 (a)	1.25	1.08	1.12	0.91	1.25	14.95		1645	
85.	Unk85 (a)	1	1.13	0.89	0.58	1	14.98		1647	
86.	Unk86 (a)	0.97	1.23	0.78	0.76	0.97	14.99		1647	
87.	Unk87	0.41	0.31	0.15	0.2	0.41	15.01		1648	
88.	Unk88 (a)	0.96	1.21	0.97	0.62	0.96	15.08		1652	
89.	Suberic acid-2TMS (a), (d)	0.8	1.3	1.33	1.96	0.8	15.11	1682	1654	92
90.	Unk90 (b)	0.83	1.73	0.6	0.29	0.83	15.12		1654	
91.	Syringaldehyde -1TMS (a)	0.88	0.92	0.9	0.67	0.88	15.15	1658	1656	88
92.	$\beta$ -D-Arabinopyranose-4TMS (a)	1.14	0.71	1.01	1.03	1.14	15.23	1692	1660	80
93.	$\beta$ -D-Xylopyranose-4TMS (d)	1.65	1.41	0.88	1.15	1.65	15.30	1694	1664	79

Supplementary Data of Khakimov et al., 2013 (submitted to JXB)

94.	Unk94	0.89	0.68	0.86	0.81	0.89	15.26		1662	
95.	Unk95 (c)	1.16	0.31	1.14	0.59	1.16	15.28		1663	
96.	Protocatechuic acid, methyl ester -2TMS (a)	0.74	0.69	0.91	0.86	0.74	15.35	1656	1667	92
97.	Unk97	1.96	1.82	0.71	1.79	1.96	15.38		1669	
98.	2,5-dimethoxymandelic acid-2TMS (e)	1.08	0.38	1.31	0.14	1.08	15.38	1867	1669	75
99.	Unk99	0.7	2.46	1.52	1.34	0.7	15.53		1677	
100.	Vanillic acid-2TMS (a), (d)	0.44	0.96	1.04	1.2	0.44	15.72	1656	1687	98
101.	Unk101 (a)	1.13	1.91	1.08	1.42	1.13	15.74		1688	
102.	Unk102 (b)	0.75	0	0.98	0.24	0.75	15.69		1686	
103.	Unk103	0.66	1.25	1.76	1.27	0.66	15.80		1692	
104.	Unk104	0.59	6.2	2.32	2.19	0.59	15.79		1691	
105.	4-hydroxycinnamic acid, methyl ester -1TMS (b)	1.47	1.09	1.2	0.78	1.47	15.88	1565	1696	97
106.	Azelaic acid-2TMS (a), (d)	0.66	2.76	1.38	5.2	0.66	15.98	1800	1802	83
107.	2,3-dihydroxyphosphoric acid, propyl ester-4TMS (a)	2	1.33	0.73	1.35	2	15.86	1708	1695	97
108.	Unk108	1.53	0.89	2.29	1.6	1.53	15.94		1699	
109.	Unk109 (d)	0.79	1.08	1.01	0.92	0.79	16.08		1808	
110.	Methyl 2-(oxy)-2-(4-(oxy)phenyl)propanoate-2TMS (d)	1.25	2.15	0.7	0.58	1.25	16.14	1757	1811	85
111.	$\alpha$ -D-Galactofuranoside, methyl-2,3,5,6-tetrakis-4TMS (a)	1.46	0.43	0.64	0.89	1.46	16.11	1845	1810	81
112.	Unk112	0.41	1.95	0.71	0.92	0.41	16.03		1805	
113.	Protocatechuic acid-3TMS (a), (d)	0.67	0.94	0.9	1.12	0.67	16.24	1826	1818	95
114.	$\alpha$ -Resorcylic acid -3TMS (a)	0.68	0.74	1.45	1.02	0.68	16.20	1826	1815	86
115.	D-Fructose-5TMS	26.5	2.17	0.37	37.7	26.5	16.41	1867	1828	90
116.	Isocitric acid-4TMS	0.61	2.81	1.71	2.48	0.61	16.34	1835	1823	91
117.	Catechin-nTMS	1.17	0.8	0.97	1.18	1.17	16.44		1830	81
118.	Homovanillic acid -2TMS (d)	1.34	2.49	1.87	1.42	1.34	16.4	1867	1827	75
119.	Unk119	1.08	1.01	0.59	0.56	1.08	16.51		1834	
120.	Unk120 (a)	2.57	2.1	2.02	1.78	2.57	16.49		1833	
121.	Unk121 (a)	1.35	1.22	1.16	0.94	1.35				
122.	Unk122	0.64	2.7	0.8	0.69	0.64				
123.	$\beta$ -D-Galactopyranoside, methyl 2,3,4,6-tetrakis-4TMS	2.07	0.86	0.8	1.41	2.07	16.68	1900	1844	85
124.	Unk124 (b), (e)	3.51	0	1	0.68	3.51	16.63		1841	
125.	Unk125 (c)	2.16	0.44	3.36	3.36	2.16	16.65		1842	
126.	Catechin-nTMS	1.32	0.66	0.91	1.24	1.32	16.77		1849	82
127.	Gentisic acid -3TMS (b)	1.19	2.92	1.2	1.34	1.19	16.78	1796	1850	87
128.	Unk128 (a)	0.81	1.21	0.57	1.21	0.81	16.72		1846	
129.	Unk129 (b)	1.31	2.23	0.48	2.59	1.31	16.75		1848	
130.	Unk130 (e)	1.71	1.02	1.22	0.65	1.71	16.76		1849	
131.	$\alpha$ -D-Glucopyranoside, methyl 2,3,4,6-tetrakis-4TMS	2.69	1.23	0.63	1.11	2.69	16.90	1928	1857	94
132.	Syringic acid-2TMS (a), (d)	0.47	1.26	0.95	1.6	0.47	16.88	1845	1856	96
133.	Unk133 (a)	0.84	0.48	0.92	0.66	0.84	16.88		1856	
134.	$\beta$ -D-Glucopyranoside, methyl 2,3,4,6-tetrakis-4TMS	3.36	1.22	0.63	1.29	3.36	17.05	1928	1866	91
135.	$\alpha$ -D-Glucopyranose, 1,2,3,4,6-pentakis-5TMS	1.5	1.19	0.41	1.65	1.5	17.02	1924	1864	93
136.	Palmitic acid, methyl ester	3.31	2.75	1.09	24.5	3.31	17.01	1870	1864	97
137.	D-Galactose, 2,3,4,5,6-pentakis-5TMS (a)	1.84	0.63	0.51	1.54	1.84	17.12	1970	1871	97
138.	Unk138 (b)	3.33	2.75	1.92	1.27	3.33	17.13		1871	
139.	Unk139 (a)	0.54	0.96	38.8	10.8	0.54	17.1		1869	
140.	p-Coumaric acid-2TMS (b), (e)	0.96	1.3	1.15	1.23	0.96	17.18	1924	1874	90
141.	Ferulic acid, methyl ester-1TMS	2.01	1.21	1.02	0.76	2.01	17.25	1765	1878	87
142.	Unk142 (b)	1.35	2.65	0.97	1.34	1.35	17.35		1884	
143.	Unk143 (b)	0.77	1.98	0.73	2.13	0.77	17.39		1887	
144.	Unk144 (b), (e)	4.84	8.48	1.91	8.02	4.84	17.40		1888	
145.	Gallic acid-4TMS (a)	0.63	0.74	1.01	1.22	0.63	17.45	1976	1890	96

## Supplementary Data of Khakimov et al., 2013 (submitted to JXB)

146.	Unk146	0.62	2.73	0.82	1.27	0.62	17.43		1889	
147.	2-hydroxymandelic acid, ethyl ester-2TMS (a), (e)	0.28	0.31	0.62	0.57	0.28	17.34	1777	1884	75
148.	4'-Cyclohexylacetophenone (b)	2.64	1.21	1.47	1.26	2.64	17.58	1703	1898	71
149.	Unk149 (b)	1.34	1.09	0.63	1.46	1.34	17.65		2003	
150.	Unk150	3.27	0.97	2.09	1.43	3.27	17.61		2000	
151.	Unk151 (b), (e)	2.1	2.23	1.39	2.26	2.1	17.69		2005	
152.	Caffeic acid methyl ester - 2TMS	1.91	2.04	0.91	0.8	1.91	17.76	1863	2010	96
153.	$\beta$ -D-Glucopyranose-5TMS (d)	1.96	1.45	0.33	2.15	1.96	17.75	1970	2009	87
154.	Unk154 (b)	0	0	0	0	0	17.76		2010	
155.	Unk155 (b)	0.53	0.36	0.76	0.68	0.53	17.9		2019	
156.	Unk156	0.32	1.44	0.31	0.61	0.32	18.00		2026	
157.	Unk157 (c)	1.61	5.63	0.64	1.87	1.61	17.97		2024	
158.	Unk158 (c), (e)	1.15	1.45	0.84	1.03	1.15	18.05		2029	
159.	Unk159 (b), (e)	1.87	1.79	0.36	1.14	1.87	18.07		2030	
160.	2-hydroxysebacic acid-3TMS (a), (d)	1.03	1.72	1.69	4.05	1.03	18.13	2059	2034	88
161.	Unk161 (b)	2.65	0	0.5	4.49	2.65	18.16		2036	
162.	Unk162	0.58	0.98	0.84	0.57	0.58	18.15		2036	
163.	Unk163 (e)	1.57	0.66	0.56	0.69	1.57	18.24		2041	
164.	Unk164 (c)	1.63	1.86	1.58	2.65	1.63	18.20		2039	
165.	Ferulic acid-2TMS (d)	1.07	1.42	1.11	1.4	1.07	18.40	2076	2052	92
166.	8,11-octadecadienoic acid, methyl ester (c)	2.11	1.91	0.83	8.33	2.11	18.35	2093	2049	93
167.	Unk167 (a)	2.8	1.13	2.01	2.98	2.8	18.33		2047	
168.	Unk168 (b)	6.84	1.3	0.64	1.22	6.84	18.62		2066	
169.	Sinapinic acid methyl ester-1TMS (d)	1.81	1.2	1.85	1.87	1.81	18.51	1943	2059	96
170.	Unk170 (b)	1.11	1.41	1.31	0.85	1.11	18.70		2072	
171.	Methyl vanillactate-2TMS (b), (e)	1.58	0.98	0.69	1.62	1.58	18.55	2030	2062	81
172.	Caffeic acid-3TMS	1.06	2.89	0.94	1.42	1.06	18.76	2114	2076	98
173.	9-methoxy-4 $\alpha$ -methyl-2,3,7-trihydroxy-4,4a-dihydro-2H-benzo[c]chromen-6(3H)-one (b)(f)	2.06	4.28	0.73	1.6	2.06	18.85		2082	76
174.	Unk174	0.89	0.87	1.08	0.82	0.89	18.96		2089	
175.	Unk175 (b), (e)	0.85	1.68	1.12	1.23	0.85	18.96		2089	
176.	Unk176	1.59	0.55	0.57	0.31	1.59	19.01		2092	
177.	Unk177 (a), (d)	1.49	1.3	1.91	3.42	1.49	19.00		2092	
178.	Unk178	1.06	0.83	0.61	0.52	1.06	19.10		2098	
179.	Unk179 (a)	1.84	1.34	2.05	3.98	1.84	19.18		2204	
180.	Unk180 (c), (e)	0.96	0.37	0.99	0.85	0.96	19.15		2201	
181.	Linoleic acid-1TMS	1.18	1.52	0.31	5.19	1.18	19.23	2202	2207	92
182.	Unk182	2.24	2.9	1.78	1.64	2.24	19.35		2216	
183.	Unk183	1.19	0.71	0.76	0.36	1.19	19.25		2209	
184.	Unk184 (b)	1.15	3.25	0.86	1.23	1.15	19.27		2210	
185.	Unk185 (c), (e)	0.85	0.3	0.8	1.18	0.85	19.38		2218	
186.	4,8-dihydroxy-2-quinolinecarboxylic acid-3TMS (a)	0.69	0.35	0.92	0.73	0.69	19.46	2265	2224	89
187.	Sinapinic acid -2TMS (d)	1	1.48	1.32	2.64	1	19.52	2221	2228	98
188.	Unk188 (b), (e)	1.51	1.89	0.92	2.23	1.51	19.53		2229	
189.	Unk189 (b)	2.35	3.55	0.86	1.67	2.35	19.52		2228	
190.	Unk190 (b)	2.19	3.55	1.63	1.55	2.19	19.68		2239	
191.	Unk191 (a)	1.04	0.34	1.04	0.45	1.04	19.71		2241	
192.	Unk192	1.12	0.57	0.82	0.9	1.12	19.65		2237	
193.	Unk193	0.83	0.96	0.93	1.51	0.83	19.63		2236	
194.	Unk194 (d)	0.87	1.47	2.36	2.91	0.87	19.81		2249	
195.	Unk195 (a), (d)	0.96	1.4	1.58	4.56	0.96	19.78		2246	
196.	Unk196 (e)	1.35	0.99	0.6	1.04	1.35	19.80		2248	
197.	Unk197 (b)	1.33	0.88	1.19	1.36	1.33	19.92		2256	
198.	Androsterone type plant sterol (b), (f)	0.85	2.11	0.68	1.34	0.85	19.89		2254	
199.	Unk199 (b)	5.07	1.37	0.68	1.5	5.07	19.90		2255	
200.	3-hydroxyandrostan-17-one-1TMS	1.19	1.28	1.02	1.34	1.19	19.98	2186	2261	78
201.	Unk201 (b)	3.56	1.36	0.63	1.33	3.56	20.02		2264	
202.	Unk202 (c), (e)	2.3	1.87	0.61	1.35	2.3	20.05		2266	

## Supplementary Data of Khakimov et al., 2013 (submitted to JXB)

203.	Unk203 (c)	0.77	0.93	0.5	1.48	0.77	20.13		2271	
204.	Unk204	2.13	5.89	2.15	1.67	2.13	20.11		2270	
205.	Unk205 (b)	1.48	2.89	1.35	1.77	1.48	20.11		2270	
206.	Unk206 (b), (e)	3.96	3.05	0.7	2.3	3.96	20.21		2277	
207.	Unk207 (b)	2.18	1.42	1.05	1.11	2.18	20.33		2286	
208.	Unk208 (c)	0.77	0.25	0.65	1.58	0.77	20.31		2284	
209.	19-Norandrosterone-3-TMS (a), (d), (f)	0.82	1	0.75	1.68	0.82	20.36	2198	2288	79
210.	Unk210	1.18	1.09	0.37	0.57	1.18	20.39		2290	
211.	Unk211 (b), (e)	1.97	1.42	0.67	1.39	1.97	20.44		2294	
212.	Unk212 (b), (e)	1.61	1.55	0.72	1.47	1.61	20.50		2298	
213.	Unk213 (b)	74.2	3.29	0.64	3.45	74.2	20.57		2403	
214.	Unk214 (a)	0.76	1.69	0.52	1.7	0.76	20.28		2282	
215.	Unk215	1.85	1.83	0.9	3.46	1.85	20.57		2403	
216.	Unk216	45.4	2.14	1.75	2.68	45.4	20.69		2412	
217.	Unk217	1.3	0.39	0.74	0.98	1.3	20.84		2424	
218.	9,10-dihydroxystearic acid-3TMS	4.18	0.43	1.5	3.7	4.18	20.87	2517	2426	85
219.	Unk219 (b), (e)	2.38	3.99	0.59	1.95	2.38	20.99		2435	
220.	3,7-di-hydroxy-androstan-17-one -2TMS (d)	0.48	1.19	0.97	3.11	0.48	21.09	2432	2443	93
221.	Unk221 (e)	1.3	0.75	0.43	2.06	1.3	21.16		2449	
222.	Unk222	1.81	0.95	0.36	1.98	1.81	21.19		2451	
223.	Unk223	51.6	3.74	1.81	1.27	51.6	21.4		2467	
224.	Unk224 (b)	1.35	1.92	0.48	1.8	1.35	21.38		2466	
225.	Unk225 (d)	1	3.73	0.25	7.1	1	21.34		2463	
226.	Unk226 (b), (e)	1.89	2.87	0.53	2.47	1.89	21.51		2476	
227.	Unk227	0.82	0.56	0.83	1.27	0.82	21.54		2478	
228.	Unk228 (b)	9.14	3.64	0.56	5.97	9.14	21.45		2471	
229.	Unk229	0.78	0.77	0.86	0.77	0.78	21.51		2476	
230.	9,10- dihydroxystearic acid, dimethyl ester-2TMS	3.09	0.99	2.8	5.27	3.09	21.49	2784	2474	70
231.	Unk231 (b), (e)	3.31	3.85	0.22	3.78	3.31	21.65		2486	
232.	2,3-dihydroxypalmitic acid, propyl ester-2TMS	2.23	1.46	1.25	10.8	2.23	21.84	2581	2601	98
233.	Unk233	3.44	1.65	0.6	6.14	3.44	21.82		2499	
234.	Unk234 (a)	0.59	0.43	1.28	1.84	0.59	21.87		2604	
235.	Unk235 (b)	3.95	2.31	1.25	2.97	3.95	21.96		2611	
236.	Unk236 (b)	1.34	1.34	0.99	2.86	1.34	22.01		2615	
237.	Unk237	0.62	0	0.04	9.94	0.62	22.42		2649	
238.	Unk238	3.93	0.6	1.12	1.53	3.93	22.59		2663	
239.	Unk239 (e)	1.34	1.04	1.07	1.08	1.34	23.22			
240.	2-Deoxy-6-phosphogluconolactone-5TMS	6.01	0.95	1.03	5.75	6.01	23.26		2820	77
241.	Unk241	11.5	1.81	0.64	13.5	11.5	23.36		2829	
242.	Unk242 (c)	1.1	0.15	1.67	2.03	1.1	23.57		2847	
243.	Unk243 (a), (d)	0.67	0.85	2.5	5.2	0.67	23.58		2848	
244.	2-hydroxytetraacosanoic acid, methyl ester-1TMS	1.64	4.91	0.63	7.86	1.64	23.69	2894	2858	98
245.	3,7-dihydroxycholest-5-ene-2TMS (c)	1.13	0.58	1.63	2.33	1.13	23.95	2900	2881	81
246.	Unk246 (b)	1.13	1.2	1.35	2.16	1.13	24.6		3041	
247.	Unk247 (b)	0.67	0.73	0.98	2.11	0.67	24.69		3050	

# Supplementary

## Paper 5

Jörg M. Augustin, Sylvia Drok, Tetsuro Shinoda, Kazutsuka Sanmiya, Jens Kvist Nielsen, **Bekzod Khakimov**, Carl Erik Olsen, Esben Halkjær Hansen, Vera Kuzina, Claus Thorn Ekstrøm, Thure Hauser, and Søren Bak

UDP-Glycosyltransferases from the UGT73C Subfamily in *Barbarea vulgaris* Catalyze Sapogenin 3-O-Glucosylation in Saponin-Mediated Insect Resistance

*Plant Physiology*, December 2012, Vol. 160, pp. 1881–1895



# UDP-Glycosyltransferases from the UGT73C Subfamily in *Barbarea vulgaris* Catalyze Sapogenin 3-O-Glucosylation in Saponin-Mediated Insect Resistance<sup>1[W][OA]</sup>

Jörg M. Augustin<sup>2</sup>, Sylvia Drok, Tetsuro Shinoda<sup>3</sup>, Kazutsuka Sanmiya<sup>4</sup>, Jens Kvist Nielsen, Bekzod Khakimov, Carl Erik Olsen<sup>5</sup>, Esben Halkjær Hansen, Vera Kuzina<sup>5</sup>, Claus Thorn Ekstrøm<sup>6</sup>, Thure Hauser<sup>5</sup>, and Søren Bak<sup>5\*</sup>

Department of Plant Biology and Biotechnology (J.M.A., S.D., B.K., V.K., S.B.), Department of Basic Science and Environment (J.K.N., C.E.O., C.T.E.), Department of Food Science (B.K.), and Department of Agriculture and Ecology (J.K.N., T.H.), University of Copenhagen, 1871 Frederiksberg, Denmark; National Institute of Vegetable and Tea Science, National Agriculture and Food Research Organization, 514–2392 Tsu, Mie, Japan (T.S., K.S.); and Evolva A/S, 2100 Copenhagen, Denmark (E.H.H.)

Triterpenoid saponins are bioactive metabolites that have evolved recurrently in plants, presumably for defense. Their biosynthesis is poorly understood, as is the relationship between bioactivity and structure. *Barbarea vulgaris* is the only crucifer known to produce saponins. Hederagenin and oleanolic acid cellobioside make some *B. vulgaris* plants resistant to important insect pests, while other, susceptible plants produce different saponins. Resistance could be caused by glucosylation of the sapogenins. We identified four family 1 glycosyltransferases (UGTs) that catalyze 3-O-glucosylation of the sapogenins oleanolic acid and hederagenin. Among these, UGT73C10 and UGT73C11 show highest activity, substrate specificity and regiospecificity, and are under positive selection, while UGT73C12 and UGT73C13 show lower substrate specificity and regiospecificity and are under purifying selection. The expression of UGT73C10 and UGT73C11 in different *B. vulgaris* organs correlates with saponin abundance. Monoglucosylated hederagenin and oleanolic acid were produced in vitro and tested for effects on *P. nemorum*. 3-O- $\beta$ -D-Glc hederagenin strongly deterred feeding, while 3-O- $\beta$ -D-Glc oleanolic acid only had a minor effect, showing that hydroxylation of C23 is important for resistance to this herbivore. The closest homolog in *Arabidopsis thaliana*, UGT73C5, only showed weak activity toward sapogenins. This indicates that UGT73C10 and UGT73C11 have neofunctionalized to specifically glucosylate sapogenins at the C3 position and demonstrates that C3 monoglucosylation activates resistance. As the UGTs from both the resistant and susceptible types of *B. vulgaris* glucosylate sapogenins and are not located in the known quantitative trait loci for resistance, the difference between the susceptible and resistant plant types is determined at an earlier stage in saponin biosynthesis.

<sup>1</sup> This work was supported by the Danish Council for Independent Research, Technology, and Production Sciences (grant nos. 09–065899/FTP and 274–06–0370), by the Villum Kann Rasmussen Foundation to Pro-Active Plants, and by a PhD stipend from the Faculty of Life Sciences, University of Copenhagen (to J.M.A.).

<sup>2</sup> Present address: Donald Danforth Plant Science Center, St. Louis, MO 63132.

<sup>3</sup> Present address: Division of Insect Sciences, National Institute of Agrobiological Sciences, Tsukuba, 305–8634 Ibaraki, Japan.

<sup>4</sup> Present address: Department of Bioresources Engineering, Okinawa National College of Technology, Nago, 905–2192 Okinawa, Japan.

<sup>5</sup> Present address: Department of Plant and Environmental Sciences, University of Copenhagen, 1871 Frederiksberg, Denmark.

<sup>6</sup> Present address: Department of Biostatistics, University of Southern Denmark, 5000 Odense C, Denmark.

\* Corresponding author; e-mail bak@life.ku.dk.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors ([www.plantphysiol.org](http://www.plantphysiol.org)) is: Søren Bak (bak@life.ku.dk).

[W] The online version of this article contains Web-only data.

[OA] Open Access articles can be viewed online without a subscription. [www.plantphysiol.org/cgi/doi/10.1104/pp.112.202747](http://www.plantphysiol.org/cgi/doi/10.1104/pp.112.202747)

Triterpenoid saponins are a heterogeneous group of bioactive metabolites found in many species of the plant kingdom. The general conception is that saponins are involved in plant defense against antagonists such as fungi (Papadopoulou et al., 1999), mollusks (Nihei et al., 2005), and insects (Dowd et al., 2011). Saponins consist of a triterpenoid aglycone (sapogenin) linked to usually one or more sugar moieties. This combination of a hydrophobic sapogenin and hydrophilic sugars makes saponins amphiphilic and enables them to integrate into biological membrane systems. There, they form complexes with membrane sterols and reorganize the lipid bilayer, which may result in membrane damage (Augustin et al., 2011).

However, our knowledge of the biosynthesis of saponins, and the genes and enzymes involved, is limited. The current conception is that the precursor 2,3-oxidosqualene is cyclized to a limited number of core structures, which are subsequently decorated with functional groups, and finally activated by adding glycosyl groups (Augustin et al., 2011). These key steps are considered to be catalyzed by three multigene families: (1) oxidosqualene

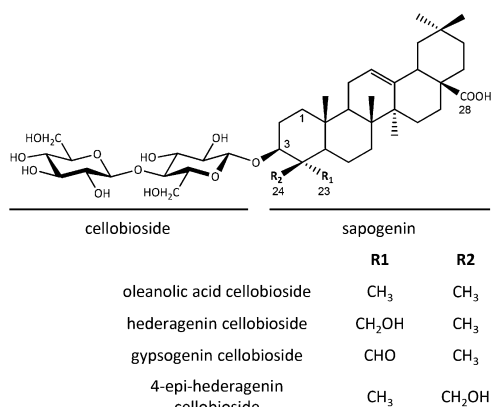


cyclases (OSCs) forming the core structures, (2) cytochromes P450 adding the majority of functional groups, and (3) family 1 glycosyltransferases (UGTs) adding sugars. This allows for a vast structural complexity, some of which probably evolved by sequential gene duplication followed by functional diversification (Osborn, 2010). A major challenge is thus to understand the processes of saponin biosynthesis, which structural variants of saponins play a role in defense against biotic antagonists, and how saponin biosynthesis evolved in different plant taxa. This knowledge is also of interest for biotechnological production and the use of saponins as protection agents against agricultural pests as well as for pharmacological and industrial uses as bactericides (De Leo et al., 2006), anticancerogens (Musende et al., 2009), and adjuvants (Sunt et al., 2009).

*Barbarea vulgaris* (winter cress) is a wild crucifer from the Cardamineae tribe of the Brassicaceae family. It is the only species in this economically important family known to produce saponins. *B. vulgaris* has further diverged into two separate evolutionary lineages (types; Hauser et al., 2012; Toneatto et al., 2012) that produce different saponins, glucosinolates, and flavonoids (Agerbirk et al., 2003b; Dalby-Brown et al., 2011; Kuzina et al., 2011). Saponins of the one plant type make plants resistant to the yellow-striped flea beetle (*Phyllotreta nemorum*), diamondback moth (*Plutella xylostella*), and other important crucifer specialist herbivores (Renwick, 2002); therefore, it has been suggested to utilize such plants as a trap crop to diminish insect damage (Badenes-Perez et al., 2005). The other plant type is not resistant to these herbivores. *B. vulgaris*, therefore, is ideal as a model species to study saponin biosynthesis, insect resistance, and its evolution, as we can contrast genes, enzymes, and their products between closely related but divergent plant types.

Insect resistance of the one plant type, called G because it has glabrous leaves, correlates with the content of especially hederagenin cellobioside, oleanolic acid cellobioside, 4-epi-hederagenin cellobioside, and gypsogenin cellobioside (Shinoda et al., 2002; Agerbirk et al., 2003a; Kuzina et al., 2009; Fig. 1). These saponins are absent in the susceptible plant type, called P because it has pubescent leaves, which contains saponins of unknown structures and function (Kuzina et al., 2011). The saponins (aglycones) of the resistance-causing saponins hederagenin and oleanolic acid cellobioside do not deter feeding by *P. nemorum*, which highlights the importance of glycosylation of saponins for resistance (Nielsen et al., 2010). Therefore, the presence or absence of saponin glycosyltransferases could be a determining factor for the difference in resistance between the insect resistant G-type and the susceptible P-type of *B. vulgaris*.

Some *P. nemorum* genotypes are resistant to the saponin defense of *B. vulgaris* (Nielsen, 1997b, 1999). Resistance is coded by dominant R genes (Nielsen et al., 2010; Nielsen 2012): larvae and adults of resistant genotypes (RR or Rr) are able to feed on G-type foliage and utilize *B. vulgaris* as host plant (de Jong et al., 2009), whereas larvae of the susceptible genotype (rr) die and adult beetles stop feeding



**Figure 1.** Chemical structures of the four known G-type *B. vulgaris* saponins that correlate with resistance to *P. nemorum* and other herbivores. The cellobioside and sapogenin parts of the saponin are underlined, and relevant carbon positions are numbered.

on G-type foliage. Larvae and adults of all known *P. nemorum* genotypes can feed on P-type *B. vulgaris* (Fig. 2).

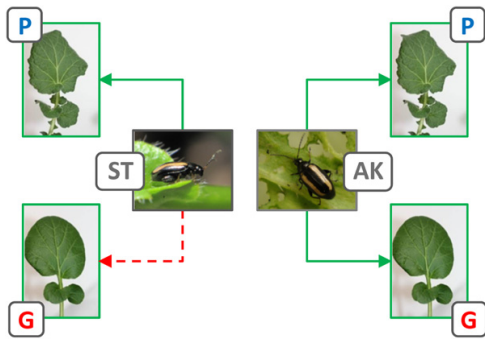
In this study, we asked which enzymes are involved in glucosylation of sapogenins in *B. vulgaris*, whether saponins with a single C3 glucosyl group are biologically active, and whether the difference between the insect resistant and susceptible types of *B. vulgaris* is caused by different glycosyltransferases.

We report the identification of two UDP-glycosyltransferases, UGT73C10 and UGT73C11, which have high catalytic activity and substrate specificity and regiospecificity for catalyzing 3-O-glucosylation of the sapogenins oleanolic acid and hederagenin. The products, 3-O-β-D-glucopyranosyl hederagenin and 3-O-β-D-glucopyranosyl oleanolic acid, are predicted precursors of hederagenin and oleanolic acid cellobioside, respectively. The expression patterns of UGT73C10 and UGT73C11 in different organs of *B. vulgaris* correlate with saponin abundance, and monoglucosylated sapogenins, especially 3-O-β-D-glucopyranosyl hederagenin, deter feeding by *P. nemorum*. Our results thus show that glucosylation with even a single glucosyl group activates the resistance function of these sapogenins. However, since the UGTs are present and active in both the insect-resistant and -susceptible types of *B. vulgaris*, we cannot explain the difference in resistance by different glucosylation abilities. Instead, the difference between the susceptible and resistant types must be determined at an earlier stage in saponin biosynthesis.

## RESULTS

### Identification of a Sapogenin UDP-Glycosyltransferase by Activity-Based Screening of a cDNA Expression Library

To identify enzymes that glycosylate sapogenins (aglycones of saponins) from *B. vulgaris*, a complementary



**Figure 2.** Feeding behavior of adult *P. nemorum* that are either susceptible (ST) or resistant (AK) toward the saponin-based defense of G-type *B. vulgaris*; the P-type produces different saponins and is not resistant against *P. nemorum*. Potential feeding is shown by green arrows, and termination of feeding briefly after initiation is indicated by a red dashed arrow. Larvae of the ST line die if fed on G-type plants.

DNA (cDNA) expression library was generated from *B. vulgaris* var *variegata*, a commercial *B. vulgaris* variety with a saponin profile similar to the insect-resistant G-type. The library was screened by activity assays using UDP-Glc and oleanolic acid as donor and acceptor substrate, respectively. A single cDNA clone was identified, of which the encoded enzyme glucosylated oleanolic acid, as evidenced by comigration with authentic 3-O-Glc oleanolic acid on thin-layer chromatography (TLC) analysis. The clone was designated *BvUGT1* and found to contain a 1,566-bp cDNA with an open reading frame (ORF) of 495 amino acids. BLAST analyses identified *Arabidopsis thaliana* UGT73C5 as its closest homolog. *BvUGT1* has 88% nucleotide identity to UGT73C5, and the encoded amino acid sequence, BvUGT1, is 83% identical to UGT73C5. In addition to oleanolic acid, BvUGT1 also glucosylated hederagenin and echinocystic acid.

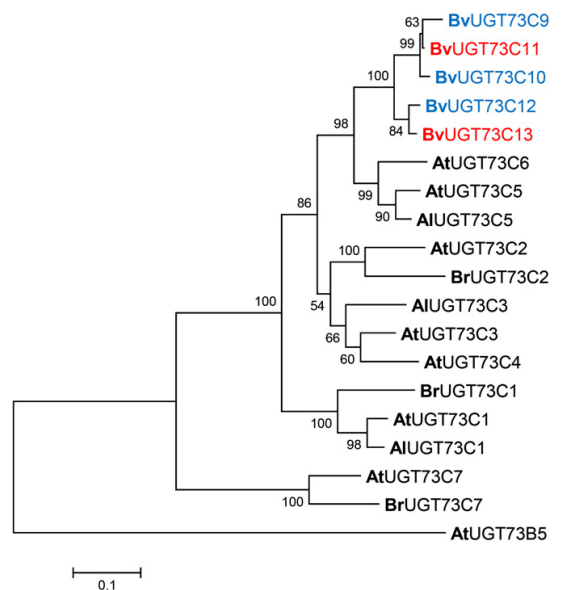
#### Identification of *BvUGT1* Homologs in G- and P-Type *B. vulgaris*

Putative *BvUGT1* homologs in the resistant G-type and susceptible P-type were searched by mining a 454 transcriptome data set from the G-type (Kuzina et al., 2011) and the P-type. Based on the identified singlets and contigs, two different full-length ORFs from G-type plants and three from P-type plants were isolated by PCR. The genomic sequences were identified by PCR and shown to be intronless, which is also the case for the seven UGT73Cs in the *A. thaliana* genome (Paquette et al., 2003). Thus, putative *BvUGT1* homologs are not only present in both the G- and P-type *B. vulgaris* genomes, but they are also expressed. The three P-type UGTs were named UGT73C9, UGT73C10, and UGT73C12, and the two G-type sequences were named UGT73C11 and UGT73C13 (Fig. 3), by the UGT

nomenclature committee (Mackenzie et al., 1997). The ORFs of the five UGTs each span 1,488 bp and encode proteins consisting of 495 amino acids.

Of the five sequences, UGT73C11 is most identical to *BvUGT1* from *B. vulgaris* var *variegata*, differing in only three nucleotides, which causes a conservative amino acid substitution of Asp-338 to Glu in UGT73C11. Based on a reconstruction of the phylogeny of the UGTs (Fig. 3), UGT73C9 and UGT73C10 from the P-type and UGT73C11 from the G-type form a discrete cluster, as does UGT73C12 from the P-type and UGT73C13 from the G-type. UGTs in the first cluster are more than 95% identical to each other, and those in the second cluster are more than 97% identical (Supplemental Table S1). Accordingly, UGT73C9/UGT73C10 from the P-type correspond to UGT73C11 from the G-type and UGT73C12 from the P-type corresponds to UGT73C13 from the G-type. In comparison with UGT73C homologs from *A. thaliana*, *Arabidopsis lyrata*, and *Brassica rapa*, the five *B. vulgaris* sequences are most closely related to *A. thaliana* UGT73C5 and UGT73C6 and a UGT73C5 homolog in *A. lyrata*.

The UGTs described in the phylogeny have been exposed to different levels of selection since they diverged, as indicated by the significantly better fit of a



**Figure 3.** Maximum likelihood phylogeny of UGT73Cs described in this study and from online databases. Species are indicated as prefixes to the UGT name: Bv, *B. vulgaris*; At, *A. thaliana*; Al, *A. lyrata*; Br, *B. rapa*. UGT73C9, UGT73C10, and UGT73C12, shown in blue, are from P-type *B. vulgaris*, while UGT73C11 and UGT73C13, shown in red, are from the G-type. AtUGT73B5 is included as an outgroup. Bootstrap values (100 iterations) are shown next to the corresponding nodes.

model with independent  $\omega$  (ratio of the number of nonsynonymous substitutions per nonsynonymous site to the number of synonymous substitutions per synonymous site [dN/dS ratios]) for each branch compared with a single common  $\omega$  ratio for all branches ( $2\Delta\ln L = 13.9$ ;  $P < 0.001$ ). Positive selection among branches was further indicated by the better fit of a model including positive selection (model M3) than a model without M0 ( $2\Delta\ln L = 304.7$ ;  $P < 0.001$ ); 4.3% of the codons were estimated to have been under positive selection. Only the branches leading to UGT73C9, UGT73C10, and UGT73C11 showed signs of positive selection; branches leading to UGT73C12 and UGT73C13 as well as *A. thaliana*, *A. lyrata*, and *B. rapa* have  $\omega < 1$ , showing that these branches are under purifying selection.

All five UGT sequences were mapped to an existing linkage map of *B. vulgaris* (Kuzina et al., 2011) and found to be located in a region that corresponds to *A. thaliana* chromosome 2 between 13.5 and 19.6 Mb. None of the UGTs lie within previously reported regions containing quantitative trait loci (QTL) for resistance toward *P. nemorum* larvae feeding (Kuzina et al., 2011). In *A. thaliana*, six out of the seven UGT73C genes are positioned in a tandem repeat cluster at 15.4 Mb on chromosome 2. Therefore, it is likely that the identified *B. vulgaris* UGT73C genes are located in a similar UGT73C cluster in the *B. vulgaris* genome.

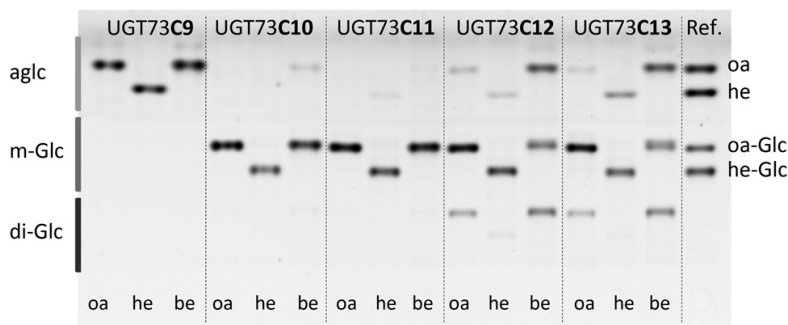
#### Heterologous Expression and in Vitro Activities of the UGT73Cs

To determine if the five UGTs isolated from G- and P-type *B. vulgaris* have similar catalytic activities as BvUGT1 from *B. vulgaris* var *variegata*, they were heterologously expressed in *Escherichia coli*. The corresponding crude protein extracts were assayed with

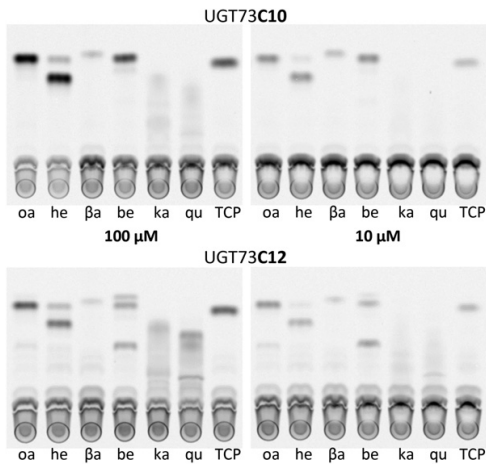
different sapogenins as putative sugar acceptors and UDP-Glc as sugar donor. UGT73C10, UGT73C11, UGT73C12, and UGT73C13 catalyzed transfer of a Glc moiety from UDP-Glc to the oleanane sapogenins oleanolic acid and hederagenin and to the lupane sapogenin betulinic acid (Fig. 4). In addition, their precursors  $\beta$ -amyrin and lupeol were glucosylated, but with lower efficiency (Fig. 5). In contrast, UGT73C9 from the P-type appeared inactive toward the compounds tested.

The glucosylation positions of the two oleanane sapogenins produced by the UGTs were determined by NMR spectroscopy. Based on one-dimensional (1-D)  $^1\text{H}$ - and  $^{13}\text{C}$ - as well as two-dimensional (2-D) Correlation Spectroscopy (COSY)-, Total Correlation Spectroscopy (TOCSY)-, and Heteronuclear Single Quantum Coherence (HSQC)-NMR analyses (Supplemental Data Set S1), the glucosides were concluded to be 3-O- $\beta$ -D-glucopyranosyl oleanolic acid and 3-O- $\beta$ -D-glucopyranosyl hederagenin. This is in agreement with these monoglucosides as predicted precursors of oleanolic acid cellobioside and hederagenin cellobioside, respectively.

In addition to the 3-O-monoglucosides, UGT73C12 and UGT73C13 also formed low amounts of diglucosides, while this activity was barely detectable for UGT73C10 and UGT73C11. Based on retention times and fragmentation patterns in liquid chromatography-mass spectrometry analyses, these diglucosides could not be oleanolic acid and hederagenin cellobioside, respectively, but represent bidesmosidic glucosylation (i.e. glycosylation at two different positions; Supplemental Fig. S1). A diglucosylated betulinic acid was, in addition to two different betulinic acid monoglucosides, produced in detectable amounts after 30 min of incubation when using betulinic acid concentrations as low as 10  $\mu\text{M}$  (Fig. 5). After alkaline hydrolysis (saponification), which cleaves the ester but not the ether bonds in



**Figure 4.** Activity of the heterologously expressed *B. vulgaris* UGT73Cs toward sapogenins. Enzyme assays contained 750 ng of recombinant UGT in 50  $\mu\text{L}$  and 50  $\mu\text{M}$  oleanolic acid (oa), hederagenin (he), or betulinic acid (be) as acceptor substrates and 1 mM UDP-Glc as donor substrate. The assays were incubated for 60 min at 30°C and analyzed by TLC. Compounds were visualized by spraying with 10% sulfuric acid in methanol and subsequent heating. The (inverted) image was taken at long-wave UV (366 nm) excitation. Migration of authentic oleanolic acid, hederagenin, 3-O- $\beta$ -Glc oleanolic acid (oa-Glc), and 3-O- $\beta$ -Glc hederagenin (he-Glc) is shown in the reference lane (Ref.). Positions of aglycones (aglc), monoglucosides (m-Glc), and diglucosides (di-Glc) are indicated on the left side.



**Figure 5.** Substrate specificity of UGT73C10 and UGT73C12. TLC analyses of activity assays with recombinant UGT73C10 or UGT73C12 using  $^{14}\text{C}$ -labeled UDP-Glc as donor substrate are shown. Substrates tested were oleanolic acid (oa), hederagenin\* (he),  $\beta$ -amyrin ( $\beta$ a), betulinic acid (be), kaempferol (ka), quercetin (qu), and TCP, applied at either 100 or 10  $\mu\text{M}$  concentration. \*The hederagenin batch contained a low amount of oleanolic acid.

glucosylated products, the betulinic acid diglucoside and one of the two betulinic acid monoglucosides were no longer detectable (Supplemental Fig. S2). Therefore, the degraded monoglucoside must be 28-*O*-glucosylated betulinic acid and the diglucoside must be 3,28-*O*-diglucosylated betulinic acid. Similarly, the diglucosidic forms of oleanolic acid and hederagenin would represent 3,28-*O*-diglucosides. Under assay conditions with high amounts of enzyme, increased incubation time, and elevated incubation temperature, UGT73C13 also produced an oleanolic acid triglucoside (Supplemental Fig. S1), which further demonstrates the lower substrate specificity and regioselectivity of UGT73C13. However, the low in vitro production of these glucosides suggests that these additional activities only play a minor role, if any, in plants.

Other members of the UGT73C subfamily have been assigned to be involved in flavonoid and brassinosteroid metabolism (Jones et al., 2003; Poppenberger et al., 2005; Modolo et al., 2007). Glycosylated flavonols derived from quercetin and kaempferol are present in *B. vulgaris* (Senatore et al., 2000; Dalby-Brown et al., 2011). Consequently, the flavonols quercetin and kaempferol, the phytosterols obtusifoliol, campesterol, sitosterol, and stigmasterol, and the brassinosteroid 24-epi-brassinolide were tested as substrates. 2,4,5-Trichlorophenol (TCP) was included as a positive control, as it can be glycosylated by several different plant UGTs (Messner et al., 2003; Brazier-Hicks and Edwards, 2005). Of the compounds tested, UGT73C9 only showed weak activity toward TCP when applied in 1 mM

concentration. In contrast, UGT73C10, UGT73C11, UGT73C12, and UGT73C13 glucosylated TCP at 10  $\mu\text{M}$  concentration (Fig. 5). The levels of oleanolic acid, hederagenin, and betulinic acid glucosides produced by these four UGTs were constantly higher than the levels of TCP glucosides, showing that sapogenins are better substrates. UGT73C10 and UGT73C11 showed weak activity toward quercetin and kaempferol at 100  $\mu\text{M}$  concentration, while at 10  $\mu\text{M}$ , glucosides could not be detected. In contrast, UGT73C12 and UGT73C13 clearly produced flavonol glucosides in assays with 100  $\mu\text{M}$  quercetin or kaempferol, while at 10  $\mu\text{M}$ , the glucosides were hardly detectable (Fig. 5). 24-Epi-brassinolide glucoside(s) were not observed with UGT73C11, whereas UGT73C13 catalyzed glucosylation of 24-epi-brassinolide to a product that comigrated with 24-epi-brassinolide glucoside, produced by *A. thaliana* UGT73C5 (Supplemental Fig. S3). None of the *B. vulgaris* UGTs glucosylated the phytosterols. *A. thaliana* UGT73B5 was included to represent a UGT73 from a different subfamily than UGT73C. UGT73B5 glucosylated TCP but neither of the sapogenins or other compounds tested (Supplemental Figs. S3 and S13).

UDP-Gal and UDP-GlcA were tested as alternative sugar donors. No glucuronides could be detected with any of the *B. vulgaris* UGTs when UDP-GlcA was used as sugar donor, but low activity was observed for UDP-Gal (Supplemental Fig. S4).  $^1\text{H-NMR}$  analysis revealed that the UDP-Gal stock contained traces of UDP-Glc, suggesting that the activity observed most likely originates from the UDP-Glc contamination (Thorsøe et al., 2005).

In summary, UGT73C10, UGT73C11, UGT73C12, and UGT73C13 preferentially glucosylate different oleanane and lupane sapogenins. Both UGT73C10 and UGT73C11 show high regioselectivity and substrate specificity by predominantly glucosylating the C3-hydroxyl group of sapogenins via an ether linkage. In comparison, UGT73C12 and UGT73C13 show lower substrate specificity and also glucosylate the sapogenin C28-carboxyl group via an ester bond. However, the ability to glucosylate at the C28-carboxyl group varied strongly: C28 glucosylation was abundant for betulinic acid and to a lesser extent for oleanolic acid and weakly for hederagenin. The similar enzymatic characteristics of UGT73C10 from the P-type and UGT73C11 from the G-type corroborate the phylogenetic reconstruction (Fig. 3), as do the characteristics of UGT73C12 from the P-type and UGT73C13 from the G-type. UGT73C9 apparently does not glucosylate any of the tested compounds besides the positive control substrate TCP, despite clustering with UGT73C10 and UGT73C11.

#### Kinetic Parameters of UGT73C11 and UGT73C13

Enzymes in the biosynthesis of plant specialized metabolism are generally characterized by low  $K_m$  and high turnover rates. To evaluate the affinity and catalytic efficiencies of the two UGT clusters (Fig. 3), the

kinetic parameters of UGT73C11 and UGT73C13 (both from the G-type) were determined toward hederagenin and oleanolic acid (Table I). Optimal assay conditions were at pH 8.6 for UGT73C11 and pH 7.9 for UGT73C13, with 1 mM dithiothreitol (DTT) as reductant. Purification of the recombinant UGTs was omitted due to decreasing specific activity upon metal chelate affinity-based purification. Instead, recombinant UGT amounts were quantified directly in crude *E. coli* protein extracts by taking advantage of an introduced N-terminal fused S-tag.

Most of the saturation curves (Supplemental Fig. S6) were hyperbolic and could be described by the Michaelis-Menten equations (for estimates, see Table I). However, for UGT73C13, the reaction velocities decreased when oleanolic acid concentrations exceeded 50  $\mu\text{M}$ , indicating that it inhibits enzyme activity beyond this concentration. Similar substrate inhibition has previously been reported for other family 1 UDP-glycosyltransferases (Luukkanen et al., 2005; Ono et al., 2010). UGT73C11 has a 7-fold lower  $K_m$  value and a 3-fold higher turnover rate ( $k_{\text{cat}}$  value) with hederagenin than UGT73C13. The two UGTs have comparable  $K_m$  values with oleanolic acid, but UGT73C11 has a 3.5-fold higher  $k_{\text{cat}}$  value. The kinetic parameters, therefore, corroborate that UGT73C10 and UGT73C11 have higher affinity for saponin and more efficiently catalyze 3-O-glucosylation of oleanolic acid and hederagenin than UGT73C12 and UGT73C13. The low  $K_m$  (less than 10  $\mu\text{M}$ ) and high  $k_{\text{cat}}$  values of UGT73C11 are in comparable ranges to flavonol UGTs with their in planta acceptor substrates (Noguchi et al., 2007; Ono et al., 2010). The 1.4-fold higher catalytic efficiency ( $k_{\text{cat}}/K_m$ ) for hederagenin than for oleanolic acid indicates that hederagenin is the preferred substrate for UGT73C11. Interestingly, UGT73C13 shows opposite substrate preference, as it has a 3-fold higher  $k_{\text{cat}}/K_m$  value for oleanolic acid than for hederagenin. The  $K_m$  for UDP-Glc was estimated to be around 95  $\mu\text{M}$  for UGT73C11 and 25  $\mu\text{M}$  for UGT73C12 (Supplemental Fig. S5).

**In Vitro Activities of the UGT73Cs toward *B. vulgaris* Saponin Mixtures**

The saponin composition of *B. vulgaris* is complex, with more than 40 putative saponins detected in liquid chromatography-mass spectrometry analyses (Supplemental

Figs. S7 and S8). The majority of these appear specific for either one of the two plant types, while others are present in variable amounts in both types. To evaluate if the UGTs can glucosylate other *B. vulgaris* saponin than oleanolic acid and hederagenin, crude saponin-containing extracts of both plant types were subjected to acidic hydrolysis to O-deglycosylate the saponins. Tandem mass spectrometry to n-fold ( $\text{MS}^n$ ) fragmentation analyses showed that the saccharide side chains of saponins in both *B. vulgaris* types consist of one to four hexosyl moieties, as concluded from the sequential loss of fragments with a mass of 162 D. The  $\text{MS}^n$  fragmentation patterns of the most intense putative saponins in the G-type extract further indicate that they are derived from saponin with masses of 456 and 472 D, corresponding to oleanolic acid and hederagenin, as well as 458 and 488 D. In addition, a few less intense putative saponins appear to be derived from saponin with masses of 470, 474, and 476 D. In metabolite extracts of the P-type, the most abundant putative saponins originate from saponin with a mass of 474 D, followed by saponins derived from 458- and 488-D saponin. Only a few putative saponins based on saponin with masses of 456 and 472 D occur in this plant type.

After acid hydrolyzation, the putative saponins could not be detected, which confirms complete deglycosylation (Supplemental Figs. S9 and S10). The hydrolyzed G-type extract contained at least 40 structurally distinct compounds that are likely to be saponin, while in the P-type extract, 13 putative saponin were detected. Incubation of these extracts with UGT73C10, UGT73C11, UGT73C12, and UGT73C13 and UDP-Glc as sugar donor yielded numerous compounds that, based on  $\text{MS}^n$  fragmentation patterns, were putative saponin monoglucosides (Supplemental Fig. S11). For both the G- and P-type saponin extracts, incubation with UGT73C10 and UGT73C11 reduced peak intensities of all putative saponin and resulted in the formation of the corresponding monoglucosides. In contrast, UGT73C12 and UGT73C13 appeared restricted to glucosylate only a subset of the putative saponin. Moreover, monoglucosides were produced at lower rates by UGT73C12 and UGT73C13 compared with UGT73C10 and UGT73C11. As expected, 3-O- $\beta$ -D-Glc hederagenin (compound G<sub>27</sub> in Supplemental Fig. S11) and 3-O- $\beta$ -D-Glc oleanolic acid

**Table I.** Kinetic parameters of UGT73C11 and UGT73C13 toward oleanolic acid and hederagenin

UGT	Saponin	$K_m$ $\mu\text{M}$	$k_{\text{cat}}$ $\text{s}^{-1}$	$k_{\text{cat}}/K_m$ $\text{s}^{-1} \mu\text{M}^{-1}$	$K_i$ $\mu\text{M}$	$V_{\text{max}}$ $\text{nmol min}^{-1} \text{mg}^{-1}$
UGT73C11	Oleanolic acid	9.7 ± 2.2	0.816	0.084		817 ± 118
	Hederagenin	3.3 ± 0.8	0.389	0.118		390 ± 38
UGT73C13	Oleanolic acid <sup>a</sup>	12.5 ± 2.1	0.231	0.019	262	231 ± 21
	Oleanolic acid <sup>b</sup>	7.6 ± 1.2	0.176	0.023		176 ± 7
	Hederagenin	22.9 ± 4.8	0.131	0.006		131 ± 10

<sup>a</sup>Kinetic parameters based on fit to the substrate inhibition equation.

<sup>b</sup>Kinetic parameters based on fit to the Michaelis-Menten equation.

(compound G<sub>35</sub> in Supplemental Fig. S11) were among the products formed from the G-type extract by UGT73C10 and UGT73C11. Surprisingly, only trace amounts of these two sapogenin monoglucosides were observed upon incubation of the G-type extract with UGT73C12 and UGT73C13. These UGTs additionally produced low amounts of diglucosides and compounds that may be kaempferol glucosides (according to their MS<sup>n</sup> fragmentation patterns). These findings corroborate that UGT73C12 and UGT73C13 have lower substrate specificity toward sapogenins than UGT73C10 and UGT73C11, which was also concluded from the *in vitro* enzyme assays (Fig. 5).

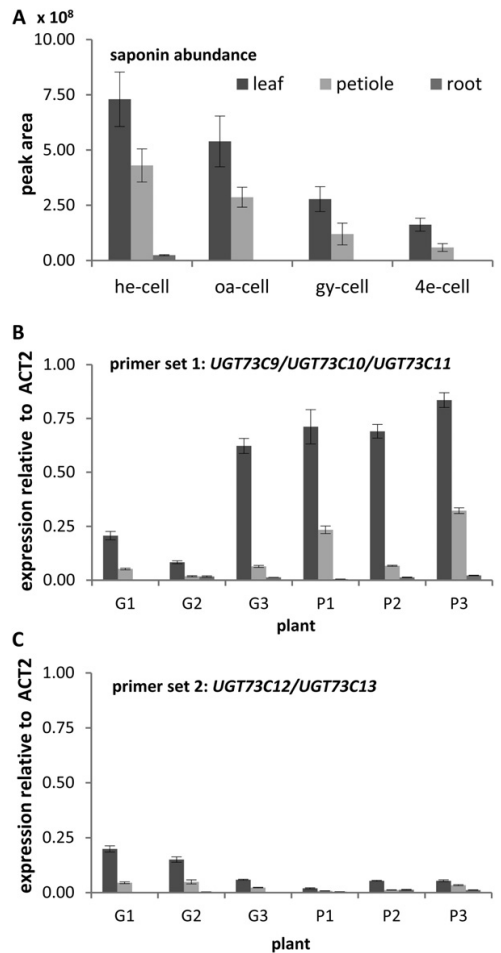
#### In Planta Saponin Accumulation Correlates with Organ-Specific Expression of the UGT73Cs

Steady-state transcript levels of the UGT73Cs were determined in leaves, petioles, and roots of 2-month-old G- and P-type *B. vulgaris* plants and compared with saponin accumulation in these organs. Metabolite extracts were evaluated by liquid chromatography-mass spectrometry and revealed a characteristic organ-specific saponin relative abundance in both plant types. Relative accumulation was highest in leaves, intermediate in petioles, and widely absent in roots (Fig. 6A; Supplemental Fig. S12). This pattern was consistent across the different plants tested.

Two primer sets were used to quantify steady-state transcription levels of the UGTs by quantitative real-time PCR (Fig. 6, B and C). Due to the high sequence identities between UGT73C11 in the G-type and UGT73C10 and UGT73C9 in the P-type, it was not possible to design a primer that could differentiate between these three genes. Accordingly, primer set 1 amplifies UGT73C11 in the G-type, while in the P-type it amplifies simultaneously UGT73C9 and UGT73C10. Similarly, primer set 2 amplifies UGT73C13 from the G-type and UGT73C12 from the P-type. All plants showed the highest expression of UGT73C11 and UGT73C9/C10 in leaves, an up to 10-fold lower expression in petioles, and up to 200-fold lower expression in roots, despite some variation among individual plants tested. A similar expression pattern was observed for UGT73C13 and UGT73C12. In general, UGT73C11 and UGT73C9/C10 were expressed at a higher level than UGT73C13 and UGT73C12. The highest expression level of UGT73C13 was observed in plants with the lowest UGT73C11 expression. Since those plants were in a more progressed developmental stage (Supplemental Fig. S12), this suggests alternating expression regulation of the two genes during plant ontogenesis.

#### 3-O- $\beta$ -D-Glc Hederagenin Is a Feeding Deterrent against *P. nemorum*

The two diglucosides hederagenin and oleanolic acid cellobioside have previously been shown to deter feeding by *P. nemorum* (Nielsen et al., 2010). To



**Figure 6.** Comparison of relative saponin abundance and expression of the UGTs in different *B. vulgaris* organs. A, Relative saponin abundance in leaf, petiole, and root extracts of three G-type plants (G1–G3), based on the mean peak areas  $\pm$  sd of the extracted ion chromatograms from liquid chromatography-mass spectrometry of the four insect resistance-correlated G-type saponins: hederagenin cellobioside (he-cell), oleanolic acid cellobioside (oa-cell), gypsogenin cellobioside (gy-cell), and 4-epi-hederagenin cellobioside (4e-cell). Overlaid base peak chromatograms of all liquid chromatography-mass spectrometry runs are provided in Supplemental Figure S12. B, Expression of UGT73C11 in the three G-type plants (G1–G3) and combined expression of UGT73C9 and UGT73C10 in three P-type plants (P1–P3), determined with primer set 1 relative to actin (ACT2). Values are means  $\pm$  sd of technical duplicates. C, Corresponding expression analysis of UGT73C13 in G1 to G3 and UGT73C12 in P1 to P3, determined with primer set 2.

determine if the corresponding monoglucosides have a similar effect, approximately 12.5 mg of 3-O- $\beta$ -D-Glc hederagenin and 8.5 mg of 3-O- $\beta$ -D-Glc oleanolic acid were produced *in vitro* with UGT73C10 (see above).

Both compounds were painted on 92-mm<sup>2</sup> radish (*Raphanus sativus*) leaf discs in doses of 3.75, 15, and 60 nmol and presented to *P. nemorum* adults of either the susceptible (ST; rr genotype) or resistant (AK; Rr genotype) line, and the area consumed was evaluated after 24 h (Fig. 1).

3-*O*-β-D-Glc hederagenin significantly reduced the leaf consumption by susceptible ST beetles, with dose-dependent reductions of 26%, 55%, and 92% in response to 3.75, 15, and 60 nmol per leaf disc, respectively (Fig. 7A; the reduction by 15 and 60 nmol was statistically significant [*P* < 0.005] when tested separately). A dose-dependent reduction of leaf consumption was also observed for the resistant AK line, with 16% and 67% reduction in response to 15 and 60 nmol, respectively (only the reduction by 60 nmol was significant when tested separately).

3-*O*-β-D-Glc oleanolic acid had a significantly weaker effect on leaf consumption for both *P. nemorum* lines (Fig. 7B). Only the high dose of 60 nmol reduced consumption by the ST line (45% reduction), whereas there was no effect on the AK line at any dose. Feeding

assays with 3.75 nmol were not conducted, as there was no significant effect with 15 nmol.

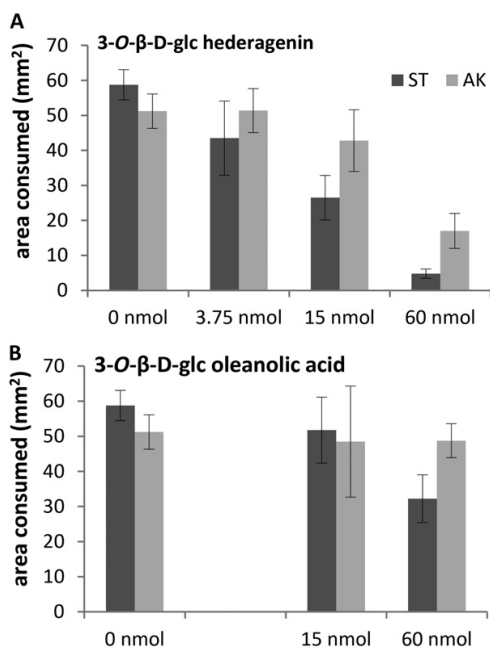
When tested in a joint linear mixed-effect model, there was a significant three-way interaction between saponenin monoglucosides, their doses, and the *P. nemorum* lines, with a significance level of *P* < 0.0001. Thus, (1) 3-*O*-β-D-Glc hederagenin is more effective than 3-*O*-β-D-Glc oleanolic acid, (2) the feeding deterrence of the saponenin monoglucosides is dose dependent, and (3) the efficacy toward the susceptible *P. nemorum* line is higher than toward the resistant line.

## DISCUSSION

Saponin biosynthesis is not fully understood, nor is the relationship between the different chemical structures and their roles in plant defense. Here, we have identified two UGTs that specifically glucosylate saponinens in the wild crucifer *B. vulgaris*. These UGTs have evolved to be specific for 3-*O*-glucosylation of saponinens. Previously, UGTs that glucosylate saponinens at the C28 carboxylic groups have been identified in *Medicago truncatula* (UGT73F3; Naoumkina et al., 2010) and in *Saponaria vaccaria* (UGT74M1; Meesapyodsuk et al., 2007). Mono-glucosylated 3-*O*-β-D-Glc hederagenin, produced in vitro by one of the UGTs identified here, UGT73C10, is a strong feeding deterrent against *P. nemorum*, demonstrating that 3-*O*-glucosylation of saponinens is essential for bioactivity. The UGTs are expressed in both a *P. nemorum* resistant and a susceptible type of *B. vulgaris*, which fits our observation that most, if not all, saponinens in the P and G-types are 3-*O*-glucosylated. The presence of UGTs in both the plant types catalyzing 3-*O*-glucosylation saponinens, and the genomic locations of genes coding for these UGTs outside QTL associated with resistance to *P. nemorum*, suggest that the difference in resistance between the two *B. vulgaris* types is determined by an earlier enzymatic step in saponin biosynthesis.

### UGT73C10/C11: Two Neofunctionalized UDP-Glc: Saponenin 3-*O*-Glucosyltransferases

Of the five UGTs we identified in *B. vulgaris* ssp. *arcuata*, UGT73C10 from the insect-susceptible P-type and UGT73C11 from the resistant G-type showed highest activity and specificity toward a wide range of saponinens. Both enzymes exhibit high regiospecificity by preferably glucosylating the C3 hydroxyl group, which is in agreement with structures of saponinens in both *B. vulgaris* types. Both enzymes, in contrast, were essentially inactive toward the flavonols and phytosterols tested. Their acceptor substrate specificity thus differs substantially from other characterized members of the UGT73C subfamily. UGT73C8 from *M. truncatula* glucosylates several (iso)flavonoids in vitro (Modolo et al., 2007). *A. thaliana* UGT73C6 was suggested to be a UDP-Glc: flavonol-3-*O*-glycoside-7-*O*-glucosyltransferase by Jones et al. (2003), based on in vitro activities and T-DNA knockout lines. Recent studies show that UGT73C6 is



**Figure 7.** Consumption of radish leaf discs painted with different amounts of 3-*O*-β-Glc hederagenin (A) and 3-*O*-β-Glc oleanolic acid (B) by susceptible ST and resistant AK lines of *P. nemorum*. Consumption is shown as mean total area consumed from two leaf discs (total area, 92 mm<sup>2</sup>) that were presented to one beetle (±1.96 SE corresponding to a confidence interval of 95%). Assays with 3.75 nmol of 3-*O*-β-Glc oleanolic acid were omitted due to the low efficacy at higher doses.

functionally similar to the well-studied UGT73C5, also from *A. thaliana*, in its ability to glucosylate brassinosteroids in overexpression lines (Husar et al., 2011). UGT73C5 in addition glucosylates numerous structurally diverse acceptor substrates (Lim et al., 2003, 2004; Poppenberger et al., 2003, 2005, 2006; Hou et al., 2004; Weis et al., 2006; Caputi et al., 2008). It was originally identified as a mycotoxin-detoxifying enzyme (Poppenberger et al., 2003), but recently, it was suggested to be involved in brassinosteroid homeostasis (Poppenberger et al., 2005). In our study, *A. thaliana* UGT73C5 also glucosylated oleanolic acid, hederagenin, and betulinic acid *in vitro*, providing further evidence for the promiscuity of this enzyme (Supplemental Fig. S13). However, it had substantially lower catalytic efficiency and regioselectivity toward oleanolic acid and hederagenin than UGT73C11 and UGT73C13 from *B. vulgaris* (Supplemental Fig. S13). *A. thaliana* is not known to produce triterpenoid saponins or sapogenins, although triterpenoids such as  $\beta$ -amyrin and lupeol accumulate in cuticular waxes of stems, siliques, and buds (Shan et al., 2008). Therefore, it is unlikely that the *in vitro* activities of UGT73C5 with sapogenins reflect an *in planta* function.

The broad substrate affinity commonly found for some UGTs has been proposed to enable flexibility in response to changes in metabolite profiles (Vogt and Jones, 2000). Specialized enzymes for new biosynthetic pathways may originate from broad progenitor enzymes and are generally characterized by having a lower  $K_m$  (and thus higher substrate specificity) and higher catalytic efficiency ( $k_{cat}/K_m$ ) than their more promiscuous progenitors (Jensen, 1976; Aharoni et al., 2005; Khersonsky and Tawfik, 2010). Ancestors of UGT73C10/C11 from *B. vulgaris* could thus have been promiscuous UGT73C5-like enzymes that evolved a more narrow specificity and higher efficiency for catalyzing sapogenin 3-O-glucosylation. Based on our analyses, UGT73C12/C13 have broader substrate and product specificities and could represent evolutionary intermediates to UGT73C10/C11 or UGTs specialized in glucosylation of yet unknown sapogenins in *B. vulgaris*.

Our phylogenetic reconstruction shows that the five *B. vulgaris* UGT73Cs indeed cluster separately from the UGT73Cs in *A. thaliana*, *A. lyrata*, and *B. rapa* (Fig. 3). It further suggests that UGT73C10, UGT73C11, and UGT73C9 originate from a gene duplication event after the split from *A. thaliana* and *B. rapa* and before the P and G-types separated. Another gene duplication separated UGT73C9 from UGT73C10, probably in the P-type after the P- and G-types split. Alternatively, this duplication occurred before the P-G bifurcation and the UGT73C9 copy was lost subsequently in the G-type.

Of the UGTs in our phylogenetic analysis, UGT73C9, UGT73C10, and UGT7311 showed clear signs of positive selection during their differentiation. This corroborates our biochemical data, which show that UGT73C10 and UGT73C11 have evolved to a new specialized function. In contrast, UGT73C12 and UGT73C13 showed no signs of selection, corroborating that they have not evolved new biochemical functions; this further suggests that they may be orthologs of *A. thaliana* UGT73C5 or

UGT73C6. The observation that UGT73C9 is under positive selection questions the function of this UGT in saponin biosynthesis. Based on our biochemical data, UGT73C9 appears as an expressed pseudogene; however, the phylogenetic analysis indicates that the gene has been under positive selection. An alternative hypothesis is that the substrate for UGT73C9 was not included in our analysis. As the saponin profiles of P- and G-type *B. vulgaris* differ, UGT73C9 could possibly be involved in the differentiation of these.

Genes for the *B. vulgaris* UGTs were located in a genomic region syntenic to a part of *A. thaliana* chromosome 2, which contains a tandem repeat cluster of UGT73Cs. Our recent genome sequencing indicates that the *B. vulgaris* UGT73Cs identified here are also part of a repetitive cluster containing several UGT-like repeats and in higher number than the corresponding UGT73C cluster in *A. thaliana*. This supports that UGT73C10/C11 evolved via gene duplications from a broad-spectrum UGT73C in a common ancestor shared with *A. thaliana*, as discussed above. It further supports the idea that the evolution of novel bioactive metabolites often occurs via gene duplication and neofunctionalization (Osborn, 2010; Weng et al., 2012) followed by increased specialization (Jensen, 1976; Aharoni et al., 2005; Khersonsky and Tawfik, 2010).

### 3-O-Glucosylation of Hederagenin Deters Feeding by *P. nemorum*

Monoglucosylation of hederagenin into 3-O- $\beta$ -D-Glc hederagenin clearly suppressed feeding by *P. nemorum*. A similar but lower suppression was found for 3-O- $\beta$ -D-Glc oleanolic acid. The diglucosylated forms of hederagenin and oleanolic acid (hederagenin cellobioside and oleanolic acid cellobioside) have previously been found to suppress feeding (Nielsen et al., 2010), in contrast to the aglycones (hederagenin and oleanolic acid). Our results now show that glucosylation with only a single glucosyl group is enough to affect herbivores. The amount of monoglucosides used in our feeding assays was comparable to natural levels of hederagenin cellobioside in *B. vulgaris* leaves (Shinoda et al., 2002), and our results thus demonstrate that 3-O- $\beta$ -D-Glc hederagenin and 3-O- $\beta$ -D-Glc oleanolic acid are biologically relevant feeding deterrents. Furthermore, the higher efficiency of hederagenin than oleanolic acid, in both their monoglucosylated and diglucosylated forms, shows that C23 hydroxylation in the hederagenin backbone increases this antifeedant effect.

The precise mechanism that enables glucosylated saponins to deter insects is not known. The dependency on glycosylation indicates that membrane perturbation plays a role, at least for *P. nemorum*. In agreement with this, saponins have been shown to damage the midgut epithelium of pea aphids (*Acyrtosiphon pisum*; De Geyer et al., 2012). Alternatively, glucosylated saponins may have a more adverse taste for insects than the corresponding sapogenins (Glendinning, 2002); however, *P. nemorum* larvae die from exposure to G-type leaves (Nielsen, 1997a).



Nielsen et al. (2010) suggested that cleavage of the  $\beta$ -1,4-glycosidic bond in the cellobiosides by  $\beta$ -glucosidases allows resistant *P. nemorum* lines to feed on G-type *B. vulgaris*. This mechanism would be similar to what has been found for fungal adaptation to saponins (Osbourn et al., 1991; Wubben et al., 1996; Pareja-Jaime et al., 2008). Our findings, however, show that the monoglucosides of the saponins are also active and that resistance must be based on the ability to hydrolyze the glycosidic bond between the aglycone and the first linked sugar at the C3 position.

The resistance of G-type *B. vulgaris* against herbivorous insects, such as *P. xylostella* and susceptible *P. nemorum*, has previously been shown to depend on the presence of saponins, and especially hederagenin and oleanolic acid cellobioside, which are absent in the susceptible P-type (Shinoda et al., 2002; Agerbirk et al., 2003a; Kuzina et al., 2009; Nielsen et al., 2010). Therefore, the synthesis of saponins was initially thought to be unique to the G-type. However, saponins were recently also discovered in the susceptible P-type (Kuzina et al., 2011), and we are now pursuing their structure and identity. The presence of closely related UGTs in the G- and P-types of *B. vulgaris*, which have the same substrate specificity and regiospecificity, strongly indicates that the difference between resistance and susceptibility of the two *B. vulgaris* types is not caused by different UGTs, despite their obvious role in activating sapogenins by glucosylation. This is further substantiated by results from our QTL analysis, where the UGTs described here do not colocalize with resistance to *P. nemorum* or saponin identity (Kuzina et al., 2011). Instead, the difference in resistance between the G- and P-types must be determined at an earlier step in saponin biosynthesis, presumably during cyclization by OSCs or backbone decoration by cytochromes P450.

### Evolution of Saponin Biosynthesis in *Barbarea* Species

The multitude of different putative sapogenins in the G- and P-types indicates that OSCs and P450s are responsible for much of the saponin diversity in this species and probably for the differences between the two plant types. The phylogeny of OSCs (Phillips et al., 2006; Augustin et al., 2011) suggests frequent changes in product spectra during evolution, which is supported by the drastic spectrum changes that may arise from only a few amino acid substitutions (Lodeiro et al., 2005). Changes in cytochrome P450 activity are also known to affect saponin profiles and activity. Carelli et al. (2011) showed that lack of a functional CYP716A12, which catalyzes C28 carboxylation of triterpenoid sapogenins, results in a complete loss of hemolytic saponins in *M. truncatula*. In contrast, nonhemolytic saponins were unaffected. The nonhemolytic saponins are derived from sapogenins that are not carboxylated at the C28 position, and MS<sup>n</sup> fragmentation of these revealed an aglycone fragment ion with a deduced mass of 474 D (Pollier et al., 2011). A similar fragmentation product was observed for P-type saponins and suggests that structurally

similar sapogenins, with four hydroxyl groups but no C28 carboxylation, are present in this plant type. Different abilities to catalyze C28 oxygenation by cytochromes P450 could thus be involved in determining the different structures of G- and P-type saponins and thus their effect on insect herbivores.

The current hypothesis for the evolution of insect resistance in *B. vulgaris* suggests that it took place after the first species of the *Barbarea* genus had emerged (Agerbirk et al., 2003b; the age of this split is unknown at present). An OSC probably mutated to be able to catalyze the conversion of oxidosqualene into saponin precursors, which is in agreement with the presence of triterpenoids in *A. thaliana*. Later, UGTs must have evolved to become specific to the novel sapogenins produced by the resistant *Barbarea* species, as we have shown here. Whether the cytochromes P450 involved in saponin biosynthesis of *Barbarea* species have also specialized is not known. Much later, *B. vulgaris* differentiated into the G- and P-types, possibly during one of the last ice ages (Hauser et al., 2012; Toneatto et al., 2012). Thus, the two plant types are genetically and geographically differentiated, reproductively somewhat incompatible, and differ for several traits apart from insect resistance and saponin structure (Toneatto et al., 2010; Dalby-Brown et al., 2011). Thus, the most likely scenario suggests that the P-type lost resistance to *P. nemorum* during this allopatric separation. Our results here clearly show that this loss of insect resistance was not caused by a loss of UGT function. Instead, we have shown that UGTs of *B. vulgaris* have adapted to the earlier evolutionary gain of saponins in this species.

## MATERIALS AND METHODS

### Activity-Based cDNA Library Screening

*Barbarea vulgaris* var *variegata* (Chiltern Seeds) leaf RNA was used for first-strand synthesis with the ZAP-cDNA Synthesis Kit (Stratagene). The resulting cDNA was digested with *Xho*I, ligated into the predigested Uni-ZAP XR vector (Stratagene), and transformed into the *Escherichia coli* strain XL1-Blue MRF' (Stratagene). After *in vivo* excision of pBluescript SK- phagemids from the Uni-ZAP XR vectors, the obtained *E. coli* colonies were combined in terrific broth (TB) medium and transferred to 96-well plates (approximately 100 colonies per well). The *E. coli* suspensions were incubated with shaking at 37°C for 3 h and then for 3 h with 0.1 mM isopropylthio- $\beta$ -galactoside (IPTG). Cultures of individual wells were combined into batches (four wells per batch), and the bacterial cells were harvested by centrifugation. The bacterial cells were resuspended in 20 mM Tris-HCl, pH 7.5, and 2 mM DTT and lysed by sonication. Enzymatic activity was tested by incubating the lysates overnight at 30°C with 200  $\mu$ M UDP-Glc and 175  $\mu$ M oleanolic acid. Ethyl acetate extracts of the activity assays were analyzed by TLC on Silica Gel 60 F<sub>254</sub> plates (5554; Merck), using chloroform:methanol:water (32:9:1) as mobile phase, and stained by spraying with 10% sulfuric acid in methanol followed by heating. Batches that showed oleanolic acid glucosylation activity were in additional screening rounds stepwise further diluted until a single active clone designated BvUGT1 was identified.

### Cloning of BvUGT1 Homologs from *B. vulgaris* ssp. *arcuata*

Contigs representing fragments of *BvUGT1* homologs were identified in a 454 pyrosequencing-generated transcriptomic G-type data set (Kuzina et al.,

2011) using local BLASTX. Total RNA was extracted from leaves of G- and P-type *B. vulgaris* using the NucleoSpin RNA Plant kit (Macherey-Nagel) and 3' RACE performed with the FirstChoice RLM-RACE kit (Ambion) according to the manufacturer's protocol. The applied primers are listed in Supplemental Table S2.

The nucleotide sequences of *UGT73C9*, *UGTC10*, *UGT73C11*, *UGT73C12*, and *UGT73C13* were cloned from genomic DNA of an F1 hybrid plant, which originated from crossings between G- and P-type plants (Kuzina et al., 2009), and ligated into pGEM-T Easy for sequencing.

PCRs for cloning were performed with Phusion High-Fidelity DNA Polymerase (Finnzymes), and PCRs for screening and A-tailing reactions were performed with Hotmaster Taq DNA Polymerase (Sprime). A-tailing reactions were set up according to the pGEM-T Easy manual (Promega). Sequencing was performed by Eurofins MWG Operon.

## Phylogenetic Analysis

UGT73 amino acid sequences were aligned (Supplemental Data Set S2) using MUSCLE and used to construct a maximum likelihood bootstrapped phylogenetic tree using MEGA (version 5.05; Jones, Taylor, and Thornton substitution model, uniform rates among sites, 100 bootstrap replications; Tamura et al., 2011). The *A. thaliana lyrata* and *Brassica rapa* UGTs, identified by BLAST searches at [www.phytozome.net](http://www.phytozome.net) and [www.brassica-rapa.org](http://www.brassica-rapa.org), have not been officially named and therefore are named here according to their grouping with *Arabidopsis thaliana*.

To test for signs of past selection on the UGTs, branch and site models were estimated using codeml in the PAML package (<http://abacus.gene.ucl.ac.uk/software/paml.html>). For positive selection between branches, the free-ratio model was compared with the one-ratio model and tested by comparing the twice log-likelihood difference between models to a  $\chi^2$  distribution with 18 degrees of freedom. Seven site models were estimated: M0 (one ratio); M1 (nearly neutral; two categories); M2 (positive selection; three categories); M3 (discrete; three categories); M5 ( $\gamma$ ; 10 categories); M7 ( $\beta$ ; 10 categories); and M8 ( $\beta$  &  $\omega > 1$ ; 11 categories); these were tested as above with degrees of freedom corresponding to the differences in the number of parameters for the models tested.

## Locating UGT73C9, UGT73C10, UGT73C11, UGT73C12, and UGT73C13 on the *B. vulgaris* Linkage Map

The five UGTs were mapped using the derived cleaved amplified polymorphic sequences or cleaved amplified polymorphic sequences technique. PCR was performed using genomic DNA of an F2 segregating population generated from a cross between P- and G-type *B. vulgaris* (Kuzina et al., 2009). PCR products obtained using primers mapPSfor and sepSrev (*UGT73C9* to *-C11*), mapPSfor and sep1Irev (*UGT73C12/C13*), or Inf and dCapsAvaII (*UGT73C11*) were digested with *EcoRV*, *BsaJI*, *AvaII*, or *PciI* to discriminate between *UGT73C9*, *UGT73C10*, *UGT73C11*, and *UGT73C13*, respectively. Data were scored and analyzed as described by Kuzina et al. (2011).

## Heterologous Expression of *B. vulgaris* UGT73Cs

N-terminally His-tagged expression constructs of *UGT73C9*, *UGT73C10*, *UGT73C11*, *UGT73C12*, and *UGT73C13* were obtained by subcloning into the *NheI* and *BamHI* restriction sites of the pET28c vector (Novagen). N-terminally S-tag expression constructs of the five *UGT73C* ORFs were achieved by Gateway cloning into pJAM1786 (Luo et al., 2007).

For heterologous expression of the His-tag and S-tag constructs, expression vectors were transformed into the *E. coli* strain XJb(DE3) (Zymo Research). Expression was carried out in 25-mL Erlenmeyer flasks and started by inoculating 2 mL of Luria-Bertani medium, containing either 50  $\mu\text{g mL}^{-1}$  kanamycin (His-tag constructs) or 100  $\mu\text{g mL}^{-1}$  carbenicillin (S-tag constructs), with a single colony. A 12-h incubation phase at 30°C and 220 rpm was followed by the addition of 4 mL of TB medium containing appropriate selection antibiotics. Ara and IPTG were added to final concentrations of 3 and 0.1 mM, respectively, and the cultures were incubated for 24 h at 15°C and 220 rpm. For expression of the S-tag constructs, 1  $\mu\text{L}$  of 50 mg  $\text{mL}^{-1}$  carbenicillin  $\text{mL}^{-1}$  culture was added approximately 12 h after the addition of TB medium.

Bacteria were harvested in aliquots corresponding to 2 mL of culture with an optical density of 8.0, resuspended in 750  $\mu\text{L}$  aliquot $^{-1}$  10 mM HEPES, pH 7.8, and stored at  $-80^\circ\text{C}$ . Bacteria were lysed by thawing aliquots at room temperature. The viscosity of lysates was lowered by incubation with DNaseI

(AppliChem) treatment (1  $\mu\text{g mL}^{-1}$ ). Cell debris were removed by centrifugation, and supernatants were used as crude protein extracts for enzyme assays. Quantification of heterologously expressed enzymes, fused to an S-tag within *E. coli* crude protein extracts, was carried out using the FRETworks S-tag assay kit (Novagen) according to the manufacturer's protocol.

## Substrate Specificity Assays

Enzyme assays to determine substrate specificity were performed in a final volume of 20  $\mu\text{L}$ , containing 2  $\mu\text{L}$  of *E. coli* crude protein extract with recombinant UGT73C9, UGT73C10, UGT73C11, UGT73C12, or UGT73C13 coupled to an S-tag. Reaction conditions were 25 mM TAPS-HCl, pH 8.6, 1 mM DTT, 7  $\mu\text{M}$  UDP-Glc (Sigma-Aldrich), and 3.31  $\mu\text{M}$  (0.74 kBq) UDP-[ $^{14}\text{C}$ ]Glc (Perkin-Elmer). Ethanol was removed from the UDP-[ $^{14}\text{C}$ ]Glc stock by evaporation prior to setting up the assays. Enzyme assays were started by addition of the acceptor substrates solubilized in dimethyl sulfoxide (DMSO) to final concentrations of 1 mM (only TCP), 100  $\mu\text{M}$ , or 10  $\mu\text{M}$  of the acceptor substrate and 6.25% to 10% (v/v) DMSO, respectively. Reactions were incubated for 30 min at 30°C and stopped by the addition of 130  $\mu\text{L}$  of methanol. Precipitated proteins were removed by centrifugation. Solvent from the supernatant was removed with a vacuum concentrator, and metabolites were dissolved in 20  $\mu\text{L}$  of 50% ethanol and analyzed by TLC. TLC plates were developed in ethyl acetate:methanol:formic acid:water (7.5:0.5:1:1), and radioactive bands were visualized using a STORM 840 PhosphorImager (Molecular Dynamics).

Acceptor substrates in this study were as follows: oleanolic acid (ICN Bio-medical), hederagenin (Carl Roth), betulonic acid (Carl Roth),  $\beta$ -amyrin (Sigma-Aldrich), lupeol (Sigma-Aldrich), quercetin (Sigma-Aldrich), kaempferol (Fluka), and obtusifolioside, campesterol, sitosterol, stigmasterol, and 2,4,5-trichlorophenol (Sigma-Aldrich).

## Determination of Enzyme Kinetic Parameters

Freshly lysed *E. coli* crude protein extracts were diluted in 10 mM TAPS-HCl, pH 8.0, and 10 mg  $\text{mL}^{-1}$  bovine serum albumin (BSA) to final concentrations of 5 ng  $\mu\text{L}^{-1}$  S-tag UGT73C11 and 45 ng  $\mu\text{L}^{-1}$  S-tag UGT73C13. The diluted crude protein extracts were applied in master mixtures with final reaction conditions as follows: 25 mM TAPS-HCl, pH 8.6 (UGT73C11) or pH 7.9 (UGT73C13), 1 mM DTT, 500  $\mu\text{M}$  UDP-Glc, 2 mg  $\text{mL}^{-1}$  BSA, and 0.5 ng  $\mu\text{L}^{-1}$  UGT73C11 or 4.5 ng  $\mu\text{L}^{-1}$  UGT73C13. Enzyme assays were performed in a volume of 20  $\mu\text{L}$ . Concentrations of UDP-[ $^{14}\text{C}$ ]Glc (Perkin-Elmer) in the total amount of UDP-Glc ranged from 3.31  $\mu\text{M}$  (0.04 kBq  $\mu\text{L}^{-1}$ ) to 33.12  $\mu\text{M}$  (0.37 kBq  $\mu\text{L}^{-1}$ ) to ensure sufficient signal intensity. Oleanolic acid and hederagenin were dissolved in 100% DMSO and assayed in duplicate in final concentrations ranging from 0.125 to 8  $\mu\text{M}$  for UGT73C11 and 1.56 to 100  $\mu\text{M}$  for UGT73C13, but with a constant final DMSO concentration of 6.25%. Reactions were preincubated for 3 min at 30°C prior to addition of the acceptor substrate. After incubation for 3 min at 30°C, enzymatic activities were stopped by the addition of 50  $\mu\text{L}$  of ethyl acetate. Assays were extracted four times with 50  $\mu\text{L}$  of ethyl acetate, and the solvent from the combined extractions was removed by evaporation in a vacuum concentrator. Metabolites were dissolved in 96% ethanol and analyzed by TLC. TLC plates were developed using dichloromethane:methanol:water (80:19:1) as mobile phase and visualized as described above. Products were quantified by codeveloping TLC plates with a defined oleanolic acid or hederagenin [ $^{14}\text{C}$ ]monoglucoside dilution series. Signal intensities were quantified using ImageQuant 5.0 (Molecular Dynamics).  $K_m$  and  $V_{max}$  values were calculated using SigmaPlot 11.0 (Systat Software) for nonlinear regression according to the Michaelis-Menten equation or the velocity equation for substrate inhibition.

$^{14}\text{C}$ -labeled monoglucosides were obtained by overnight incubation of 20 nmol of oleanolic acid and hederagenin with UGT73C11 at reaction conditions similar to those applied for the actual enzyme assays (500  $\mu\text{M}$  UDP-Glc including 33.12  $\mu\text{M}$  UDP-[ $^{14}\text{C}$ ]Glc [0.37 kBq  $\mu\text{L}^{-1}$ ]). Complete conversion of the aglycones was confirmed by TLC analysis of aliquots of these reactions.

## Plant Material

*B. vulgaris* ssp. *arcuata* seeds were collected in natural populations in Denmark: G-type (Amager; 55°38'N, 12°34'E) and P-type (Tisse; 55°36'N, 11°18'E). Plants were grown at 20°C, 16 h of light/8 h of darkness, and 70% to 75% air humidity, fertilized once a week, and the soil was treated with Bac-timos L (Abbott Laboratories) whenever necessary.

## Comparison of Saponin Levels and in Planta Expression of UGT73Cs

To determine saponin levels, metabolites were extracted from 20 to 30 mg of ground, lyophilized leaf, petiole, and root material by boiling for 10 min with 37.5  $\mu$ L of 55% ethanol per mg of tissue powder. Samples were cooled on ice and centrifuged to remove insoluble particles. Supernatants were kept for more than 2 h at  $-20^{\circ}\text{C}$  and centrifuged to remove precipitates. Extracts were filtered (polyvinylidene difluoride; 0.45  $\mu\text{m}$ ) and transferred to glass sample vials for liquid chromatography-mass spectrometry analysis. An Agilent 1100 Series LC device (Agilent Technologies), equipped with a Gemini NX column (35 $^{\circ}\text{C}$ ; 2.0  $\times$  150 mm, 3.5  $\mu\text{m}$ ; Phenomenex) and coupled to a Bruker HCT-Ultra ion-trap mass spectrometer (Bruker Daltonics), was used for spectrometry analysis. Mobile phases were eluent A, water with 0.1% (v/v) formic acid, and eluent B, acetonitrile with 0.1% (v/v) formic acid. The gradient program was as follows: 0 to 1 min, isocratic 12% B; 1 to 33 min, linear gradient 12% to 80% B; 33 to 35 min, linear gradient 80% to 99% B; 35 to 38 min, isocratic 99% B; 38 to 45 min, isocratic 12% B at a constant flow rate of 0.2 mL  $\text{min}^{-1}$ . The detector was operated in negative electrospray mode and included tandem mass spectrometry to two stages ( $\text{MS}^2$ ) and three stages ( $\text{MS}^3$ ). Chromatograms were analyzed with DataAnalysis 4.0 (Bruker Daltonics), and saponin abundance was calculated based on summed extracted ion chromatograms of all adduct ions.

RNA was extracted from 100 to 150 mg of ground leaf, petiole, and root material by incubation for 10 min with 900  $\mu\text{L}$  of prewarmed hexadecyltrimethylammonium bromide extraction buffer (Chang et al., 1993) at 65 $^{\circ}\text{C}$  and 660 rpm. After 2-fold extraction with 900  $\mu\text{L}$  of chloroform-isoamyl alcohol, RNA was precipitated overnight (4 $^{\circ}\text{C}$ ) from the supernatant by the addition of LiCl to a final concentration of 2 M. Pellets were dissolved in 500  $\mu\text{L}$  of sodium chloride-Tris-EDTA buffer (le Provost et al., 2007; prewarmed to 65 $^{\circ}\text{C}$ ) containing 0.1% SDS. RNA was extracted with chloroform-isoamyl alcohol and precipitated from the aqueous phase by adjusting the NaCl concentration to 0.67 M, adding 1 volume of isopropanol, and subsequent incubation for 5 h at  $-20^{\circ}\text{C}$ . RNA pellets were washed with 70% ethanol ( $-20^{\circ}\text{C}$ ), dried, and redissolved in 30  $\mu\text{L}$  of diethyl pyrocarbonate-treated water. The remaining genomic DNA was removed by on-column DNase treatment using the RNeasy Mini Kit (Qiagen). RNA extracts were assessed for purity and quantified with a NanoDrop ND-1000 (NanoDrop Technologies) and a 2100 Bioanalyzer (Agilent Technologies).

Reference gene sequences were obtained by mapping the 454 pyrosequencing-derived reads of G- and P-type leaf RNA preparations (V. Kuzina and S. Bak, unpublished data) to a data set consisting of all *A. thaliana* cDNA sequences (TAIR9\_cdna\_20090619) using the CLC Genomics Workbench (CLC bio). Two primer pairs, ACT2\_for1/ACT2\_rev1 and ACT2\_for2/ACT2\_rev2, were designed from reads mapped to *A. thaliana* ACT2 (AT3G18780). With the exception of four single-nucleotide polymorphisms in an intron region of the ACT2\_for1/ACT2\_rev1 product from the P-type, sequences derived for each primer set from the two plant types were 100% identical. The sequence identity of the two PCR products to the *A. thaliana* ACT2 ORF were 91% and 96%, respectively, while the encoded protein sequences were 100% identical to *A. thaliana* ACT2. Threshold cycle values of the two primer sets were almost identical in quantitative real-time PCR tests on leaf, petiole, and root tissues from a single G-type plant ( $\pm 0.08$ – $0.26$ ). In addition, threshold cycle values across the three investigated tissues were found widely constant, with a range of  $\pm 0.31$ .

Five micrograms of RNA from each leaf, petiole, and root extract was applied in 100- $\mu\text{L}$  reactions for cDNA synthesis using the iScript cDNA Synthesis Kit (Bio-Rad) according to the manufacturer's instructions. Quantitative real-time PCR experiments were performed with the DyNAmo Flash SYBR Green quantitative real-time PCR Kit (Finnzymes) in 20- $\mu\text{L}$  reactions according to the manufacturer's instructions by adding 1  $\mu\text{L}$  of the cDNA preparations as template per reaction. Primer pairs were RTS\_for and RTS\_rev (UGT73C9 to -C11), RTIL\_for and RTIL\_rev (UGT73C12/C13), as well as ACT2\_for1 and ACT2\_rev1 (ACT2). Duplicates of each setup were run on a Qiagen Rotor-Gene Q Real-Time PCR cyclor with settings for melting, annealing, extension, and acquiring of 10 s at 95 $^{\circ}\text{C}$ , 10 s at 65 $^{\circ}\text{C}$ , 20 s at 72 $^{\circ}\text{C}$ , and 1 s at 76 $^{\circ}\text{C}$ , respectively.

Quantitative real-time PCR experiments were analyzed using LinRegPCR (version 12.7; Ramakers et al., 2003; Ruijter et al., 2009). Relative expression values were calculated as the ratios of the starting concentrations (N0) given for the ACT2 reference and the corresponding UGT73C primer sets in the LinRegPCR output.

## Extraction and Reglucosylation of *B. vulgaris* Sapogenins

Crude saponin extracts from the G- and P-type were obtained by boiling freshly harvested leaves for 10 min with 5 mL of 55% ethanol  $\text{g}^{-1}$  fresh leaf material. Extracts were cooled on ice, centrifuged to remove insoluble particles, and the cleared supernatant was stored at  $-20^{\circ}\text{C}$  for more than 4 h. Precipitates were removed by centrifugation, and HCl was added to a final concentration of 1 M followed by incubation for 24 h at 99 $^{\circ}\text{C}$  and 1,400 rpm. A 1.2-fold volume of 1 M Tris base was added to shift the pH to basic conditions, and ethanol concentrations were adjusted to 14%. Polyvinylpyrrolidone and BSA were added to final concentrations of 50 mg  $\text{mL}^{-1}$  and 10 mg  $\text{mL}^{-1}$ , respectively, followed by six extractions each with one-tenth volume of ethyl acetate. The ethyl acetate fractions were combined, and solvent was removed in a vacuum concentrator. Metabolites were redissolved in 96% ethanol, and the polyvinylpyrrolidone/BSA-based purification step was repeated in one-tenth scale. Finally, the sapogenin-containing extracts were dissolved in 1 mL of 96% ethanol per initially applied 2.5 mL of hydrolyzed leaf extract.

Enzymatic activity assays were performed in a volume of 50  $\mu\text{L}$  with reaction conditions of 25 mM TAPS, pH 8.6 (UGT73C9 to -C11), pH 7.9 (UGT73C12/C13), or pH 8.2 (combination of UGT73C9, UGT73C10, or UGT73C11 with UGT73C12 or UGT73C13), 1 mM DTT, 1 mM UDP-Glc, and with diluted *E. coli* crude protein extracts containing in total 750 ng of the recombinant UGT73C(s). Aliquots of the sapogenin-containing extracts were dried in a vacuum concentrator and redissolved in 1  $\mu\text{L}$  of DMSO per 6.4  $\mu\text{L}$  of the initial sapogenin-containing ethanol solution. Addition of 3.13  $\mu\text{L}$  of the sapogenin-containing DMSO solutions was used to start reactions after 3 min of preincubation at 30 $^{\circ}\text{C}$ . Reactions were incubated for 30 or 120 min at 30 $^{\circ}\text{C}$ , and enzymatic activities were subsequently stopped by the addition of 325  $\mu\text{L}$  of ice-cold methanol. Precipitated proteins were removed by centrifugation, and the supernatant was evaporated to dryness in a vacuum concentrator. The dried extracts were redissolved in 60  $\mu\text{L}$  of 50% methanol, filtered (polyvinylidene difluoride; 0.45- $\mu\text{m}$  pore diameter), and subjected to liquid chromatography-mass spectrometry analysis (see above).

## Production of Hederagenin and Oleanolic Acid Monoglucosides for NMR and Bioassays

For large-scale production of hederagenin and oleanolic acid monoglucoside, four 2-L Erlenmeyer flasks, containing 250 mL of TB medium with 50  $\mu\text{g}$   $\text{mL}^{-1}$  kanamycin, were inoculated with fresh XJb(DE3) colonies harboring the pET28::UGT73C10 plasmid and incubated for 12 h at 30 $^{\circ}\text{C}$  and 180 rpm. Addition of 500 mL of TB medium and adjustment of the final concentrations of kanamycin, Ara, and IPTG to 50  $\mu\text{g}$   $\text{mL}^{-1}$ , 3 mM, and 0.1 mM, respectively, were followed by further incubation at 15 $^{\circ}\text{C}$  and 140 rpm for 24 h. The bacteria were harvested by centrifugation, resuspended in 10 mM HEPES, pH 7.9, and frozen at  $-80^{\circ}\text{C}$ . Lysis was achieved by thawing bacteria in a water bath at room temperature. DNA was degraded by treatment with DNase I (0.01 mg  $\text{mL}^{-1}$ , 5 mM  $\text{MgCl}_2$ , and 1 mM  $\text{CaCl}_2$ ). Cell debris were removed by centrifugation, and the supernatant was adjusted to 20 mM HEPES, pH 7.9, and 500 mM NaCl prior to the addition of 3 mL of equilibrated HIS-Select Nickel Affinity Gel (Sigma-Aldrich). One hour of incubation at 4 $^{\circ}\text{C}$  was followed by removal of the supernatant and three times washing of the affinity gel with 20 mM HEPES, pH 7.9, and 500 mM NaCl and once with 25 mM TAPS, pH 8.6, and 1 mM DTT. Enzymatic reactions were set up in 100-mL glass flasks at a final volume of 50 mL. The reaction conditions were 25 mM TAPS, pH 8.6, 1 mM DTT, and 750  $\mu\text{M}$  UDP-Glc. Approximately 1.5 mL of UGT73C10-loaded affinity gel was added to each reaction mixture, and enzymatic reactions were started by the addition of 10 mg of hederagenin (Extrasynthese) and oleanolic acid (Extrasynthese) dissolved in 3.125 mL of DMSO. Reaction mixtures were incubated at 37 $^{\circ}\text{C}$  and 150 rpm, and progressing glucosylation of the two sapogenins was monitored by TLC analysis of 20- $\mu\text{L}$  aliquots.

Hederagenin and oleanolic acid monoglucosides were extracted with ethyl acetate and, after evaporation of the solvent in a vacuum concentrator, dissolved in 60% to 70% DMSO prior to application to preparative HPLC for further purification. An Agilent 1200 series preparative HPLC system (Agilent Technologies), fitted with a Phenomenex Synergi 4 $\mu$  Hydro-RP column (21.2  $\times$  250 mm, 4  $\mu\text{m}$ , 80  $\text{\AA}$ ; Phenomenex), was used for this. Elution was carried out using a mobile phase containing acetonitrile and water with 0.01% trifluoroacetic acid. The gradient protocol was as follows: 5% acetonitrile for 5 min, linear gradient from 5% to 30% acetonitrile for 5 min, linear gradient from 30% to 100% acetonitrile for 50 min, and 100% acetonitrile for 5 min, at a constant flow rate of 15 mL  $\text{min}^{-1}$ . A diode array detector was used to monitor the elution of compounds by their UV absorption at 200 nm. Fractions

containing oleanolic acid and hederagenin glucosides were collected and evaporated to dryness using a vacuum concentrator.

The purified hederagenin and oleanolic acid monoglucosides were dissolved in NMR-suitable methanol-d<sub>4</sub> (Sigma-Aldrich), and NMR spectra were recorded at room temperature on a Bruker Avance DSX 500-MHz NMR spectrometer (Bruker Daltonics) equipped with a broadband inverse probe. Acquired data were calibrated according to the residual solvent peaks at 3.31 ppm for <sup>1</sup>H spectra and 49.01 ppm for <sup>13</sup>C spectra. For structural elucidation of the two monoglucosides, 1-D <sup>1</sup>H and <sup>13</sup>C as well as 2-D COSY, TOCSY, and HSQC experiments were performed and compared with corresponding spectra of oleanolic acid and hederagenin and reported NMR data of structurally related compounds (Supplemental Data Set S1).

### Phyllotreta nemorum Feeding Assays

Nonchoice feeding assays were performed as described previously by Nielsen et al. (2010). Briefly, purified 3-O-β-D-Glc hederagenin and 3-O-β-D-Glc oleanolic acid were in final concentrations of 2, 0.5, and 0.125 mM dissolved in 75% ethanol. Sapogenin monoglucoside solution (15 μL) was painted on both sides of 95-mm<sup>2</sup> radish (*Raphanus sativus*) leaf discs, which resulted in doses of 60 nmol (632 pmol mm<sup>-2</sup>), 15 nmol (158 pmol mm<sup>-2</sup>), and 3.75 nmol (39 pmol mm<sup>-2</sup>) sapogenin monoglucoside per leaf disc. Control leaf discs were treated with solvent only. Two identically treated leaf discs were exposed to one beetle for 24 h. Consumed leaf area was measured with a stereomicroscope. For the origin and maintenance of the two flea beetle (*P. nemorum*) lines, see Nielsen et al. (2010).

Results were analyzed using the R software package (www.r-project.org). The linear effect model allowed for a possible correlation between measurements from the same beetle. The starting model included a three-way interaction between beetle line, compound type, and dose; a 5% significance level was used for model reduction tests.

Sequence data from this article can be found in the GenBank/EMBL data libraries under accession numbers JQ291611 (*BvUGT1*), JQ291612 (*UGT73C9*), JQ291613 (*UGT73C10*), JQ291614 (*UGT73C11*), JQ291615 (*UGT73C12*), and JQ291616 (*UGT73C13*).

### Supplemental Data

The following materials are available in the online version of this article.

**Supplemental Figure S1.** Comparison of oleanolic acid glucosylation products after long-term incubation of oleanolic acid with UGT73C10, UGT73C11, UGT73C12, and UGT73C13 with oleanolic acid cellobioside.

**Supplemental Figure S2.** Alkaline hydrolysis (saponification) of betulinic acid glucosylation products derived from UGT73C13 activity.

**Supplemental Figure S3.** Comparison of UGTs from *B. vulgaris* (UGT73C11, UGT73C13, and UGT73C9) and *A. thaliana* (UGT73C5 and UGT73B5) in their activity toward hederagenin, 24-epi-brassinolide, and TCP.

**Supplemental Figure S4.** Comparison of UDP-Glc and UDP-Gal as sugar donor substrates of UGT73C10.

**Supplemental Figure S5.** Determination of *K<sub>m</sub>* values of UDP-Glc for UGT73C11 and UGT73C12.

**Supplemental Figure S6.** Kinetics of UGT73C11 and UGT73C13 with oleanolic acid and hederagenin as acceptor substrates.

**Supplemental Figure S7.** Liquid chromatography-mass spectrometry analysis of a G-type *B. vulgaris* metabolite extracted with 55% ethanol.

**Supplemental Figure S8.** Liquid chromatography-mass spectrometry analysis of a P-type *B. vulgaris* metabolite extracted with 55% ethanol.

**Supplemental Figure S9.** Liquid chromatography-mass spectrometry analysis of an acidic hydrolyzed G-type *B. vulgaris* metabolite extract.

**Supplemental Figure S10.** Liquid chromatography-mass spectrometry analysis of an acidic hydrolyzed P-type *B. vulgaris* metabolite extract.

**Supplemental Figure S11.** Glucosylation activity of UGT73C9 to UGT73C13 toward G-type and P-type *B. vulgaris* sapogenin extracts.

**Supplemental Figure S12.** Overlaid Liquid chromatography-mass spectrometry analyses of metabolite extracts from the *B. vulgaris* plants

used for the saponin abundance and UGT73C9 to -C13 expression correlation analysis.

**Supplemental Figure S13.** Comparison of UGTs from *B. vulgaris* (UGT73C11, UGT73C13, and UGT73C9) and *A. thaliana* (UGT73C5 and UGT73B5) in their activity toward sapogenins.

**Supplemental Table S1.** Amino acid and nucleotide sequence identities of UGT73s used in the phylogenetic analysis.

**Supplemental Table S2.** Primers used in this study.

**Supplemental Data Set S1.** Structure elucidation of hederagenin and oleanolic acid monoglucosides based on 1-D <sup>1</sup>H- and <sup>13</sup>C- and 2-D TOCSY-, COSY-, and HSQC-NMR data.

**Supplemental Data Set S2.** Multiple sequence alignment, amino acid sequences, and nucleotide sequences used for the phylogenetic analysis.

### ACKNOWLEDGMENTS

We are grateful to Rubini Kannagara for helpful discussions and commenting on the manuscript. Mohammed Saddik Motawie and Henrik Toft Simonsen are thanked for consulting in chemical aspects, and Tamara van Mølken is thanked for *P. nemorum* images. Peter McKenzie is acknowledged for naming the UGTs according to the UGT nomenclature. Mika Zagrobelny is thanked for help and discussions on the use of codeml in the PAML package.

Received June 28, 2012; accepted September 30, 2012; published October 1, 2012.

### LITERATURE CITED

- Agerbirk N, Olsen CE, Bibby BM, Frandsen HO, Brown LD, Nielsen JK, Renwick JAA (2003a) A saponin correlated with variable resistance of *Barbarea vulgaris* to the diamondback moth *Plutella xylostella*. *J Chem Ecol* **29**: 1417–1433
- Agerbirk N, Ørsgaard M, Nielsen JK (2003b) Glucosinolates, *P. nemorum* resistance, and leaf pubescence as taxonomic characters in the genus *Barbarea* (Brassicaceae). *Phytochemistry* **63**: 69–80
- Aharoni A, Gaidukov L, Khersonsky O, McQ Gould S, Roodveldt C, Tawfik DS (2005) The ‘evolvability’ of promiscuous protein functions. *Nat Genet* **37**: 73–76
- Augustin JM, Kuzina V, Andersen SB, Bak S (2011) Molecular activities, biosynthesis and evolution of triterpenoid saponins. *Phytochemistry* **72**: 435–457
- Badenes-Perez FR, Shelton AM, Nault BA (2005) Using yellow rocket as a trap crop for diamondback moth (Lepidoptera: Plutellidae). *J Econ Entomol* **98**: 884–890
- Brazier-Hicks M, Edwards R (2005) Functional importance of the family 1 glucosyltransferase UGT72B1 in the metabolism of xenobiotics in *A. thaliana thaliana*. *Plant J* **42**: 556–566
- Caputi L, Lim E-K, Bowles DJ (2008) Discovery of new biocatalysts for the glycosylation of terpenoid scaffolds. *Chemistry* **14**: 6656–6662
- Carelli M, Biazzi E, Panara F, Tava A, Scaramelli L, Porceddu A, Graham N, Odoardi M, Piano E, Arcioni S, et al (2011) *Medicago truncatula* CYP716A12 is a multifunctional oxidase involved in the biosynthesis of hemolytic saponins. *Plant Cell* **23**: 3070–3081
- Chang S, Puryear J, Cairney J (1993) A simple and efficient method for isolating RNA from pine trees. *Plant Mol Biol Rep* **11**: 113–116
- Dalby-Brown L, Olsen CE, Nielsen JK, Agerbirk N (2011) Polymorphism for novel tetraglycosylated flavonols in an eco-model crucifer, *Barbarea vulgaris*. *J Agric Food Chem* **59**: 6947–6956
- De Geyter E, Smaghe G, Rahbé Y, Geelen D (2012) Triterpene saponins of *Quillaja saponaria* show strong aphicidal and deterrent activity against the pea aphid *Acyrtosiphon pisum*. *Pest Manag Sci* **68**: 164–169
- de Jong PW, Breuker CJ, Vos H, Vermeer KMCA, Oku K, Verbaarschot P, Nielsen JK, Brakefield PM (2009) Genetic differentiation between resistance phenotypes in the phytophagous *P. nemorum*, *Phyllotreta nemorum*. *J Insect Sci* **9**: 1–8
- De Leo M, De Tommasi N, Sanogo R, D’Angelo V, Germanò MP, Bisignano G, Braca A (2006) Triterpenoid saponins from *Pteleopsis suberosa* stem bark. *Phytochemistry* **67**: 2623–2629

- Dowd PF, Berhow MA, Johnson ET (2011) Differential activity of multiple saponins against omnivorous insects with varying feeding preferences. *J Chem Ecol* **37**: 443–449
- Glendinning JI (2002) How do herbivorous insects cope with noxious secondary plant compounds in their diet? *Entomol Exp Appl* **104**: 15–25
- Hauser TP, Toneatto F, Nielsen JK (2012) Genetic and geographic structure of an insect resistant and a susceptible type of *Barbarea vulgaris* in western Europe. *Evol Ecol* **26**: 611–624
- Hou B, Lim E-K, Higgins GS, Bowles DJ (2004) N-Glucosylation of cytokinins by glycosyltransferases of *A. thaliana* thaliana. *J Biol Chem* **279**: 47822–47832
- Husar S, Berthiller F, Fujioka S, Rozhon W, Khan M, Kalaivanan F, Elias L, Higgins GS, Li Y, Schuhmacher R, et al (2011) Overexpression of the UGT73C6 alters brassinosteroid glucoside formation in *A. thaliana* thaliana. *BMC Plant Biol* **11**: 51
- Jensen RA (1976) Enzyme recruitment in evolution of new function. *Annu Rev Microbiol* **30**: 409–425
- Jones P, Messner B, Nakajima J, Schäffner AR, Saito K (2003) UGT73C6 and UGT78D1, glycosyltransferases involved in flavonol glucoside biosynthesis in *A. thaliana* thaliana. *J Biol Chem* **278**: 43910–43918
- Khersonsky O, Tawfik DS (2010) Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu Rev Biochem* **79**: 471–505
- Kuzina V, Ekström CT, Andersen SB, Nielsen JK, Olsen CE, Bak S (2009) Identification of defense compounds in *Barbarea vulgaris* against the herbivore *Phyllotreta nemorum* by an ecometabolomic approach. *Plant Physiol* **151**: 1977–1990
- Kuzina V, Nielsen JK, Augustin JM, Torp AM, Bak S, Andersen SB (2011) *Barbarea vulgaris* linkage map and quantitative trait loci for saponins, glucosinolates, hairiness and resistance to the herbivore *Phyllotreta nemorum*. *Phytochemistry* **72**: 188–198
- le Provost G, Herrera R, Paiva JA, Chaumeil P, Salin F, Plomion C (2007) A micromethod for high throughput RNA extraction in forest trees. *Biol Res* **40**: 291–297
- Lim E-K, Ashford DA, Hou B, Jackson RG, Bowles DJ (2004) *A. thaliana* glycosyltransferases as biocatalysts in fermentation for regioselective synthesis of diverse quercetin glucosides. *Biotechnol Bioeng* **87**: 623–631
- Lim E-K, Baldauf S, Li Y, Elias L, Worrall D, Spencer SP, Jackson RG, Taguchi G, Ross J, Bowles DJ (2003) Evolution of substrate recognition across a multigene family of glycosyltransferases in *A. thaliana*. *Glycobiology* **13**: 139–145
- Lodeiro S, Schulz-Gasch T, Matsuda SPT (2005) Enzyme redesign: two mutations cooperate to convert cycloartenol synthase into an accurate lanosterol synthase. *J Am Chem Soc* **127**: 14132–14133
- Luo J, Nishiyama Y, Fuell C, Taguchi G, Elliott K, Hill L, Tanaka Y, Kitayama M, Yamazaki M, Bailey P, et al (2007) Convergent evolution in the BAHF family of acyl transferases: identification and characterization of anthocyanin acyl transferases from *A. thaliana* thaliana. *Plant J* **50**: 678–695
- Luukkanen L, Taskinen J, Kurkela M, Kostiaainen R, Hirvonen J, Finel M (2005) Kinetic characterization of the 1A subfamily of recombinant human UDP-glucuronosyltransferases. *Drug Metab Dispos* **33**: 1017–1026
- Mackenzie PI, Owens IS, Burchell B, Bock KW, Bairoch A, Bélanger A, Fournel-Gigleux S, Green M, Hum DW, Iyanagi T, et al (1997) The UDP glycosyltransferase gene superfamily: recommended nomenclature update based on evolutionary divergence. *Pharmacogenetics* **7**: 255–269
- Meesapyodsuk D, Balsevich J, Reed DW, Covello PS (2007) Saponin biosynthesis in *Saponaria vaccaria*: cDNAs encoding  $\beta$ -amyrin synthase and a triterpene carboxylic acid glycosyltransferase. *Plant Physiol* **143**: 959–969
- Messner B, Thulke O, Schäffner AR (2003) *A. thaliana* glycosyltransferases with activities toward both endogenous and xenobiotic substrates. *Planta* **217**: 138–146
- Modolo LV, Blount JW, Achnine L, Naoumkina MA, Wang X, Dixon RA (2007) A functional genomics approach to (iso)flavonoid glycosylation in the model legume *Medicago truncatula*. *Plant Mol Biol* **64**: 499–518
- Musende AG, Eberding A, Wood C, Adomat H, Fazli L, Hurtado-Coll A, Jia W, Bally MB, Guns ET (2009) Pre-clinical evaluation of Rh2 in PC-3 human xenograft model for prostate cancer in vivo: formulation, pharmacokinetics, biodistribution and efficacy. *Cancer Chemother Pharmacol* **64**: 1085–1095
- Naoumkina MA, Modolo LV, Huhman DV, Urbanczyk-Wochniak E, Tang Y, Sumner LW, Dixon RA (2010) Genomic and coexpression analyses predict multiple genes involved in triterpene saponin biosynthesis in *Medicago truncatula*. *Plant Cell* **22**: 850–866
- Nielsen JK (1997a) Variation in defences of the plant *Barbarea vulgaris* and in counteradaptations by the *P. nemorum* *Phyllotreta nemorum*. *Entomol Exp Appl* **82**: 25–35
- Nielsen JK (1997b) Genetics of the ability of *Phyllotreta nemorum* larvae to survive in an atypical host plant, *Barbarea vulgaris* ssp. *arcuata*. *Entomol Exp Appl* **82**: 37–44
- Nielsen JK (1999) Specificity of a Y-linked gene in the *P. nemorum* *Phyllotreta nemorum* for defences in *Barbarea vulgaris*. *Entomol Exp Appl* **91**: 359–368
- Nielsen JK (2012) Non-random segregation of an autosomal gene in males of the *P. nemorum*, *Phyllotreta nemorum*: implications for colonization of a novel host plant. *Entomol Exp Appl* **143**: 301–312
- Nielsen JK, Nagao T, Okabe H, Shinoda T (2010) Resistance in the plant, *Barbarea vulgaris*, and counter-adaptations in *P. nemorum* mediated by saponins. *J Chem Ecol* **36**: 277–285
- Nihei K-I, Ying B-P, Murakami T, Matsuda H, Hashimoto M, Kubo I (2005) Pachyelasides A-D, novel molluscicidal triterpene saponins from *Pachyelasma tessmannii*. *J Agric Food Chem* **53**: 608–613
- Noguchi A, Saito A, Homma Y, Nakao M, Sasaki N, Nishino T, Takahashi S, Nakayama T (2007) A UDP-glucose:isoflavone 7-O-glycosyltransferase from the roots of soybean (*Glycine max*) seedlings: purification, gene cloning, phylogenetics, and an implication for an alternative strategy of enzyme catalysis. *J Biol Chem* **282**: 23581–23590
- Ono E, Homma Y, Horikawa M, Kunikane-Doi S, Imai H, Takahashi S, Kawai Y, Ishiguro M, Fukui Y, Nakayama T (2010) Functional differentiation of the glycosyltransferases that contribute to the chemical diversity of bioactive flavonol glycosides in grapevines (*Vitis vinifera*). *Plant Cell* **22**: 2856–2871
- Osborn A (2010) Gene clusters for secondary metabolic pathways: an emerging theme in plant biology. *Plant Physiol* **154**: 531–535
- Osborn AE, Clarke BR, Dow JM, Daniels MJ (1991) Partial characterization of avenacinase from *Gaeumannomyces graminis* var. *avenae*. *Physiol Mol Plant Pathol* **38**: 301–312
- Papadopoulou K, Melton RE, Leggett M, Daniels MJ, Osborn AE (1999) Compromised disease resistance in saponin-deficient plants. *Proc Natl Acad Sci USA* **96**: 12923–12928
- Paquette S, Møller BL, Bak S (2003) On the origin of family 1 plant glycosyltransferases. *Phytochemistry* **62**: 399–413
- Pareja-Jaime Y, Roncero MIG, Ruiz-Roldán MC (2008) Tomatinase from *Fusarium oxysporum* f. sp. *lycopersici* is required for full virulence on tomato plants. *Mol Plant Microbe Interact* **21**: 728–736
- Phillips DR, Rasbery JM, Bartel B, Matsuda SPT (2006) Biosynthetic diversity in plant triterpene cyclization. *Curr Opin Plant Biol* **9**: 305–314
- Pollier J, Morreel K, Geelen D, Goossens A (2011) Metabolite profiling of triterpene saponins in *Medicago truncatula* hairy roots by liquid chromatography Fourier transform ion cyclotron resonance mass spectrometry. *J Nat Prod* **74**: 1462–1476
- Poppenberger B, Berthiller F, Bachmann H, Lucyshyn D, Peterbauer C, Mitterbauer R, Schuhmacher R, Krska R, Glössl J, Adam G (2006) Heterologous expression of *A. thaliana* UDP-glycosyltransferases in *Saccharomyces cerevisiae* for production of zearalenone-4-O-glucoside. *Appl Environ Microbiol* **72**: 4404–4410
- Poppenberger B, Berthiller F, Lucyshyn D, Sieberer T, Schuhmacher R, Krska R, Kuchler K, Glössl J, Leuschnic G, Adam G (2003) Detoxification of the *Fusarium* mycotoxin deoxynivalenol by a UDP-glycosyltransferase from *A. thaliana* thaliana. *J Biol Chem* **278**: 47905–47914
- Poppenberger B, Fujioka S, Soeno K, George GL, Vaistij FE, Hiranuma S, Seto H, Takatsuto S, Adam G, Yoshida S, et al (2005) The UGT73C5 of *A. thaliana* thaliana glucosylates brassinosteroids. *Proc Natl Acad Sci USA* **102**: 15253–15258
- Ramakers C, Ruijter JM, Deprez RHL, Moorman AFM (2003) Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. *Neurosci Lett* **339**: 62–66
- Renwick JAA (2002) The chemical world of crucifers: lures, treats and traps. *Entomol Exp Appl* **104**: 35–42
- Ruijter JM, Ramakers C, Hoogaars WMH, Karlen Y, Bakker O, van den Hoff MJB, Moorman AFM (2009) Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. *Nucleic Acids Res* **37**: e45
- Senatore F, D'Agostino M, Dini I (2000) Flavonoid glycosides of *Barbarea vulgaris* L. (Brassicaceae). *J Agric Food Chem* **48**: 2659–2662
- Shan H, Wilson WK, Phillips DR, Bartel B, Matsuda SPT (2008) Trinorlupeol: a major nonsterol triterpenoid in *A. thaliana*. *Org Lett* **10**: 1897–1900

- Shinoda T, Nagao T, Nakayama M, Serizawa H, Koshioka M, Okabe H, Kawai A** (2002) Identification of a triterpenoid saponin from a crucifer, *Barbarea vulgaris*, as a feeding deterrent to the diamondback moth, *Plutella xylostella*. *J Chem Ecol* **28**: 587–599
- Sun H-X, Xie Y, Ye Y-P** (2009) Advances in saponin-based adjuvants. *Vaccine* **27**: 1787–1796
- Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S** (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**: 2731–2739
- Thorsoe KS, Bak S, Olsen CE, Imberty A, Breton C, Lindberg Møller B** (2005) Determination of catalytic key amino acids and UDP sugar donor specificity of the cyanohydrin glycosyltransferase UGT85B1 from *Sorghum bicolor*: molecular modeling substantiated by site-specific mutagenesis and biochemical analyses. *Plant Physiol* **139**: 664–673
- Toneatto F, Hauser TP, Nielsen JK, Ørgaard M** (2012) Genetic diversity and similarity in the *Barbarea vulgaris* complex (Brassicaceae). *Nord J Bot* **30**: 506–512
- Toneatto F, Nielsen JK, Ørgaard M, Hauser TP** (2010) Genetic and sexual separation between insect resistant and susceptible *Barbarea vulgaris* plants in Denmark. *Mol Ecol* **19**: 3456–3465
- Vogt T, Jones P** (2000) Glycosyltransferases in plant natural product synthesis: characterization of a supergene family. *Trends Plant Sci* **5**: 380–386
- Weis M, Lim E-K, Bruce N, Bowles D** (2006) Regioselective glucosylation of aromatic compounds: screening of a recombinant glycosyltransferase library to identify biocatalysts. *Angew Chem Int Ed Engl* **45**: 3534–3538
- Weng JK, Philippe RN, Noel JP** (2012) The rise of chemodiversity in plants. *Science* **336**: 1667–1670
- Wubben JP, Price KR, Daniels MJ, Osbourn AE** (1996) Detoxification of oat leaf saponins by *Septoria avenae*. *Phytopathology* **86**: 986–992



***Popular Science Paper***  
***&***  
***Conference***  
***Presentations***





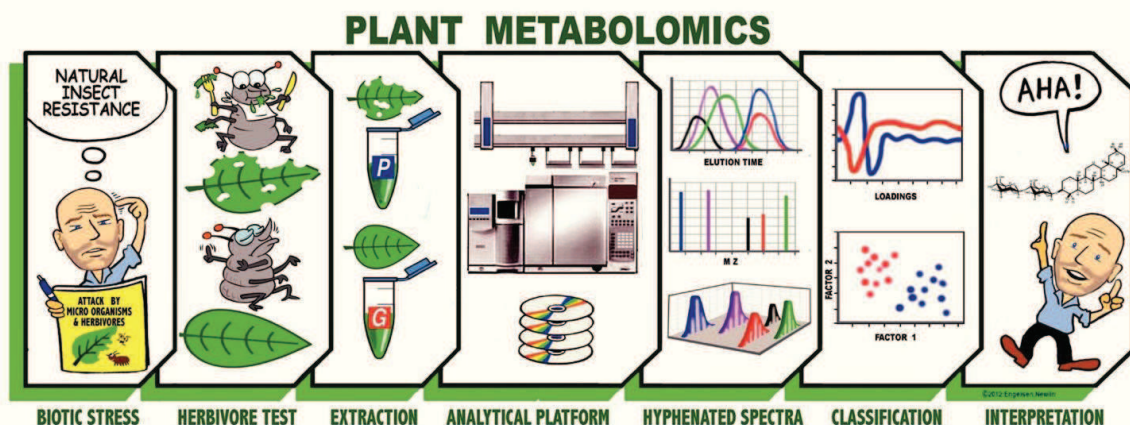
# Plante-metabolomics: opdagelse af nye bioaktive stoffer med PARAFAC2

PARAFAC2 tillader automatisk peak-detektion af lavintensitet og stærkt overlappende toppe i LC-MS og kan derfor være et nyttigt eksplorativt redskab for opdagelse af nye bioaktive stoffer i planter.

Af Bekzod Khakimov, Søren Balling Engelsen, Rasmus Bro, Institut for Fødevarevidenskab, Københavns Universitet og Lars Nørgaard, FOSS

Plante-metabolomics handler om kvantitativ og/eller kvalitativ analyse af metaboliter fra plantevæv [1]. Fremskridt indenfor udviklingen af analytisk udstyr (GC, LC, MS and NMR) og

rende vis have opstillet en hypotese om tørkeresistens, som er et "hot-topic" pga. de globalt set stigende klimaudfordringer for vores cerealieproduktion. Vi har i dette tilfælde behov for et bioaktivitets-assay, hvilket i dette tilfælde er enkelt. Man udsætter et antal biller på blade fra de to genotyper af *Barbarea vulgaris*, og efter et passende stykke tid, bestemmes den spiste bladmasse ved gravimetri.



Figur 1. Typisk "work-flow" inden for plante-metabolomics.

udviklingen af nye multivariate data-teknikker har ført til store fremskridt i viden om plante-metabolomet og dets variationer, når det bliver udsat for indre og ydre perturbationer som f.eks. insektangreb og klimaforandringer. Metabolomics kan opfattes som det phenotypiske endepunkt af sekvensen *genomics - transcriptomics - proteomics - metabolomics*, der reflekterer dynamikken i planten. Derfor er plante-metabolomics blevet en nøgleteknologi i forståelsen af kompleksiteten af plante-metabolismen, hvordan den kontrolleres og som forbindelsesled mellem genotype og phenotype. Figur 1 viser et typisk "work-flow" i plante-metabolomics.

I dette tilfælde ønsker vi at undersøge, hvordan det kan være, at én genotype af planten *Barbarea vulgaris* (vinterkarse) er resistent mod billen med det flotte navn *Phyllotreta nemorum* (jordlopper), der er et betydeligt skadedyr i f.eks. rapsmarker, mens en anden genotype ikke er det [2]. Man kunne på tilsva-

Derefter ekstraheres plante-metabolomet over i en væskefase, da ingen af de analytiske teknikker i metabolomics er særlig gode til at studere fast fase. Denne ekstraktion er kritisk for resultatet af plante-metabolomics, da den vil introducere en be- ▶

**SKANLAB**      **Retsch**  
Solutions in Milling & Sieving

[www.retsch.dk](http://www.retsch.dk)  
[birte@skanlab.com](mailto:birte@skanlab.com)

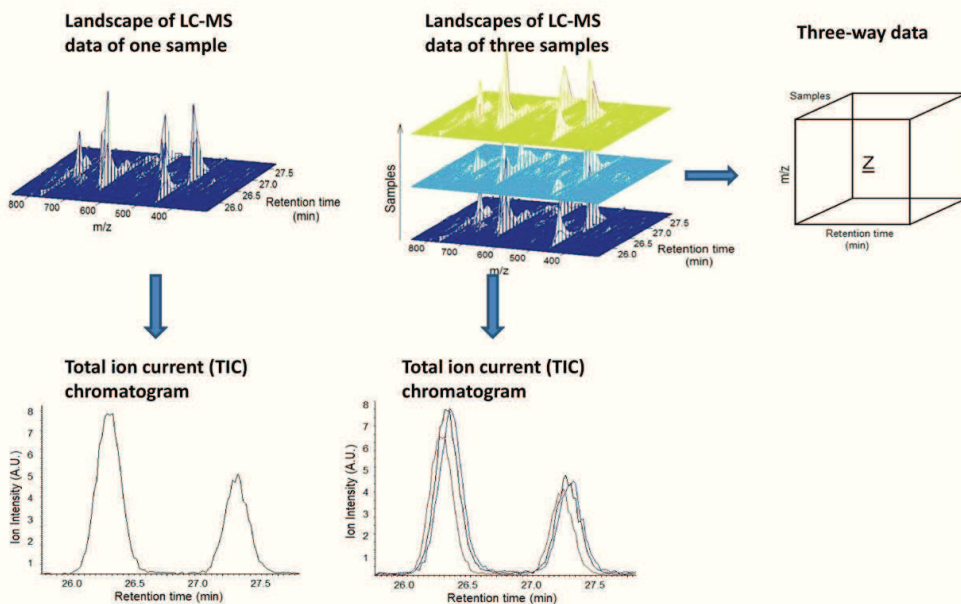
# ■ DET KEMOMETRISKE RUM

tydelig bias, som skyldes, at ikke alle plante-metabolitter deler netop de fysiske-kemiske egenskaber som optimeres ved valg af solvent, temperatur, mixing-tid etc. I vores tilfælde har vi optimeret proceduren for ekstraktion af triterpenoide, som plante-biologerne allerede ved er stærkt bioaktive over for insekter.

Til at undersøge hvilke stoffer, der er ansvarlige for *Barbarea vulgaris* plantens insektresistens, ønsker vi at foretage en metabolom-profilering af 127 planter fra en segregeret population, der stammer fra en krydsning af en resistent G-type og en modtagelig P-type plante. Ved krydsning opnås en segregeret population, der spænder over hele spektret fra den fulde resistens i G-typen til den fulde modtagelighed i P-typen. Ligeledes vil metabolit-profilerne i populationen blive randomiseret. Til at undersøge vores plante-metabolom benytter vi LC-MS som analytisk platform med en påkrævet høj følsomhed. Der benyttes endvidere en eksperimentel metode, der er optimeret for triterpenoide og som har en relativ hurtig scanningshastighed.

PARAFAC2 [3] kan resolve de ubehandlede LC-MS-landskaber op i tre modes: en score-vektor, der indeholder koncentrationen af de resolvede metabolitter, et massespektrum for hver af de resolvede metabolitter samt en elueringsprofil for hver prøve. Desværre går det ikke at foretage PARAFAC2-modellering på det komplette LC-MS-datasæt, da kompleksiteten bliver for høj. Det er derfor nødvendigt at opdele problemet i en række mindre delproblemer. De komplette LC-MS kromatografiske profiler af vores planteekstrakter deles op i 17 (elution time) intervaller (se figur 4). Hvert interval er baselinje-separeret fra de tilstødende intervaller, således at man ikke opdeler kromatografiske toppe, der hører til samme stof.

PARAFAC2-modellering udføres nu på hvert af de 17 intervaller. Figur 5 viser et eksempel på PARAFAC2's fantastiske evne til at resolve stærkt overlappende toppe. I dette elueringsinterval, som umiddelbart ser fuldstændigt uoverskueligt ud, er PARAFAC2-algoritmen i stand til at identificere hele 6

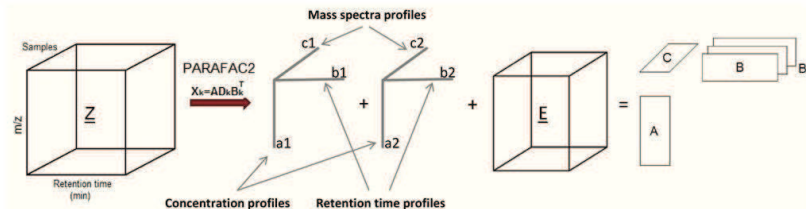


Figur 2. Struktur af LC-MS-data: (Venstre) et LCMS-landskab målt på en enkelt prøve samt (forneden) total ionkromatogram af samme prøver. (midt) Samme måling men flere prøver. (Højre) Tre-vejs-datastrukturen som opnås ved at stable de rå LC-MS-data fra flere forskellige prøver.

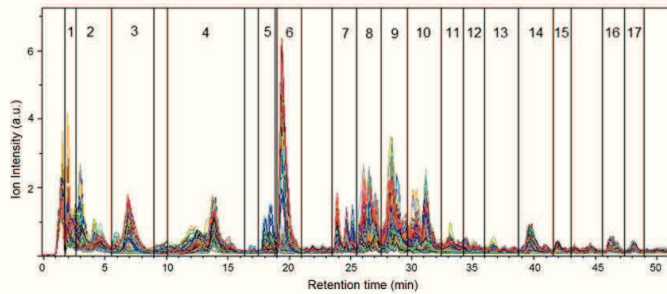
Figur 2 viser strukturen af LC-MS-data og hvordan disse, når man har mange prøver, kan stablet til en trevejs datastruktur. Vi ser også, at data i elueringsprofilen har betydelige eluerings-skift fra prøve til prøve. Som vi har vist i forrige klumme (Dansk Kemi, 93 (11) 2012) passer denne type datastruktur perfekt til PARAFAC2 algoritmen (figur 3).

underliggende toppe. Det er lidt som at åbne sin julekalender og gå på opdagelse efter hidtil ukendte plante-metabolitter! I dette tilfælde viste det sig, at en af de 6 underliggende toppe repræsenterer en hidtil ukendt bioaktiv metabolit: en glycosyleret saponin med en trisaccharid bundet til aglycon-skelettet.

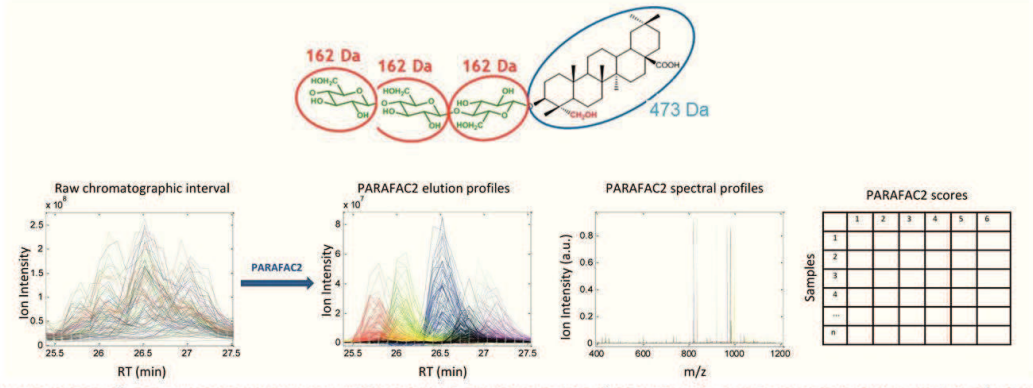
Ved at dele vores LC-MS-problem op i 17 intervaller er PARAFAC2-algoritmen i stand til at identificere samlet set 71 peaks. Hvert af disse peaks svarer til en plante-metabolit, og for hver af disse metabolitter finder PARAFAC2-algoritmen en individuel relativ koncentration, som efterfølgende kan korreleres eller kalibreres til den målte bioaktivitet. I dette tilfælde niveauet af resistens målt som mængde spist bladmasse.



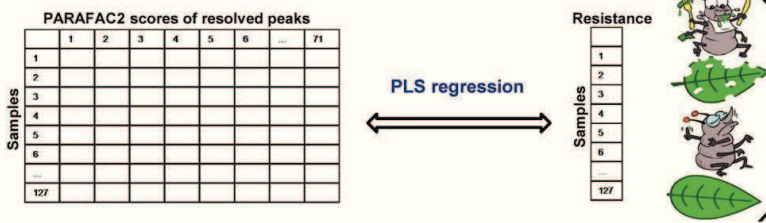
Figur 3. En skematisk oversigt over PARAFAC2-modellen.



Figur 4. For at reducere kompleksiteten af de 127 LC-MS-data opdeles disse i 17 baselinje-separerede intervaller i elueringsretningen.



Figur 5. Eksempel på en syv komponent PARAFAC2-model af et interval af de rå LC-MS-data. (a) Superimoseret plot af de rå LC-MS-kromatogrammer, (b) de resolvede elueringsprofiler, (c) de tilhørende massespektre og (d) koncentrationsprofilerne af de resolvede plante-metabolitter.



Figur 6. Korrelations- og PLS-regressionsanalyse mellem PARAFAC2-koncentrationer af de 71 resolvede metabolit-toppe mod resistanceniveau af de 127 planter. Resistance-data for de 127 planters er defineret til at have værdier mellem 0 og 5, hvor de som er mest resistente får tildelt værdien 0, mens de mest modtagelige planter får værdien 5.

Figur 6 viser, hvordan vi kan opstille en metabolit-tabel med PARAFAC2-scores og via PLS lave regression til bioaktivitet.

**Outro**

Vi håber med dette eksempel at have vist, at PARAFAC2 med fordel kan anvendes som supplement til eksisterende metoder, specielt når man har LC-MS-data med vanskeligt identificerbare toppe. PARAFAC2 er overraskende god til detektion af stærkt overlappede, eluerings-shiftede toppe selv med meget lavt signal-støj-forhold. Derudover har PARAFAC2-algoritmen den fordel, at den kan modellere direkte på de rå data uden nogen form for forbehandling, at dens løsninger er unikke, og at den giver relative koncentrationer, der kun behøver en skalering for at give absolutte koncentrationer.

**Vi vil gerne sige stor tak til professor Søren Bak, Institut for Plante- og Miljøvidenskab, Københavns Universitet, for**

**at have introduceret os til dette spændende problem med bioaktive stoffer i *Barbera vulgaris*.**

**E-mail**

Bekzod Khakimov: bzo@life.ku.dk  
 Søren Balling Engelsen: se@life.ku.dk  
 Rasmus Bro: rb@life.dk.  
 Lars Nørgaard: lno@foss.dk

**Referencer**

- O. Fiehn, *Metabolomics - the link between genotypes and phenotypes*, *Plant Molecular Biology*. **48** (2002) 155-171.
- B. Khakimov, J.M. Amigo, S. Bak, S.B. Engelsen. *Plant metabolomics: Resolution and quantification of elusive peaks in liquid chromatography-mass spectrometry profiles of complex plant extracts using multi-way decomposition methods*. *Journal of Chromatography A*, **1266** (2012):84-94
- R. Bro, C. A. Andersson, H. A. L. Kiers. *PARAFAC2 - Part II. Modeling chromatographic data with retention time shifts*. *Journal of Chemometrics*, **13** (1999) 295-309.

**Plant metabolomics – PARAFAC2 resolution of bioactive triterpenoid saponins in LC-MS profiles from *Barbarea vulgaris* and implications for plant-insect interactions**

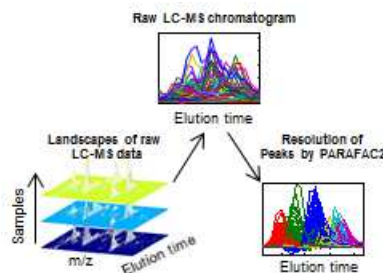
Bekzod Khakimov<sup>a,b</sup>, José Manuel Amigo<sup>a</sup>, Søren Bak<sup>b</sup>, Søren Balling Engelsen<sup>a</sup>

<sup>a</sup> Quality & Technology, Department of Food Science, Faculty of Science, University of Copenhagen, Rolighedsvej 30, 1958 Frederiksberg C, Denmark

<sup>b</sup> Plant Biochemistry, Department of Plant Biology and Biotechnology, Faculty of Science, University of Copenhagen, Rolighedsvej 30, 1958 Frederiksberg C, Denmark

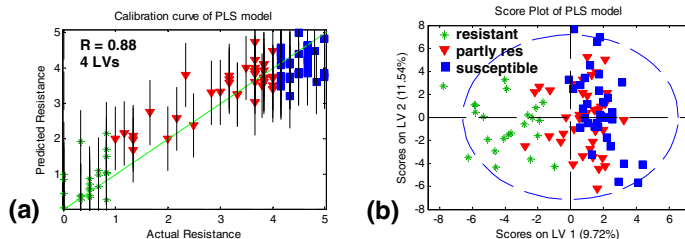
Previous studies on metabolomic profiling of 127 F2 *Barbarea vulgaris* plants derived from a cross between a parental glabrous (G) and pubescent (P) type by LC-MS, revealed four bioactive triterpenoid saponins (hederagenin cellobioside, oleanolic acid cellobioside, epihederagenin cellobioside, and gypsogenin cellobioside) that correlated with resistance against the insect herbivore, *Phyllotreta nemorum* [1]. Our work demonstrates the application of the multi-way technique PARAFAC2 [2] for resolving complex LC-MS data obtained from the 127 F2 *Barbarea vulgaris* plants. PARAFAC2 enabled resolution and quantification of several elusive (e.g. overlapped, elution time shifted and low S/N ratio) chromatographic peaks, which could not be detected and quantified by conventional chromatographic data analysis.

The score values obtained from PARAFAC2 models correspond to relative amounts of the resolved chromatographic peaks. This enabled a precise relative quantification of resolved peaks. A total of 71 peaks (including baselines and tails of neighboring peaks) were resolved from all the PARAFAC2 models developed for 17 different chromatographic intervals. Correlation analysis showed that 9 out of 71 resolved peaks significantly correlated with resistance against *P. nemorum* larvae herbivore. Subsequent partial least squares (PLS) regression analyses showed that the four previously identified bioactive saponins and five unknown saponin like chromatographic peaks were highly correlated to the resistance level of F2 plants. The method also enabled a good separation between resistant and susceptible plants [3].



**Figure 1.** PARAFAC2 model developed for selected region of the raw LC-MS data shows resolution of overlapped peaks.

**Figure 2. (a)** Regression curve of the PLS model. This curve shows the correlation between the actual resistance level of the F2 plants against insect herbivore and the predicted resistance found by PLS. **(b)** The score plot (LV1 vs LV2) of the PLS model shows separation between resistant (green) and susceptible (blue) F2 plants.



Complex and problematic chromatographic peaks with elution time shifts, strong overlaps and low S/N ratio were successfully modeled by PARAFAC2, without need for preprocessing the raw data. Spectral loadings resolved by PARAFAC2 matched well with the experimentally obtained mass spectra of peaks. All these features of PARAFAC2 elegantly illustrated its performance for quantitative and qualitative analysis of complex LC-MS metabolomic data.

**References**

1. Kuzina, V., Ekstrom, C.T., Andersen, S.B., Nielsen, J.K., Olsen, C.E., and Bak, S. 2009. *Plant Physiology* 151: 1977-1990.
2. Bro, R., Andersson, C.A., and Kiers, H.A.L. 1999. *Journal of Chemometrics* 13: 295-309.
3. Khakimov, B., Amigo J.M., Bak, S., and Engelsen, S.B., PARAFAC2 resolution of bioactive triterpenoid saponins in LC-MS profiles from *Barbarea vulgaris* and implications for plant-insect interactions, 2012, *submitted*.

Comprehensive Metabolomic Profiling of Phenolic and Organic Acids of Cereals using gas chromatography-mass spectrometry (GC-MS) and advanced chemometrics



Bekzod Khakimov\*, Birthe Møller Jespersen & Søren Balling Engelsen

Quality & Technology, Dept. Food Science, University of Copenhagen

\*Presenting author. E-mail: bzo@life.ku.dk, www.models.life.ku.dk



**Purpose**

Most metabolomic studies performed on cereal grains focus on primary metabolites. There is a need to improve existing methods for metabolomic profiling of primary and secondary metabolites with low concentrations. This study employs, for the first time, an improved protocol for extracting total phenolics, new GC-MS derivatization method and PARAFAC2 analysis for comprehensive metabolomics of phenolic and organic acids from grain flour samples.

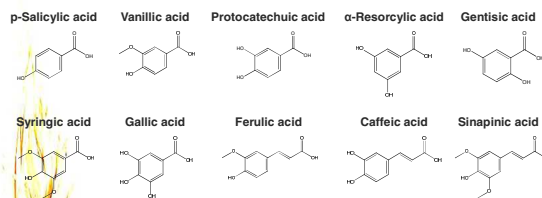
**Motivation**

Beyond the main bulk chemical components of cereals, low concentration metabolites, polyphenols and organic acids, contribute significantly to the quality and health beneficial properties. Cereals and cereal products are one of the richest sources of total polyphenol intake in human diet. The main bioactive polyphenols of cereals are phenolic compounds derived from hydroxybenzoic and hydroxycinnamic acids. Phenolic acids has been shown to be important texturizing agents in cooking-extrusion of cereals and they have been recognized as the main antioxidant constituents. Phenolics of cereals are mainly present in conjugated forms with sugars and other cell membrane components that alter their solubility and in turn their bioavailability. Holistic evaluation of bioactive metabolites of cereals require comprehensive, unbiased, sensitive and high-throughput analytical approaches.

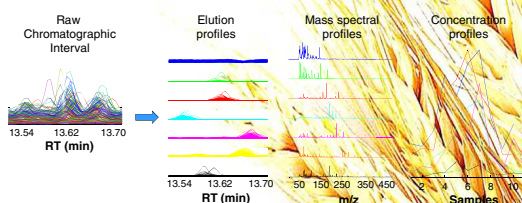
**Methodology**

The study include winter wheat (Hereward), Purple jasmine wheat, commercial rye and oat and jasmine rice. 50 mg of milled grain samples were extracted in 80% MeOH and hydrolyzed using hydrochloric acid. Phenolic extracts of cereals were trimethylsilylated using state-of-the-art derivatization method developed for unbiased GC-MS analysis [1]. Integrated Agilent GC-MS - GERSTEL MPS autosampler enabled complete automation of sample derivatization and injection. Complex raw GC-MS data was processed by multi-way decomposition method, Parallel Factor Analysis 2 (PARAFAC2) [2,3], which enable deconvolution of more than 200 metabolites, resolution of their pure mass spectra and precise quantification of the relative peak abundances. Combination of PARAFAC2 resolved mass spectra and GC-MS libraries, Wiley08 and NIST05, allowed identification of nearly one hundred metabolites.

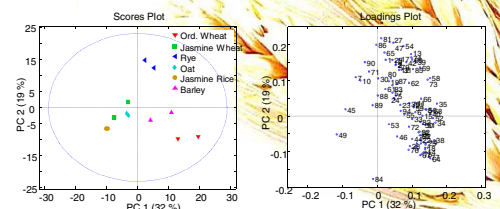
**Structures of 10 most dominant phenolics identified from cereal samples**



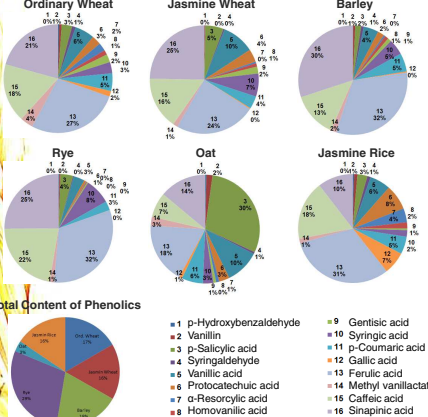
**PARAFAC2 based processing of raw GC-MS data**



**PCA analysis of Identified Phenolic and Organic Acids of Cereals**



**Relative concentrations of 16 identified most abundant PHENOLIC ACIDS**



**Main Findings**

- ✓ 247 metabolites with unique retention indices (RI) and electron impact-mass spectra (EI-MS) were resolved and quantified by PARAFAC2 modeling of raw GC-MS data
- ✓ 92 metabolites were identified based on their RI and EI-MS, that compiled phenolic acids, phenolic acid esters, small molecular organic acids, aldehydes, alcohols, sugar alcohols, fatty acids and sterols
- ✓ Ferulic acid was the most abundant phenolic compound in most of the samples
- ✓ Phenolic profiles of Oat and Jasmine Rice were significantly different compared to Wheat, Barley and Rye
- ✓ The relative concentration of Salicylic acid was highest in Oat
- ✓ Rye showed the highest content of total phenolics (sum of 16 most abundant phenolic acids), while Oat possessed the lowest concentration

**Conclusion**

The protocol developed for metabolomic profiling of phenolic and organic acids in grain flour samples show good reproducibility and excellent sensitivity and allows for quantitative detection of more than 200 metabolites from only 1 µl GC-MS injection. Detailed qualitative and quantitative information is gained on phenolic and organic acid profiles of the main cereals. The approach can easily be adopted in rapid screening of other food matrices

**References**

1. Bekzod Khakimov, Mohammed Saddik Motawia, Søren Bak, Søren Balling Engelsen. Submitted.
2. Khakimov B, Amigo JM, Bak S, Engelsen SB. 2012. *Journal of chromatography.A* 1266, 84-94.
3. Bro R, Andersson CA, Kiers HAL. 1999. *Journal of Chemometrics* 13, 295-309.



BEKZOD KHAKIMOV BAHROMOVICH  
**Metabolomics and bioactive substances in plants**



Metabolomic analysis of plants broadens understanding of how plants may benefit humans, animals and the environment, provide sustainable food and energy, and improve current agricultural, pharmacological and medicinal practices in order to bring about healthier and longer life. The quality and amount of the extractible biological information is largely determined by data acquisition, data processing and analysis methodologies of the plant metabolomics studies. This PhD study focused mainly on the development and implementation of new metabolomics methodologies for improved data acquisition and data processing. The study mainly concerned the three most commonly applied analytical techniques in plant metabolomics, GC-MS, LC-MS and NMR. In addition, advanced chemometrics methods e.g. PARAFAC2 and ASCA have been extensively used for development of complex metabolomics data processing and analysis methods. The first study (*Journal of Chromatography A*, 1266 (2012) 84–94) demonstrated how the application of a multi-way decomposition method, PARAFAC2, can help in providing maximum extraction of metabolite features from the raw LC-MS data obtained from complex plant extracts. The second study (*Analytical and Bioanalytical Chemistry*, In Press, DOI: 10.1007/s00216-013-7341-z) outlines a novel GC-MS derivatization method using TMSCN for trimethylsilylation for improved analysis of complex biological mixtures. A review paper (*Journal of Cereal Science*, Accepted, DOI: 10.1016/j.jcs.2013.10.002) written for the *Journal of Cereal Science* comprises current analytical challenges and perspectives of cereal metabolomics with emphasis on new development in the use of multivariate data analysis methods for exploitation of the full information level in the analytical platforms. The fourth study (*Journal of Experimental Botany*, Submitted) combined the knowledge gained from the first and second studies and applied cutting-edge chemometric methods in a real case biological question related to barley breeding. This study revealed several biological questions associated with plant- environment, plant-gene mutation relationships and alterations of the plants' physiology during their development stages.

*The ultimate goal of metabolomics method development studies can be reached when the detectible part of the metabolome will be equal or close to the actual metabolome of the investigated sample matrix.*