



universität
wien

DISSERTATION

Titel der Dissertation

Novel tools for structural elucidation, pathway reconstruction and functional interpretation of high-throughput HRMS metabolomics data

Verfasser

Mag. Hannes Dörfler

angestrebter akademischer Grad

Doctor of Philosophy (PhD)

Wien, 2014

Studienkennzahl lt. Studienblatt:

A 794 685 437

Dissertationsgebiet lt. Studienblatt:

Biologie

Betreuer:

Univ.-Prof. Dr. Wolfram Weckwerth

This PhD thesis is based on the following first-author publications, which are presented in their original format:

Doerfler, H., et al., *Granger causality in integrated GC-MS and LC-MS metabolomics data reveals the interface of primary and secondary metabolism*. Metabolomics : Official journal of the Metabolomic Society, 2013. **9**(3): p. 564-574.

Doerfler, H., et al., *mzGroupAnalyzer-Predicting Pathways and Novel Chemical Structures from Untargeted High-Throughput Metabolomics Data*. PloS one, 2014. **9**(5): p. e96188.

In context to these publications, a co-author manuscript currently in revision is presented as full-text:

Targeted MAPA for quantification of proteotypic peptides in complex shotgun proteomics samples – a case study on mature pollen of Tomato under heat stress, Palak Chaturvedi, **Hannes Doerfler**, M^a Angeles Castillejo, Stefanie Wienkoop, Volker Egelhofer and Wolfram Weckwerth

Acknowledgements

I would like to express my gratitude to **Univ.-Prof. Dr. Wolfram Weckwerth** for giving me the opportunity to work in the exciting scientific field of systems biology, for granting me access to top-tier mass spectrometers, for a lot of insightful scientific discussions, and for supervising my thesis.

Dipl.-Biol. Dr. Stefanie Wienkoop for fruitful discussions and support especially in regards to technical challenges with Orbitrap-MS.

Mag. David Lyon for introducing me to the LC-MS setup, data handling, and many fruitful discussions and scientific cooperations.

Mag. Lena Fragner for the introduction to the GC-QqQ-MS and GC-TOF-MS platform, her continued support and patience and her effort to keep the MoSys lab a safe and clean environment for everyone working there.

Dr. Xiaoliang Sun for our very fruitful collaboration, his patience, and his unmatched ability to identify and remedy user problems for our bioinformatic tools.

Dipl.-Biol. Dr. Thomas Nägele for interesting discussions about metabolomics, metabolic modelling, and fruitful scientific cooperations.

Univ.-Prof. Dr. Franz Hadacek and **Dr. Vladimir Chobot** for the introduction into the area of redox chemistry of natural compounds and for supervising me in the electrochemistry lab.

Thomas Joch and **Andreas Schröfl** for excellent *Arabidopsis* cultivation and providing the lab with exquisite plant material.

No consistent moral decision can be made without understanding the consequences of our actions...and science is the way we determine the consequences of our actions. It is the prerequisite to a consistent moral philosophy.

Lawrence M. Krauss

Table of contents

I. Introduction	9
1. High-resolution metabolomics	9
1.1 Why mass spectrometry?	9
1.2 GC-MS-based metabolomics	12
1.3 LC-MS-based metabolomics	16
1.4 Data integration in metabolomics	19
2. Metabolomics in life sciences	20
2.1 Metabolomics and systems biology	20
2.2 Metabolomics in plant research	23
2.3 Plant stress and ROS metabolism	26
3. Structure elucidation in MS-based metabolomics	28
3.1 Mass spectrometric fragment information	28
3.2 High-resolution MS and sum formula annotation of novel metabolites	32
3.3 Structure elucidation and metabolic pathway reconstruction	36
3.4 m/z feature extraction of LC-MS data	39
4. References	41
II. Publications	50
III. Concluding remarks	121
IV. Abstract	123
V. Zusammenfassung	125
VI. Curriculum vitae	127

I. Introduction^{*}

1. High-resolution metabolomics

1.1 Why mass spectrometry?

Biological processes in tissues and fluids are tied to the presence and activity of a multitude of small intermediates called metabolites. Metabolites carry out a vast number of vital physiological functions in the organism like signalling, protection, structure synthesis and much more. Approaches to reliably detect those molecules in the frame of a biological question, such as a clinical investigation or fundamental research experiments, must include their unambiguous identification and ideally absolute quantification. With ongoing improvements in analytical methods regarding selectivity, sensitivity and sample throughput, today it is possible to simultaneously detect a high number of metabolites in a biological sample in a single analytical step.

Owing to these advancements, *metabolomics* – defined as the comprehensive and quantitative analysis of all metabolites under a given set of conditions – could establish itself as a major bioanalytical technique in bioanalytics in recent years (Goodacre, 2005; Goodacre et al., 2004). Consequently, the metabolome is defined as the complete set of metabolites present in an organism at a distinct point of time. However, at the present day it is methodologically not possible to analytically capture the entity of the metabolome like its definition would demand (Dettmer et al., 2007). Therefore it is practice to preliminary characterize potential metabolites rather than to perform a complete structural elucidation including stereochemistry (Dunn and Ellis, 2005; Goodacre et al., 2004), or to narrow down the analytical scope towards certain classes of compounds in a profiling or target approach (Dunn and Ellis, 2005; Goodacre et al., 2004). Metabolomic techniques are already successfully incorporated in many areas of the “life sciences”, which contain medical research fields like clinical diagnostics (Collino et al., 2013; Trifonova et al., 2013), toxicological risk assessment (Munoz and Albores, 2010; Wilson et al., 2013), biomarker and drug discovery (Anderson and Kodukula, 2013; Robertson and Frevert, 2013; van Ravenzwaay et al., 2007), as well as in applications for the food and feed sector (Boggess et al., 2013), crop biotechnology (Hall et al., 2008), biofuel (Zhu et al., 2013) and plant research in general (Guy

^{*} Own work and citations of own work are marked in bold.

et al., 2008; Hall, 2006; Schauer and Fernie, 2006; Weckwerth, 2011a). The inherent difficulty with metabolomic analyses though arises from the fact that the metabolome contains biochemical compounds of highly differing physico-chemical properties and analytes deriving from the same biological background and thus display enormous heterogeneity regarding carbon structure, size, polarity, charge, heteroatom content and abundance (Goodacre, 2005). For metabolomic analyses only analytical methods can be considered which comply with the task of detecting such diverse biomolecules, and this within a reasonable timeframe. The following table depicts a compilation of instrumental techniques suited for the analysis of small molecules and their respective chemical preference towards analyte properties (Table 1).

Analyte property	Method
Volatility	Gas chromatography (GC)
Polarity	Liquid chromatography (LC)
Charge	Capillary electrophoresis (CE)
Size	Size exclusion chromatography (SEC)
Mass	Mass spectrometry (MS)
Rotation, Vibration	Infrared spectroscopy (IR)
Spin	Nuclear magnetic resonance spectroscopy (NMR)
h ν absorbance	UV/VIS spectroscopy
h ν emission	Fluorescence spectroscopy
Chirality	Stereoselective chromatography
Redox potential	Cyclic voltammetry

Table 1: Selected physico-chemical properties of small molecules and their corresponding analytical method of choice (Fiehn 2007, modified)

Several analytical techniques can be used for experiments on the “metabolomics”-level today, however for the best unbiased and parallel detection of numerous low-abundant compounds in a biological sample, high-resolution mass spectrometric detectors are currently the most powerful analytical devices today. Mass spectrometers can provide both qualitative and quantitative information and can be applied to a wide range of analytical questions. Featuring high dynamic range, detection limits lie in the picomol area; depending on the detector and sample pre-treatment, they can be lowered down to the femtomol or attomol level (Dettmer et al., 2007). The general setup of a mass spectrometer consists of an ion source, where charged species are produced, which are subsequently separated according to their mass-to-charge (m/z) ratio in the mass filter and eventually detected at the ion detector (Figure 1). The

relative abundance of the detected m/z values are then assembled in a mass spectrum. Ionization and transition into the gas phase of the analyte is pivotal for the detection in the mass spectrometer and varies depending on the instrument type, as will be discussed in the upcoming chapters. Figure 1 depicts the basic layout of a mass spectrometer.

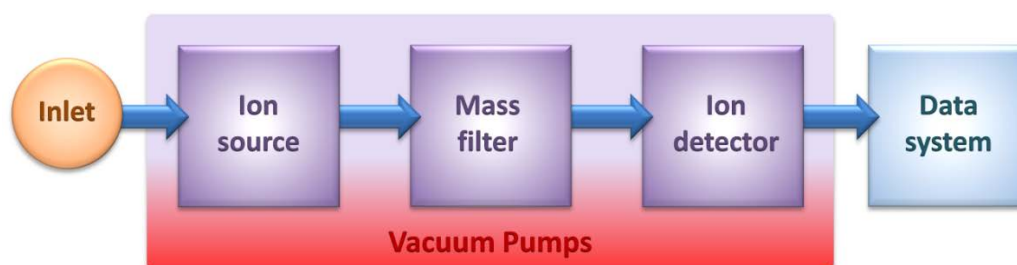


Figure 1: Schematic arrangement of a mass spectrometer

The performance of a mass spectrometer is largely determined by its accuracy and its resolution. The accuracy is to a large extent a function of the resolution itself and describes the ability to accurately provide mass-to-charge (m/z) information. For instance, an instrument with a 0.005% accuracy can provide information about a 1000 Da molecule ± 0.05 Da, or ± 50 ppm. For consistent mass accuracy, proper calibration is needed. Mass spectrometric resolution accordingly is the ability to distinguish between ions of different m/z ratio. Two predominant definitions exist for mass spectrometric resolving power, depending on the instrument used: Two peaks of similar height can be referred to as resolved when the valley between their apexes Δm has fallen to 10% of their intensity (10% valley definition; equivalent to peak width at 5% peak height); for single peaks, Δm may be defined as the full width of the peak at its half maximum (FWHM definition). FWHM resolution is usually double the 10% valley resolution. High resolution MS allows for the detection of the exact mass of a compound, as well as characterization of the isotopic distribution in high-molecular species like proteins (Figure 2).

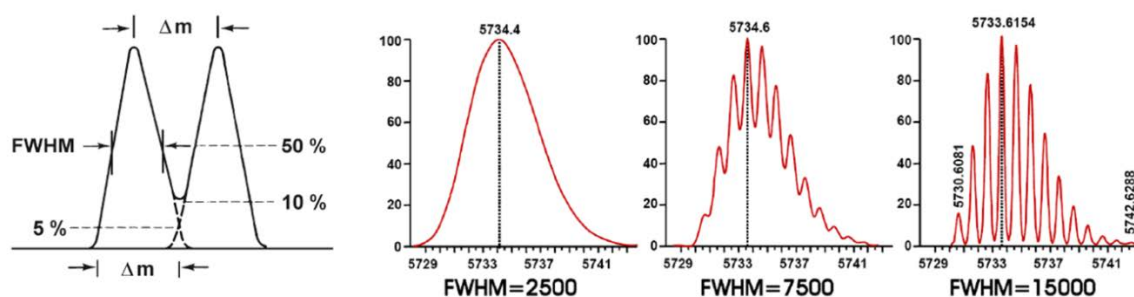


Figure 2. Left: Different definitions of mass spectrometric resolution. Right: Isotopic pattern of bovine insulin at selected FWHM resolution values. Only an FWHM resolution value of 15 000 and above allows for base-line separation of the isotopic peaks (NIH, Washington University in St. Louis).

The importance of high resolution MS in metabolomic experiments becomes evident considering the possibility to constitute a sum formula for an unknown compound due to its accurately measured m/z ratio, as will be discussed in detail in chapter 3.

Especially in view of complex biological samples, it is a popular and well-established approach to couple mass spectrometers to chromatographic columns. This so-called hyphenated arrangement entails several advantages: not only serves the separation step as additional dimension for the identification of analytes; also, detection interferences by co-eluting analytes or matrix compounds are reduced, potentially enabling the detection of substances not distinguishable by MS alone. The main methods in metabolomics approaches are gas chromatography coupled to mass spectrometry (*GC-MS*) and liquid chromatography coupled to mass spectrometry (*LC-MS*), which are described in the upcoming sections.

1.2 GC-MS-based metabolomics

GC-MS is a robust and convenient technique for the analysis of small and/or volatile compounds. The separation principle lies in the different interaction of the analyte molecules between a stationary viscose phase or (to a much lesser extent employed) solid phase and a mobile gas phase. The principles of gas-chromatographic separation were formulated in the mid of the 20th century and significant proceedings on column chemistry were already reported several years later (Rohrschneider, 1966; Rohrschneider, 1969). The idea of using an MS detector coupled to GC for metabolite analysis dates back approximately 40 years (Fiehn, 2008; Thompson et al., 1975) and has since been used for clinical diagnostics (Fiehn, 2008; Niwa, 1986) and plant research (Fiehn, 2008; Furuya, 1965; Kellogg et al., 1964). Key advantages of gas chromatographic analyses are their reproducibility, and also their comparability between similar setups of different labs; combined with its excellent performance regarding separation efficiency, selectivity and sensitivity, this method has been attributed the status of a “gold standard” in metabolomic experiments (Antoniewicz, 2013; Fiehn, 2008; Koek et al., 2011; Krone et al., 2010).

A very helpful concept to classify compounds in a gas chromatographic analysis is the retention index (RI) proposed by Kovats in 1958 (Kovats, 1958): The relation of carbon content to retention time in a sequence of homologous n-alkanes can be used to characterize unknown compounds in a sample:

$$I = 100 * (y - x) \frac{\log\left(\frac{t_{sample}}{t_x}\right)}{\log\left(\frac{t_y}{t_x}\right)} + 100$$

where x = C number of species eluting before the sample

y = C number of species eluting after sample

t = respective retention times

I = retention index

Every n-alkane is assigned the number of its carbon atoms times 100. Figure 3a shows that the log RT vs. carbon count relation is linear. However, Kovats-indices may only be used for isothermal gas chromatography. For analyses with temperature programs, an equation was formulated by Van den Dool and Kratz (Vandendool and Kratz, 1963). Figure 3b depicts a GC-derived chromatogram with a superimposed run of a series of n-alkanes.

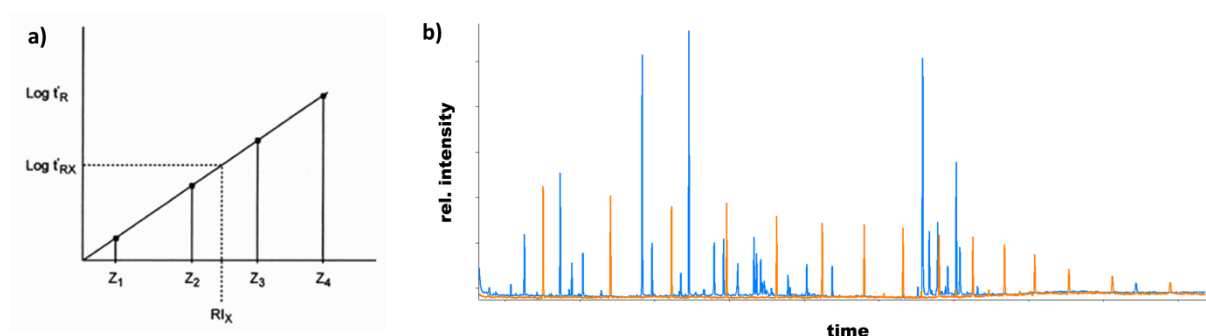


Figure 3. a) Plotting log retention time (t'_R) versus a homologous alkane series (z) yields a linear slope under isothermal GC conditions (IUPAC). b) A series of linear alkanes in a GC chromatogram (orange) serves as reference point for unknown signals in a sample (blue) (Doerfler, 2014).

Even though GC-MS is highly applicable for experiments in the „metabolomics“-style, this method is inherently restricted to volatile analytes and compounds which may not exceed a molecular weight of approximately 800 Da. For polar molecules, the possibility of derivatization exists, rendering them accessible for analysis in the gas phase (Halket et al., 2005). Several derivatization methods have been developed for GC-MS measurements, while the most prominent agent for polar metabolites is N-Methyl-N-(trimethylsilyl)trifluoroacetamide (MSTFA). Due to its high reactivity, most analytes with the functional groups -OH, -COOH, =NH, -NH₂ and -SH can have their hydrogen atoms replaced by a trimethylsilyl group. Trimethylsilylation can lower the boiling point of polar substances or make them vaporizable, thus amenable to GC in the first place, as is often the case with

amino acids or sugars (Halket et al., 2005) (Figure 4). At any rate, it has to be considered that every derivatization step before the actual measurement demands additional validation.

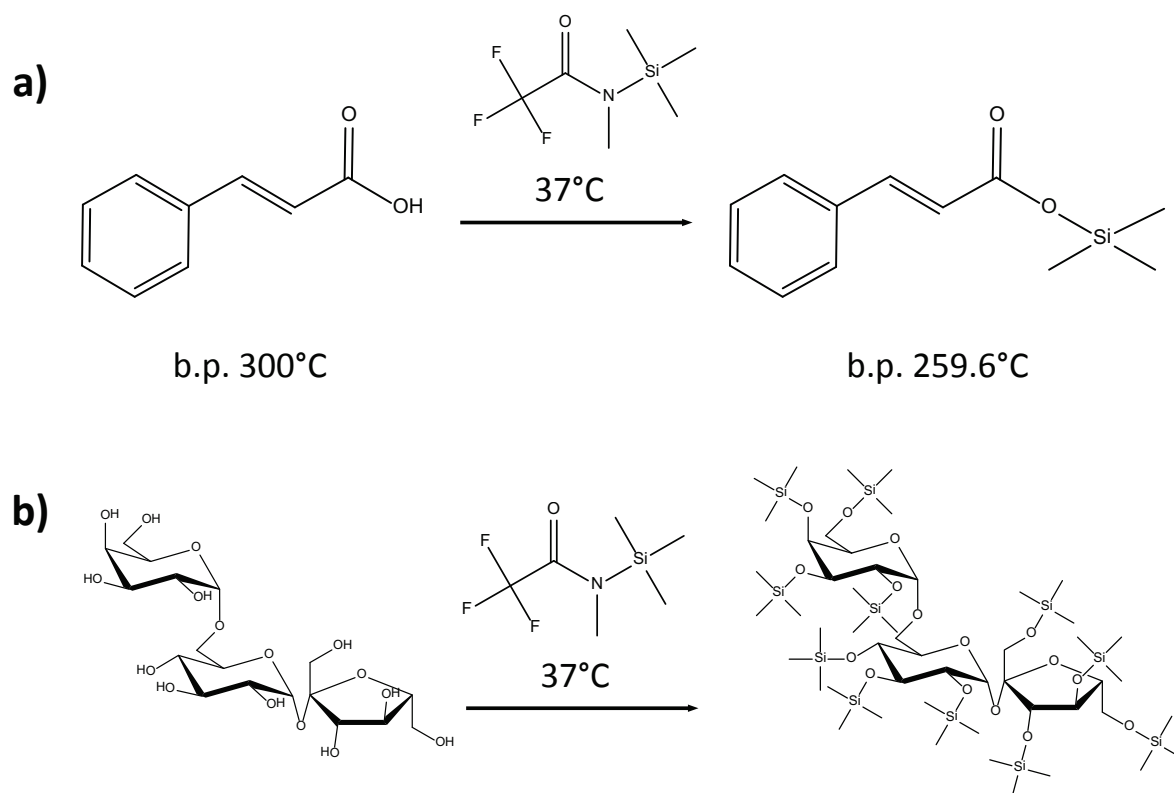


Figure 4. Derivatization in GC with MSTFA. a) Exchange of the hydrogen atom in the carboxyl group by a trimethylsilyl group in cinnamic acid lowers the boiling point by about 40°C (ChemSpider). b) Silylation of the 11 hydroxy-groups in raffinose allows for vaporization and analysis by GC (Doerfler, 2014).

Ionization in a GC-MS setup happens through electron impact (EI): Compounds eluting from the column are conveyed into the evacuated ion source where they are met with a stream of accelerated electrons. This impact happens at 70 eV and causes both ionization in form of radical cation formation, as well as fragmentation of the molecule. Even though the precursor is often not visible, EI spectra are highly specific and reproducible and thus well suited for database storage (Brown et al., 2009; Kind et al., 2009). As both retention time and fragmentation spectra are characteristic for a compound, identification in GC-MS can be achieved by searching unknowns against pre-existing databases (Kind et al., 2009; Kopka et al., 2005; Steinhauser et al., 2004). Figure 5 shows an EI spectrum of a typical metabolite detected in a GC-MS run and its identification by comparison with a library entry.

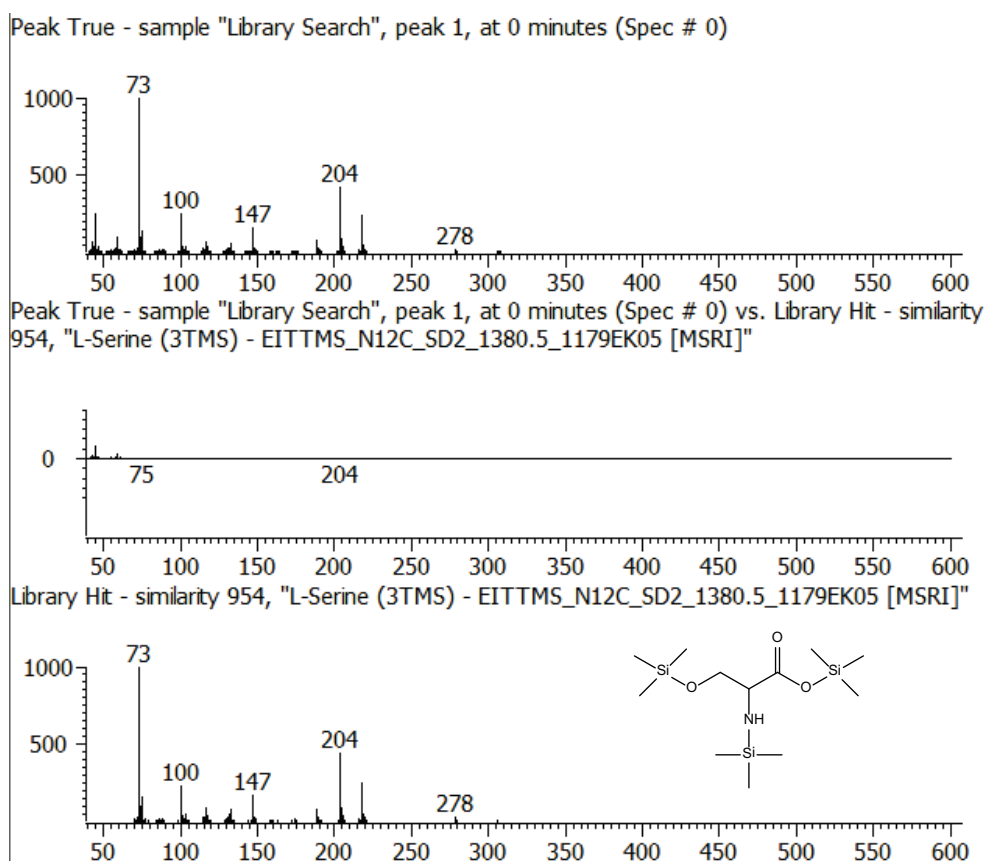


Figure 5. Identification of serine (3TMS) in a plant extract through library search after GC-MS analysis. Top: spectrum of detected peak. Middle: Subtracted spectrum. Bottom: library reference spectrum. The two spectra show a high agreement, resulting in a highly probable database hit (**Doerfler, 2014**).

Mass spectrometric detectors for GC-MS are mainly *quadrupole*, *triple quadrupole* (*QqQ*) or *time-of-flight* (*TOF*) analyzers. Triple quadrupole mass spectrometers have the ability to perform special ion separation experiments like selected reaction monitoring for highly sensitive quantification. TOF mass spectrometers profit from a faster scanning rate and higher resolution, while not having MS/MS capability. The installation of a second, polar column after the main column in GC, and with it the introduction of a second retention index, can help reduce the number of co-eluting compounds in metabolomics experiments and thus increase the number of detected metabolites in a so-called GCxGC-MS setup (Kempa et al., 2009; Weckwerth, 2011b).

At the moment, only a fraction of all the small metabolites present in a biological sample can be reliably identified and quantified, while their actual number lies in the range of hundred thousands to millions (Weckwerth, 2011a). For future growth and effectiveness of GC-MS based metabolomics, more database building, method development and exchange between laboratories is essential (Hummel et al., 2010; Koek et al., 2011; Kopka, 2006).

1.3 LC-MS-based metabolomics

Compounds exceeding the molecular weight threshold for GC can be analyzed by a liquid – chromatography–mass spectrometry (LC-MS, or HPLC for *high pressure liquid chromatography* or *high performance liquid chromatography*) setup. Chromatographic separation of analyte molecules here happens between an immobile phase of various chemical properties and a mobile liquid phase. Conventionally, stationary phases in liquid chromatography are either immobilized liquids or chemically bound phases. Immobilized liquids only have historical significance today due to several disadvantages, one of them being the incapability to perform gradient elution on such a system. Solid phases consisting of bound groups are the prevalent method for chromatographic setups and two major strategies can be distinguished here: *Normal-phase chromatography* (NPLC) and *reversed-phase chromatography* (RPLC). NPLC stationary phases consist of a silica gel or aluminium oxide structure and are of polar nature, while RPLC phases are so-called end-capped silica gels, meaning they are chemically modified to carry apolar structures like C8- or C18-alkane groups (Figure 6).

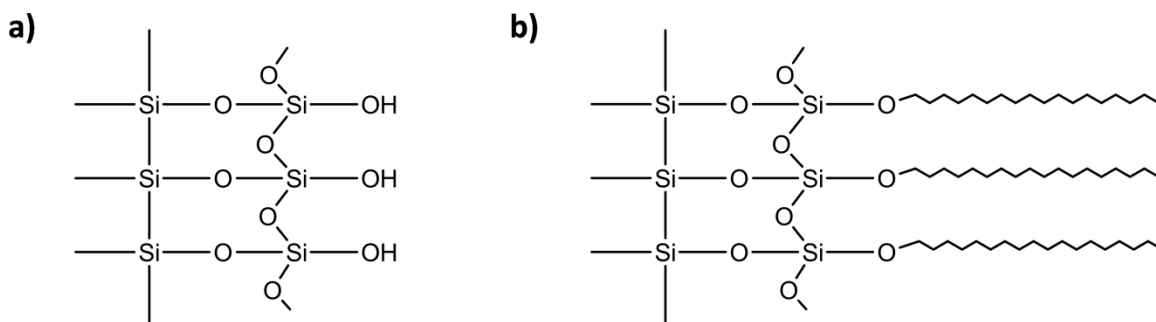


Figure 6. a) Core structure of a stationary NPLC silica gel phase. b) C18 end-capped siloxane phase.

Reversed-phase chromatographic techniques make up the largest part of liquid chromatographic separations today due to their more universe applicability for metabolites of biological origin, especially larger, hydrophobic compounds. Column chemistry for HPLC in general is a constantly evolving field. Latest developments of interest for bioanalytics include monolithic columns, which support low pressure and good mass transfer efficiency while operated under high flow rates (Rigobello-Masini et al., 2013; Tolstikov et al., 2007; Tolstikov et al., 2003), as well as HILIC (*hydrophilic interaction chromatography*) columns, which facilitate separation of small polar molecules under RPLC eluent conditions via

adsorption on polar stationary phases (Lammerhofer, 2010). In the last years, HPLC has been advanced to UPLC (*ultra high pressure liquid chromatography* or *ultra high performance liquid chromatography*), featuring improved separation efficiency and peak capacity due to the use of particles smaller than 1.7 μm .

Ionization of analytes in LC-MS has long been a technical challenge. In GC-MS, ion generation in vacuum is practicable in the gaseous state, while analytes eluting from an LC column have to be rid of their solvent before ionization and transfer into the mass analyzer can happen. Current ionization methods for LC-MS include *electrospray ionization* (ESI), *atmospheric pressure chemical ionization* (APCI) and *atmospheric pressure photo ionization* (APPI). For metabolomic and proteomic studies, ESI is the most widely used technique as it allows for ionization of low molecular weight compounds such as secondary metabolites and peptides. Below, the principle of the electrospray ionization process is depicted (Figure 7).

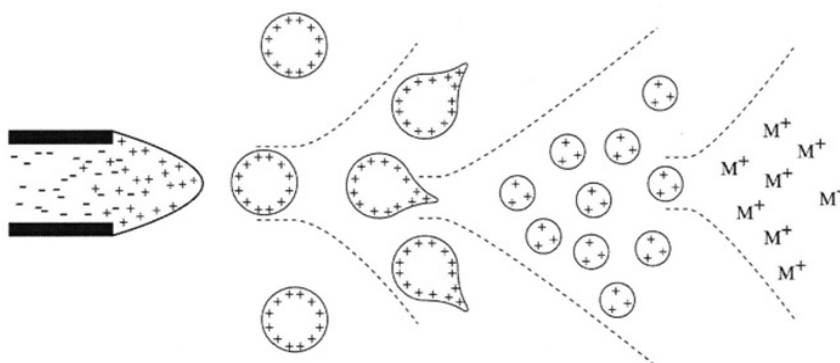


Figure 7. Principle of the electrospray ionization process (positive mode) (Wilm, 2011). An electric field between the needle and the ion transfer capillary causes positively charged molecules to accumulate on the needle point, forming a so-called Taylor cone. From its tip, a spray of droplets is emitted, which will subsequently form Taylor cones of their own and give rise to several new smaller droplets due to Coulomb repulsion. This process continues until, due to solvent evaporation, only charged particles remain, which are then transferred into the mass spectrometer. In practice, evaporable eluents like methanol, acetonitrile and water are chosen as eluents while formic acid is used as proton donor (openstax).

While the ideal electrospray process would yield singly or multiply protonated analytes molecules $[M+H]^+$ and $[M+nH]^{n+}$, a characteristic trait of ESI is the formation of additional molecule adducts during the flight process, which can complicate the spectrum. Hence, in full-scan ESI spectra one can often observe sodium and potassium adducts $[M+Na]^+$ and $[M+K]^+$ instead of protonated species, as well as formations with water, acetic acid, ammonia, and alike. While for peptides multiple charging is observed regularly, there have been no reports of metabolites carrying more than a single charge in electrospray ionization as of now (Rodriguez-Aller et al., 2013). Electrospray ionization is a soft ionization technique,

meaning that the precursor ion is conveyed into the mass analyzer without undergoing fragmentation and thus an unscathed MS¹ precursor peak is visible in the spectrum, accessible for further fragmentation experiments. In 2002, the inventors of electrospray ionization and matrix-assisted laser desorption/ionization (MALDI), John B. Fenn (ESI) and Koichi Tanaka (MALDI), together with Kurt Wüthrich (“Bio-NMR”), received the Nobel prize in chemistry “for their development of soft desorption ionisation methods for mass spectrometric analyses of biological macromolecules” (Fenn, 2002; Fenn, 2003; Fenn et al., 1989).

Suitable mass spectrometers for LC-MS-based metabolomics are often instruments capable of obtaining the exact mass of a precursor compound, as is the case with *FTICR* (*Fourier transform ion cyclotron resonance*) and *Orbitrap* mass analyzers, the latter also having the ability for ion trapping. These types of mass spectrometers can perform MS² as well as MSⁿ experiments on ions of selected m/z ratios, which, in combination with the acquisition of accurate precursor masses, allows for structural elucidation of unknowns in metabolomic experiments (see chapter 3 for an in-depth discussion). Figure 8 shows a typical UPLC-Orbitrap chromatogram of a crude plant extract.

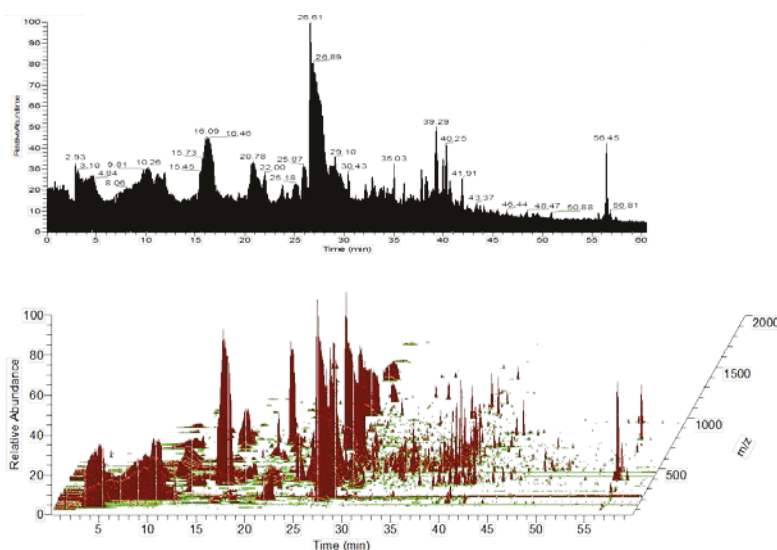


Figure 8. A methanolic extract of *Arabidopsis* leaves analyzed by UPLC-Orbitrap-MS.

Above: 2-dimensional total ion chromatogram (TIC) with retention time on the x- and relative intensity on the y-axis. Below: an additional z-axis visualizes m/z ratio of measured compounds at corresponding times, including MSⁿ scan events (Doerfler, 2014).

LC-ESI-MS setups can generally be optimized on various levels towards the analytes of interest. Parameters for method development on the chromatographic side include choice of column, mobile phase composition, flow rate and gradient, while on the MS side spray polarity, resolution, fragmentation type, collision energy, MS²/MSⁿ triggering thresholds and several other instrument specific factors must be taken into consideration.

1.4 Data integration in metabolomics

As has been shown, instrumental analysis techniques have their individual set of preferences towards certain analyte characteristics, while at the same time will be not be able to detect some compound classes at all. Reports of GC-MS-, as well as LC-MS-based investigations of metabolic processes are widespread in the literature (Barrett, 2012; Beecher, 2011; Milne et al., 2013; Monteiro et al., 2013; Nicholls, 2012; Sarker and Nahar, 2012; Want and Masson, 2011), however, by their nature, they can only highlight a fraction of the metabolome of an organism. Figure 9 shows established instrumental methods for metabolomics within their field of application concerning compound classes.

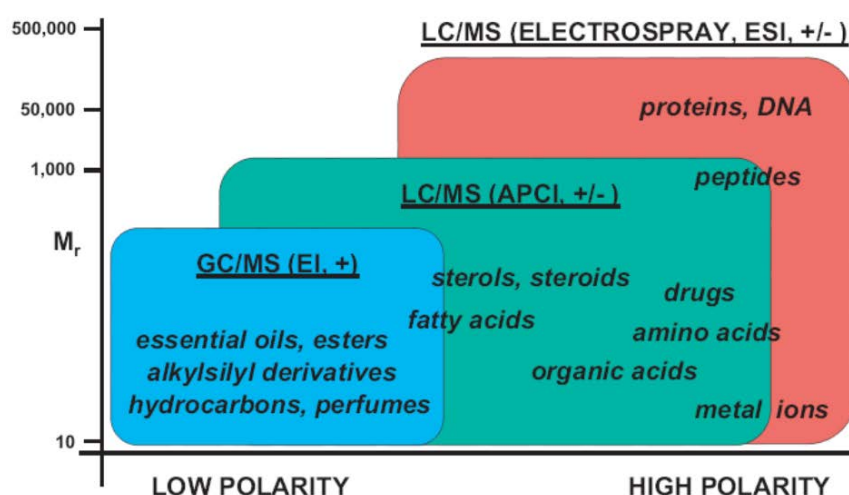


Figure 9. Current analytical platforms for metabolomics-style experiments. Chemical derivatization can help shift the boundaries between instrument types for facilitated analysis. Image modified according to (Halket et al., 2005).

Integration of differential analytical platforms for increased metabolite coverage is thus a core challenge of metabolomic approaches. One strategy to address this issue was shown by the study of Scherling et al. within a large-scale grassland biodiversity study on various herbs employing analysis both at the GC-MS and LC-MS platform, effectively broadening the range of detected masses from 40 to 2000 m/z (Scherling et al., 2010). The present work demonstrates the effectiveness of such an integrated approach by a truly combined GC-MS and high-resolution LC-Orbitrap-MS analysis of time-dependent metabolic profiles of *Arabidopsis thaliana* under cold stress by extraction of the same sample with subsequent data merging and functional interpretation by application of multivariate statistics and modelling approaches (II.1) (Doerfler et al., 2013).

2. Metabolomics in life sciences

2.1 Metabolomics and systems biology

As has been described in chapter 1, modern bioanalytical techniques like GC-MS and HPLC-MS facilitate simultaneous detection of numerous metabolites in a biological sample, be it knowns or unknowns. A metabolomic analysis of an organism at a distinct point of time therefore represents a metabolic snapshot of the system under the current conditions. Because a plethora of metabolites is captured in such an analysis, it is possible to infer biological information like alterations in the underlying metabolic network by comparing the metabolome of organisms under different conditions, an advantage which is unique to the metabolomic approach (Weckwerth, 2003). Metabolomic datasets are featuring hundreds to thousands of values, which necessitates the use of multivariate statistical methods (Eliasson et al., 2011; Kemsley et al., 2007; Okada et al., 2010) (Figure 10).

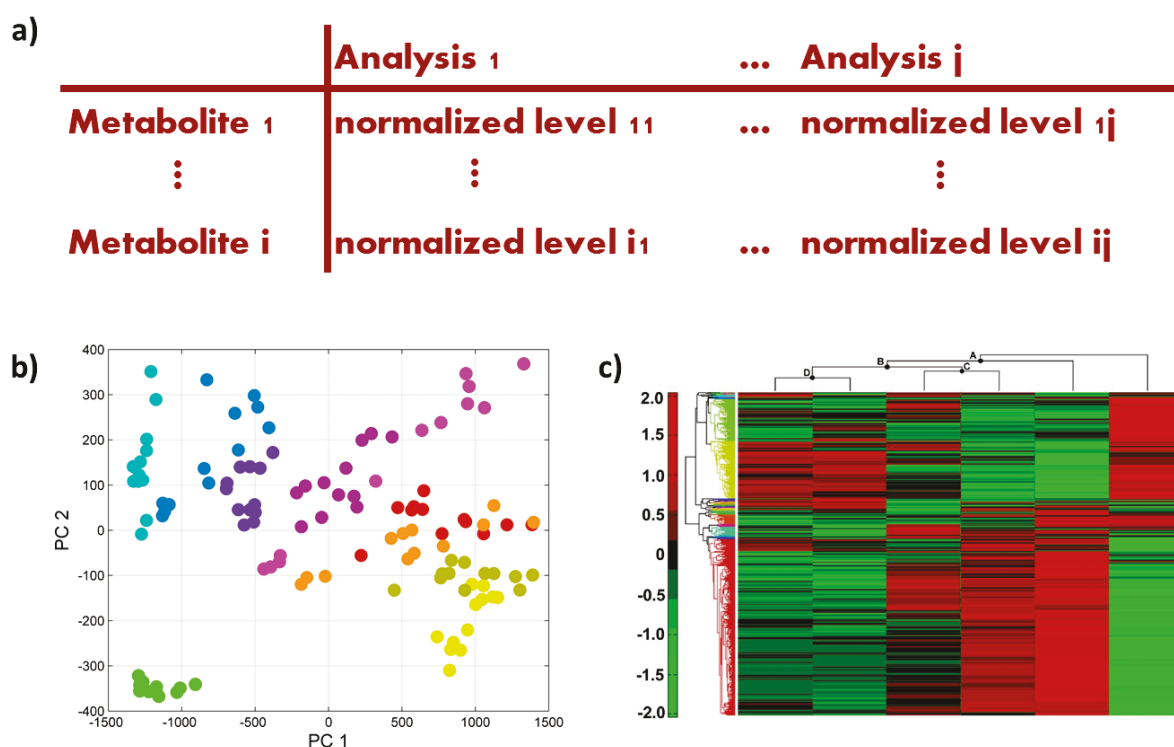


Figure 10. Tools for visualization of multivariate analysis methods. a) Metabolite covariance matrix for multivariate analysis. b) Principal component analysis allows for exploring biological information by clustering of related compounds. c) Data from a function can be visualized through colours in a heat map.

Combined with the other major “omic” sciences – genomics, transcriptomics and proteomics – metabolomics is thus outstandingly well suited for the investigation of biological organisms on a systems scale (Goodacre, 2005; Weckwerth, 2003). Owing to latest breakthroughs in molecular biology, especially in view of the human genome project (Schuler et al., 1996; Venter et al., 2001; Venter et al., 1996), it has become apparent that the classic approach of looking at single genes and proteins must be extended to a stage where the interaction between the different levels of biological organization is taken into consideration (Ideker et al., 2001). These recent insights also helped strengthen the view of biology as informational science (Hood et al., 2004). Figure 11 depicts the hierarchy of biological information within an „omics“ background.

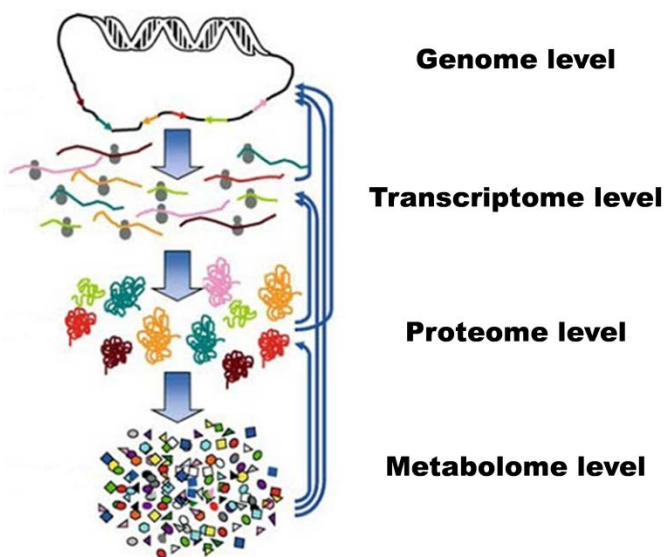


Figure 11. The pathway of biological information. The genome, transcriptome, proteome and metabolome are interconnected and mutually influencing each other. The metabolome is the endpoint of this relation and represents the physiological readout of the organism with respect to current environmental conditions. Image modified according to (Goodacre, 2005).

With ongoing progress in bioanalytical tools, particularly regarding genome sequencing and transcript and protein analysis (Kell, 2004; Kitano, 2002), efforts to describe biological processes on a functional level by integration of these different layers of information have become feasible (May et al., 2008). The metabolome holds a special place in this context as it represents the end product of the biochemical machinery and thus encrypts information from all the preceding biochemical steps in a cell. Changes or perturbations of any of those levels will result in an altered metabolic profile, which can be highlighted by metabolomic analyses. The properties of the metabolome can thus be traced back to the gene level, enabling the elucidation of gene functions in an approach called functional genomics (Fiehn, 2002; Rhee, 2013). Within the background of environmental effects, functional genomics approaches are now used to systematically study the genotype-phenotype relationship of organisms

(Weckwerth, 2008). In this context, it is important to note that only a true, untargeted metabolomic approach can yield useful results, as due to genetic crosstalk and pleiotropic effects, unpredictable results must be considered (Fiehn, 2002). Since methods to perturb the biological system on a genetic level have become more robust, it is possible to explore whole gene-metabolite networks (Ideker et al., 2001), which subsequently solidifies the role of metabolomics for both investigating the modes of action of transcripts and proteins as well as a viable technique for the identification of biomarkers (Krumsiek et al., 2012). Because metabolites are the result of enzymatic activity, while enzymes themselves are encoded in the genome, annotation of genes and enzymes is a pivotal process for a fundamental understanding of the interrelation of biological processes (May et al., 2011; May et al., 2008).

To fit biological data from a metabolomics experiment into genome-scale metabolite networks, modeling approaches are necessary (May et al., 2011). In order to gather information for modeling efforts, extensive untargeted data mining of several replicates under different conditions is needed beforehand (Wienkoop et al., 2008). In combination with multivariate statistics and pattern recognition like independent component analysis (ICA), this method has been shown to be an effective approach to unveil biochemical relationships (Wienkoop et al., 2008). Assembling measured metabolite levels and their changes over a time profile or under different conditions in a covariance matrix allows for application of statistical methods which have the potential to unveil metabolic perturbations (Weckwerth, 2011b). Modelling methods in systems biology include kinetic modelling and flux balance analysis (FBA) (Nägele and Weckwerth, 2012; Sweetlove and Ratcliffe, 2011; Weckwerth, 2011b) as well as application of Granger cause-effect theory (**Doerfler et al., 2013**; Fujita et al., 2010). When treating individual metabolite levels as steady states, inverse approaches may unfold key points of regulatory instances and facilitate reconstruction of metabolic networks (**Doerfler et al., 2013**; Nägele and Weckwerth, 2012; Weckwerth, 2011b).

Metabolic modelling has already proven to be a cornerstone of reconstructing biochemical pathways and improved knowledge in this area will eventually enable researchers to move toward prediction of biological systems. However, several issues exist to date which in future have to be solved for a better integration and understanding of metabolomic data, among them improved identification rates of metabolites and optimized extraction protocols in order to minimize the overlap of the same metabolites deriving from different cell compartments (Weckwerth, 2003).

2.2 Metabolomics in plant research

Due to the manifold metabolic pathways found in higher plants and their ability to synthesize natural products of human interest, the plant community has played a major role in pushing metabolomic techniques forward (Saito and Matsuda, 2010). Plants are not only the basics for food and feed production on a global scale and are sources for biofuel (Banerjee, 2011); they also represent a rich source of interesting substances for medicine, health care and nutrition (Ngo et al., 2013; Nguyen et al., 2012). Especially in view of climate change and a steadily growing world population, plant biotechnology is shaping up to be a cornerstone for worldwide human nutrition more than ever before (Weckwerth, 2011a). Plant metabolomics and related disciplines will thus play a major role in the industry in the future, while in fundamental research they will unveil new insights into biochemical processes, together with genomics and proteomics help determine the taxonomy of undescribed organisms, and answer questions about evolutionary origin (Govindaraghavan et al., 2012; Reynolds, 2007). Because metabolite dynamics respond to environmental fluctuations, metabolic analyses of plants provide an immediate picture of the biochemical state of the organism and can also be used, among others, for unveiling plant-environment interactions (Scherling et al., 2010).

Historically, the metabolism of plants is divided in the *primary metabolism* and the *secondary metabolism*. The primary metabolism is universal and in general comprises compounds of low molecular weight, which are involved in all central processes like growth and energy metabolism, e.g. amino acids, sugars, nucleic acids, organic acids and others. The basic catabolic and anabolic principles of the central plant metabolism, such as glycolysis, citric acid and Calvin cycle have been elucidated more than 50 years ago (Calvin and Benson, 1948; Krebs and Johnson, 1937). Nowadays, many of the compounds in the primary metabolism can be detected simultaneously within the GC-MS setup, which has developed as an established method in plant metabolomics (Cerdan-Calero et al., 2012; Lisec et al., 2006; Schauer et al., 2005; Shuman et al., 2011; Weckwerth, 2003). The identification and quantification of plant primary metabolites is an established method in the metabolomics field, as it allows both for functional gene studies, as well as explorative experiments, especially in the field of plant-environment interactions (Morgenthal et al., 2006; Weckwerth and Morgenthal, 2005). Apart from the central metabolic routes which grant the plant survival and reproduction, the so-called *secondary metabolism* exists. The plant secondary metabolism is a highly interconnected system of metabolic pathways, capable of synthesizing specialized metabolites on demand, for example in times of biotic or abiotic stress (Nascimento and Fett-

Neto, 2010). Although long observed as unnecessary by-product in the plant kingdom, secondary metabolites are mediators during environmental stress (Debnath et al., 2011), and newer findings indicate a vital role of them in the central signalling metabolism (Hadacek et al., 2011). Therefore, the term *specialized metabolites* was recently proposed to use for biomolecules that fall into this category (Nakabayashi and Saito, 2013). The boundary between primary and secondary metabolism is blurred, as there is often only very little structural difference in the molecules of both categories (Figure 12).

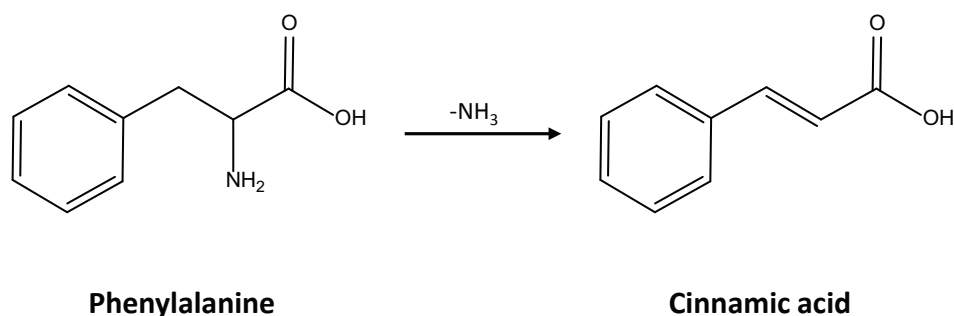


Figure 12. Biosynthesis of cinnamic acid from phenylalanine. The amino acid phenylalanine is a typical primary metabolite, while cinnamic acid is considered a secondary metabolite (specialized metabolite) (Croteau et al., 2000; Zhang et al., 2007).

In general, the secondary plant metabolism comprises compounds of highly diverse chemistry. These biomolecules may vary in their basic carbon structure, type and amount of functional groups, heteroatom content, bond chemistry, etc. Many such natural products have been used by humans since centuries, for example as dyes, waxes or cosmetics. Others were recognized to possess physiological activity in the human body and have found applications as drugs or toxins. Several typical secondary metabolites show cytostatic activity and are being used as antitumor therapeutics, like taxol, a diterpene alkaloid (Wani et al., 1971). Apart from uses for improving life quality, several plant compounds are essential for the human body to function and have to be part of a healthy nutrition, like vitamin B1 (thiamine).

Secondary metabolite synthesis is tied to specific nodes of the primary metabolism, and can be activated by the plant if required; for instance members of the anthocyanin family are synthesized during flowering from aromatic amino acids, which themselves derive from the well-investigated shikimic acid pathway (Maeda and Dudareva, 2012). The richness and diversity of plant secondary metabolites has long been subject of discussion, and their role is still not completely elucidated, however in latest years, they have been recognized as pivotal signalling and stress mediators, as will be discussed in the upcoming chapter. Figure 13 shows selected prominent compounds of major secondary metabolite classes and their origin in the primary metabolism.

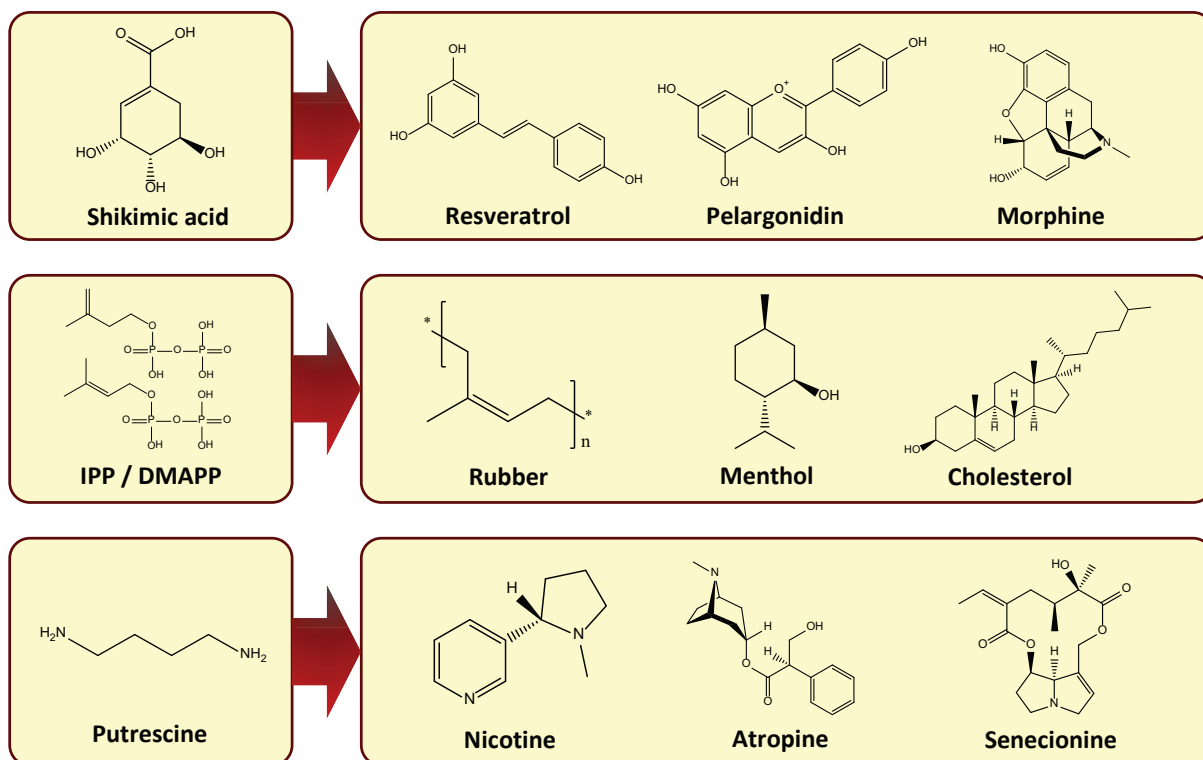


Figure 13. Selected secondary metabolites and their starting point in the primary metabolism (Doerfler, 2014). Shikimic acid is the building block for aromatic amino acids and represents the precursor for phenylpropanoids, stilbenes, flavonoids and isoquinoline alkaloids. Isopentenyl pyrophosphate (IPP) together with its isomer form dimethylallyl pyrophosphate (DMAPP) are the subunits for isoprene, terpenes and terpenoids, comprising larger carbon structures, for instance the steroid family. Putrescine (tetramethylenediamine) is an N-containing primary metabolite and involved in the biosynthesis of heterocyclic compounds like nicotine and atropine, but also in as precursor for pyrrolizidine alkaloids like senecionine (Kanehisa and Goto, 2000; Ogata et al., 1999).

Apart from immediate economic usage, the investigation of the plant secondary metabolism is crucial for answering several biological questions: First, because secondary metabolites are highly specific to their taxon, plant metabolic profiling can help determine the origin of plants based on differences in their growing locations or habitat (Lee et al., 2013; Zhang et al., 2014). Also, plant secondary metabolite studies are a valuable tool in functional genomics (Spiering et al., 2014), for example to trace back functions of genes when comparing knockout modifications to wild type plants, as has lately been shown by Tohge et al. (Tohge et al., 2005; Yonekura-Sakakibara et al., 2007). Analysis of compounds derived from the secondary metabolism can be achieved employing the liquid chromatography-mass spectrometry (LC-MS) setup, as most of those compounds exceed the molecular weight for GC-MS. Studies performing high resolution LC-MS experiments of plant secondary metabolites have already been shown to have the potential to reveal new insights into the complex molecular networks present in higher plants and identify novel metabolites in an untargeted way (Allwood and Goodacre, 2010; De Vos et al., 2007; Nakabayashi et al., 2009;

von Roepenack-Lahaye et al., 2004). Because the ongoing elucidation of the plant metabolome is directly tied to technical limitations concerning accuracy, robustness, reproducibility and database availability, major progressions in the area of plant secondary metabolomics may be expected for the near future by employing differential state-of-the-art analytical methods. The role of high-resolution mass measurement for plant metabolomics is discussed in chapter 3.

2.3 Plant stress and ROS metabolism

Plants are immobile organisms and as such cannot evade biotic or abiotic stressors. In order to maintain their metabolic homeostasis, their metabolism has to be adjusted to changing environmental conditions (Obata and Fernie, 2012). From an evolutionary perspective, this situation led to compartmentalization of vital processes into organelles and to the development of highly regulated metabolic pathways necessary for adaptation to environmental stress. This metabolic reprogramming ushers in up- and downregulation of certain intermediates apt to deal with specific stresses, or the synthesis of compounds or structures which are not produced under everyday conditions. Metabolomic techniques are very convenient tools to investigate the metabolic fluxes and dissect signalling pathways happening during environmental stress, because metabolites react more immediate to changing environmental conditions than transcriptional processes (Less and Galili, 2008).

The most common abiotic stresses for plants encountered in nature are drought, heat, cold, light, salt, heavy metal stress as well as nutritional limitation. Such stresses severely alter both the morphology as well as the molecular phenotype of the organism. In figure 14, *Arabidopsis thaliana* specimen under selected types of abiotic stress are depicted.



Figure 14. Different *Arabidopsis thaliana* specimen. From left to right: unstressed, high light stress (Doerfler, 2014), and freezing stress (Gai et al., 2011). Changes in the morphological phenotype are accompanied with alterations in the molecular phenotype.

The stress metabolism of model plants like *Arabidopsis* has been heavily investigated in the past already and as metabolic profiling techniques are adding substantially (Shulaev et al., 2008), a lot of knowledge was created in recent years in this research field. Several primary metabolites have been identified as abiotic stress markers in independent studies; for instance proline has been shown to act as important osmoprotectant during salt, cold and drought stress (Shulaev et al., 2008; Skirycz et al., 2010; Urano et al., 2009), while sugars and sugar derivatives like maltose, raffinose, erythritol and inositol have been detected in elevated levels during general temperature (Cook et al., 2004; Kaplan et al., 2004; Rizhsky et al., 2004) and light stress (Caldana et al., 2011). Also intermediates from the tricarboxylic acid cycle were reported to be involved in abiotic stresses (Wulff-Zottele et al., 2010).

In recent years, the production of reactive oxygen species (ROS) in plants under stressful conditions and their impact on metabolic homeostasis has grasped the attention of plant physiologists (Apel and Hirt, 2004). Reactive oxygen species such as hydroxyl radicals ($\cdot\text{OH}$) or superoxide anions ($\text{O}_2^{\cdot-}$) are unavoidable by-products in aerobic organisms deriving from enhanced light reaction (Lockhart, 2013) and can potentially be damaging to the cell when scavenging mechanisms are not present (Nishiyama et al., 2001; Pero et al., 1990). Many compounds of the secondary metabolism have been identified as effective radical scavengers as they can stabilize harmful free radicals and prevent them from inflicting damage to vital molecules or cell compartments (Asada, 2006). Figure 15 shows the process of radical scavenging by a phenolic species which is often found in typical secondary metabolites.

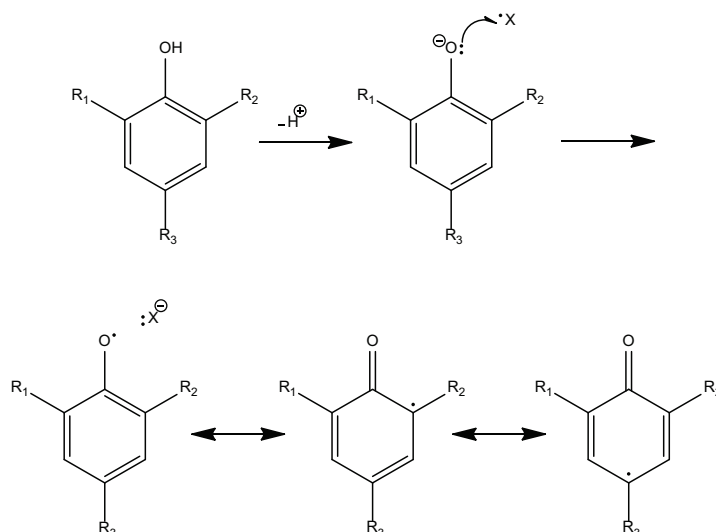


Figure 15. Process of scavenging of a reactive radical species ($\cdot\text{X}$) by a phenolic ring system. After deprotonation, the oxygen donates a single electron to the radical, quenching its reactivity, and itself becoming a stabilized radical. The aromatic system may then proceed to scavenge another unstable radical, or form polymerization products (Chobot, 2014).

However, despite their harmful effects, reactive oxygen species have been observed as important signalling mediators lately (Choudhury et al., 2013; Jaspers and Kangasjarvi, 2010; Pitzschke et al., 2006). Newer results indicate that the interaction of ROS with bioactive plant secondary metabolites play a crucial role in maintaining the redox homeostasis in cells by balancing pro- and antioxidative effects (Chobot et al., 2013; Hadacek et al., 2011).

3. Structure elucidation in MS-based metabolomics

3.1 Mass spectrometric fragment information

Mass spectrometric analysis is a method to characterize molecules by determining their masses (depending on the machine ranging from nominal to accurate) and isotopic patterns. Instruments capable of delivering precursor fragment information however, facilitate attempts to assign a chemical structure to a measured mass. Structure elucidation by investigation of fragmentation patterns is thus an established identification approach in MS-based bioanalytics. This holds especially true for molecules whose fragmentation spectra are reproducible and easy to interpret, as is the case for biopolymers like peptides, which – depending on instrument parameters – have their covalent bonds break in a characteristic fashion, and thus subsequently allow for the determination of their amino acid sequence, eventually leading to protein identification (figure 16). Because this process is highly reproducible, it can be automated and is successfully used in proteomics for protein identification and de-novo sequencing (Deutzmann, 2004; Savitski et al., 2005; Shen et al., 2011).

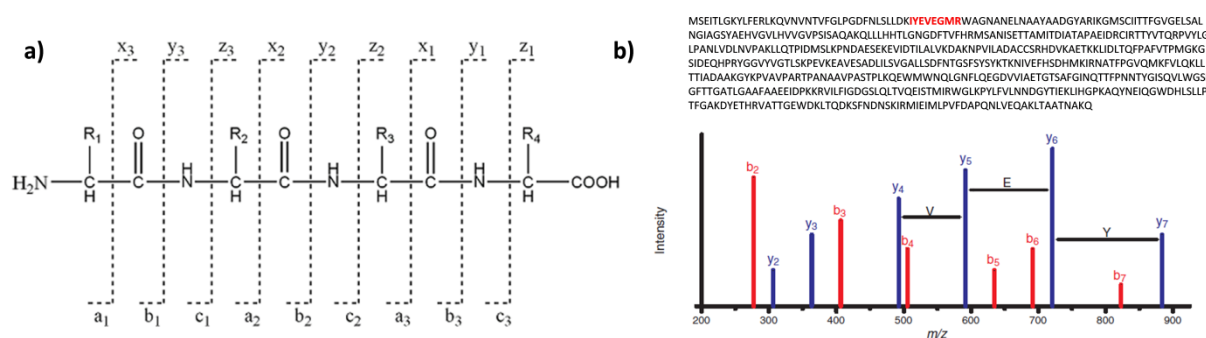


Figure 16. a) Fragmentation scheme of polypeptides. Scission of the peptidic bond results in b and y-ions, which are used to identify the amino acid sequence of a peptide. b) A characteristic sequence of the amino acids V, E and Y in the spectrum leads to the identification of the protein IYVEGMR (Pyruvate decarboxylase isozyme 1). Image modified (Sadygov et al., 2004).

Unlike proteins, metabolites come in vastly differing chemical structures, containing a plethora of functional groups and heteroatoms. This not only has severe impact on the separation and ionization mode of choice as explained above, but also results in a very complex fragmentation behaviour. In general, covalent bonds with low electron density are most prone to breaking when met with an electron particle stream as in GC-MS ion sources, or with an inert collision gas after electrospray ionization in LC-MS. Manifold tables and rules for mass spectrometric fragmentation patterns of small molecules exist, however, as there are several unobservable physicochemical processes happening during a fragmentation event, like gas-phase reactions of molecule rearrangement, de-novo peak annotation of metabolite fragment peaks is an exhaustive process and requires in-depth chemical knowledge (Kind and Fiehn, 2010). Figure 17 shows a fragmentation pathway of a typical plant secondary metabolite with annotated MS² spectrum.

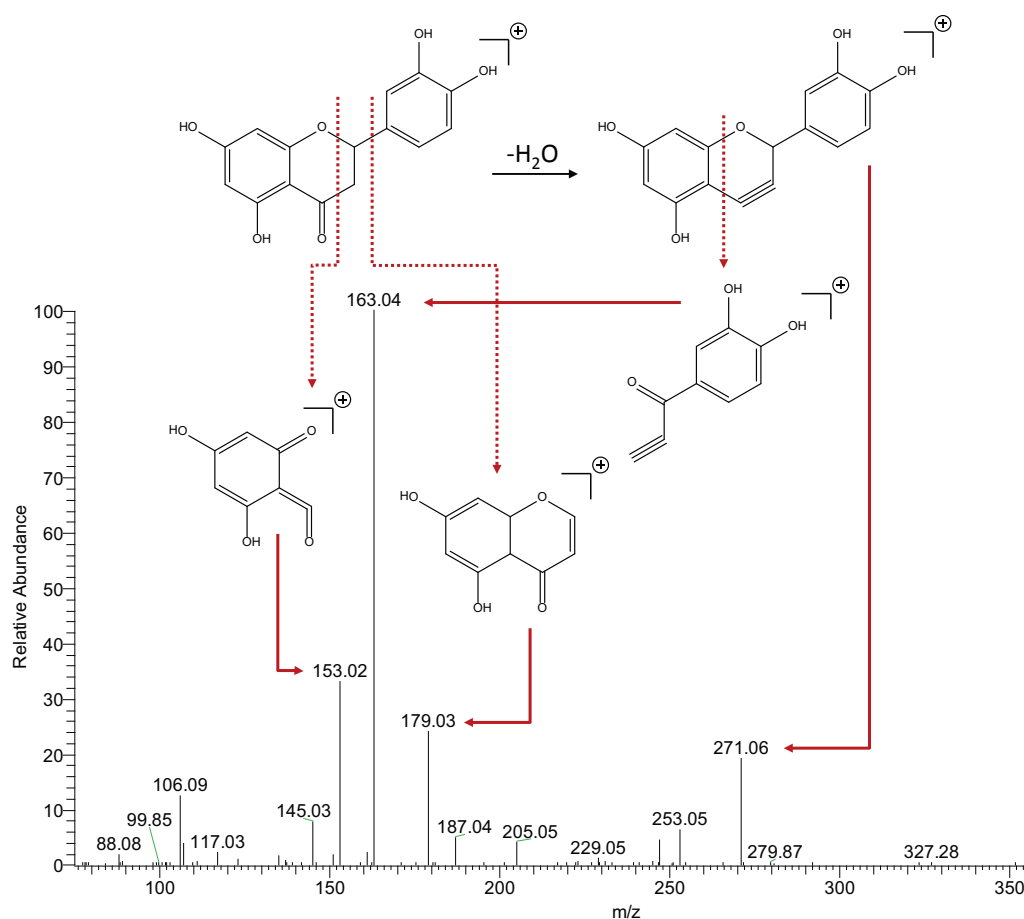


Figure 17. Putative fragmentation pathway of the flavonoid eriodictyol. After dehydration and cleavage of the A ring, the main fragment m/z 163 arises, while scission of the B ring results in the fragment m/z 179. m/z 153 results from the breaking of the two electron-deficient bonds of the C ring due to two adjacent oxygen atoms and is a characteristic peak in flavanone fragment spectra. Data was acquired from a standard compound on Orbitrap-MS (Doerfler, 2014) and matches with literature (Tsimogiannis et al., 2007). Data recorded in positive mode facilitating collision induced dissociation (CID).

Compared to GC-EI-MS, where ionization and fragmentation happen in one step and at a fixed energy level, resulting in very reproducible mass spectra, state-of-the-art mass spectrometers for LC-coupling offer a variety of different fragmentation types. For metabolomics, *collision induced dissociation* (CID) and *high energy collision dissociation* (HCD, available in Orbitrap detectors) are used. CID fragmentation benefits from high speed and sensitivity, while HCD is able to deliver exact fragment masses at the cost of reduced speed (Jedrychowski et al., 2011). An advantage of ion traps compared to TOF or Triple quadrupole mass analyzers is the ability to isolate daughter ions and perform additional of fragmentation steps, termed MSⁿ ability. MSⁿ experiments are required to distinguish masses which show the same spectrum even after initial fragmentation (figure 18).

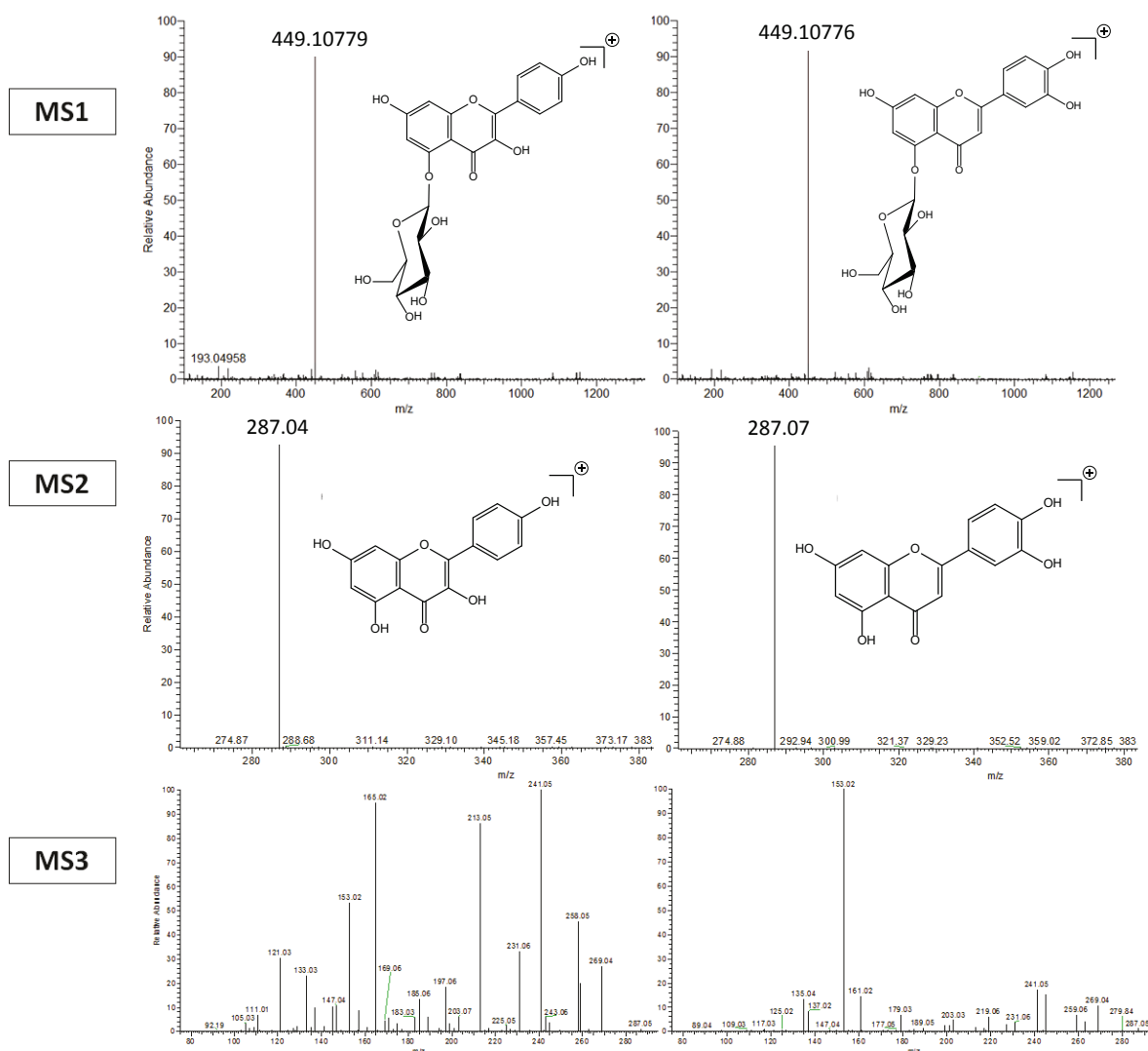


Figure 18. MS¹, MS² and MS³ spectra of kaempferol-5-O-glucoside (left) and luteolin-5-O-glucoside (right). The structures are constitutional isomers and display the same masses in the full scan spectrum (MS¹). MS² fragmentation results in the cleavage of the glycosidic bond, leaving behind the aglycons at both m/z 287. Only at MS³ level the compounds are distinguishable by their mass spectrum. Data acquired from standard compounds on Orbitrap-MS and recorded in positive mode facilitating collision induced dissociation (CID).

Unlike standardized GC-MS workflows, the extent of molecule breakdown is controlled and has to be optimized by the instrument operator. In the case of ion trap machines, desired precursor masses are isolated and subdued to a controlled amount of collision energy with the aim of generating an optimal fragment spectrum. With growing collision energies, the precursor structure depletes in favour of its stable daughter ions. Figure 19 shows the change in the MS² spectrum of kaempferol at different collision energies.

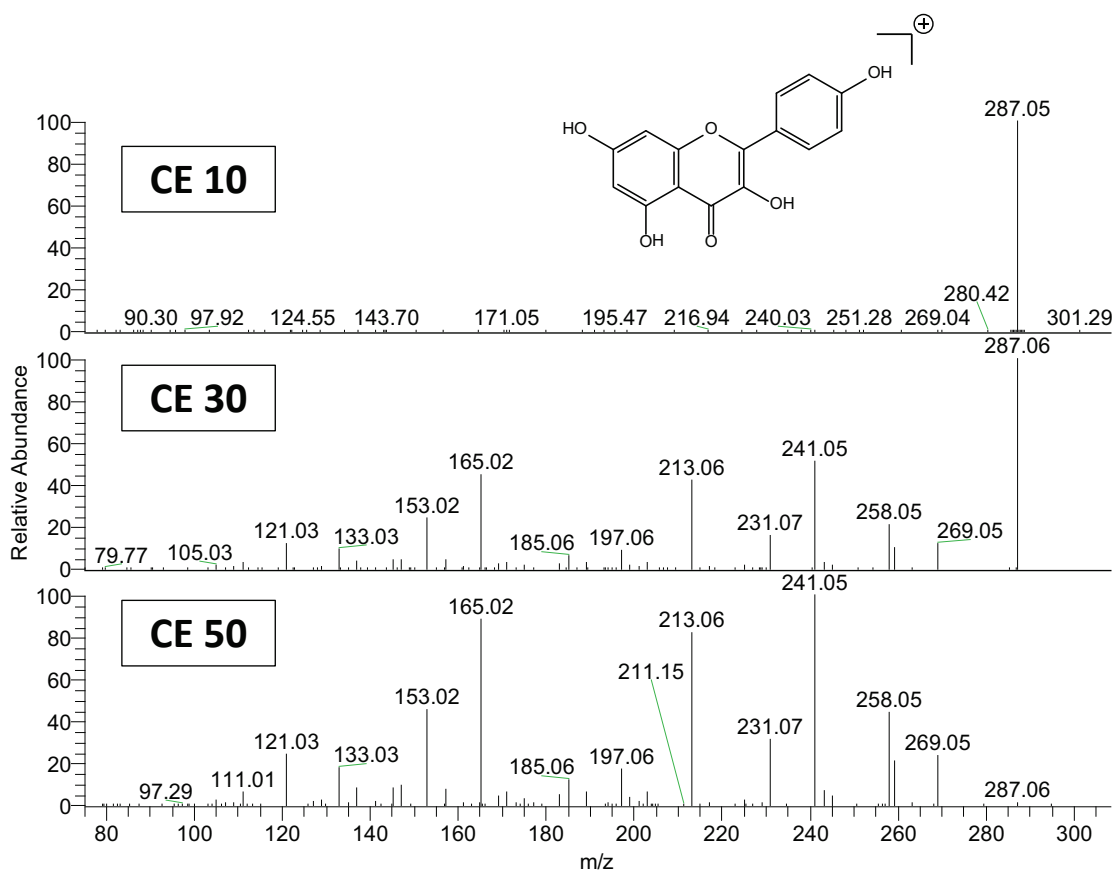


Figure 19. MS² product ions of the plant secondary metabolite kaempferol at different normalized collision energies. At CE 10, the unscathed precursor is visible because due to the low collision energy there is no fragmentation happening. At CE 30, several fragments arise, while the mother ion is still displaying the highest intensity peak in the spectrum. At CE 50, full fragmentation is observed, with the precursor molecule being barely visible in the spectrum.

At the moment there is still no accordance in the literature which fragmentation type and collision energy should be used in a regular fashion for reporting metabolomics data, because data may vary due to experimental layout and available instruments (Neumann and Bocker, 2010). However, efforts are being made to unite spectral data from various sources and in the near future supply scientists with comprehensive metabolite information platforms (Kanehisa, 2002; Posma et al., 2013; Sheldon et al., 2009).

3.2 High-resolution MS and sum formula annotation of novel metabolites

It is often the case in metabolomic studies that authors report recorded metabolite signals immediately as identifications, while actually the term “annotations” should be used, because reliable identification of novel compounds requires thorough structural evidence by a multitude of analytic techniques including extraction and purification (Neumann and Bocker, 2010). As this practice is not always feasible in metabolomic experiments time-wise, and also due to the fact that not all laboratories exhibit such a comprehensive range of analytical platforms, different levels of identification have been formulated (Neumann and Bocker, 2010) and minimum standards for reporting metabolite data have been proposed (Sumner et al., 2007). At any rate, the arrival of mass spectrometers capable of recording accurate compound masses in the low to sub-ppm levels (like FT-ICR and Orbitrap detectors) represents a significant step forward for the identification of unknowns in metabolomics-style experiments: once the monoisotopic mass of a metabolite is obtained, a sum formula can be formulated and via a database query a putative structure can be assigned (Dunn et al., 2012). Even though sum formula annotation of unknowns is theoretically also possible with less accurate mass spectrometers, it quickly becomes apparent that exact mass measurements are essential, because poorly resolved masses give no useful information about possible structures (figure 20).

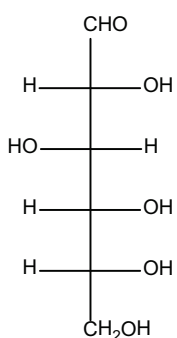
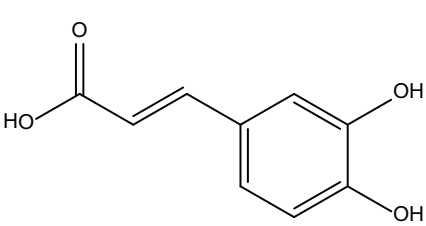
			
D-Glucose		Caffeic acid	
sum formula	C ₆ H ₁₂ O ₆	sum formula	C ₉ H ₈ O ₄
nominal mass	180	nominal mass	180
molecular weight	180.15588	molecular weight	180.15742
monoisotopic mass	180.06339	monoisotopic mass	180.04226

Figure 20. Different mass definitions for D-glucose and caffeic acid. The molecular weight of these substances is not distinguishable until the third decimal place. High-resolution MS is able to obtain the monoisotopic mass within a sub-ppm mass window and thus can differentiate between the two compounds.

Metabolites are limited to a mass range of about 50 to 2000 Da, thus the terms *exact mass* (the calculated sum of the masses of the individual isotopes of a molecule) and *monoisotopic mass* (the sum of the most abundant isotope masses of a molecule) are the same in most of the cases, compared to proteins and other macromolecules, where, due to the abundance of heavier isotopes, the most abundant mass is not necessarily the monoisotopic mass. Figure 21 shows a highly resolved Orbitrap mass spectrum of the main anthocyanin compound of *Arabidopsis thaliana* at 1181.29736 m/z (exact mass 1181.29800) (Bloor and Abrahams, 2002).

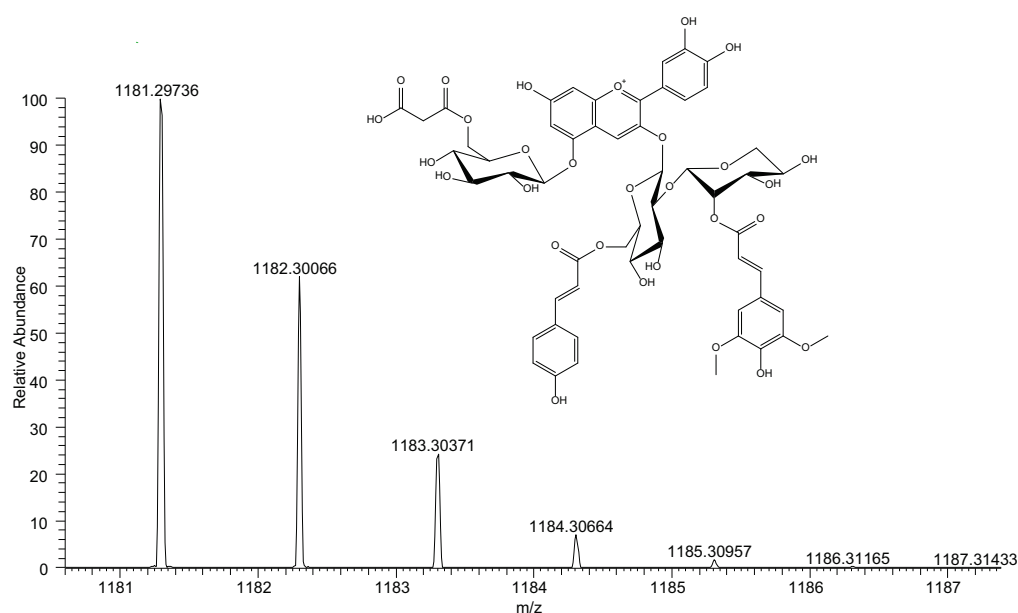


Figure 21. A prominent secondary metabolite of *A. thaliana* in a full scan spectrum at m/z 1181. With a resolution of 60 000 and mass accuracy of lower than 1 ppm, the monoisotopic mass and the adjacent isotopes can be obtained (Doerfler, 2014).

Even though high-resolution mass spectrometry renders sum formula annotations practicable in the first place, even at sub-ppm mass accuracy several chemically sound compositions are possible, and this number increases with mass and the variety of elements contained (Kind and Fiehn, 2006). Figure 22 shows how the sum formula amount for selected elements is depending on mass accuracy.

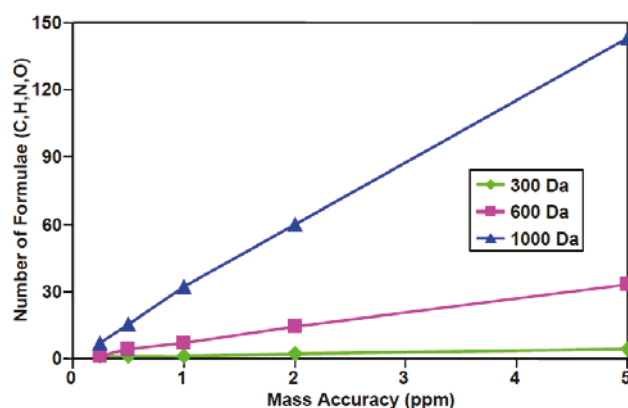


Figure 22. Total possible sum formulas for unknowns at 300, 600 and 1000 Da with different allowed accuracy window for sum formula annotation using the elements C, H, O and N. It is visible that even with sub-ppm mass accuracy, several possible sum formulas are existent for a certain mass (Quenzer, 2002).

Hence, applying the constraints of using the elements C, H, O and N with an allowed accuracy limit of 1 ppm in a sum formula calculator for m/z 1181, 12 possible formulas are obtained for this compound, from which the correct one comes only in the 9th spot in a list ranked by growing ppm delta (Table 2). Considering that metabolites often also contain more heteroatoms like S and P (and sometimes even halogens), and that adducts with Na⁺ or K⁺, which tend to emerge during electrospray ionization, must also be factored in, this number can increase exponentially.

m/z	theo. Mass	delta (ppm)	composition
1181.29724	1181.29717	0.06	C40 H63 O39 N
	1181.29717	0.06	C39 H57 O34 N8
	1181.29748	-0.21	C67 H43 O13 N9
	1181.29749	-0.21	C68 H49 O18 N2
	1181.29698	0.22	C81 H41 O7 N4
	1181.29666	0.49	C53 H55 O28 N3
	1181.29665	0.5	C52 H49 O23 N10
	1181.29800	-0.64	C54 H51 O24 N7
	1181.29800	-0.64	C55 H57 O29
	1181.29831	-0.91	C82 H37 O3 N8
	1181.29832	-0.91	C83 H43 O8 N
	1181.29615	0.93	C66 H47 O17 N5

Table 2. A list of possible sum formulas for the measured precursor mass 1181.29724. Choosing the monoisotopic masses of C (100), H (100), O (100) and N (10), and 1 ppm error window for sum formula generation, the correct formula for the exact mass 1181.29800 comes in the 9th place (Doerfler, 2014). Sum formulas were calculated with the Xcalibur sum formula calculator (XcaliburTM 2.2, Thermo Scientific).

Consequently, as the correct sum formula offer will not reliably come at the first place, using the lowest ppm value as deciding factor for sum formula annotation attempts is a futile approach. To tackle this issue, several strategies have been proposed, using mass spectrometric data alone:

For one, the lot of generated sum formulas for one m/z signal can be filtered and ultimately narrowed down by applying rules for chemical plausibility; putative compositions will as such be checked for their valences in respect to their elements contained. Kind and Fiehn showed that the classic chemical restraints for sum formula annotation like ring double bond equivalents (RDBE) or the nitrogen rule are insufficient for dealing with the highly complex molecules that can be detected in a mass spectrometry-based metabolomics experiment and proposed seven golden rules for filtering mass spectrometric data (Kind and Fiehn, 2007).

An additional method, utilizing pure mass spectrometric information, is to investigate isotopic abundances and ion patterns of the recorded data (Stoll et al., 2006). Isotope ratios from the mass spectrum can help determine the elements contained in the compound, while ions observed with multiple adducts (e.g. recurring formulas with Na^+ or K^+) can hint to the correct sum formula. This approach, however, requires mass spectrometers sensitive enough to record even low abundant isotopic peaks (Dunn et al., 2012).

Apart from pure computational approaches, isotopic labelling can be an efficient way to annotate new compounds in metabolomics-style experiments. Giavalisco et al. showed that elemental composition annotation can be improved largely by comparing whole metabolomes of ^{12}C vs. ^{13}C -grown *Arabidopsis* plants and aligning their m/z signals according to their retention time, which at the same time led to the discovery of a plethora of previously undescribed substances (Giavalisco et al., 2008; Giavalisco et al., 2009).

Another strategy in this context is to account for potential substrate-product pairs in combination with sum formula generation. Considering that metabolites derive from enzyme-mediated networks, it is often the case to find compounds which differ by characteristic structures or functional groups, like methyl- or hydroxy-moieties (Dunn et al., 2012). Rogers et al. showed how these biological relations can be exploited for improved annotation of elemental compositions (Rogers et al., 2009). In combination with sub-ppm mass measurements and additional fragmentation scans, the present work shows how complete metabolic pathways can be tracked down and structures of new compounds may be proposed when taking such biochemical conversions into consideration, within the frame of an automated data processing algorithm, as will be shown in 3.3 (Doerfler et al., 2014).

A helpful concept to deal with sum formula annotations of unknown peaks in high resolution mass spectrometry is the construction of Kendrick diagrams and van Krevelen plots (Reemtsma, 2009). Kendrick diagrams were first proposed by Kendrick in 1963 for the computation of alkane sequences in petro chemistry (Kendrick, 1963) and assume a Kendrick Mass of exactly 12 u for each ^{12}C atom of a compound instead of the official IUPAC mass. This allows for the calculation of a specific Kendrick mass defect (KMD) and the construction of KMD vs. Kendrick Mass diagrams, which can highlight structural similarities of compounds on a horizontal line. Van Krevelen plots were invented by van Krevelen in the 1950s to characterize the chemical composition of coals and mineral oils. Such diagrams feature the H/C ratio of substances on the y-axis, and their respective O/C ratios on the x-axis and can be useful to investigate clustering of chemically familiar substances. Because compounds of biogenic origin tend to create specific patterns on Kendrick and van Krevelen

diagrams, they have been rediscovered for the determination of sum formulas in NOM (natural organic matter) research (Kim et al., 2003; Wu et al., 2004). Figure 23 shows a van Krevelen plot for NOM analysis.

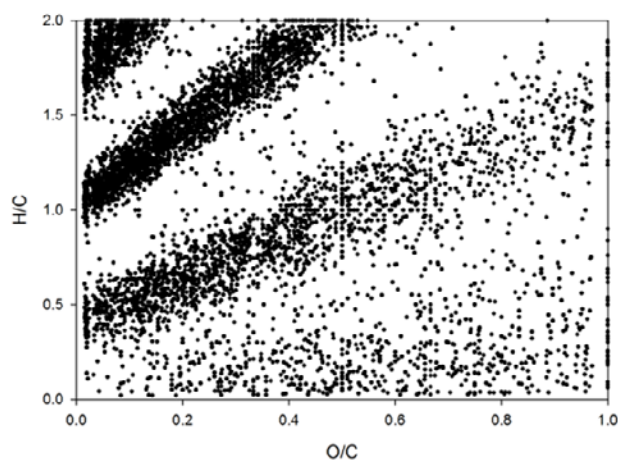


Figure 23. Van Krevelen plot of a typical NOM analysis. Clustering of specific compound families due to their H/C and O/C ratio is visible (Doerfler, 2013).

Van Krevelen plots can be used to visualize typical enzyme-catalysed reactions, and particularly oxidative processes in metabolomics-style experiments (Kai et al., 2011). Additional axes for N or S atoms, as occurring in secondary metabolites, can be added to the plot as a third dimension. In combination with sub-ppm mass measurements, this allows for large-scale pathway elucidation, as will be discussed in the upcoming chapter.

3.3 Structure elucidation and metabolic pathway reconstruction

The identification of unknown signals is momentarily the bottleneck in metabolomics-style experiments and will be in the focus of the metabolomics society for the upcoming years (Dunn et al., 2012; Schuhmacher et al., 2013), as for the moment, only a fraction of the compounds in a complex sample can be reliably identified. While for targeted analysis, identification and quantification of groups of certain substances is already possible with the help of databases (especially in GC-MS), untargeted metabolomic analyses to date suffer from diverse shortcomings like insufficient database availability and lack of data processing procedures (Schuhmacher et al., 2013). For characterization of data generated by an untargeted, high-resolution metabolomics experiment, it is thus imperative to extract as much information as possible. This includes retention times and intensities on the chromatogram side, as well as accurate precursor masses, and spectral information generated by fragmentation scans on the mass spectrometric side.

Especially for metabolome data derived from a timed sample drawing process or experiments focusing on comparing metabolomes under different stress or disease conditions,

it can be expected to detect compounds which are in biochemical relation to each other; these metabolites represent educts and products mirroring the metabolic change happening in the plants. The present work focuses on a method to automatically detect these metabolic linkages between acquired m/z features from the precursor level, thus alleviating sum formula annotation. An example of this process is showcased in figure 24. Compound m/z 1181, the main anthocyanin in *Arabidopsis thaliana* (see above), has a related metabolite in m/z 1197, with a net difference of one oxygen in their sum formulas (Saito et al., 1995). By narrowing down the sum formula predictions of detected compounds to a restricted list of possible biochemical reactions occurring in nature, whole metabolic pathways and putatively new substances can be extracted from untargeted metabolomics data. Subsequent validation of these compounds with recognized biological familiarity by MS^2 and MS^3 scan events can enable the identification of unknowns and allows for large-scale reconstruction of metabolic networks (II.2) (Doerfler et al., 2014).

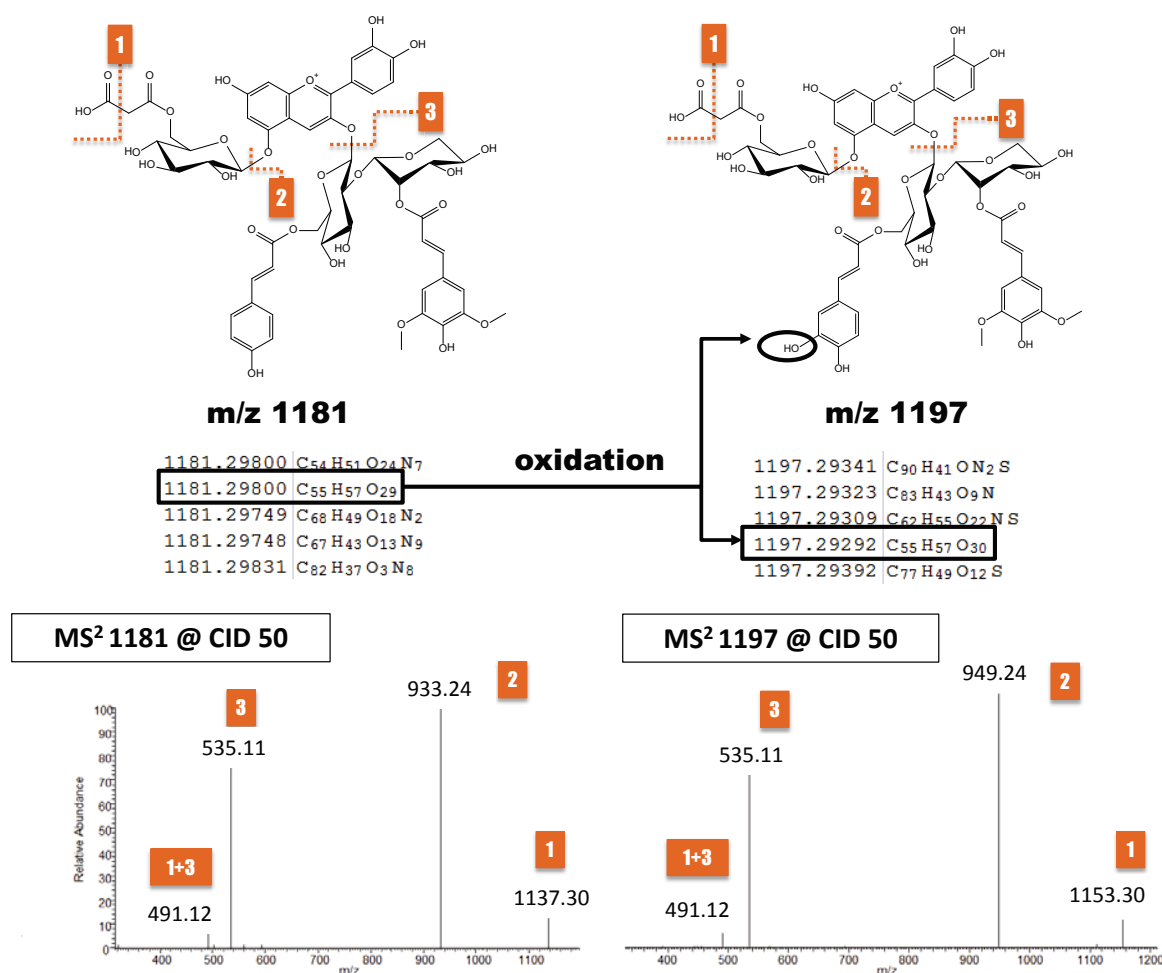


Figure 24. Correct sum formula annotation of a compound by searching for biochemical familiarity. Anthocyanin m/z 1181 differs one hydroxy group on the *p*-coumaroyl group from compound m/z 1197. The corresponding sum formula is automatically found from various offers by an algorithm termed *mzGroupAnalyzer*. The fragmentation scheme and MS^2 product ion scans help validate the new structure (Doerfler et al., 2014).

Investigating educt – product relationships in metabolome data can give insights into pathway regulation and ultimately help identify unknowns and construct metabolite networks (Doerfler et al., 2013). A helpful concept in this regard is the application of the Granger causality test to time-dependent metabolome data (Doerfler et al., 2013; Sriyudthsak et al., 2013). Granger causality analysis, originating from causal relations in economic models (Granger, 1969), is a correlation analysis for time series data which – in a metabolomics context – can identify how compounds are influencing others due to a time-lagged change in their concentrations. Figure 25 depicts a typical Granger causation, where a specific metabolite acts as a precursor for its glycosylated form during the *Arabidopsis* cold-stress experiment described in chapter 2.

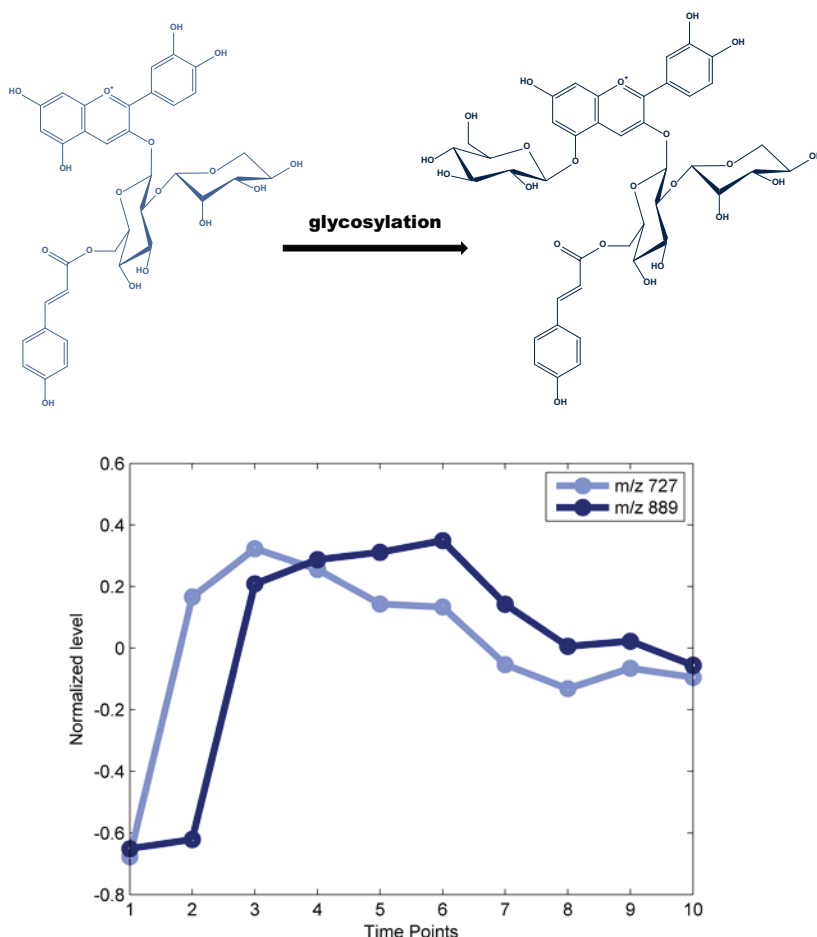


Figure 25. Granger relation between two secondary metabolites in *Arabidopsis thaliana*. Compound “m/z 727” is getting glycosylated to compound “m/z 889” during high cold and light stress. After a sharp increase in the beginning, its relative concentration is falling, giving rise to its successor molecule with a time lag of 1 sampling point. A Granger cause is called significant with a p significance value of under 0.05. The presented Granger relation displays a p value of 0.003549.

Both the *mzGroupAnalyzer* algorithm, as well as the Granger causality test, and many other tools for statistical treatment of metabolomics data, have been included in the metabolomics toolbox COVAIN (Sun and Weckwerth, 2012).

3.4 m/z feature extraction of LC-MS data

Extraction and alignment of m/z features from untargeted metabolomics or proteomics analyses is a crucial step in “omics”-style experiments, as only reliable selection of each the same precursor ions for numerous biological and/or technical replicates allows for meaningful statistical data treatment and subsequent biological interpretation. This issue has been addressed successfully by the development of the so-called mass accuracy precursor alignment (MAPA) algorithm (Hoehenwarter et al., 2008) which is realized within the ProtMAX program (Egelhofer et al., 2013). This algorithm will extract and align m/z precursors according to user preferences over a number of mass spectrometric data files in .mzXML format. Depending on the choice of the mass spectrometric acquisition method – which itself is largely dependent on the type of biological samples to analyze – two major ways to extract mass spectrometric precursor information are being used: The classical “spectral count” approach (the summed number of all MS² spectra which are triggered for a specific m/z value) is useful for alignment and relative quantification of proteomics data. An alternative to this is using the “ion intensity count” option, which was already being successfully used by extraction and relative quantification of precursors in metabolomics data (Doerfler et al., 2013). Figure 26 depicts the difference between these two methods showcasing a MS¹ spectrum of a typical plant secondary metabolite.

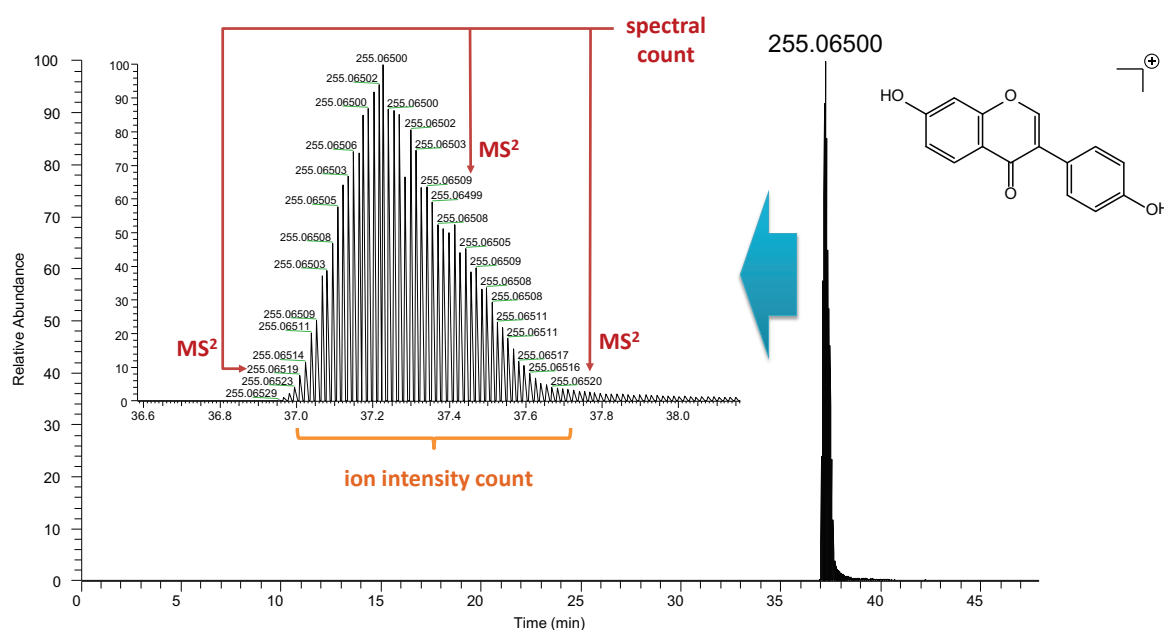


Figure 26. A zoom-in on the MS¹ level of the compound daidzein with the singly protonated mass 255.06519 in a highly resolved precursor mass spectrum reveals the individual profile-mode scans. In a typical MS² method, only a few product scans happen for a selected precursor mass, which is the basis for relative quantification in spectral counting. An ion intensity count method within a retention time window enables summation of

individual scans of a precursor up to the third digit, which results in an improved quantification of precursors of the same m/z ratio (Doerfler, 2014).

A recognized downside to spectral counting has to date been the partly false binning of precursors, for instance in the quantification of proteins due to protein isoforms. To remedy this issue, a targeted MAPA (TMAPA) approach was developed, which can extract and align proteotypic peptide precursor ions from shotgun proteomics data by facilitating the “target list” option in ProtMAX. This strategy was shown to yield overall better results in protein quantification (II.3) and is also expected to be suitable for extraction and relative quantification for future experiments in metabolomics and direct-infusion-MS (DIMS) experiments.

4. References

- Allwood, J.W., and R. Goodacre. 2010. An introduction to liquid chromatography-mass spectrometry instrumentation applied in plant metabolomic analyses. *Phytochem Anal.* 21:33-47.
- Anderson, D.C., and K. Kodukula. 2013. Biomarkers in pharmacology and drug discovery. *Biochem Pharmacol.*
- Antoniewicz, M.R. 2013. Tandem mass spectrometry for measuring stable-isotope labeling. *Curr Opin Biotechnol.* 24:48-53.
- Apel, K., and H. Hirt. 2004. Reactive oxygen species: metabolism, oxidative stress, and signal transduction. *Annu Rev Plant Biol.* 55:373-399.
- Asada, K. 2006. Production and scavenging of reactive oxygen species in chloroplasts and their functions. *Plant Physiol.* 141:391-396.
- Banerjee, A. 2011. Food, feed, fuel: transforming the competition for grains. *Dev Change.* 42:529-557.
- Barrett, D. 2012. Advances in metabolic profiling. *Bioanalysis.* 4:643-644.
- Beecher, C.W. 2011. Metabolomic studies at the start and end of the life cycle. *Clin Biochem.* 44:518-519.
- Bloor, S.J., and S. Abrahams. 2002. The structure of the major anthocyanin in *Arabidopsis thaliana*. *Phytochemistry.* 59:343-346.
- Boggess, M.V., J.D. Lippolis, W.J. Hurkman, C.K. Fagerquist, S.P. Briggs, A.V. Gomes, P.G. Righetti, and K. Bala. 2013. The need for agriculture phenotyping: "Moving from genotype to phenotype". *J Proteomics.*
- Brown, M., W.B. Dunn, P. Dobson, Y. Patel, C.L. Winder, S. Francis-McIntyre, P. Begley, K. Carroll, D. Broadhurst, A. Tseng, N. Swainston, I. Spasic, R. Goodacre, and D.B. Kell. 2009. Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. *Analyst.* 134:1322-1332.
- Caldana, C., T. Degenkolbe, A. Cuadros-Inostroza, S. Klie, R. Sulpice, A. Leisse, D. Steinhauser, A.R. Fernie, L. Willmitzer, and M.A. Hannah. 2011. High-density kinetic analysis of the metabolomic and transcriptomic response of *Arabidopsis* to eight environmental conditions. *Plant J.* 67:869-884.
- Calvin, M., and A.A. Benson. 1948. The Path of Carbon in Photosynthesis. *Science.* 107:476-480.
- Cerdan-Calero, M., J.M. Sendra, and E. Sentandreu. 2012. Gas chromatography coupled to mass spectrometry analysis of volatiles, sugars, organic acids and aminoacids in Valencia Late orange juice and reliability of the Automated Mass Spectral Deconvolution and Identification System for their automatic identification and quantification. *J Chromatogr A.* 1241:84-95.
- ChemSpider. <http://www.chemspider.com/Chemical-Structure.4518630.html>.
- Chobot, V. 2014. *personal communication*.
- Chobot, V., L. Kubicova, G. Bachmann, and F. Hadacek. 2013. Versatile redox chemistry complicates antioxidant capacity assessment: flavonoids as milieu-dependent anti- and pro-oxidants. *Int J Mol Sci.* 14:11830-11841.
- Choudhury, S., P. Panda, L. Sahoo, and S.K. Panda. 2013. Reactive oxygen species signaling in plants under abiotic stress. *Plant Signal Behav.* 8:e23681.
- Collino, S., F.P. Martin, and S. Rezzi. 2013. Clinical metabolomics paves the way towards future healthcare strategies. *Br J Clin Pharmacol.* 75:619-629.
- Cook, D., S. Fowler, O. Fiehn, and M.F. Thomashow. 2004. A prominent role for the CBF cold response pathway in configuring the low-temperature metabolome of *Arabidopsis*. *Proc Natl Acad Sci U S A.* 101:15243-15248.
- Croteau, R., T.M. Kutchan, and N.G. Lewis. 2000. Natural Products (Secondary Metabolites). *Biochemistry and Molecular Biology of Plants*:1250-1318.

- De Vos, R.C.H., S. Moco, A. Lommen, J.J.B. Keurentjes, R.J. Bino, and R.D. Hall. 2007. Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. *Nat Protoc.* 2:778-791.
- Debnath, M., M. Pandey, and P.S. Bisen. 2011. An omics approach to understand the plant abiotic stress. *Omics.* 15:739-762.
- Dettmer, K., P.A. Aronov, and B.D. Hammock. 2007. Mass spectrometry-based metabolomics. *Mass Spectrom Rev.* 26:51-78.
- Deutzmann, R. 2004. Structural characterization of proteins and peptides. *Methods Mol Med.* 94:269-297.
- Doerfler, H. 2013. personal communication with Chobot, V.
- Doerfler, H. 2014. *own work.*
- Doerfler, H., D. Lyon, T. Nagele, X. Sun, L. Fagner, F. Hadacek, V. Egelhofer, and W. Weckwerth. 2013. Granger causality in integrated GC-MS and LC-MS metabolomics data reveals the interface of primary and secondary metabolism. *Metabolomics.* 9:564-574.
- Doerfler, H., X. Sun, L. Wang, D. Engelmeier, D. Lyon, and W. Weckwerth. 2014. mzGroupAnalyzer- Predicting Pathways and Novel Chemical Structures from Untargeted High-Throughput Metabolomics Data. *PLoS One.* 9:e96188.
- Dunn, W.B., and D.I. Ellis. 2005. Metabolomics: Current analytical platforms and methodologies. *Trac-Trend Anal Chem.* 24:285-294.
- Dunn, W.B., A. Erban, R.J.M. Weber, D.J. Creek, M. Brown, R. Breitling, T. Hankemeier, R. Goodacre, S. Neumann, J. Kopka, and M.R. Viant. 2012. Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics.* 9:44-66.
- Egelhofer, V., W. Hoehenwarter, D. Lyon, W. Weckwerth, and S. Wienkoop. 2013. Using ProtMAX to create high-mass-accuracy precursor alignments from label-free quantitative mass spectrometry data generated in shotgun proteomics experiments. *Nat Protoc.* 8:595-601.
- Eliasson, M., S. Rannar, and J. Trygg. 2011. From data processing to multivariate validation--essential steps in extracting interpretable information from metabolomics data. *Curr Pharm Biotechnol.* 12:996-1004.
- Fenn, J.B. 2002. Electrospray ionization mass spectrometry: How it all began. *J Biomol Tech.* 13:101-118.
- Fenn, J.B. 2003. Electrospray wings for molecular elephants (Nobel lecture). *Angew Chem Int Ed Engl.* 42:3871-3894.
- Fenn, J.B., M. Mann, C.K. Meng, S.F. Wong, and C.M. Whitehouse. 1989. Electrospray ionization for mass spectrometry of large biomolecules. *Science.* 246:64-71.
- Fiehn, O. 2002. Metabolomics--the link between genotypes and phenotypes. *Plant Mol Biol.* 48:155-171.
- Fiehn, O. 2008. Extending the breadth of metabolite profiling by gas chromatography coupled to mass spectrometry. *Trends Analyt Chem.* 27:261-269.
- Fujita, A., P. Severino, J.R. Sato, and S. Miyano. 2010. Granger Causality in Systems Biology: Modeling Gene Networks in Time Series Microarray Data Using Vector Autoregressive Models. *Lect N Bioinformat.* 6268:13-24.
- Furuya, T. 1965. Gas-Liquid Chromatography of Plant Glycosides. *J Chromatogr.* 18:152-156.
- Gai, Y.P., X.L. Ji, W. Lu, X.J. Han, G.D. Yang, and C.C. Zheng. 2011. A novel late embryogenesis abundant like protein associated with chilling stress in *Nicotiana tabacum* cv. bright yellow-2 cell suspension culture. *Mol Cell Proteomics.* 10:M111 010363.
- Giavalisco, P., J. Hummel, J. Lisec, A.C. Inostroza, G. Catchpole, and L. Willmitzer. 2008. High-resolution direct infusion-based mass spectrometry in combination with whole ¹³C metabolome isotope labeling allows unambiguous assignment of chemical sum formulas. *Anal Chem.* 80:9417-9425.
- Giavalisco, P., K. Kohl, J. Hummel, B. Seiwert, and L. Willmitzer. 2009. ¹³C isotope-labeled metabolomes allowing for improved compound annotation and relative quantification in liquid chromatography-mass spectrometry-based metabolomic research. *Anal Chem.* 81:6546-6551.

- Goodacre, R. 2005. Metabolomics – the way forward. *Metabolomics*. 1:1-2.
- Goodacre, R., S. Vaidyanathan, W.B. Dunn, G.G. Harrigan, and D.B. Kell. 2004. Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol.* 22:245-252.
- Govindaraghavan, S., J.R. Hennell, and N.J. Sucher. 2012. From classical taxonomy to genome and metabolome: towards comprehensive quality standards for medicinal herb raw materials and extracts. *Fitoterapia*. 83:979-988.
- Granger, C.W.J. 1969. Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. *Econometrica*. 37:414-426.
- Guy, C., J. Kopka, and T. Moritz. 2008. Plant metabolomics coming of age. *Physiol Plant*. 132:113-116.
- Hadacek, F., G. Bachmann, D. Engelmeier, and V. Chobot. 2011. Hormesis and a Chemical Raison D'etre for Secondary Plant Metabolites. *Dose Response*. 9:79-116.
- Halket, J.M., D. Waterman, A.M. Przyborowska, R.K. Patel, P.D. Fraser, and P.M. Bramley. 2005. Chemical derivatization and mass spectral libraries in metabolic profiling by GC/MS and LC/MS/MS. *J Exp Bot*. 56:219-243.
- Hall, R.D. 2006. Plant metabolomics: from holistic hope, to hype, to hot topic. *New Phytol*. 169:453-468.
- Hall, R.D., I.D. Brouwer, and M.A. Fitzgerald. 2008. Plant metabolomics and its potential application for human nutrition. *Physiol Plant*. 132:162-175.
- Hoehenwarter, W., J.T. van Dongen, S. Wienkoop, M. Steinfath, J. Hummel, A. Erban, R. Sulpice, B. Regierer, J. Kopka, P. Geigenberger, and W. Weckwerth. 2008. A rapid approach for phenotype-screening and database independent detection of cSNP/protein polymorphism using mass accuracy precursor alignment. *Proteomics*. 8:4214-4225.
- Hood, L., J.R. Heath, M.E. Phelps, and B. Lin. 2004. Systems biology and new technologies enable predictive and preventative medicine. *Science*. 306:640-643.
- Hummel, J., N. Strehmel, J. Selbig, D. Walther, and J. Kopka. 2010. Decision tree supported substructure prediction of metabolites from GC-MS profiles. *Metabolomics*. 6:322-333.
- Ideker, T., T. Galitski, and L. Hood. 2001. A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet*. 2:343-372.
- IUPAC. <http://goldbook.iupac.org/R05360.html>. <http://goldbook.iupac.org/R05360.html>.
- Jaspers, P., and J. Kangasjarvi. 2010. Reactive oxygen species in abiotic stress signaling. *Physiol Plant*. 138:405-413.
- Jedrychowski, M.P., E.L. Huttlin, W. Haas, M.E. Sowa, R. Rad, and S.P. Gygi. 2011. Evaluation of HCD- and CID-type fragmentation within their respective detection platforms for murine phosphoproteomics. *Mol Cell Proteomics*. 10:M111 009910.
- Kai, K., H. Takahashi, H. Saga, T. Ogawa, S. Kanaya, and D. Ohta. 2011. Metabolomic characterization of the possible involvement of a Cytochrome P450, CYP81F4, in the biosynthesis of indolic glucosinolate in Arabidopsis. *Plant Biotechnology*. 28:379-385.
- Kanehisa, M. 2002. The KEGG database. *Novartis Found Symp*. 247:91-101; discussion 101-103, 119-128, 244-152.
- Kanehisa, M., and S. Goto. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 28:27-30.
- Kaplan, F., J. Kopka, D.W. Haskell, W. Zhao, K.C. Schiller, N. Gatzke, D.Y. Sung, and C.L. Guy. 2004. Exploring the temperature-stress metabolome of Arabidopsis. *Plant Physiol*. 136:4159-4168.
- Kell, D.B. 2004. Metabolomics and systems biology: making sense of the soup. *Curr Opin Microbiol*. 7:296-307.
- Kellogg, H.M., E. Brochmann-Hanssen, and A.B. Svendsen. 1964. Gas Chromatography of Esters of Plant Acids and Their Identification in Plant Materials. *J Pharm Sci*. 53:420-423.
- Kempa, S., J. Hummel, T. Schwemmer, M. Pietzke, N. Strehmel, S. Wienkoop, J. Kopka, and W. Weckwerth. 2009. An automated GCxGC-TOF-MS protocol for batch-wise extraction and alignment of mass isotopomer matrixes from differential ¹³C-labelling experiments: a case study for photoautotrophic-mixotrophic grown Chlamydomonas reinhardtii cells. *J Basic Microb*. 49:82-91.

- Kemsley, E.K., G. Le Gall, J.R. Dainty, A.D. Watson, L.J. Harvey, H.S. Tapp, and I.J. Colquhoun. 2007. Multivariate techniques and their application in nutrition: a metabolomics case study. *Br J Nutr.* 98:1-14.
- Kendrick, E. 1963. A Mass Scale Based on $\text{CH}_2 = 14.0000$ for High Resolution Mass Spectrometry of Organic Compounds. *Anal Chem.* 35:2146-2154.
- Kim, S., R.W. Kramer, and P.G. Hatcher. 2003. Graphical method for analysis of ultrahigh-resolution broadband mass spectra of natural organic matter, the van Krevelen diagram. *Anal Chem.* 75:5336-5344.
- Kind, T., and O. Fiehn. 2006. Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics.* 7:234.
- Kind, T., and O. Fiehn. 2007. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics.* 8:105.
- Kind, T., and O. Fiehn. 2010. Advances in structure elucidation of small molecules using mass spectrometry. *Bioanal Rev.* 2:23-60.
- Kind, T., G. Wohlgemuth, Y. Lee do, Y. Lu, M. Palazoglu, S. Shahbaz, and O. Fiehn. 2009. FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Anal Chem.* 81:10038-10048.
- Kitano, H. 2002. Systems biology: a brief overview. *Science.* 295:1662-1664.
- Koek, M.M., R.H. Jellema, J. van der Greef, A.C. Tas, and T. Hankemeier. 2011. Quantitative metabolomics based on gas chromatography mass spectrometry: status and perspectives. *Metabolomics.* 7:307-328.
- Kopka, J. 2006. Current challenges and developments in GC-MS based metabolite profiling technology. *J Biotechnol.* 124:312-322.
- Kopka, J., N. Schauer, S. Krueger, C. Birkemeyer, B. Usadel, E. Bergmuller, P. Dormann, W. Weckwerth, Y. Gibon, M. Stitt, L. Willmitzer, A.R. Fernie, and D. Steinhauser. 2005. GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics.* 21:1635-1638.
- Kovats, E. 1958. Gas-Chromatographische Charakterisierung Organischer Verbindungen .1. Retentionsindices Aliphatischer Halogenide, Alkohole, Aldehyde Und Ketone. *Helv Chim Acta.* 41:1915-1932.
- Krebs, H.A., and W.A. Johnson. 1937. Metabolism of ketonic acids in animal tissues. *Biochem J.* 31:645-660.
- Krone, N., B.A. Hughes, G.G. Lavery, P.M. Stewart, W. Arlt, and C.H. Shackleton. 2010. Gas chromatography/mass spectrometry (GC/MS) remains a pre-eminent discovery tool in clinical steroid investigations even in the era of fast liquid chromatography tandem mass spectrometry (LC/MS/MS). *J Steroid Biochem Mol Biol.* 121:496-504.
- Krumsiek, J., K. Suhre, A.M. Evans, M.W. Mitchell, R.P. Mohney, M.V. Milburn, B. Wagele, W. Romisch-Margl, T. Illig, J. Adamski, C. Gieger, F.J. Theis, and G. Kastenmuller. 2012. Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information. *PLoS Genet.* 8:e1003005.
- Lammerhofer, M. 2010. HILIC and mixed-mode chromatography: the rising stars in separation science. *J Sep Sci.* 33:679-680.
- Lee, D.K., M.H. Yoon, Y.P. Kang, J. Yu, J.H. Park, J. Lee, and S.W. Kwon. 2013. Comparison of primary and secondary metabolites for suitability to discriminate the origins of Schisandra chinensis by GC/MS and LC/MS. *Food Chem.* 141:3931-3937.
- Less, H., and G. Galili. 2008. Principal transcriptional programs regulating plant amino acid metabolism in response to abiotic stresses. *Plant Physiol.* 147:316-330.
- Lisec, J., N. Schauer, J. Kopka, L. Willmitzer, and A.R. Fernie. 2006. Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nat Protoc.* 1:387-396.
- Lockhart, J. 2013. Frenemies: antagonistic bHLH/bZIP transcription factors integrate light and reactive oxygen species signaling in Arabidopsis. *Plant Cell.* 25:1483.
- Maeda, H., and N. Dudareva. 2012. The shikimate pathway and aromatic amino Acid biosynthesis in plants. *Annu Rev Plant Biol.* 63:73-105.

- May, P., N. Christian, O. Ebenhoh, W. Weckwerth, and D. Walther. 2011. Integration of proteomic and metabolomic profiling as well as metabolic modeling for the functional analysis of metabolic networks. *Methods Mol Biol.* 694:341-363.
- May, P., S. Wienkoop, S. Kempa, B. Usadel, N. Christian, J. Rupprecht, J. Weiss, L. Recueno-Munoz, O. Ebenhoh, W. Weckwerth, and D. Walther. 2008. Metabolomics- and proteomics-assisted genome annotation and analysis of the draft metabolic network of *Chlamydomonas reinhardtii*. *Genetics*. 179:157-166.
- Milne, S.B., T.P. Mathews, D.S. Myers, P.T. Ivanova, and H.A. Brown. 2013. Sum of the parts: mass spectrometry-based metabolomics. *Biochemistry*. 52:3829-3840.
- Monteiro, M.S., M. Carvalho, M.L. Bastos, and P. Guedes de Pinho. 2013. Metabolomics analysis for biomarker discovery: advances and challenges. *Curr Med Chem*. 20:257-271.
- Morgenthal, K., W. Weckwerth, and R. Steuer. 2006. Metabolomic networks in plants: Transitions from pattern recognition to biological interpretation. *Biosystems*. 83:108-117.
- Munoz, B., and A. Albores. 2010. The role of molecular biology in the biomonitoring of human exposure to chemicals. *Int J Mol Sci*. 11:4511-4525.
- Nägele, T., and W. Weckwerth. 2012. Mathematical Modeling of Plant Metabolism—From Reconstruction to Prediction. *Metabolites*. 2:553-566.
- Nakabayashi, R., M. Kusano, M. Kobayashi, T. Tohge, K. Yonekura-Sakakibara, N. Kogure, M. Yamazaki, M. Kitajima, K. Saito, and H. Takayama. 2009. Metabolomics-oriented isolation and structure elucidation of 37 compounds including two anthocyanins from *Arabidopsis thaliana*. *Phytochemistry*. 70:1017-1029.
- Nakabayashi, R., and K. Saito. 2013. Metabolomics for unknown plant metabolites. *Anal Bioanal Chem*. 405:5005-5011.
- Nascimento, N.C., and A.G. Fett-Neto. 2010. Plant secondary metabolism and challenges in modifying its operation: an overview. *Methods Mol Biol*. 643:1-13.
- Neumann, S., and S. Bocker. 2010. Computational mass spectrometry for metabolomics: identification of metabolites and small molecules. *Anal Bioanal Chem*. 398:2779-2788.
- Ngo, L.T., J.I. Okogun, and W.R. Folk. 2013. 21st century natural product research and drug development and traditional medicines. *Nat Prod Rep*. 30:584-592.
- Nguyen, Q.T., M.E. Merlo, M.H. Medema, A. Jankevics, R. Breitling, and E. Takano. 2012. Metabolomics methods for the synthetic biology of secondary metabolism. *FEBS Lett*. 586:2177-2183.
- Nicholls, A.W. 2012. Realising the potential of metabolomics. *Bioanalysis*. 4:2195-2197.
- NIH. Washington University in St. Louis. <http://msr.dom.wustl.edu/tof-ms-resolution-mass-measurement-accuracy/>.
- Nishiyama, Y., H. Yamamoto, S.I. Allakhverdiev, M. Inaba, A. Yokota, and N. Murata. 2001. Oxidative stress inhibits the repair of photodamage to the photosynthetic machinery. *Embo J*. 20:5587-5594.
- Niwa, T. 1986. Metabolic profiling with gas chromatography-mass spectrometry and its application to clinical medicine. *J Chromatogr*. 379:313-345.
- Obata, T., and A.R. Fernie. 2012. The use of metabolomics to dissect plant responses to abiotic stresses. *Cell Mol Life Sci*. 69:3225-3243.
- Ogata, H., S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 27:29-34.
- Okada, T., F.M. Afendi, M. Altaf-Ul-Amin, H. Takahashi, K. Nakamura, and S. Kanaya. 2010. Metabolomics of medicinal plants: the importance of multivariate analysis of analytical chemistry data. *Curr Comput Aided Drug Des*. 6:179-196.
- openstax. http://cnx.org/contents/925e204d-d85f-4db8-be50-3e4d60057b37@1/Principles_of_Mass_Spectrometr.
- Pero, R.W., M.W. Anderson, G.A. Doyle, C.H. Anna, F. Romagna, M. Markowitz, and C. Bryngelsson. 1990. Oxidative stress induces DNA damage and inhibits the repair of DNA lesions induced by N-acetoxy-2-acetylaminofluorene in human peripheral mononuclear leukocytes. *Cancer Res*. 50:4619-4625.

- Pitzschke, A., C. Forzani, and H. Hirt. 2006. Reactive oxygen species signaling in plants. *Antioxid Redox Signal.* 8:1757-1764.
- Posma, J.M., S.L. Robinette, E. Holmes, and J.K. Nicholson. 2013. MetaboNetworks, an interactive Matlab-based toolbox for creating, customizing and exploring sub-networks from KEGG. *Bioinformatics.*
- Quenzer, T. 2002. Automated accurate mass analysis using FTICR mass spectrometry, Proceedings of the 50th Annual Conference on Mass Spectrometry and Allied Topics, Orlando, FL.
- Reemtsma, T. 2009. Determination of molecular formulas of natural organic matter molecules by (ultra-) high-resolution mass spectrometry: status and needs. *J Chromatogr A.* 1216:3687-3701.
- Reynolds, T. 2007. The evolution of chemosystematics. *Phytochemistry.* 68:2887-2895.
- Rhee, K. 2013. Minding the gaps: metabolomics mends functional genomics. *EMBO Rep.* 14:949-950.
- Rigobello-Masini, M., J.C. Penteado, and J.C. Masini. 2013. Monolithic columns in plant proteomics and metabolomics. *Anal Bioanal Chem.* 405:2107-2122.
- Rizhsky, L., H. Liang, J. Shuman, V. Shulaev, S. Davletova, and R. Mittler. 2004. When defense pathways collide. The response of Arabidopsis to a combination of drought and heat stress. *Plant Physiol.* 134:1683-1696.
- Robertson, D.G., and U. Frevert. 2013. Metabolomics in drug discovery and development. *Clin Pharmacol Ther.* 94:559-561.
- Rodriguez-Aller, M., R. Gurny, J.L. Veuthey, and D. Guillarme. 2013. Coupling ultra high-pressure liquid chromatography with mass spectrometry: constraints and possible applications. *J Chromatogr A.* 1292:2-18.
- Rogers, S., R.A. Scheltema, M. Girolami, and R. Breitling. 2009. Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics.* 25:512-518.
- Rohrschneider, I. 1966. [A method of characterization of the liquids used for separation in gas chromatography]. *J Chromatogr.* 22:6-22.
- Rohrschneider, L. 1969. [A method for the characterization of stationary liquids in gas chromatography. II. Calculation of retention ratios]. *J Chromatogr.* 39:383-397.
- Sadygov, R.G., D. Cociorva, and J.R. Yates, 3rd. 2004. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat Methods.* 1:195-202.
- Saito, K., and F. Matsuda. 2010. Metabolomics for functional genomics, systems biology, and biotechnology. *Annu Rev Plant Biol.* 61:463-489.
- Saito, N., F. Tatsuzawa, A. Nishiyama, M. Yokoi, A. Shigihara, and T. Honda. 1995. Acylated cyanidin 3-sambubioside-5-glucosides in *Matthiola incana*. *Phytochemistry.* 38:1027-1032.
- Sarker, S.D., and L. Nahar. 2012. Hyphenated techniques and their applications in natural products analysis. *Methods Mol Biol.* 864:301-340.
- Savitski, M.M., M.L. Nielsen, F. Kjeldsen, and R.A. Zubarev. 2005. Proteomics-grade de novo sequencing approach. *J Proteome Res.* 4:2348-2354.
- Schauer, N., and A.R. Fernie. 2006. Plant metabolomics: towards biological function and mechanism. *Trends Plant Sci.* 11:508-516.
- Schauer, N., D. Steinauer, S. Strelkov, D. Schomburg, G. Allison, T. Moritz, K. Lundgren, U. Roessner-Tunali, M.G. Forbes, L. Willmitzer, A.R. Fernie, and J. Kopka. 2005. GC-MS libraries for the rapid identification of metabolites in complex biological samples. *FEBS Lett.* 579:1332-1337.
- Scherling, C., C. Roscher, P. Giavalisco, E.D. Schulze, and W. Weckwerth. 2010. Metabolomics unravel contrasting effects of biodiversity on the performance of individual plant species. *PLoS One.* 5:e12569.
- Schuhmacher, R., R. Krska, W. Weckwerth, and R. Goodacre. 2013. Metabolomics and metabolite profiling. *Anal Bioanal Chem.* 405:5003-5004.
- Schuler, G.D., M.S. Boguski, E.A. Stewart, L.D. Stein, G. Gyapay, K. Rice, R.E. White, P. Rodriguez-Tome, A. Aggarwal, E. Bajorek, S. Bentolila, B.B. Birren, A. Butler, A.B. Castle, N. Chiannikulchai, A. Chu, C. Clee, S. Cowles, P.J. Day, T. Dibling, N. Drouot, I. Dunham, S. Duprat, C. East, C. Edwards, J.B. Fan, N. Fang, C. Fizames, C. Garrett, L. Green, D. Hadley, M.

- Harris, P. Harrison, S. Brady, A. Hicks, E. Holloway, L. Hui, S. Hussain, C. Louis-Dit-Sully, J. Ma, A. MacGilvery, C. Mader, A. Maratukulam, T.C. Matise, K.B. McKusick, J. Morissette, A. Mungall, D. Muselet, H.C. Nusbaum, D.C. Page, A. Peck, S. Perkins, M. Piercy, F. Qin, J. Quackenbush, S. Ranby, T. Reif, S. Rozen, C. Sanders, X. She, J. Silva, D.K. Slonim, C. Soderlund, W.L. Sun, P. Tabar, T. Thangarajah, N. Vega-Czarny, D. Vollrath, S. Voyticky, T. Wilmer, X. Wu, M.D. Adams, C. Auffray, N.A. Walter, R. Brandon, A. Dehejia, P.N. Goodfellow, R. Houlgatte, J.R. Hudson, Jr., S.E. Ide, K.R. Iorio, W.Y. Lee, N. Seki, T. Nagase, K. Ishikawa, N. Nomura, C. Phillips, M.H. Polymeropoulos, M. Sandusky, K. Schmitt, R. Berry, K. Swanson, R. Torres, J.C. Venter, J.M. Sikela, J.S. Beckmann, J. Weissenbach, R.M. Myers, D.R. Cox, M.R. James, et al. 1996. A gene map of the human genome. *Science*. 274:540-546.
- Sheldon, M.T., R. Mistrik, and T.R. Croley. 2009. Determination of ion structures in structurally related compounds using precursor ion fingerprinting. *J Am Soc Mass Spectrom*. 20:370-376.
- Shen, Y., N. Tolic, F. Xie, R. Zhao, S.O. Purvine, A.A. Schepmoes, R.J. Moore, G.A. Anderson, and R.D. Smith. 2011. Effectiveness of CID, HCD, and ETD with FT MS/MS for degradomic-peptidomic analysis: comparison of peptide identification methods. *J Proteome Res*. 10:3929-3943.
- Shulaev, V., D. Cortes, G. Miller, and R. Mittler. 2008. Metabolomics for plant stress response. *Physiol Plant*. 132:199-208.
- Shuman, J.L., D.F. Cortes, J.M. Armenta, R.M. Pokrzywa, P. Mendes, and V. Shulaev. 2011. Plant metabolomics by GC-MS and differential analysis. *Methods Mol Biol*. 678:229-246.
- Skirycz, A., S. De Bodt, T. Obata, I. De Clercq, H. Claeys, R. De Rycke, M. Andriankaja, O. Van Aken, F. Van Breusegem, A.R. Fernie, and D. Inze. 2010. Developmental stage specificity and the role of mitochondrial metabolism in the response of Arabidopsis leaves to prolonged mild osmotic stress. *Plant Physiol*. 152:226-244.
- Spiering, M.J., B. Kaur, J.F. Parsons, and E. Eisenstein. 2014. Systems approaches to unraveling plant metabolism: identifying biosynthetic genes of secondary metabolic pathways. *Methods Mol Biol*. 1083:253-273.
- Sriyudthsak, K., F. Shiraishi, and M.Y. Hirai. 2013. Identification of a metabolic reaction network from time-series data of metabolite concentrations. *PLoS One*. 8:e51212.
- Steinhauser, D., B. Usadel, A. Luedemann, O. Thimm, and J. Kopka. 2004. CSB.DB: a comprehensive systems-biology database. *Bioinformatics*. 20:3647-3651.
- Stoll, N., E. Schmidt, and K. Thürow. 2006. Isotope pattern evaluation for the reduction of elemental compositions assigned to high-resolution mass spectral data from electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *J Am Soc Mass Spectrom*. 17:1692-1699.
- Sumner, L.W., A. Amberg, D. Barrett, M.H. Beale, R. Beger, C.A. Daykin, T.W. Fan, O. Fiehn, R. Goodacre, J.L. Griffin, T. Hankemeier, N. Hardy, J. Harnly, R. Higashi, J. Kopka, A.N. Lane, J.C. Lindon, P. Marriott, A.W. Nicholls, M.D. Reilly, J.J. Thaden, and M.R. Viant. 2007. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics*. 3:211-221.
- Sun, X., and W. Weckwerth. 2012. COVAIN: a toolbox for uni- and multivariate statistics, time-series and correlation network analysis and inverse estimation of the differential Jacobian from metabolomics covariance data. *Metabolomics*. 8:81-93.
- Sweetlove, L.J., and R.G. Ratcliffe. 2011. Flux-balance modeling of plant metabolism. *Front Plant Sci*. 2:38.
- Thompson, J.A., S.P. Markey, and P.V. Fennessey. 1975. Gas-chromatographic/mass-spectrometric identification and quantitation of tetrone and deoxytetrone acids in urine from normal adults and neonates. *Clin Chem*. 21:1892-1898.
- Tohge, T., Y. Nishiyama, M.Y. Hirai, M. Yano, J. Nakajima, M. Awazuhara, E. Inoue, H. Takahashi, D.B. Goodenowe, M. Kitayama, M. Noji, M. Yamazaki, and K. Saito. 2005. Functional genomics by integrated analysis of metabolome and transcriptome of Arabidopsis plants over-expressing an MYB transcription factor. *Plant J*. 42:218-235.
- Tolstikov, V.V., O. Fiehn, and N. Tanaka. 2007. Application of liquid chromatography-mass spectrometry analysis in metabolomics: reversed-phase monolithic capillary chromatography

- and hydrophilic chromatography coupled to electrospray ionization-mass spectrometry. *Methods Mol Biol.* 358:141-155.
- Tolstikov, V.V., A. Lommen, K. Nakanishi, N. Tanaka, and O. Fiehn. 2003. Monolithic silica-based capillary reversed-phase liquid chromatography/electrospray mass spectrometry for plant metabolomics. *Anal Chem.* 75:6737-6740.
- Trifonova, O., P. Lokhov, and A. Archakov. 2013. Postgenomics diagnostics: metabolomics approaches to human blood profiling. *Omics.* 17:550-559.
- Tsimogiannis, D., M. Samiotaki, G. Panayotou, and V. Oreopoulou. 2007. Characterization of flavonoid subgroups and hydroxy substitution by HPLC-MS/MS. *Molecules.* 12:593-606.
- Urano, K., K. Maruyama, Y. Ogata, Y. Morishita, M. Takeda, N. Sakurai, H. Suzuki, K. Saito, D. Shibata, M. Kobayashi, K. Yamaguchi-Shinozaki, and K. Shinozaki. 2009. Characterization of the ABA-regulated global responses to dehydration in Arabidopsis by metabolomics. *Plant J.* 57:1065-1078.
- van Ravenzwaay, B., G.C. Cunha, E. Leibold, R. Looser, W. Mellert, A. Prokoudine, T. Walk, and J. Wiemer. 2007. The use of metabolomics for the discovery of new biomarkers of effect. *Toxicol Lett.* 172:21-28.
- Vandendool, H., and P.D. Kratz. 1963. A Generalization of the Retention Index System Including Linear Temperature Programmed Gas-Liquid Partition Chromatography. *J Chromatogr.* 11:463-471.
- Venter, J.C., M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt, J.D. Gocayne, P. Amanatides, R.M. Ballew, D.H. Huson, J.R. Wortman, Q. Zhang, C.D. Kodira, X.H. Zheng, L. Chen, M. Skupski, G. Subramanian, P.D. Thomas, J. Zhang, G.L. Gabor Miklos, C. Nelson, S. Broder, A.G. Clark, J. Nadeau, V.A. McKusick, N. Zinder, A.J. Levine, R.J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A.E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T.J. Heiman, M.E. Higgins, R.R. Ji, Z. Ke, K.A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G.V. Merkulov, N. Milshina, H.M. Moore, A.K. Naik, V.A. Narayan, B. Neelam, D. Nusskern, D.B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, et al. 2001. The sequence of the human genome. *Science.* 291:1304-1351.
- Venter, J.C., H.O. Smith, and L. Hood. 1996. A new strategy for genome sequencing. *Nature.* 381:364-366.
- von Roepenack-Lahaye, E., T. Degenkolb, M. Zerjeski, M. Franz, U. Roth, L. Wessjohann, J. Schmidt, D. Scheel, and S. Clemens. 2004. Profiling of Arabidopsis secondary metabolites by capillary liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry. *Plant Physiol.* 134:548-559.
- Wani, M.C., H.L. Taylor, M.E. Wall, P. Coggon, and A.T. McPhail. 1971. Plant antitumor agents. VI. The isolation and structure of taxol, a novel antileukemic and antitumor agent from *Taxus brevifolia*. *J Am Chem Soc.* 93:2325-2327.
- Want, E., and P. Masson. 2011. Processing and analysis of GC/LC-MS-based metabolomics data. *Methods Mol Biol.* 708:277-298.
- Weckwerth, W. 2003. Metabolomics in systems biology. *Annu Rev Plant Biol.* 54:669-689.
- Weckwerth, W. 2008. Integration of metabolomics and proteomics in molecular plant physiology--coping with the complexity by data-dimensionality reduction. *Physiol Plant.* 132:176-189.
- Weckwerth, W. 2011a. Green systems biology - From single genomes, proteomes and metabolomes to ecosystems research and biotechnology. *J Proteomics.* 75:284-305.
- Weckwerth, W. 2011b. Unpredictability of metabolism--the key role of metabolomics science in combination with next-generation genome sequencing. *Anal Bioanal Chem.* 400:1967-1978.
- Weckwerth, W., and K. Morgenthal. 2005. Metabolomics: from pattern recognition to biological interpretation. *Drug Discov Today.* 10:1551-1558.

- Wienkoop, S., K. Morgenthal, F. Wolschin, M. Scholz, J. Selbig, and W. Weckwerth. 2008. Integration of metabolomic and proteomic phenotypes: analysis of data covariance dissects starch and RFO metabolism from low and high temperature compensation response in *Arabidopsis thaliana*. *Mol Cell Proteomics*. 7:1725-1736.
- Wilm, M. 2011. Principles of electrospray ionization. *Mol Cell Proteomics*. 10:M111 009407.
- Wilson, V.S., N. Keshava, S. Hester, D. Segal, W. Chiu, C.M. Thompson, and S.Y. Euling. 2013. Utilizing toxicogenomic data to understand chemical mechanism of action in risk assessment. *Toxicol Appl Pharmacol*. 271:299-308.
- Wu, Z., R.P. Rodgers, and A.G. Marshall. 2004. Two- and three-dimensional van krevelen diagrams: a graphical analysis complementary to the kendrick mass plot for sorting elemental compositions of complex organic mixtures based on ultrahigh-resolution broadband fourier transform ion cyclotron resonance mass measurements. *Anal Chem*. 76:2511-2516.
- Wulff-Zottele, C., N. Gatzke, J. Kopka, A. Orellana, R. Hoefgen, J. Fisahn, and H. Hesse. 2010. Photosynthesis and metabolism interact during acclimation of *Arabidopsis thaliana* to high irradiance and sulphur depletion. *Plant Cell Environ*. 33:1974-1988.
- Yonekura-Sakakibara, K., T. Tohge, R. Niida, and K. Saito. 2007. Identification of a flavonol 7-O-rhamnosyltransferase gene determining flavonoid pattern in *Arabidopsis* by transcriptome coexpression analysis and reverse genetics. *J Biol Chem*. 282:14932-14941.
- Zhang, X.A., X.S. Cheng, C.Q. Wang, Z.C. Xue, L.W. Yang, and Z. Xi. 2007. Studies on interactions between plant secondary metabolites and glutathione transferase using fluorescence quenching method. *Indian J Biochem Biophys*. 44:101-105.
- Zhang, Y., F. Li, F. Huang, G. Xie, R. Wei, T. Chen, J. Liu, A. Zhao, and W. Jia. 2014. Metabolomics Analysis Reveals Variation in *Schisandra Chinensis* Metabolites from Different Origins. *J Sep Sci*.
- Zhu, H., X. Ren, J. Wang, Z. Song, M. Shi, J. Qiao, X. Tian, J. Liu, L. Chen, and W. Zhang. 2013. Integrated OMICS guided engineering of biofuel butanol-tolerance in photosynthetic *Synechocystis* sp. PCC 6803. *Biotechnol Biofuels*. 6:106.

Ich habe mich bemüht, sämtliche Inhaber der Bildrechte ausfindig zu machen und ihre Zustimmung zur Verwendung der Bilder in dieser Arbeit eingeholt. Sollte dennoch eine Urheberrechtsverletzung bekannt werden, ersuche ich um Meldung bei mir.

II. Publications

1. **Doerfler, H., et al.**, *Granger causality in integrated GC-MS and LC-MS metabolomics data reveals the interface of primary and secondary metabolism*. Metabolomics : Official journal of the Metabolomic Society, 2013. **9**(3): p. 564-574.
2. **Doerfler, H., et al.**, *mzGroupAnalyzer-Predicting Pathways and Novel Chemical Structures from Untargeted High-Throughput Metabolomics Data*. PloS one, 2014. **9**(5): p. e96188.
3. *Targeted MAPA for quantification of proteotypic peptides in complex shotgun proteomics samples – a case study on mature pollen of Tomato under heat stress*, Palak Chaturvedi, **Hannes Doerfler**, M^a Angeles Castillejo, Stefanie Wienkoop, Volker Egelhofer and Wolfram Weckwerth (publishable full-text)

Granger causality in integrated GC–MS and LC–MS metabolomics data reveals the interface of primary and secondary metabolism

Hannes Doerfler · David Lyon · Thomas Nägele ·
Xiaoliang Sun · Lena Fragner · Franz Hadacek ·
Volker Egelhofer · Wolfram Weckwerth

Received: 9 August 2012 / Accepted: 28 September 2012
© The Author(s) 2012. This article is published with open access at Springerlink.com

Abstract *Metabolomics* has emerged as a key technique of modern life sciences in recent years. Two major techniques for metabolomics in the last 10 years are gas chromatography coupled to mass spectrometry (GC–MS) and liquid chromatography coupled to mass spectrometry (LC–MS). Each platform has a specific performance detecting subsets of metabolites. GC–MS in combination with derivatisation has a preference for small polar metabolites covering primary metabolism. In contrast, reversed phase LC–MS covers large hydrophobic metabolites predominant in secondary metabolism. Here, we present an integrative metabolomics platform providing a mean to reveal the interaction of primary and secondary metabolism in plants and other organisms. The strategy combines GC–MS and LC–MS analysis of the same sample, a novel alignment tool MetMAX and a statistical toolbox COVAIN for data integration and linkage of Granger Causality with metabolic modelling. For metabolic modelling we have implemented the combined GC–LC–MS metabolomics data covariance matrix and a

stoichiometric matrix of the underlying biochemical reaction network. The changes in biochemical regulation are expressed as differential Jacobian matrices. Applying the Granger causality, a subset of secondary metabolites was detected with significant correlations to primary metabolites such as sugars and amino acids. These metabolic subsets were compiled into a stoichiometric matrix N . Using N the inverse calculation of a differential Jacobian J from metabolomics data was possible. Key points of regulation at the interface of primary and secondary metabolism were identified.

Keywords Plant systems biology · Metabolomics · Cold acclimation · Granger causality · Mass spectrometry · Differential Jacobian

1 Introduction

The interaction of primary and secondary metabolism in plants and other organisms is probably one of the most active regulatory circuits balancing biotic and abiotic environmental pressures to the system. Secondary metabolites therefore serve as important functional units to cope with these stresses and at the same time provide the richest resource of natural products in medicine and nutrition. Besides their obvious interconnectivity, in most metabolomics studies either primary or secondary metabolites are analysed to reveal the metabolic response of the system to a specific perturbation. However, by analysing complex reprogramming of metabolism in response to environmental changes it becomes clear that a comprehensive interpretation is hardly possible without integration of the data as recently shown by combining GC–MS and LC–MS metabolomics data in a long-term biodiversity experiment

Hannes Doerfler, David Lyon, Thomas Nägele and Xiaoliang Sun contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s11306-012-0470-0) contains supplementary material, which is available to authorized users.

H. Doerfler · D. Lyon · T. Nägele · X. Sun · L. Fragner ·
V. Egelhofer · W. Weckwerth (✉)
Department of Molecular Systems Biology, University of
Vienna, Althanstrasse 14, 1090 Vienna, Austria
e-mail: wolfram.weckwerth@univie.ac.at

F. Hadacek
Department of Chemical Ecology and Ecosystem Research,
University of Vienna, Vienna, Austria

(Scherling et al. 2010). In another study metabolic cross-talk during the final ripening process in melon fruit (*Cucumis melo*) was revealed by the identification of large metabolic association networks and global patterns of coordinated compositional changes of primary and secondary metabolism (Moing et al. 2011). However, due to the complexity of interactions between various pathways it is hardly possible to unambiguously trace back changes in metabolism to regulatory cues. The study of such complex interactions is focused by the research field of systems biology attempting to resolve the relationship between individual entities, for example molecules or genes, in a complex system in order to understand the resulting system behaviour. Numerous experimental and mathematical approaches to comprehensively analyse plant metabolic networks have been proposed relying on iterative processes of model development, model simulation and experimental validation (Giersch 2000; Morgan and Rhodes 2002; Rios-Estepa and Lange 2007; Nägele et al. 2010). In addition to approaches of mathematical modelling, systems biology also comprises multidimensional data analysis focusing on interpretation of the results of experiments on transcriptomics, proteomics and metabolomics (Weckwerth 2011a). Recently, we developed a toolbox, called COVAIN, which provides statistical methods allowing for the comprehensive analysis of high-dimensional metabolomics data (Sun and Weckwerth 2012). The Granger causality analysis, which is amongst other methods also implemented in COVAIN, is a time-series correlation analysis, which allows for the identification of variables being controlled by time-lagged values of other variables. This method originates from the investigation of causal relations within econometric models (Granger 1969), and recently it was also applied in a study of yeast metabolism (Walther et al. 2010). Granger causality analysis considers the time-series of variable X and Y, which can be expressed as follows (Eq. 1):

$$\begin{aligned} X(t) &= \sum_{i=1}^d C_{X,i} X(t-i) + \sum_{i=1}^d C_{XY,i} Y(t-i) + R_X(t) \\ Y(t) &= \sum_{i=1}^d C_{YX,i} X(t-i) + \sum_{i=1}^d C_{Y,i} Y(t-i) + R_Y(t) \end{aligned} \quad (1)$$

$C_{X,i}$ is the regression coefficient between $X(t)$ and $X(t-i)$, and $C_{XY,i}$ is the regression coefficient between $X(t)$ and $Y(t-i)$. $X(t)$ and $Y(t)$ represent the conditions at time point t , R is the residual error, and d is the maximal time lag between the variables. An association between X and Y is assumed to exist if the p value of the F test on the cross-coefficients is less than 0.01 (Sun and Weckwerth 2012). Hence, Granger causality between variables may be identified if a time series of variables is available which shows a dynamical behaviour and allows for the robust estimation of regression coefficients. Besides this pair-wise analysis of

variables, Granger causality is also applicable to more than two variables using a Granger model of the n -th order (Granger 1969).

Each single point of a time series at which variables are determined describes a quasi steady state of the considered system such as the metabolite contents describe the metabolism of a plant leaf cell at a certain time point. A so-called Jacobian matrix characterizes the local dynamics around such a steady state. In this context, the dynamic representation of a metabolic pathway can be described by a system of differential equations where changes of metabolite concentrations over time are expressed as functions of all metabolite concentrations considered within the system. The corresponding Jacobian is the matrix of all first-order partial derivatives of all functions on all metabolites. Hence, the Jacobian describes the influence of the change of each metabolite upon the changes of other metabolites.

Applying an approach that links the Jacobian with the covariance of the involved metabolite concentrations (Steuer et al. 2003; Weckwerth 2011b, 2003), statistical features of the data are being connected to dynamical properties of the system (Eq. 2):

$$JC + CJ^T = -2D \quad (2)$$

Here, C is the covariance matrix of metabolites, J is the Jacobian and D represents a fluctuation matrix taking into account the apparent stochasticity of the data. If the stoichiometric matrix N of the underlying metabolic system is exploited this equation can be used for inverse calculation of the Jacobian from metabolomics covariance data (Weckwerth 2011b). As it was described previously (Sun and Weckwerth 2012), the solution of J cannot be obtained directly due to under-determined equations. To circumvent this problem, reversibility and irreversibility of the reactions within a metabolic network are integrated in the “directed stoichiometric matrix” and non-zero entries of J can be calculated (Sun and Weckwerth 2012). In cases when J contains less non-zero entries than C , an over-determined problem exists, which can be solved, e.g. by minimizing total least squares.

To reveal perturbation sites between two different metabolic states we recently introduced the differential Jacobian matrix (Sun and Weckwerth 2012). The differential Jacobian matrix, dJ_{ij} , is defined by the relative change between the Jacobian matrices a and b , representing two metabolic states (Eq. 3):

$$dJ_{ij} = \log_2 \left(\text{abs} \left(\frac{J_{a,ij}}{J_{b,ij}} \right) \right) \quad (3)$$

The entries of the differential Jacobian describe the relative changes between Jacobian a and b for every element ij .

Summarizing both methods of Granger causality analysis and the differential Jacobian, it becomes obvious that neither statistical correlation analysis nor mathematical modelling of metabolic networks is capable of providing a comprehensive functional interpretation on their own. This is due to the fact that knowledge of metabolite interaction is needed for model development while unknown interactions can only be estimated by statistical methods like correlation analysis. On the other hand, statistical correlation analysis does not provide adequate tools for far-reaching analysis of metabolite interaction as they are represented by enzymatic interconversions. To overcome this limitation, we developed an approach for integrated analysis of primary and secondary metabolism in *Arabidopsis thaliana* during exposure to low temperature from the same sample by combined use of GC–MS and LC–MS techniques. Merging methods of correlation analysis and mathematical modelling indicated key points of regulation at the interface of primary and secondary metabolism during cold exposure in *A. thaliana*. For the first time, the inverse calculation of a differential biochemical Jacobian from metabolomics data is demonstrated.

2 Materials and methods

2.1 Chemicals

Methanol (HPLC-grade), Chloroform (anhydrous, >99 %, p.a.), Acetonitrile (UHPLC-grade) and Pyridine (anhydrous, >99,8 %) were purchased from Sigma-Aldrich (Vienna, Austria). Formic acid (98–100 %) was purchased from Merck (Vienna, Austria). *N*-methyl-*N*-(trimethylsilyl)trifluoroacetamide (95–100 %) was purchased from Macherey–Nagel (Düren, Germany). Chloramphenicol (>98 %) and Ampicillin trihydrate (analytical standard) were purchased from Fluka (Vienna, Austria). $^{13}\text{C}_6$ -Sorbitol (99 %) was purchased by Campro Scientific (Berlin, Germany).

2.2 Plant material and harvest

Arabidopsis thaliana plants Col-0 (wild type) were cultivated in a growth chamber under controlled conditions. The substrate for plant growth was composed of Einheitserde® ED63 and perlite. Plants were fertilized once with NPK fertilization solution (WUXAL®Super; MANNA°-Dünger, Ammerbuch, Germany). Light intensity was $250 \mu\text{mol m}^{-2} \text{s}^{-1}$ for 8 h followed by 16 h darkness, relative humidity was 60 % with a temperature of 22 °C. Of 120 *A. thaliana* specimen, 12 plants were harvested in a non cold acclimated state directly from the growth chamber; the remaining plants were put to 4 °C with the same light intensity and humidity applied as described above.

Every 48 h, 2 h after the onset of the light period the plants were harvested randomly resulting in a total number of ten time points including time point “0” of the non cold acclimated state. Leaves were sampled in three biological replicates, representing pools of four plants each. Immediately after cutting leaves from the plants, they were put in aluminium bags and quenched in liquid nitrogen. Plant material was ground to a fine powder using mortar and pestle with liquid nitrogen. Sample material was stored at –80 °C between all steps until extraction.

2.3 Extraction procedure and sample preparation for primary and secondary metabolite analysis

For GC–MS analysis a protocol according to Weckwerth et al. was used (Weckwerth et al. 2004). Deep frozen plant material was ground to a fine powder using a mortar and pestle under constant adding of liquid nitrogen. About 45 mg of each replicate was transferred to pre-cooled reaction tubes. For the extraction process, 1 ml of ice cold extraction mixture (methanol:chloroform:water, 5:2:1, v:v:v) was subsequently added. Additionally, 10 μl of internal $^{13}\text{C}_6$ -Sorbitol standard were added into each tube. Tubes were vortexed for several seconds and incubated on ice for 8 min to achieve a good extraction. Hereupon, the samples were centrifuged for 4 min at $14,000\times g$, separating the soluble compounds from remaining cell structure components. For phase separation, the supernatant was then carried over into a new tube containing 500 μl deionized water and 200 μl chloroform. After 2 min of centrifugation at $14,000\times g$, the water/methanol phase, containing the polar metabolites, was separated from the subjacent chloroform phase and completely dried out overnight.

Samples were derivatised by dissolving the dried pellet in 20 μl of a 40 mg methoxyamine hydrochloride per 1 ml pyridine solution and incubation on a thermoshaker at 30 °C for 90 min. After adding of 80 μL of *N*-methyl-*N*-(trimethylsilyl)trifluoroacetamid (MSTFA), the mixture was again incubated at 37 °C for 30 min with strong shaking.

A solution of even-numbered alkanes (Decane C10, Dodecane C12, Tetradecane C14, Hexadecane C16, Octadecane C18, Eicosane C20, Docosane C22, Tetracosane C24, Hexacosane C26, Octacosane C28, Triacontane C30, Dotriacontane C32, Tetratriacontane C34, Hexatriacontane C36, Octatriacontane C38, Tetracontane C40) was spiked into the derivatized sample before GC–MS analysis in order to infer the retention time and create the retention index.

For LC–MS analysis, frozen plant leaf material was ground as for GC–MS sample preparation, followed by addition of 1 ml pre-chilled 80/20 v:v MeOH/H₂O extraction solution containing each 1 μg of the internal standards Ampicillin and Chloramphenicol per 50 mg of fresh weight. Samples were hereupon centrifuged at $15,000\times g$ for 15 min

and the supernatant was placed into a new tube and completely dried out overnight. The resulting pellet was then dissolved in 100 µl of a 50/50 v:v MeOH/H₂O solution and centrifuged again for 15 min at 20,000×g. The remaining supernatant was then filtered through a STAGE tip (Empore/Disk C18, diameter 47 mm) into a vial with a micro insert tip. Before analysis lipid components were removed by adding 500 µl of chloroform, centrifugation and separation of the non-polar-phase to avoid contamination of the ESI ion transfer capillary.

2.4 GC-TOF/MS analysis

GC-MS measurements were carried out on an Agilent 6890 gas chromatograph coupled to a LECO Pegasus® 4D GCxGC-TOF mass spectrometer. Injection was performed splitless with a 4 mm inner diameter tapered liner containing deactivated glass wool at an injection temperature of 230 °C. Components were separated on an Agilent HP-5MS column (30 m length, 0.25 mm diameter, 0.25 µm film). The initial oven temperature was 70 °C hold for one minute, followed by a ramp of 9 °C per minute with 310 °C target temperature, which was held for 5 min, followed by a 20 °C jump to 330 °C, also being held for 5 min. Data acquisition rate on the mass spectrometer was 15 spectra/second with a detector voltage of 1600 V. Length of a run was 35 min, after an acquisition delay of 7.5 min, and the mass range was 40–600 *m/z*. The obtained raw data were processed by the on-board LECO Chroma-TOF® software capable of spectrum deconvolution, base line correction and automated peak searching. Compounds were manually annotated with the help of a retention index and mass spectra comparison to a spectra library (Kopka et al. 2005) and, with a minimum match factor of 850, arranged into a reference table (Supplement 1). Subsequently, chromatograms acquired under the same conditions were matched against the reference compound list. For relative quantification, peak areas from selected unique fragment ions for every identified compound were used. The obtained data matrix was directly exported from the Pegasus® software into an Excel worksheet.

2.5 LC-MS metabolomics

5 µL of sample were injected on a Waters HSST3 column (100 mm length × 100 µm I.D.) using a nano LC ultra 1D pump from Eksigent and a HTC PAL autosampler. Mobile phase A consisted of H₂O with 0.1 % formic acid (FA) and mobile phase B of 90 % acetonitrile (ACN) with 0.1 % FA. A constant flow rate of 500 nl/min was used with the following nonlinear gradient:

Time (min)	% A	% B
0	95	5
3	95	5
20	85	15
22	83	17
27	80	20
42	68	32
57	54	46
68	25	75
72	5	95
90	5	95
95	95	5
115	95	5

A nano ESI source from Thermo Scientific was used. All data were acquired in positive ionisation mode. Each Full Scan, resolution 60,000, was followed by a data dependent MS² scan, resolution 7,500, of the most abundant ion, which was subjected to Collision Induced Dissociation (CID) using a normalized collision energy of 50 %. The Orbitrap was used to acquire MS spectra, ranging from 100 to 2,000 *m/z*, as well as MS/MS spectra. Only recognized charge states were allowed to trigger MS² spectra generation. Non-peptidic precursor selection was enabled, and the dynamic exclusion list was set to 500 values, with a duration of 90 s, and a repeat count of one. Minimum signal threshold was set to 1,000 (absolute value). The temperature of the heated capillary and electrospray voltage were 180 °C and 2 kV, respectively. After MS analysis, mzXML files were created using MassMatrix MS Data File Conversion (v3.9c) from raw files and analyzed using the MetMAX algorithm which is based on PROTMAX (v2.7) (Hoehenwarter et al. 2008) with the following settings: Ion Count, Intensity, Decimals: two cut, all charge states, Environment 10 min, Unite neighbours, Intensity expected one, no retention time filter. Results from this data extraction process were arranged in an Excel file of the same order as the GC-MS obtained data.

2.6 Statistical data analysis

GC-MS as well as LC-MS obtained data were normalized to fresh weight and internal standards (¹³C₆-Sorbitol for GC-MS, Ampicillin for LC-MS). To reduce the high variation in the LC-MS data, the data set was filtered as follows: the coefficient of variation (CV) of each time point was calculated, as well as the average CV of all ten time points. All values equal or lower than 30 % were disregarded. Data matrices from both measurements were combined. Data pre-processing, principal component

analysis (PCA), Granger causality analysis as well as calculation of the differential Jacobian matrices were performed in the Matlab® Toolbox COVAIN (Sun and Weckwerth 2012). After filling of missing values and adjustment of outliers, as well as log10 and z-transformation, Granger analysis was applied. Granger parameters were set to time lag 1 with a significance p value of 0.05. The Granger causality analysis was performed pair-wise on metabolite concentrations applying linear regression.

3 Results and discussion

3.1 Cold-induced reprogramming of primary metabolism in *Arabidopsis thaliana*

Primary metabolite content of leaf tissue of *A. thaliana*, accession Columbia (Col-0), were identified by GC-TOF/MS before and after 2, 4, 6, 8, 10, 12, 14, 16 and 18 days of exposure to 4 °C. Fast, intermediate and slow steps of metabolic readjustment could be distinguished and results are explicitly summarized in Supplement 2. To shortly summarize the findings, components of primary carbohydrate metabolism displayed a fast increase due to cold exposure. Besides the increase of carbohydrate contents, the fastest response to cold exposure was the significant decrease of aromatic amino acids, phenylalanine, tyrosine and tryptophan. Significant accumulation of organic acids, like ascorbic acid, citrate or malic acid, were observed after 4 to 8 days of cold exposure. Accumulation of pyruvic acid was found to be part of late metabolic readjustment after 12 days at 4 °C. Changes in polyamine content were not unique, as putrescine accumulated within the first 2 days of cold exposure while spermidine accumulated significantly after 12 days.

3.2 Analysis of the interaction between primary and secondary metabolism

Applying the alignment tool MetMAX to metabolomics data, which was developed on the basis of the PROTMAX algorithm (Hoehenwarter et al. 2008), and using *ion count* and *intensity* as quantitative parameters in the algorithm to correctly bin the m/z ratios, about 3,000 m/z ratios were acquired by each chromatographic run (explicit description of settings are provided in [Materials and methods](#)). From this data set 349 m/z values in total were filtered and reliably identified for all time points of cold exposure by statistical analysis and calculation of the coefficient of variance (CV) (Hoehenwarter et al. 2008). In a next step this LC-MS data set was integrated with the GC-MS data set from the same samples using a function in COVAIN for data integration (Sun and Weckwerth 2012). The

interactions between primary and secondary metabolites were investigated by Granger causality analysis—another function of COVAIN. With a p value <0.05 approximately 15,000 Granger causations were determined, which were either describing time series correlation between metabolites within the GC-MS data matrix, within the LC-MS data matrix or between components of these two matrices. The results are summarized in Supplement 3. As a consequence of the experimental design, which was intended to stimulate flavonoid accumulation due to low temperature and elevated light intensity, Granger causations were predominantly detected in flavonoid biosynthesis. (Fig. 1). Putative metabolic interaction sites were identified revealing the synthesis of the molecule A17 (m/z 1151) from shikimic acid and phenylalanine (Fig. 1). Additionally, precursor molecules of A17 could be identified within the LC-MS data set allowing for the reconstruction of the synthesis pathway: the molecule $[\text{Cy} + \text{Glc} + \text{Mal}]^+$ (m/z 535) is substrate for synthesis of A8 (m/z 1137) which is subsequently methylated to molecule A17 (m/z 1151). In addition to molecule A17, most cyanidin derivatives were putatively identified to be associated to molecule A1 (m/z 743) which was previously termed cyanidin 3-O-[2''-O-(xylosyl)glucoside] 5-O-glucoside (Tohge et al. 2005). Accurate mass-to-charge-ratios of metabolites were used for calculation of sum formulas and putative identifications were confirmed by comparison with existing literature, as well as tracking the specific MS^n fragmentation patterns, which are described for flavonoids (Matsuda et al. 2009; Waridel et al. 2001) (Table 1).

3.3 Analysis of cold-induced metabolic perturbation sites—calculation of a differential Jacobian from metabolomics data

Based on the data set derived from GC-MS analysis a simplified metabolic network structure was derived comprising interconversions of primary metabolism. The prominent interaction with secondary metabolism via phenylalanine, which was identified by Granger causation, was also included in the network structure (Fig. 2). We focused on the phenylalanine-derived synthesis of putative flavonoids because this is one of the most prominent examples of interaction described in the literature (Winkel-Shirley 2002; Tzin et al. 2012).

The underlying stoichiometric matrix of this network was compiled for the inverse calculation of a differential Jacobian matrix using metabolomics covariance data according to (Sun and Weckwerth 2012). Metabolic states a and b were defined by time points of differentially cold acclimated plants (state a) and, as a reference state, of non-acclimated plants (state b). With reference to metabolite levels of non-acclimated plants, calculation was performed

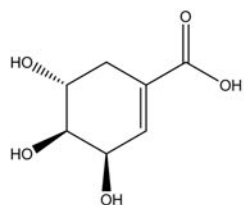
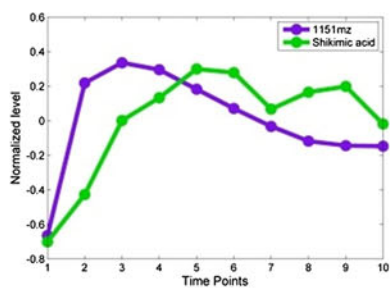
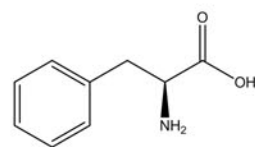
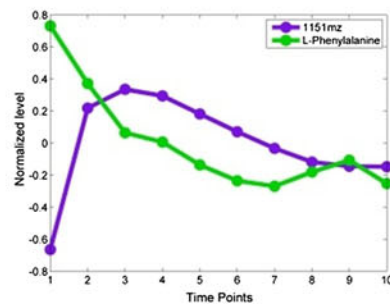
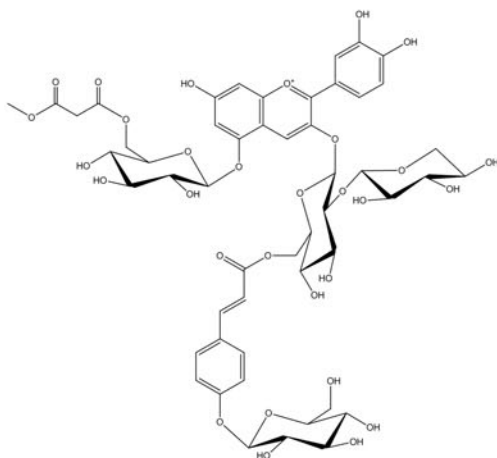
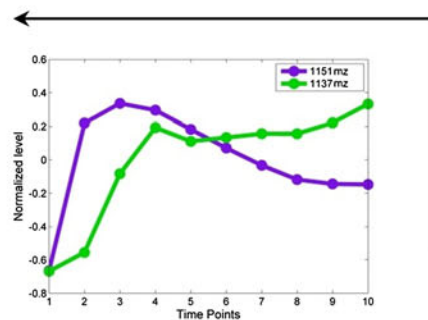
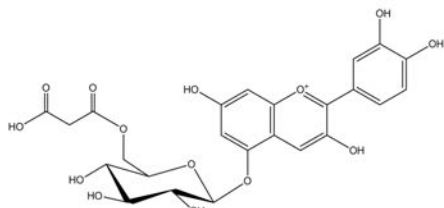
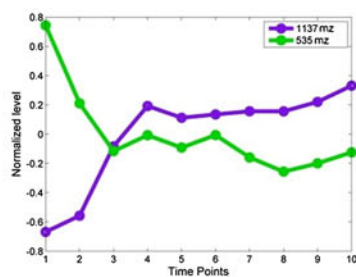
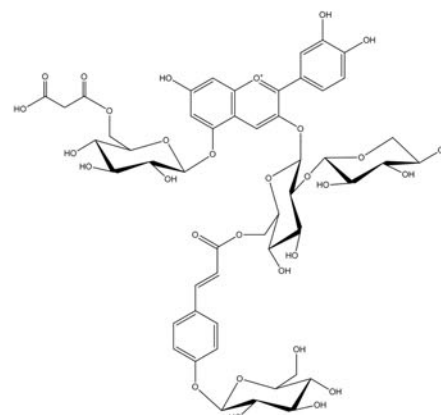
**Shikimic acid** $p = 0.00369$ **L-Phenylalanine** $p = 0.00010$ **A17**
m/z 1151 $p = 0.00115$ **[Cy+Glc+Mal]⁺**
m/z 535 $p = 0.00369$ **A8**
m/z 1137

Fig. 1 Granger causality analysis between molecules of GC–MS and LC–MS measurements from primary and secondary metabolism. Key metabolites for phenylpropanoid synthesis identified by GC–TOF/MS from *Arabidopsis* leaves have Granger correlations with compound A17 (Shi and Xie 2010) obtained by LC–MS, either by upregulation (shikimic acid) or downregulation (phenylalanine) over 18 days of cold stress. Also, *m/z* 1137, A8, is shown to be a precursor for its methyl ester form, A17, while itself being caused by *m/z* 535, Cyanidin 5-O-(6'''-O-malonyl)glucoside. Corresponding *p* values are depicted in the figure

for plants after 2 days at 4 °C (Fig. 3 a), 8 days at 4 °C (Fig. 3 b), 14 days at 4 °C (Fig. 3 c), and 18 days at 4 °C (Fig. 3 d) to reveal short-term, intermediate and long-term effects of cold exposure on metabolism. Mean values of 30 calculations were built and the ratios of mean values to standard errors of calculation are given in Supplement 4.

Resulting mean values of entries of the differential Jacobian matrices indicated a progressive perturbation of metabolism during exposure to 4 °C. Interactions of soluble sugars and pyruvic acid-derived metabolites were affected strongly after 2 days at 4 °C (Fig. 3a). After 8 days, the metabolic perturbation was only of alleviated intensity (Fig. 3b) and after 14 and 18 days it became even

smaller than before cold exposure (Fig. 3c, d). Relative changes of the flavonoid pool (*Flavonoids*), being induced by relative changes of its substrate pool phenylalanine, became maximal after 8 days of cold exposure (Fig. 3b) and were dampened until 18 days at 4 °C (Fig. 3c, d).

3.4 Integration of GC–MS and LC–MS data for a comprehensive understanding of plant–environment interactions

Cold-induced reprogramming of primary metabolism in *A. thaliana* is a prominent example of plant–environment interaction. Like many previous studies, our results show that levels of various metabolites are affected significantly by low temperature. In a comprehensive analysis of primary metabolism by GC–MS technique we were able to distinguish fast from slow metabolic reactions induced by cold exposure. Contents of putative cryoprotective compounds like sucrose, galactinol and putrescine showed a fast significant increase, thus proving their involvement in the immediate response to abiotic stress. However, in contrast to previous studies where asparagine content was

Table 1 Associated molecules to A1 (*m/z* 743) identified by Granger causality analysis

Molecule abbreviation	Mass-to-charge ratio (<i>m/z</i>)	Name of molecule	Reference	Experimental mass deviation (ppm)	CAS-no.
A2	829.2034 [M + H] ⁺	Cyanidin 3-O-[2''-O-(xylosyl)glucoside]5-O-(6'''-O-malonyl)glucoside	Tohge et al. 2005	1.0854	866259-91-8
A3	889.2307 [M + H] ⁺	Cyanidin 3-O-[2''-O-(xylosyl) 6''-O-(<i>p</i> -coumaroyl)glucoside]5-O-glucoside	Tohge et al. 2005	−0.5623	866259-92-9
A6	1051.2926 [M + H] ⁺	Cyanidin 3-O-[2''-O-(xylosyl)-6''-O-(<i>p</i> -O-(glucosyl)- <i>p</i> -coumaroyl)glucoside]5-O-glucoside	Tohge et al. 2005	0.5707	906811-94-7
A7	1095.2977 [M + H] ⁺	Cyanidin 3-O-[2''-O-(2'''-O-(sinapoyl) xylosyl)6''-O-(<i>p</i> -coumaroyl)glucoside]5-O-glucoside	Tohge et al. 2005	−0.2739	866259-94-1
A8	1137.2930 [M + H] ⁺	Cyanidin 3-O-[2''-O-(xylosyl)6''-O-(<i>p</i> -O-(glucosyl) <i>p</i> -coumaroyl)glucoside]5-O-[6'''-O-(malonyl)glucoside]	Tohge et al. 2005	−0.4396	475163-06-5
A9	1181.2981 [M + H] ⁺	Cyanidin 3-O-[2''-O-(2'''-O-(sinapoyl)xylosyl)6''-O-(<i>p</i> -O-coumaroyl)glucoside]5-O-[6'''-O-(malonyl) glucoside]	Tohge et al. 2005	−0.4233	864155-73-7
A10	1257.3530 [M + H] ⁺	Cyanidin 3-O-[2''-O-(2'''-O-(sinapoyl) xylosyl)6''-O-(<i>p</i> -O-(glucosyl) <i>p</i> -coumaroyl) glucoside]5-O-glucoside	Tohge et al. 2005	1.5906	n.a.
A11	1343.3509 [M + H] ⁺	Cyanidin 3-O-[2''-O-(6'''-O-(sinapoyl) xylosyl)6''-O-(<i>p</i> -O-(glucosyl)- <i>p</i> -coumaroyl)glucoside]5-O-(6'''-O-malonyl)glucoside	Tohge et al. 2005	−0.8188	475163-04-3
A17	1151.3086 [M + H] ⁺	Cyanidin 3-O-[2''-O-(xylosyl)6''-O-(<i>p</i> -O-(glucosyl) <i>p</i> -coumaroyl)glucoside] 5-O-[6'''-O-(methyl-malonyl)glucoside]	Shi and Xie 2010	0.7817	n.a.

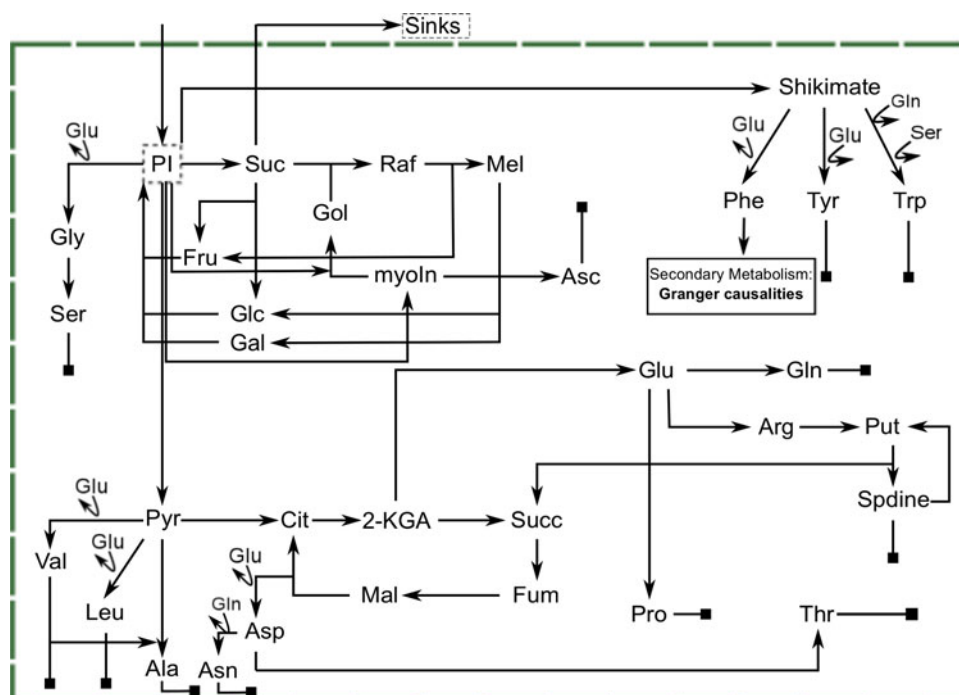


Fig. 2 Schematic representation of the primary metabolism in leaf cells of *A. thaliana*. Secondary metabolites identified by Granger causalities are exemplarily integrated derived from phenylalanine. *PI* phosphorylated intermediates, *Glu* glutamate, *Gln* glutamine, *Gly* glycine, *Ser* serine, *Suc* sucrose, *Fru* fructose, *Glc* glucose, *Gol* galactinol, *Raf* raffinose, *Mel* melibiose, *myoin* myo-Inositol, *Asc*

ascorbic acid, *Gal* galactose, *Phe* phenylalanine, *Tyr* tyrosine, *Trp* tryptophan, *Pyr* pyruvic acid, *Val* valine, *Leu* leucine, *Ala* alanine, *Cit* citric acid, *2-KGA* 2-ketoglutaric acid, *Succ* succinic acid, *Fum* fumaric acid, *Mal* malic acid, *Asp* aspartic acid, *Asn* asparagine, *Arg* arginine, *Put* putrescine, *Spd*ine spermidine, *Pro* proline, *Thr* threonine

found to increase significantly during cold exposure (Klotke et al. 2004; Usadel et al. 2008), we found that asparagine content decreases significantly within the first two days of cold exposure. This discrepancy may be explained by the different light intensities, which were used in the experiments. In the present study, we applied a light intensity of $250 \mu\text{mol m}^{-2} \text{s}^{-1}$, which was significantly higher than in studies of (Klotke et al. 2004) and (Usadel et al. 2008) to stimulate biosynthesis of secondary metabolites. Elevated light was shown to repress the transcription of asparagine synthetase genes (Tsai and Coruzzi 1991), and may therefore explain the observed decrease in asparagine content in the present study.

Besides those findings, levels of tryptophan and phenylalanine were significantly decreased during the first two days of cold exposure. These aromatic amino acids (AAAs) are central metabolic precursors for synthesis of secondary metabolites (Tzin and Galili 2010). However, in contrast to primary metabolites like sugars or amino acids, plant secondary metabolites cannot be analysed by GC–MS, but LC–MS has to be applied. Although numerous approaches already provided evidence for the usefulness of a combined GC–MS and LC–MS approach (Tzin et al. 2009, 2012), these approaches were driven by the available knowledge about certain interactions between pathways of primary and

secondary metabolism. The mass-to-charge (m/z) ratios, which represent the primary results of LC–MS analysis, are identified by comparison to available data bases or libraries. Although such an approach is very powerful because it allows for the simultaneous analysis of hundreds of metabolites, it is limited by the a priori knowledge about metabolic pathways. To overcome this limitation, we developed and applied statistical Granger causality analysis to unravel putative interactions of primary and secondary metabolism. Thus, shikimic acid as well as AAAs were correlated with a set of m/z ratios from LC–MS measurements which could afterwards be identified as members of the cyanidin family representing the predominant flavonoids in *A. thaliana* (Bloor and Abrahams 2002; Tohge et al. 2005). It is known from literature that accumulation of anthocyanins in leaves is stress-inducible, protecting against photoinhibitory damage caused by high irradiance (Havaux and Kloppstech 2001; Page et al. 2012). We also identified a correlation between ascorbic acid and anthocyanins, which was previously described by Page and co-workers who compared six *Arabidopsis* accessions under high light conditions (Page et al. 2012). The authors concluded from their experiments that the ability to accumulate anthocyanins in *Arabidopsis* is tuned by the status of ascorbic acid. Although we are not yet able to give a

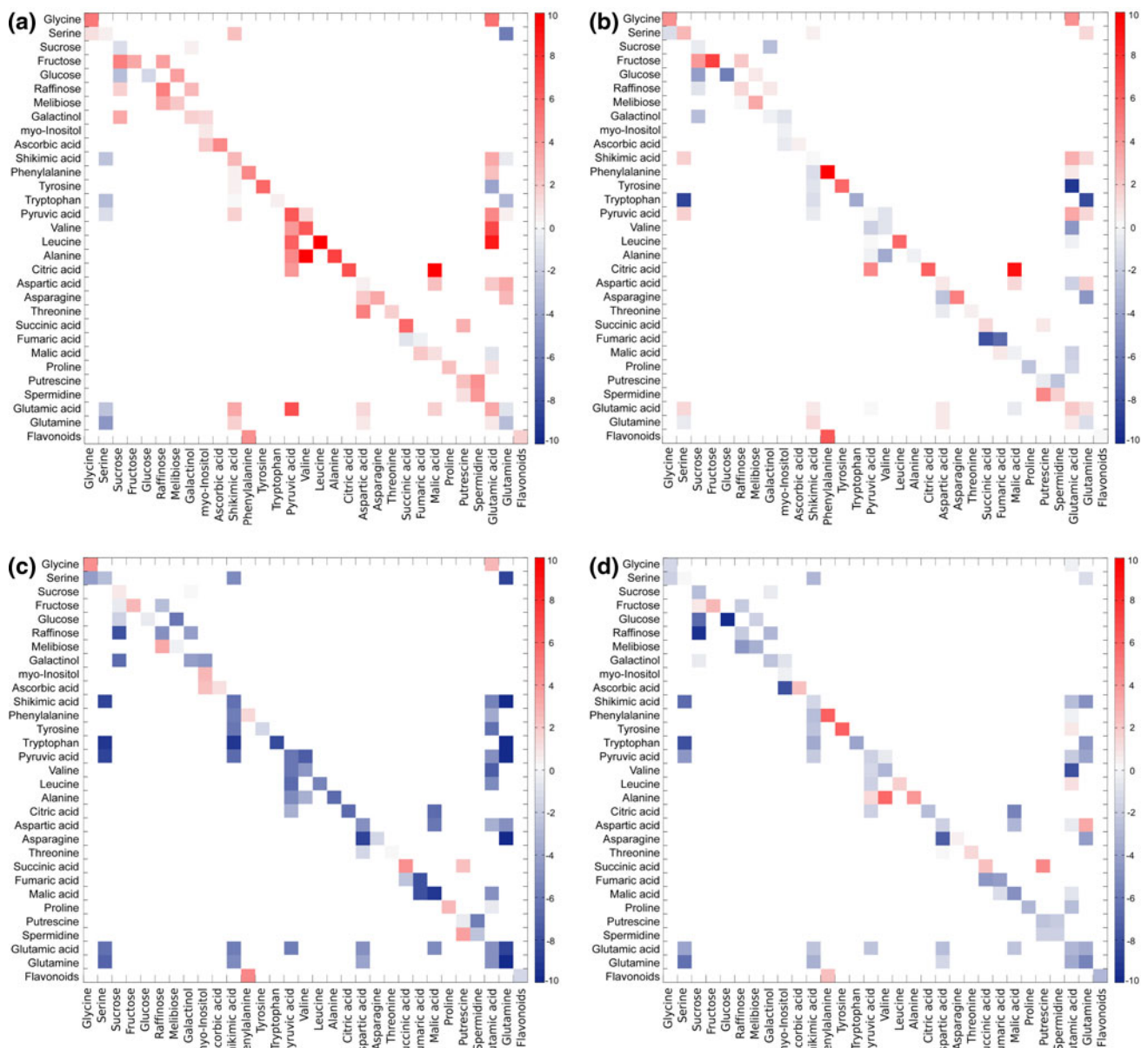


Fig. 3 Entries of the differential Jacobian after 2d(a), 8d(b), 14d(c) and 18d(d) of cold exposure relative to the non-acclimated condition are visualized by the heat map. Red colours indicate an increase of putative metabolic interaction while blue colours indicate

physiological interpretation of all the metabolic correlations we found by Granger causality analysis, we are now able to derive possible interactions and test them by further experimental investigation. We exemplified this by describing the metabolic network, which is represented by our GC–MS data, and expanded this network by the phenylalanine-derived synthesis of secondary metabolites identified by Granger causality analysis. Applying the inverse calculation of the differential Jacobian (Sun and Weckwerth 2012), the synthesis of secondary metabolites, termed as *Flavonoids*, were indicated to become maximally

a decrease relative to the non-acclimated plants. Entries of the differential Jacobian matrices represent mean values of 30 calculations

dependent on changes in phenylalanine content after 8 days of cold exposure (Fig. 3b). Here, the term *perturbation* describes the change in flavonoid content due to a relative change in phenylalanine content. Additionally, the calculation of the differential Jacobian allowed for the estimation of system behaviour after perturbation by changing environmental conditions. While entries of the differential Jacobian became positive after 2 days at 4 °C, most of them got negative after 14 days at 4 °C pointing to a change in dynamical system behaviour around the metabolic steady state. Because positive entries result from a ratio of greater

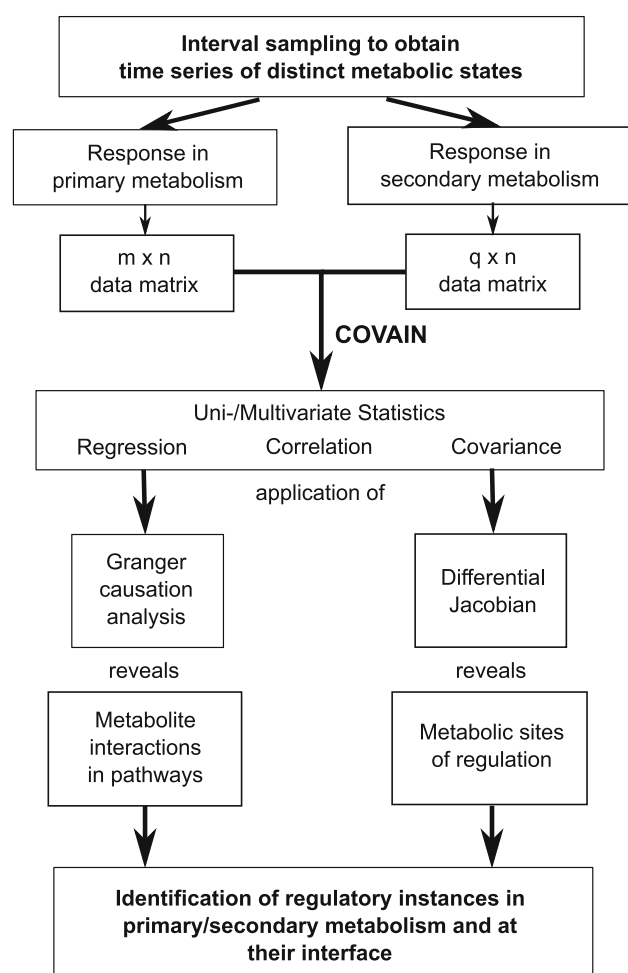


Fig. 4 Granger causality analysis and the differential Jacobian for comprehensive analysis of GC–MS and LC–MS data sets to evaluate pathway interactions and regulatory instances in metabolism

than 1 this finding might indicate an increased reactivity of primary metabolites during the first 2 days of cold exposure. However, due to thermodynamic effects on enzymatic interconversions of primary metabolism at 4 °C (Nägele et al. 2012) this putative increase might be dampened and rather represents a compensation of thermodynamic effects on metabolic homeostasis than an increase in reactivity. Although this shows clearly that we cannot explicitly characterize rates of metabolic interconversions by this covariance-based approach, we are now able to detect relative changes in metabolic homeostasis due to changing environmental conditions. Additionally, deriving the Jacobian from covariance data represents a novel and convenient method to predict biochemical changes in multidimensional data sets, which is hardly feasible by classical biochemical experiments. Localizing biochemical *hot spots* from metabolomics data provides the basis for eventually understanding the perturbation dynamics in a whole metabolic network. Granger causality analysis is

applied to reveal significant co-variance within the metabolic network and thereby used to extend its stoichiometric matrix (Fig. 4). As we have shown in the present study, this enables the interpretation of metabolic constitutions within a physiological context, which is fundamental for a comprehensive understanding of plant-environment interactions (Weckwerth 2011a; Nägele and Weckwerth, 2012).

4 Conclusions

Based on our findings of cold-induced changes in primary and secondary metabolism of *A. thaliana*, we conclude that the identification of Granger causalities offers a novel method to comprehensively analyse GC- and LC–MS data from the same sample. Particularly, interfaces of complex biochemical networks can be characterized providing new insights in pathway regulation. The direct linkage of statistical (i.e. Granger causality analysis) with mathematical methods (differential Jacobian) is demonstrated in the present study as depicted in Fig. 4. All the described features from integration of different data sets such as GC–MS and LC–MS data to statistical methods such as Granger causality analysis and metabolic modelling using an inverse calculation of the differential Jacobian are implemented in the metabolomics toolbox COVAIN (Sun and Weckwerth 2012). The calculation of the differential Jacobian from metabolomics data provides hints to pathway regulation, however, these predictions need to be tested by classical biochemical methods. We propose the presented strategy as a fundamental concept to link genome-scale metabolic reconstruction and metabolomics data (Weckwerth 2011b). The approach can be systematically used for genotype-metabo-phenotype studies.

Acknowledgments We would like to thank the gardeners for the excellent plant cultivation and the members of the Department Molecular Systems Biology for fruitful discussions. We thank the anonymous reviewers for their valuable comments. We thank the WWTF (Vienna Science and Technology Fund) for financial support of XS. We thank the EU-Marie-Curie-ITN MERIT for financial support of TN.

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

References

- Bloor, S. J., & Abrahams, S. (2002). The structure of the major anthocyanin in *Arabidopsis thaliana*. *Phytochemistry*, 59, 343–346.
- Giersch, C. (2000). Mathematical modelling of metabolism. *Current Opinion in Plant Biology*, 3, 249–253.
- Granger, C. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37, 414–426.

- Havaux, M., & Kloppstech, K. (2001). The protective functions of carotenoid and flavonoid pigments against excess visible radiation at chilling temperature investigated in *Arabidopsis npq* and *tr* mutants. *Planta*, 213, 953–966.
- Hoehenwarter, W., van Dongen, J. T., Wienkoop, S., Steinfath, M., Hummel, J., Erban, A., et al. (2008). A rapid approach for phenotype-screening and database independent detection of cSNP/protein polymorphism using mass accuracy precursor alignment. *Proteomics*, 8, 4214–4225.
- Klotke, J., Kopka, J., Gatzke, N., & Heyer, A. G. (2004). Impact of soluble sugar concentrations on the acquisition of freezing tolerance in accessions of *Arabidopsis thaliana* with contrasting cold adaptation—evidence for a role of raffinose in cold acclimation. *Plant, Cell and Environment*, 27, 1395–1404.
- Kopka, J., Schauer, N., Krueger, S., Birkemeyer, C., Usadel, B., Bergmüller, E., et al. (2005). GMD@CSB.DB: the Golm Metabolome database. *Bioinformatics*, 21(8), 1635–1638.
- Matsuda, F., Yonekura-Sakakibara, K., Niida, R., Kuromori, T., Shinozaki, K., & Saito, K. (2009). MS/MS spectral tag-based annotation of non-targeted profile of plant secondary metabolites. *The Plant Journal*, 57, 555–577.
- Moing, A., Aharoni, A., Biais, B., Rogachev, I., Meir, S., Brodsky, L., et al. (2011). Extensive metabolic cross-talk in melon fruit revealed by spatial and developmental combinatorial metabolomics. *New Phytologist*, 190, 683–696.
- Morgan, J. A., & Rhodes, D. (2002). Mathematical modeling of plant metabolic pathways. *Metabolic Engineering*, 4, 80–89.
- Nägele, T., Henkel, S., Hörmiller, I., Sauter, T., Sawodny, O., Ederer, M., et al. (2010). Mathematical modelling of the central carbohydrate metabolism in *Arabidopsis thaliana* reveals a substantial regulatory influence of vacuolar invertase on whole plant carbon metabolism. *Plant Physiology*, 153, 260–272.
- Nägele, T., Stutz, S., Hörmiller, I. I., & Heyer, A. G. (2012). Identification of a metabolic bottleneck for cold acclimation in *Arabidopsis thaliana*. *The Plant Journal*, doi:10.1111/j.1365-3113X.2012.05064.x.
- Nägele, T., & Weckwerth, W. (2012). Mathematical modeling of plant metabolism—from reconstruction to prediction. *Metabolites*, 2(3), 553–566.
- Page, M., Sultana, N., Paszkiewicz, K., Florance, H., & Smirnov, N. (2012). The influence of ascorbate on anthocyanin accumulation during high light acclimation in *Arabidopsis thaliana*: further evidence for redox control of anthocyanin synthesis. *Plant, Cell and Environment*, 35, 388–404.
- Rios-Estapa, R., & Lange, B. M. (2007). Experimental and mathematical approaches to modeling plant metabolic networks. *Phytochemistry*, 68, 2351–2374.
- Scherling, C., Roscher, C., Giavalisco, P., et al. (2010). Metabolomics unravel contrasting effects of biodiversity on the performance of individual plant species. *PLoS ONE*, 5, e12569.
- Shi, M.-Z., & Xie, D.-Y. (2010). Features of anthocyanin biosynthesis in *pap1-D* and wild-type *Arabidopsis thaliana* plants grown in different light intensity and culture media conditions. *Planta*, 231(6), 1385–1400.
- Steuer, R., Kurths, J., Fiehn, O., & Weckwerth, W. (2003). Observing and interpreting correlations in metabolomic networks. *Bioinformatics*, 19, 1019–1026.
- Sun, X., & Weckwerth, W. (2012). COVAIN: a toolbox for uni- and multivariate statistics, time-series and correlation network analysis and inverse estimation of the differential Jacobian from metabolomics covariance data. *Metabolomics*, 8, 81–93. doi:10.1007/s11306-012-0399-3.
- Tohge, T., Nishiyama, Y., Hirai, M. Y., Yano, M., Nakajima, J., Awazuhara, M., et al. (2005). Functional genomics by integrated analysis of metabolome and transcriptome of *Arabidopsis* plants over-expressing an MYB transcription factor. *The Plant Journal*, 42, 218–235.
- Tsai, F., & Coruzzi, G. (1991). Light represses transcription of asparagine synthetase genes in photosynthetic and nonphotosynthetic organs of plants. *Molecular and Cellular Biology*, 11, 4966–4972.
- Tzin, V., & Galili, G. (2010). New insights into the shikimate and aromatic amino acids biosynthesis pathways in plants. *Molecular Plant*, 3, 956–972.
- Tzin, V., Malitsky, S., Aharoni, A., & Galili, G. (2009). Expression of a bacterial bi-functional chorismate mutase/prephenate dehydratase modulates primary and secondary metabolism associated with aromatic amino acids in *Arabidopsis*. *The Plant Journal*, 60, 156–167.
- Tzin, V., Malitsky, S., Ben Zvi, M. M., Bedair, M., Sumner, L., Aharoni, A., et al. (2012). Expression of a bacterial feedback-insensitive 3-deoxy-D-arabino-heptulosonate 7-phosphate synthase of the shikimate pathway in *Arabidopsis* elucidates potential metabolic bottlenecks between primary and secondary metabolism. *New Phytologist*, 194(2), 430–439.
- Usadel, B., Blasing, O. E., Gibon, Y., Poree, F., Höhne, M., Günter, M., et al. (2008). Multilevel genomic analysis of the response of transcripts, enzyme activities and metabolites in *Arabidopsis* rosettes to a progressive decrease of temperature in the non-freezing range. *Plant, Cell and Environment*, 31, 518–547.
- Walther, D., Strassburg, K., Durek, P., & Kopka, J. (2010). Metabolic pathway relationships revealed by an integrative analysis of the transcriptional and metabolic temperature stress-response dynamics in yeast. *OMICS*, 14, 261–274.
- Waridel, P., Wolfender, J.-L., Ndjoko, K., Hobby, K. R., Major, H. J., & Hostettmann, K. (2001). Evaluation of quadrupole time-of-flight tandem mass spectrometry and ion-trap multiple-stage mass spectrometry for the differentiation of C-glycosidic flavonoid isomers. *Journal of Chromatography A*, 926, 29–41.
- Weckwerth, W. (2003). Metabolomics in systems biology. *Annual Review of Plant Biology*, 54, 669–689.
- Weckwerth, W. (2011a). Green systems biology—from single genomes, proteomes and metabolomes to ecosystems research and biotechnology. *Journal of Proteomics*, 75, 284–305.
- Weckwerth, W. (2011b). Unpredictability of metabolism—the key role of metabolomics science in combination with next-generation genome sequencing. *Analytical and Bioanalytical Chemistry*, 400, 1967–1978.
- Weckwerth, W., Wenzel, K., & Fiehn, O. (2004). Process for the integrated extraction identification, and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks. *Proteomics*, 4, 78–83.
- Winkel-Shirley, B. (2002). Biosynthesis of flavonoids and effects of stress. *Current Opinion in Plant Biology*, 5, 218–223.



mzGroupAnalyzer-Predicting Pathways and Novel Chemical Structures from Untargeted High-Throughput Metabolomics Data

Hannes Doerfler[‡], Xiaoliang Sun[‡], Lei Wang, Doris Engelmeier, David Lyon, Wolfram Weckwerth*

Department of Ecogenomics and Systems Biology, University of Vienna, Vienna, Austria

Abstract

The metabolome is a highly dynamic entity and the final readout of the genotype x environment x phenotype (GxExP) relationship of an organism. Monitoring metabolite dynamics over time thus theoretically encrypts the whole range of possible chemical and biochemical transformations of small molecules involved in metabolism. The bottleneck is, however, the sheer number of unidentified structures in these samples. This represents the next challenge for metabolomics technology and is comparable with genome sequencing 30 years ago. At the same time it is impossible to handle the amount of data involved in a metabolomics analysis manually. Algorithms are therefore imperative to allow for automated *m/z* feature extraction and subsequent structure or pathway assignment. Here we provide an automated pathway inference strategy comprising measurements of metabolome time series using LC-MS with high resolution and high mass accuracy. An algorithm was developed, called *mzGroupAnalyzer*, to automatically explore the metabolome for the detection of metabolite transformations caused by biochemical or chemical modifications. Pathways are extracted directly from the data and putative novel structures can be identified. The detected *m/z* features can be mapped on a van Krevelen diagram according to their H/C and O/C ratios for pattern recognition and to visualize oxidative processes and biochemical transformations. This method was applied to *Arabidopsis thaliana* treated simultaneously with cold and high light. Due to a protective antioxidant response the plants turn from green to purple color via the accumulation of flavonoid structures. The detection of potential biochemical pathways resulted in 15 putatively new compounds involved in the flavonoid-pathway. These compounds were further validated by product ion spectra from the same data. The *mzGroupAnalyzer* is implemented in the graphical user interface (GUI) of the metabolomics toolbox COVAIN (Sun & Weckwerth, 2012, *Metabolomics* 8: 81–93). The strategy can be extended to any biological system.

Citation: Doerfler H, Sun X, Wang L, Engelmeier D, Lyon D, et al. (2014) *mzGroupAnalyzer*-Predicting Pathways and Novel Chemical Structures from Untargeted High-Throughput Metabolomics Data. *PLoS ONE* 9(5): e96188. doi:10.1371/journal.pone.0096188

Editor: Akos Vertes, The George Washington University, United States of America

Received: October 29, 2013; **Accepted:** April 4, 2014; **Published:** May 20, 2014

Copyright: © 2014 Weckwerth et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The authors thank the Austrian Science Fund (FWF) for financial support to Hannes Dörfler and David Lyon (FWF project P25488 and P23441, respectively). Lei Wang was supported by a grant of the China Scholarship Council. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: wolfram.weckwerth@univie.ac.at

‡ These authors contributed equally to this work.

Introduction

Metabolomic techniques have been recently established and refined to characterize the widely heterogeneous small molecules present at a specific time in a living tissue. Several analytical approaches exist for the application of metabolomics to various biological questions, for instance gas chromatography coupled to mass spectrometry (GC-MS) for the detection of small and volatile compounds [1] or capillary electrophoresis coupled to mass spectrometry (CE-MS) for charged compounds [2]. The method with the highest preference for larger and more hydrophobic metabolites and complementary to GC-MS is liquid chromatography coupled to mass spectrometry (LC-MS) [3,4]. Recently, we showed that the GC-MS and LC-MS techniques can be integrated into a combined platform to increase the total coverage of the metabolome, as well as to provide insights into the mutual regulation of both primary and secondary metabolism by analysis of the same sample and subsequent data merging and processing [4–6]. Metabolomics-via-LC-MS approaches have yet to hit the

ranks of other analytical techniques with respect to their robustness and database availability, but no other method has the potential to achieve a better coverage of the metabolome whilst maintaining high resolution and inferring additional structural information in the process. In contrast to metabolite profiling on the GC-MS platform, where standard operating procedures and large databases are present and being improved continuously, there exists little standardization on how to approach the analysis of larger metabolites using LC-MS [7]. A significant step forward in the field of untargeted metabolomics is the advent of instruments capable of sub-ppm mass accuracy measurements as well as precursor fragmentation, enabling the acquisition of the exact mass as well as obtaining molecule fragment information in order to construct a meaningful sum formula [8,9]. Owing to these technological advances, metabolomics has already proven to be a valuable tool in fields like biomarker discovery and functional genomics [10–12]. In this study, we introduce an algorithm called *mzGroupAnalyzer* to provide characterization and identification of data signals acquired by an LC-Orbitrap-FT-MS system utilizing

sub-ppm mass measurements and intelligent sum formula query. We applied this strategy to plant secondary metabolism. Plants are most important resources for natural products, providing the highest diversity of chemical structures in the range of 200,000–400,000 different compounds and a high *in vivo* plasticity in response to environmental conditions. Due to this diversity of chemical structures ranging from simple hydrocarbons to complex heteroatomic molecules, the elucidation of the metabolome of higher plants and in general of natural products of any origin poses a challenge for metabolomic techniques. Traditionally, the metabolism of higher plants is subclassified into the primary metabolism, which comprises molecules with a low molecular weight that are involved in the central energy conversion cycles of the organism, and the secondary metabolism, which is not per se involved in energy homeostasis but rather involved in the chemical communication with the environment. Only recently, the field of plant secondary metabolites has begun moving more and more into the focus of new bio-analytical techniques and their stereotypic role as simple chemical weapons is being revised as numerous findings indicate their ability to stimulate vital processes in the cell by regulating the concentration of reactive oxygen species *in vivo* [13,14]. Also, secondary metabolites are of vast interest to the area of medicine and nutrition, as these phytochemicals often possess physiological activity in the human body, for example antioxidant activity or cancer chemoprevention [15].

To provide a suitable experimental system, we stressed *Arabidopsis thaliana* plants by parallel cold and light treatment introducing high oxidative stress over a 3-week interval. This led to a comprehensive switch of the vegetative growth metabolism to protective accumulation of secondary metabolites. We show a way of embedding the multitude of signals into a biochemical context through automated detection of metabolic steps between the acquired *m/z* features. Putative structures are inferred by analysis of the product ions from the same data. By applying *mzGroupAnalyzer* to LC-MS raw data, we are able to prove the existence and relation of known molecules as well as propose novel compounds and pathways, in the present case molecules arising during oxidative stress within the secondary metabolism of *Arabidopsis thaliana*. This strategy can be applied systematically and conveniently to any kind of LC-MS data set and is expected to improve identification and structural elucidation in complex metabolomics data, which is currently the limiting step in large-scale metabolomics studies.

Materials and Methods

Plant material and harvest

Arabidopsis thaliana Col-0 was cultivated in a growth chamber under controlled conditions: light intensity was $280 \mu\text{mol m}^{-2} \text{s}^{-1}$ in an 8-hour light/16-hour dark day cycle; relative humidity was 60% with an average temperature of 22°C. Time point “zero” of cold stress consisted of replicates of non-stressed plants, while every 2 days another sample batch of cold-acclimated plants in a 4°C cold chamber was taken. Rosette leaves were harvested approximately 2 hours after the beginning of the light period. Metabolic activity in the leaves was quenched by immediately putting the plant material into liquid nitrogen after harvesting. Deep-frozen leaf material was ground to a fine powder with a pestle and mortar under steady addition of liquid nitrogen and subsequently stored at -80°C before measurement.

Chemicals

Methanol (HPLC-grade), chloroform (anhydrous, >99%, p.a.) and acetonitrile (UHPLC-grade) were purchased from Sigma-Aldrich (Vienna, Austria). Formic acid (98–100%) was purchased from Merck (Vienna, Austria). Chloramphenicol (>98%) and Ampicillin trihydrate (analytical standard) were purchased from Fluka (Vienna, Austria).

Extraction procedure and sample preparation for secondary metabolite analysis

For LC-MS analysis, about 50 mg of frozen plant-leaf material was extracted by 1 ml pre-chilled 80/20 v:v MeOH/H₂O solution containing 1 μg of the internal standards Ampicillin and Chloramphenicol. Samples were centrifuged at 15,000 g for 15 minutes. The supernatant was dried out overnight in a new tube and re-dissolved in 100 μl of 50/50 v:v MeOH/H₂O solution and centrifuged again for 15 minutes at 20,000 g. The supernatant was then filtered through a STAGE tip (Empore/Disk C18, diameter 47 mm) before it was conveyed into a GC vial with a micro insert tip. Plant extracts were further extracted with 500 μl of chloroform to remove the highly abundant lipid components. The LC-MS method for secondary metabolite analysis has been described before [6].

Data processing, *mzGroupAnalyzer* and pathway viewer

Both *mzGroupAnalyzer* and *Pathway Viewer* have been integrated into the GUI of the COVAIN toolbox [16]. The standalone version of COVAIN can be downloaded at <http://www.univie.ac.at/mosys/software.html>. The data processing strategy and subsequent analysis of the data using *mzGroupAnalyzer* and *Pathway viewer* are explained in the *mzGroupAnalyzer*-Tutorial (Figure S1). *m/z*-values acquired by LC-MS were exported to Excel data sheets using the Xcalibur software. Elemental composition determination was enabled, with a maximum of 10 possible sum formulas for each compound and a ppm deviation of 1.

The single excel-sheets can be uploaded to *mzGroupAnalyzer* via the GUI of COVAIN. Furthermore, a user-defined rules-file is uploaded and the folder for storage of the result-files is provided. By starting the *mzGroupAnalyzer*, the lists of *m/z* values with associated chemical compositions from Xcalibur output are read and the atomic differences between *m/z*-pairs are calculated and compared with the putative chemical modifications provided by the rules-file.

Based on all chemical modifications provided by the rules-file, *mzGroupAnalyzer* searches pathways between pairs of *m/z* features. Regarding each *m/z* feature as a node in a metabolic network, an edge connects two nodes if a chemical modification exists. These edges and nodes generate large networks. A pathway can therefore be constructed by searching the shortest path between two nodes. Redundant paths that are included in other longer paths are removed. The pathway searching algorithm can deal with time series data by filtering out pathways that do not reflect the correct chronological order of the given measurements. For example, there is no path from *m/z* feature *A* occurring on Day 2 to *m/z* feature *B* occurring on Day 1, although theoretically a chemical modification from *A* to *B* is possible. Finally, to better visualize the pathways, we further developed *Pathway Viewer*, which is integrated in *mzGroupAnalyzer*, and is able to plot the pathways after a series of user-defined filtering options like *m/z* range or time points. The following result files will be created, exported and saved:

1. Transformations corresponding to the rules-file.
2. A ranking of frequently found chemical modifications.

- Putative transformations that were not listed in the rules-file.
- A mzStructure file for *Pathway viewer*.
- A Pajek-file (<http://vlado.fmf.uni-lj.si/pub/networks/pajek/>) for network visualization.

Next the *Pathway viewer* is started and a table of transformations and the possibility for visualization of pathways is provided (see Figure S1 – *mzGroupAnalyzer*-Tutorial). The above process is summarized in Figure 1. The list of chemical and biochemical reactions searched by *mzGroupAnalyzer* (currently comprising 56 metabolic reaction steps in the provided rules-file) is shown in Table S1. This list can be easily extended to novel transformations.

For the construction of the van Krevelen plots, sum formulas with *mzGroupAnalyzer*-predicted metabolic transformations were assorted in an Excel sheet and their O/C and H/C ratios were calculated. These values were exported to SigmaPlot 12.3 and mapped to a multiple scatter plot within the boundaries 0 to 1 for the O/C ratio and 0 to 2 for H/C ratio.

Results and Discussion

Development of the *mzGroupAnalyzer* algorithm to identify biochemical and chemical transformations in non-targeted metabolomics data

The development and application of algorithms is essential to systematically search for biochemical and chemical transformations of compounds and to find putative pathways in highly complex LC-MS based metabolomics data. We have established an algorithm which is able to extract putative chemical transformations from high mass accuracy metabolome data. The algorithm generates multiple potential pathways directly from raw-LC-MS data and visualizes these as networks. The entire approach is implemented as a graphical user interface (GUI) in COVAIN (Figure 1). COVAIN is a toolbox for statistical data mining in metabolomics and other OMICS approaches [16] in the Matlab environment. In order to evaluate the algorithm, we measured various reference compounds of typical plant secondary metabolites with nano-UPLC-Orbitrap-MS. Subsequently, sum formulas were generated in Xcalibur 2.0 using the integrated sum formula

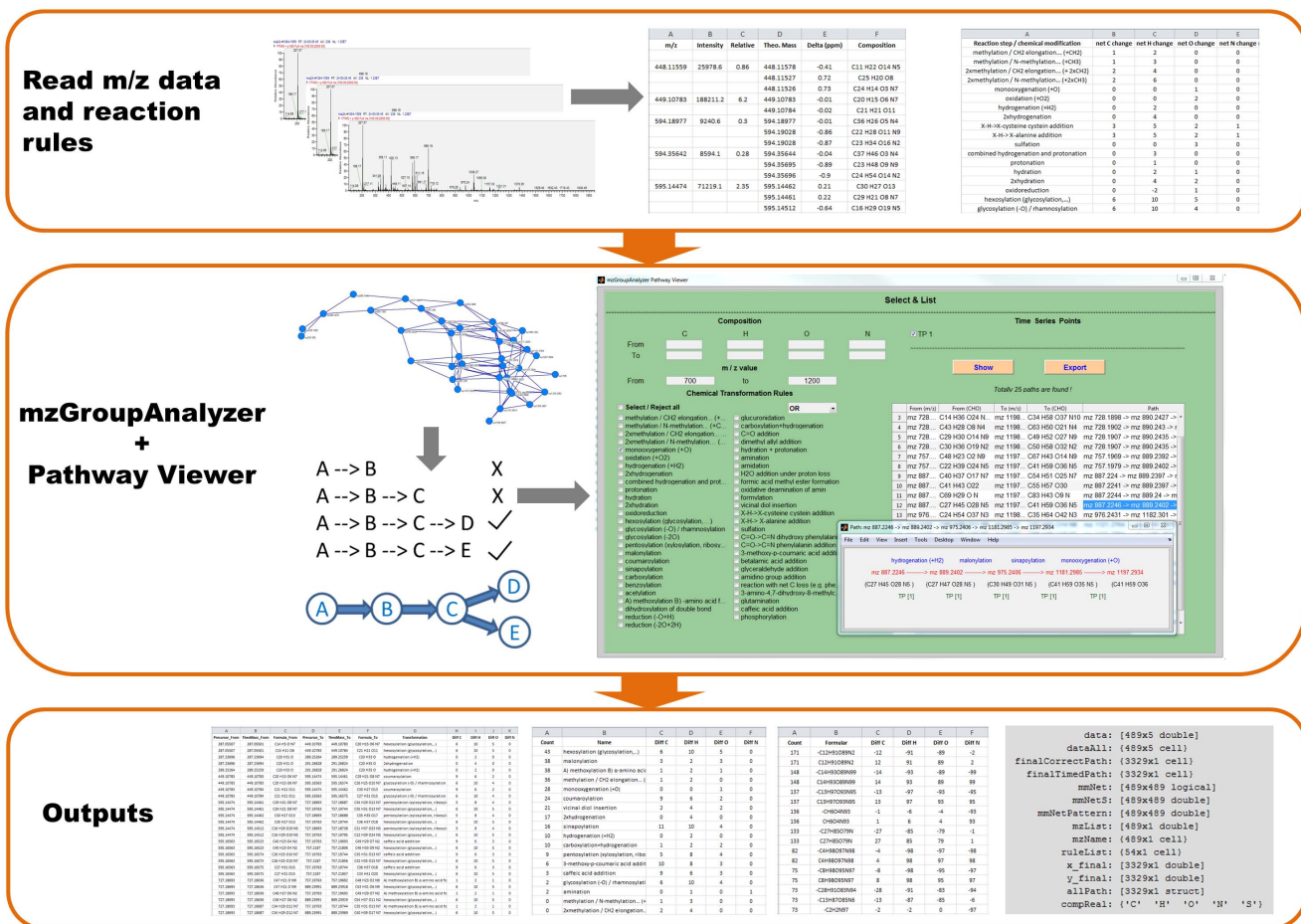


Figure 1. Scheme of the *mzGroupAnalyzer* and *Pathway Viewer* algorithm and GUI implementation. The program reads the *m/z* features which are extracted from Xcalibur, as well as the user predefined reaction rules. Then it finds transformations between all pairs of *m/z* features, and reports the frequency of transformations for the listed and not listed but potentially existing rules. Next, the program starts searching pathways inside the *m/z* features' network. A shorter path existing in other longer paths is removed, thereby non-redundant pathways are obtained. Then, *mzGroupAnalyzer* opens the *Pathway Viewer*, in which pathways satisfying user-defined filtering options will be displayed on the panel. The pathway diagram, which consists of reaction rules, *m/z* feature names, compositions and time points, can be plotted by clicking the table. Finally, all the results, including the frequency table of transformations, the interconnected network visualization file (in Pajek's format), the inferred pathways and a Matlab workspace (suffixed with *mzStruct.mat*) containing all results, will be exported to the user-specified folder.

doi:10.1371/journal.pone.0096188.g001

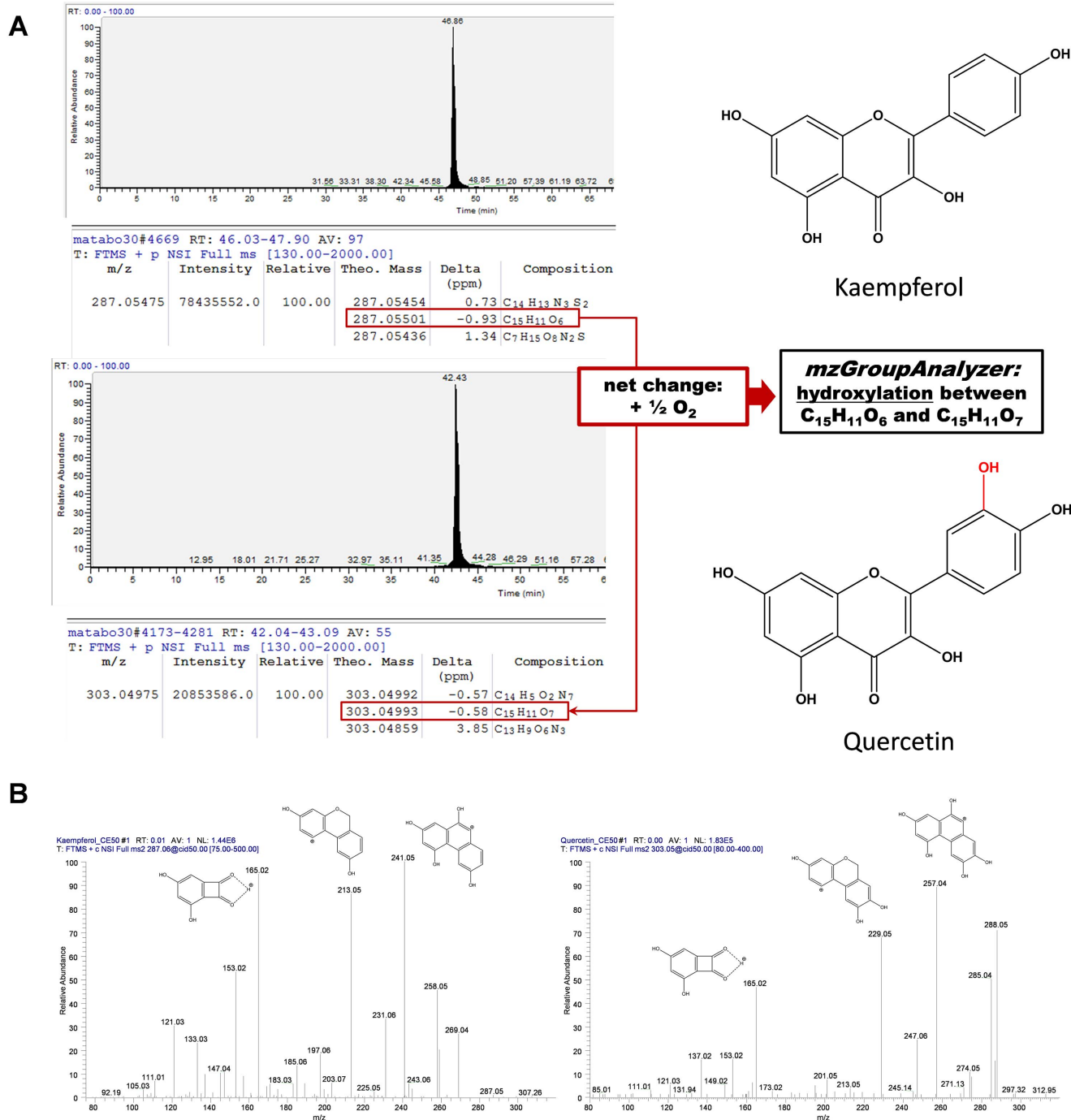


Figure 2. *mzGroupAnalyzer* is able to detect possible metabolic steps out of various proposed sum formulas for a measured m/z feature. **A** Kaempferol and Quercetin standards measured by LC-MS result into several sum formula suggestions for the measured mass-to-charge-ratio (m/z). Because a low ppm deviation of assigned elemental composition to the mass is not the decisive factor, the correct sum formula might not be the first one proposed and thus several must be looked at, which is handled by the program automatically. If *mzGroupAnalyzer* finds a possible reaction step (out of a list of reactions which can be altered manually), it is reported to the user. **B** MS² spectra of Kaempferol (left) and Quercetin (right). The fragmentation schemes are in accordance with published literature [42,43]. The difference of one oxygen atom (nominal mass 16) is visible in the fragments m/z 213→229 and 241→257, while m/z 165 occurs in both product scans. doi:10.1371/journal.pone.0096188.g002

calculator. Within a 1 ppm mass accuracy window and using the monoisotopic masses of several common elements, possible sum formulas were obtained (Figure 2 and Figure S1 – *mzGroupAnalyzer*-Tutorial). Many of the sum formulas did not result in feasible chemical structures but had a very high mass accuracy according

to the measured compound, thus leading to false positives. As a consequence the correct sum formula prediction was not among the top first hits. The fact that high mass accuracy alone is not sufficient for correct sum formula prediction has been shown before [17,18]. This is especially true for metabolites which often

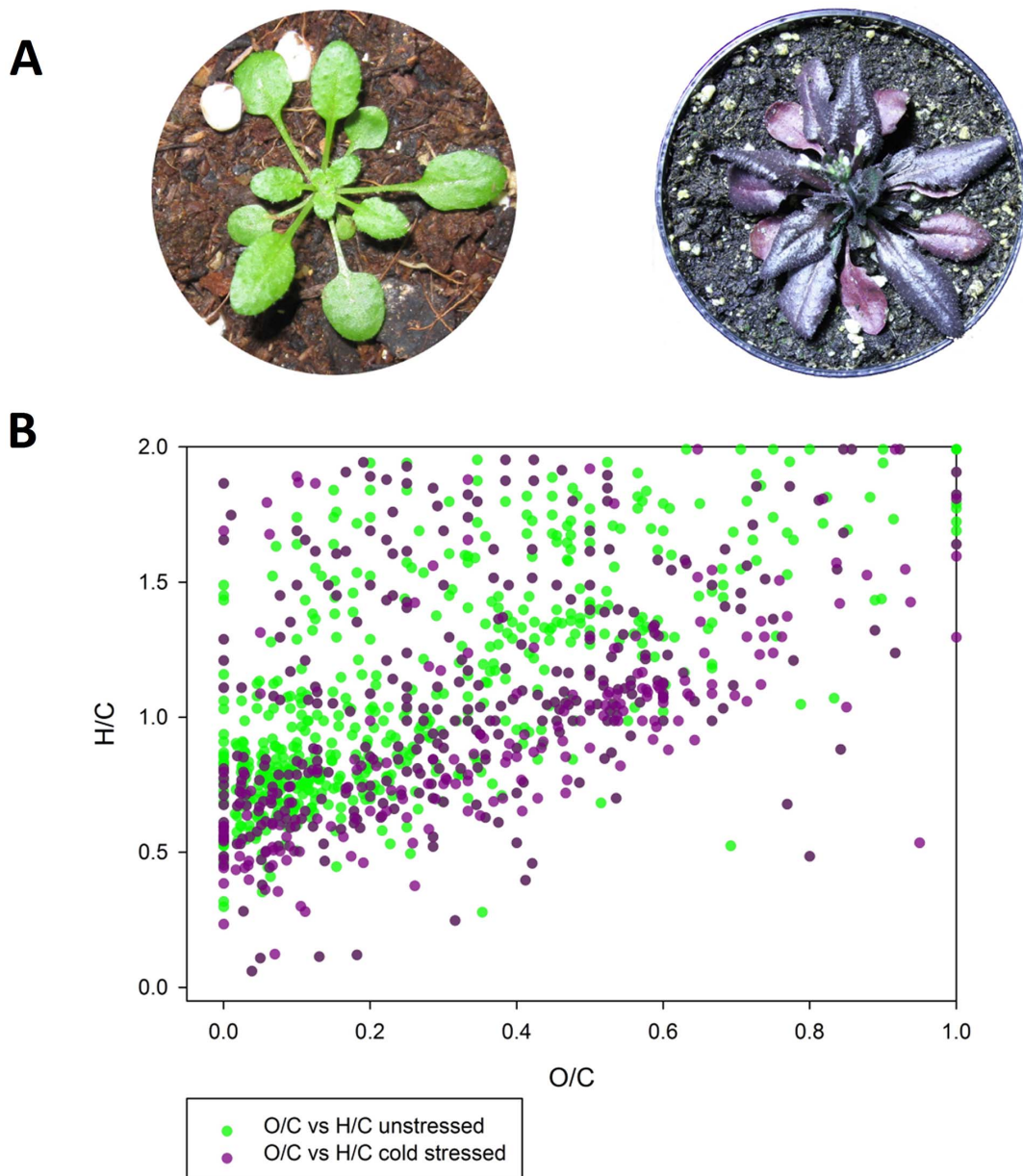


Figure 3. After oxidative stress the *Arabidopsis thaliana* plants turn from green into purple indicating a dramatic shift in metabolism, specifically elevated flavonoid biosynthesis involved in oxidative stress protection [6]. **A** Plants turn from green to purple under high light and cold temperature treatment. **B** Van Krevelen diagram of the most abundant m/z values of unstressed (green dots) and 20-day cold stressed (purple dots) *Arabidopsis* plants. A clear shift of metabolism in the stressed plants is visible.
doi:10.1371/journal.pone.0096188.g003

contain elements like S and P, or even halogens. The utilization of electrospray ionization tends to produce Na^+ , K^+ and other adducts in positive ionization mode. Because *mzGroupAnalyzer* looks for putative chemical transformations of compounds the considered number of sum formulas can be reduced. The application of *mzGroupAnalyzer* revealed metabolic conversions of the protonated reference compound such as the addition of a hydroxyl group to Kaempferol leading to Quercetin (Figure 2). Due to this chemical transformation the sum formula pair of Kaempferol and Quercetin is automatically detected. Indeed, this reaction occurs in the flavonol biosynthetic pathway catalyzed by a flavonoid 3'-monooxygenase. The constraints for detectable metabolic reactions (e.g. +1C+2H denotes a net methylation) are uploaded to the

program before performing the analysis and can be customized by the user (Figure S1 – *mzGroupAnalyzer*-Tutorial). *mzGroupAnalyzer* is also able to recognize the frequency of equidistant steps between m/z values and exports these data as suggestions for novel modifications (see Methods section). Furthermore, whole series of reaction steps in the data can be detected and analyzed in the context of metabolic pathways. This will be discussed in the next sections.

Cold- and light-induced stress has a dramatic effect on the *Arabidopsis thaliana* metabolome

To test the performance of the *mzGroupAnalyzer* algorithm we designed the following experiment. *Arabidopsis thaliana* plants were

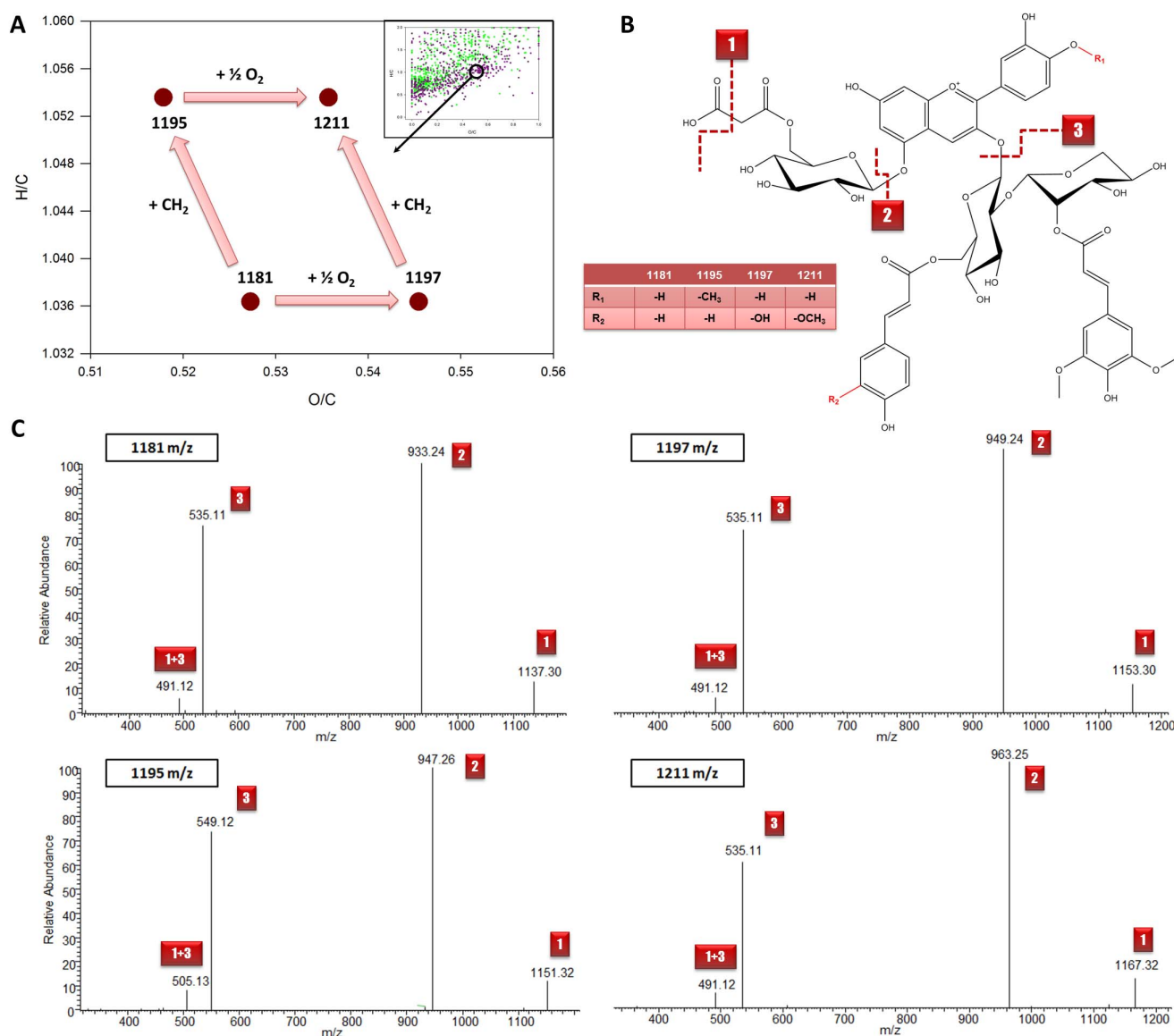


Figure 4. Exploration of the van Krevelen diagram created by sum formulas with chemical transformations detected by *mzGroupAnalyzer*. **A** m/z 1181, 1195, 1197 and 1211 are interconnected with net shifts of $\frac{1}{2} O_2$ and a CH_2 group and form a rhombic pattern. **B** Proposed fragmentation scheme of these compounds under the chosen conditions. **C** Product ion scans show similar fragmentation behavior of the polysubstituted anthocyanins. The spectrum of m/z 1195 shows a peak at m/z 549, pointing to a methyl group at the cyanidin core. A putative methylation site is shown.

doi:10.1371/journal.pone.0096188.g004

exposed to excess irradiation in a 4°C environment. This treatment is described in the literature to produce high levels of reactive oxygen species (ROS) [19–21]. ROS, such as hydrogen peroxide (H_2O_2), hydroxyl radicals (OH), superoxide anion (O_2^-) and singlet oxygen (1O_2), are unavoidable by-products of photosynthetic organisms occurring in organelles with a high oxidative turnover rate during normal metabolic activity and can be highly damaging for cells and tissues under stress [13,22,23]. These stress conditions require an effective scavenging system in order to prevent the organism from being damaged by free radicals [24]. Especially biomolecules from the phenylpropanoid family, such as flavonoids and anthocyanins, have been recognized as effective radical-scavenging compounds [25,26]. After several days of stress in our experimental systems, the plants began to produce violet pigments in the rosette leaves (Figure 3A). The

purple color is the result of the accumulation of anthocyanidins as a response to oxidative stress [25–28]. The enhanced production of anthocyanidins under these stress conditions requires a large-scale metabolic reprogramming as recently described by us [6]. Leaf extracts of the cold/light-treated *Arabidopsis thaliana* plants were analyzed with LC/MS. From these analyses sum formulas of putative metabolites were generated based on the acquired m/z values focusing only on C, H and O elements (see also Figure S1 – *mzGroupAnalyzer*-Tutorial). The generated sum formulas were loaded into a van Krevelen plot to visualize the chemical and biochemical transformations of metabolites during cold and light stress (Figure 3B). Oxidative shifts in the plot induced by cold-related stress might be explained by the incorporation of oxygen, as well as radical scavenging by redox-active compounds, such as aromatic hydroxyl groups which can stabilize radicals after

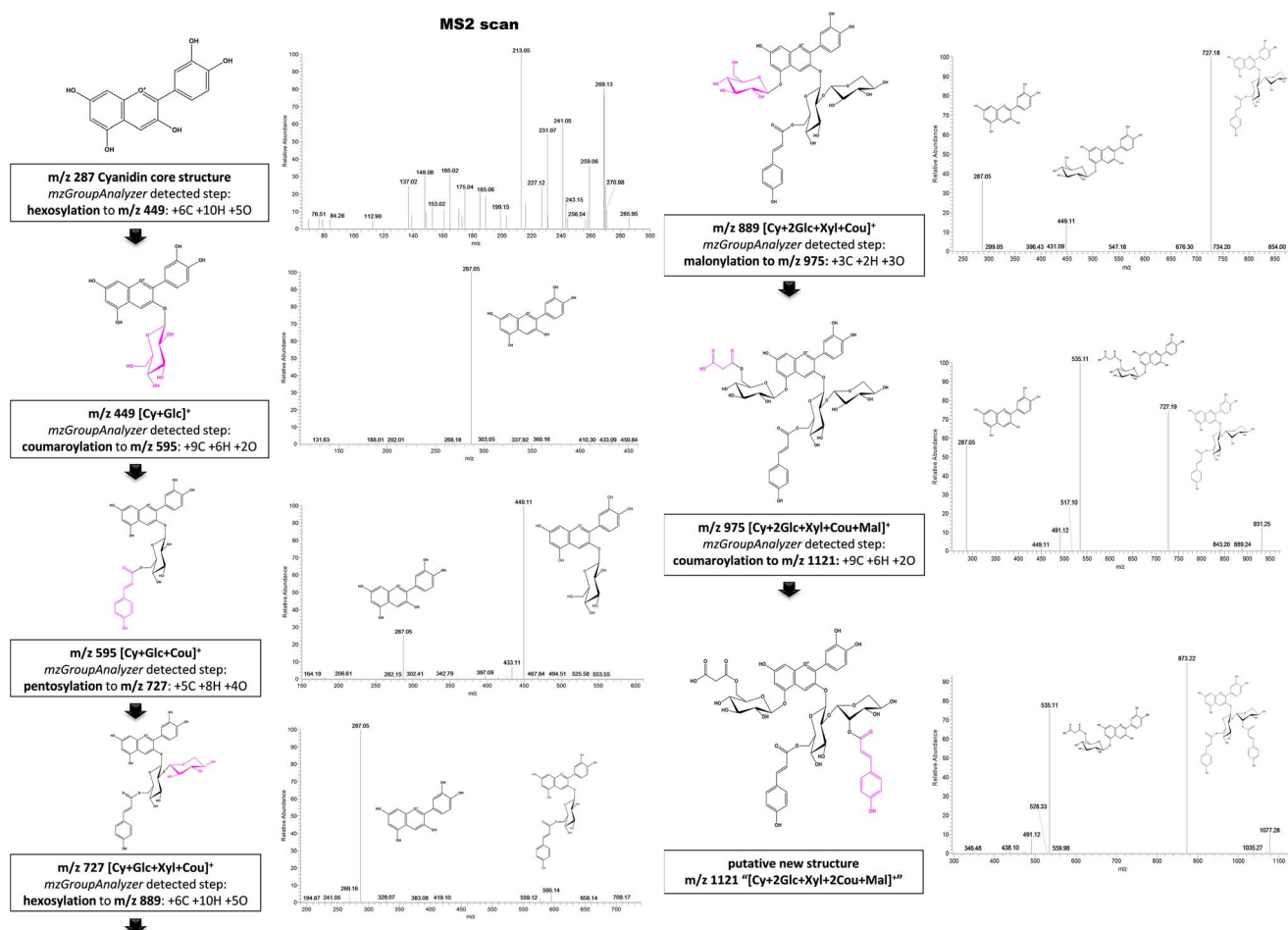


Figure 5. Identification of biochemical transformations of *in vivo* data using *mzGroupAnalyzer*. A metabolic pathway leading to a putative new compound *m/z* 1121 is revealed. Amongst several hundreds of other interlinked *m/z* values in the data, the figure shows metabolic transitions derived from sub-ppm accuracy measurements on the left side and their corresponding MS² product ion scans on the right. Comparison of the spectral information from step to step reveals the possible location of metabolic structural changes. Stereochemistry is assumed due to literature findings [33].

doi:10.1371/journal.pone.0096188.g005

deprotonation. Van Krevelen plots were originally introduced to characterize carbon-based resources like mineral oils and coal according to their possible chemical composition acquired by high-resolution mass spectrometry [29,30]; only recently, van Krevelen plots have been applied as useful tools for visualization of metabolic processes and pathways [31] as well as for sum formula annotation of natural organic matter (NOM) [32].

The investigation of van Krevelen plots composed of untargeted metabolomics data can validate structural familiarity between compounds through a metabolic pattern as depicted in Figure 4. In the O/C ratio range from 0.51 to 0.55 and H/C range of 1.036 to 1.056, the chemical relation of compounds *m/z* 1181, 1195, 1197 and 1211 is visible (Figure 4A), while their structures are validated by MS² product ion scans (Figure 4B and 4C). These compounds were also automatically detected by the *mzGroupAnalyzer* approach when investigating the putative chemical and biochemical transformations from the generated sum formula hits (see also Figure S1 – *mzGroupAnalyzer*-Tutorial). In the following section we explored the potential of *mzGroupAnalyzer* to reveal full pathways leading to novel structures.

Broad-scale analysis of metabolic conversions and novel structure prediction by *mzGroupAnalyzer* in the cold/light stress metabolome of *Arabidopsis thaliana*

Time-dependent sampling of *Arabidopsis thaliana* leaf samples in cold stress yielded strong alterations in the metabolic profiles (see above). All precursor ions from the LC/MS analysis obtained from one sampling time point were assigned to ten possible sum formulas with the following parameters: 100 max C, 200 max H, 50 max O, 10 max N, 1 ppm maximum deviation from suggested sum formula. Data from all time points were exported from Xcalibur 2.0 software as single Excel sheets and uploaded into *mzGroupAnalyzer* using the graphical user interface (see above and methods; see Figure S1 – *mzGroupAnalyzer*-Tutorial). Further, a user-defined rules-file of the molecular shifts of chemical and biochemical transformations needs to be uploaded via the GUI. Here, we provide a rules-file with 56 reactions. By starting the *mzGroupAnalyzer* via the GUI, the algorithm detects metabolic transformations between pairs of *m/z* values in the uploaded list of suggested sum formulas. A table of detected pathways is generated using the GUI option “Pathway Viewer”. Individual pathways can be visualized by clicking on the corresponding cell. Furthermore, the whole pathway network can be exported as a Pajek-file for

Table 1. Putative compounds including their *mzGroupAnalyzer*-predicted sum formulas, the corresponding exact mass as well as dominant MS² product ion fragments.

name or <i>m/z</i> value	sum formula detected by <i>mzGroupAnalyzer</i>	exact mass	main fragments	mass accuracy (ppm)	reference
A1	C ₃₂ H ₃₉ O ₂₀	743.20292	287, 581	0.35	Tohge et al. [33]
A2	C ₃₅ H ₄₁ O ₂₃	829.20331	287, 535, 581, 785	0.29	Tohge et al.
A3	C ₄₁ H ₄₅ O ₂₂	889.23970	287, 449, 727	−0.11	Tohge et al.
A4	C ₄₃ H ₄₉ O ₂₄	949.26083	287, 449, 787	0.48	Tohge et al.
A5	C ₄₄ H ₄₇ O ₂₅	975.24009	491, 535, 727, 931	−0.23	Tohge et al.
A6	C ₄₇ H ₅₅ O ₂₇	1051.29252	449, 889	−0.27	Tohge et al.
A7	C ₅₂ H ₅₅ O ₂₆	1095.29761	449, 933	0.41	Tohge et al.
A8	C ₅₀ H ₅₇ O ₃₀	1137.29292	491, 535, 889	−0.13	Bloor & Abrahams [40]
A9	C ₅₅ H ₅₇ O ₂₉	1181.29800	491, 535, 933, 1137	−0.28	Bloor & Abrahams
A10	C ₅₈ H ₆₅ O ₃₁	1257.35043	449, 1095	−0.07	Tohge et al.
A11	C ₆₁ H ₆₇ O ₃₄	1343.35083	491, 535, 1095, 1299	0.51	Bloor & Abrahams
A12	C ₂₆ H ₂₉ O ₁₅	581.15010	287, 449	0.43	Tohge et al.
A13	C ₃₅ H ₃₅ O ₁₇	727.18688	287, 581	0.23	Tohge et al.
A14	C ₃₇ H ₃₉ O ₁₉	787.20801	287	0.21	Tohge et al.
A15	C ₄₁ H ₄₅ O ₂₂	889.23970	287	−0.14	Tohge et al.
A16	C ₄₆ H ₄₅ O ₂₁	933.24478	287, 727	0.28	Tohge et al.
A17	C ₅₂ H ₅₅ O ₂₆	1095.29761	449, 933	0.56	Tohge et al.
919	C ₄₂ H ₄₇ O ₂₃	919.25026	287, 757	0.16	-
991	C ₄₄ H ₄₇ O ₂₆	991.23501	491, 535, 743, 947	0.17	-
1005	C ₄₅ H ₄₉ O ₂₆	1005.25066	491, 535, 757, 961	0.28	-
1035	C ₄₆ H ₅₁ O ₂₇	1035.26122	491, 535, 787, 991	0.31	-
1065	C ₅₁ H ₅₃ O ₂₅	1065.28704	449, 903	0.29	-
1081	C ₅₁ H ₅₃ O ₂₆	1081.28196	449, 919	0.68	-
1111	C ₅₂ H ₅₅ O ₂₇	1111.29252	449, 949	0.36	-
1121	C ₅₃ H ₅₃ O ₂₇	1121.27687	491, 535, 873, 1077	0.13	-
1125	C ₅₃ H ₅₇ O ₂₇	1125.30817	449, 963	0.51	Saito et al. [39]
1151	C ₅₄ H ₅₅ O ₂₈	1151.28744	491, 535, 903, 1107	0.48	-
1167	C ₅₄ H ₅₅ O ₂₉	1167.28235	491, 535, 919, 1123	0.54	Kasai et al. [41]
1195	C ₅₆ H ₅₉ O ₂₉	1195.31365	505, 549, 947, 1151	0.49	-
1197	C ₅₅ H ₅₇ O ₃₀	1197.29292	491, 535, 949, 1153	0.28	Saito et al.
1211	C ₅₆ H ₅₉ O ₃₀	1211.30857	491, 535, 963, 1167	0.05	Saito et al.
1313	C ₆₀ H ₆₅ O ₃₃	1313.34026	491, 535, 1065, 1269	0.13	-
1329	C ₆₀ H ₆₅ O ₃₄	1329.33518	491, 535, 1081, 1285	0.43	-
1359	C ₆₁ H ₆₇ O ₃₅	1359.34574	491, 535, 1111, 1315	0.58	-
1373	C ₆₂ H ₆₉ O ₃₅	1373.37156	491, 535, 1125, 1329	0.67	-
1549	C ₇₂ H ₇₇ O ₃₈	1549.40873	535, 1301, 1505	0.57	-

The nomenclature is according to [33]. Compounds *m/z* 1125, 1197 and 1211 were found in *Matthiola incana* by [39].
doi:10.1371/journal.pone.0096188.t001

visualization (see Figure S1 – *mzGroupAnalyzer*-Tutorial and Figure 1). Over the course of the cold and light stress, several highly frequent reactions were identified. Overall, the 10 predominant reactions were mono-oxygenations, followed by hydrogenations, hydrations, methoxylations, methylations, oxidoreductions, oxidations, malonylations, hexosylations, and dihydroxylations of double bonds. Due to possible false positives within the detected metabolic steps and reactions, these reaction lists have to be carefully validated manually and present preliminary results. More evaluation of the *mzGroupAnalyzer* and further proof-of-concept studies are needed in the future.

Nevertheless, *mzGroupAnalyzer* reported many reactions, which were validated to be metabolic pathways: By investigating the information from the MS² spectra as well as comparing them with published literature (e.g. [33], see Table 1), the presence of several compounds of the anthocyanin family in the *Arabidopsis* plants was revealed. Additionally, the program reported metabolic steps from these compounds to previously undescribed *m/z* features, which are suspected to be new anthocyanins. In Figure 5, it is shown how a metabolic pathway is extracted by *mzGroupAnalyzer*. The prediction of the pathway is validated by corresponding MS² product ion spectra from the same data set. This pathway leads to

Ara3A #1400 RT: 31.79 AV: 1 NL: 1.36E5
T: FTMS + c NSI d Full ms2 1121.28@cid50.00 [295.00-1135.00]

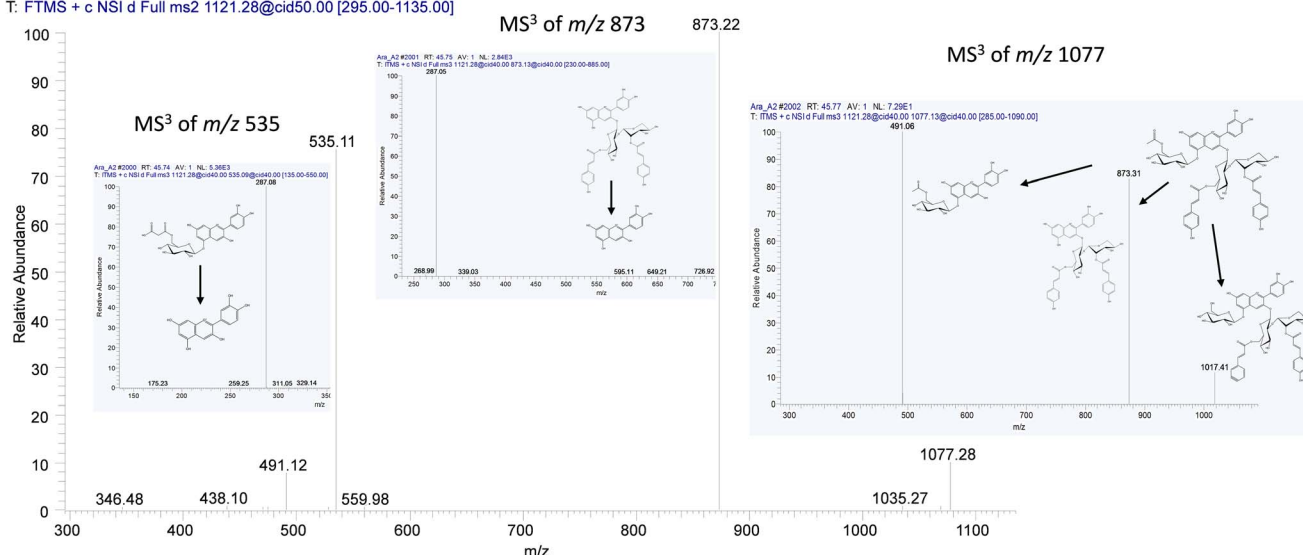


Figure 6. Structure validation of m/z 1121 by MS^3 product ion scans. Both MS^2 fragments m/z 535 and 873 result in the core cyanidin structure by undergoing MS^3 fragmentation. m/z 1077, the putatively decarboxylized form of 1121, yields m/z 491, as observed in the MS^2 spectrum already, by scission of the 3-O-glycosidic bond. Fragment m/z 873 arises again from the breaking of the glycosidic bond at 5-O. m/z 1017 would comply with the complete removal of the rest of the former malonyl group together with a water loss (-60 u). A putative structure is given. doi:10.1371/journal.pone.0096188.g006

a putatively novel compound m/z 1121 which was detected automatically by *mzGroupAnalyzer*.

mzGroupAnalyzer detected reactions between the m/z signals 287, 449, 595, 727, 889, 975 and 1121. m/z 287, the core cyanidin structure, is hexosylated to structure m/z 449, which itself shows the intact cyanidin structure in its MS^2 product ion scan and which has been reported as cyanidin 5-O-glycoside [33]. A coumaroyl group is then added to the sugar group in m/z 449, resulting in compound m/z 595, while showing the previous two compounds in the MS^2 spectrum. To m/z 595, a 5-C sugar is then added to the already existing hexose group – judged from the MS^2 spectrum – forming compound m/z 727. In the MS^2 spectrum of m/z 727 (“A13”) only m/z 595 is visible, indicating that the sugar-sugar bond is more prone for collision-induced dissociation (CID) than the hexose-coumaroyl bond. Next, another metabolic shift of $+6C$, $+10H$ and $+5O$ from m/z 727 to m/z 889 (“A3”) was detected by *mzGroupAnalyzer*, which indicates a hexosylation reaction. Indeed, MS^2 fragmentation scans again showed peaks at m/z 287 and m/z 449, as well as m/z 727 itself in the spectrum, proving our assumptions that CID is happening in both positions, 3-O as well as 5-O. From m/z 889 to m/z 975, a malonylation step was detected. In the product ion scans, the fragment m/z 727 is again present, as well as a new peak at m/z 535, with m/z 449 nearly disappearing. m/z 975 has been reported as molecule “A5” before [33], and coincides with all our findings both in the metabolic route as well as in the MS^2 spectra. The last step in this reaction list was detected to be a coumaroylation from m/z 975 to m/z 1121. Again, fragment 535 is found in the MS^2 scan, while now a higher peak, m/z 873, emerges. We assume this peak is the structure of m/z 727 with another coumaroyl group in position 2 of the xylose, as this position tends to carry further groups. The peaks at m/z 1077 and m/z 931 in the MS^2 product ion scan of m/z 1121 and m/z 975, respectively, correspond to a decarboxylation of the malonyl group in the native molecule. Furthermore the m/z 491 in these two MS^2 product ion scans supports the decarboxylation reaction of m/z 535. Figure 6 shows additional MS^3

product ion scans of the isolated MS^2 product ions m/z 535, 873 and 1077. Both MS^2 peaks m/z 535 and 873 result in the core cyanidin structure m/z 287 by undergoing MS^3 fragmentation. m/z 1077, the putatively decarboxylized form of m/z 1121, yields m/z 491, as observed in the MS^2 spectrum already, by CID-cleavage of the 3-O-glycosidic bond. Fragment m/z 873 arises again from the CID-cleavage of the glycosidic bond at 5-O. m/z 1017 is a putative structure generated by CID cleavage of an acetyl-group together with a water loss (-60 u). These findings lead us to propose m/z 1121 as compound “[Cy+2Glc+Xyl+2Cou+Mal] $^{+}$ ”, or systematically, cyanidin 3-O-[2"-O-(6"-O-(p-coumaroyl) xylosyl) 6"-O-(p-O-(glucosyl)-p-coumaroyl) glucoside] 5-O-(6"-O-malonyl) glucoside. The resulting putative structure is shown in Figure 5.

mzGroupAnalyzer detected a differential appearance of various anthocyanidins (see Table 1) during the different time points. m/z 287 was detected in all time points, m/z 449 after 2 days, and 595, 727, 889, 975 and 1121 only after 4 days of oxidative stress. Following this strategy, 15 new compounds in *Arabidopsis thaliana* could be proposed by investigating the *mzGroupAnalyzer* pathway suggestions together with their product ion spectra m/z . Putative compounds with their *mzGroupAnalyzer*-predicted sum formulas, the corresponding exact mass as well as dominant MS^2 product ion fragments are summarized in Table 1.

Using these new substances in combination with confirmed compounds from the literature, a network of the anthocyanin family starting with the KEGG pathway was reconstructed (Figure 7). Product ion scans of the new compounds and their reconstructed structures can be found in the Figure S2.

Conclusions

In this study, we showed that the application of *mzGroupAnalyzer* – a novel algorithm for untargeted identification of chemical modifications in metabolome data – on time-dependent, high-throughput LC-Orbitrap-FT-MS metabolomic profiles can give new insights into biochemical pathways, and combined with MS^n scans has the power to validate known compounds and predict

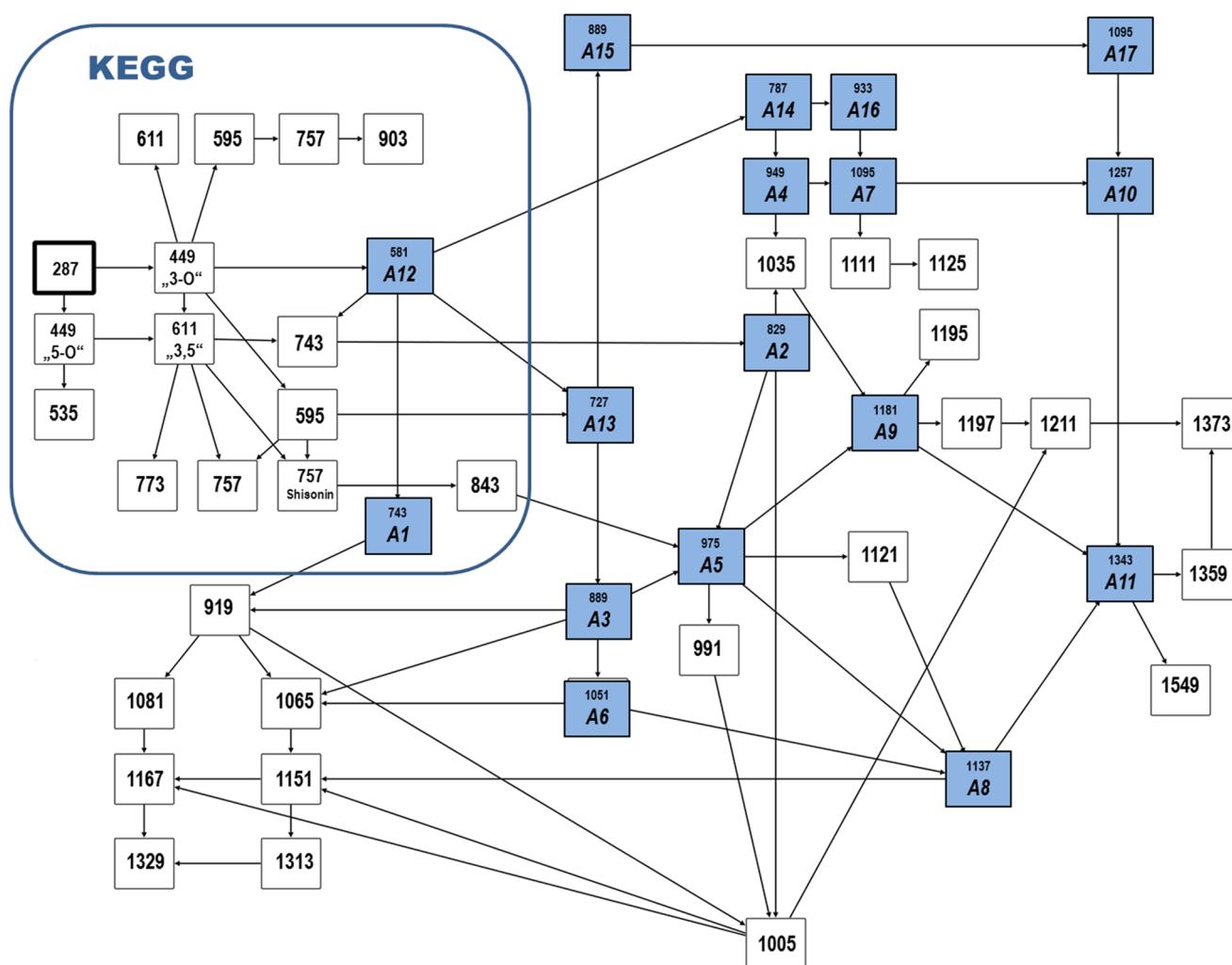


Figure 7. A proposed network of the detected anthocyanin family featuring putatively novel compounds as well as known structures including the KEGG pathway of anthocyanin biosynthesis [38,44]. Only compounds from the KEGG anthocyanin pathway are depicted, for which a suitable precursor mass was found in the data. Exact masses, sum formulas, and main MS^2 fragments of the new compounds are compiled in Table 1; reconstructed structures together with MS^2 scans are in the supporting information. The network was created with VANTED [45,46].

doi:10.1371/journal.pone.0096188.q007

new structures. Attempts at assigning sum formulas to highly resolved metabolomic data have been done before and have proven to be very fruitful [34,35], yet it is clear that those approaches ultimately have to be automatized. The visualization of the elemental composition of metabolites on a van Krevelen diagram is useful for recognition of metabolic patterns, which can point to a structural similarity between those molecules.

The unambiguous structural assignment and stereochemical elucidation of a compound is always a time-intensive process and demands further validation from other analytical platforms, such as NMR [7]. In our approach, we circumvent some of these issues by capturing metabolic intermediates of a compound in the dataset, thus creating an additional layer of information. The complementary analysis of product ion spectra associated with the predicted chemical modifications of the precursor m/z -ratios introduces novel aspects for structural elucidation of unknown metabolites, and enables further validation. While the issues of unambiguous sum formula annotation in high-resolution LC-MS metabolomics still remain, and further application and validation

of other analytical techniques are needed, *mzGroupAnalyzer* proves to be a convenient tool for tracking metabolic changes, thus sum formulas, and inferring metabolic pathways from time-series data, leading to the prediction of entirely novel, hitherto undescribed compounds. Furthermore, *mzGroupAnalyzer* is able to handle time-series data and is thus able to identify time-dependent chemical modifications. Finally, *mzGroupAnalyzer* is connected with the *Pathway Viewer* which plots corresponding pathways in a user friendly way.

Several strategies will be implemented in future to further improve the algorithm. First, the transformation frequency will be used in future to rank sum formulas corresponding to the same m/z feature. Secondly, more strict sum formula filtering criteria will be applied in selecting correct sum formulas from m/z features [36]. Thirdly, more reaction rules, and how these rules are connected in a pathway, will be learned from metabolic pathway databases ([37,38] for KEGG and MetaCyc).

Supporting Information

Figure S1 “S1 mzGA tutorial.pptx”; tutorial for mzGroupAnalyzer in pptx format. (PPTX)

Figure S2 “S2 MS2 putative cyanidins.pptx”; recorded MS2 product ion scans of the putative new cyanins in pptx format. (PPTX)

Table S1 Rules file for chemical and biochemical transformations. (XLSX)

References

- Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN, et al. (2000) Metabolite profiling for plant functional genomics. *Nature Biotechnology* 18: 1157–1161.
- Soga T (2007) Capillary electrophoresis-mass spectrometry for metabolomics. *Methods Mol Biol* 358: 129–137.
- Weckwerth W (2011) Unpredictability of metabolism—the key role of metabolomics science in combination with next-generation genome sequencing. *Analytical and Bioanalytical Chemistry* 400: 1967–1978.
- Scherling C, Roscher C, Giallisco P, Schulze ED, Weckwerth W (2010) Metabolomics unravel contrasting effects of biodiversity on the performance of individual plant species. *Plos One* 5: e12569.
- Mari A, Lyon D, Fragner L, Montoro P, Piacente S, et al. (2013) Phytochemical composition of *L. analyzed* by an integrative GC-MS and LC-MS metabolomics platform. *Metabolomics* 9: 599–607.
- Doerfler H, Lyon D, Nagele T, Sun X, Fragner L, et al. (2013) Granger causality in integrated GC-MS and LC-MS metabolomics data reveals the interface of primary and secondary metabolism. *Metabolomics: Official journal of the Metabolomic Society* 9: 564–574.
- Dunn WB, Erban A, Weber RJM, Creek DJ, Brown M, et al. (2012) Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics*.
- Olsen JV, de Godoy LM, Li G, Macek B, Mortensen P, et al. (2005) Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Molecular & cellular proteomics: MCP* 4: 2010–2021.
- Kind T, Fiehn O (2010) Advances in structure elucidation of small molecules using mass spectrometry. *Bioanalytical reviews* 2: 23–60.
- Weckwerth W (2011) Green systems biology - From single genomes, proteomes and metabolomes to ecosystems research and biotechnology. *Journal of Proteomics* 75: 284–305.
- Weckwerth W, Morgenthal K (2005) Metabolomics: from pattern recognition to biological interpretation. *Drug Discov Today* 10: 1551–1558.
- van der Greef J, Hankemeier T, McBurney RN (2006) Metabolomics-based systems biology and personalized medicine: moving towards n = 1 clinical trials? *Pharmacogenomics* 7: 1087–1094.
- Mittler R, Vanderauwera S, Gollery M, Van Breusegem F (2004) Reactive oxygen gene network of plants. *Trends in plant science* 9: 490–498.
- Croteau R, Kuchan TM, Lewis NG (2000) Natural Products (Secondary Metabolites). *Biochemistry and Molecular Biology of Plants*: 1250–1318.
- Lee KW, Lee HJ, Lee CY (2004) Vitamins, phytochemicals, diets, and their implementation in cancer chemoprevention. *Crit Rev Food Sci Nutr* 44: 437–452.
- Sun X, Weckwerth W (2012) COVAIN: a toolbox for uni- and multivariate statistics, time-series and correlation network analysis and inverse estimation of the differential Jacobian from metabolomics covariance data. *Metabolomics: Official journal of the Metabolomic Society* 8: 81–93.
- Kind T, Fiehn O (2006) Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics* 7: 234.
- Quenzer T (2002) Automated accurate mass analysis using FTICR mass spectrometry. *Proceedings of the 50th Annual Conference on Mass Spectrometry and Allied Topics*, Orlando, FL.
- Kimura M, Yamamoto YY, Seki M, Sakurai T, Sato M, et al. (2003) Identification of Arabidopsis genes regulated by high light-stress using cDNA microarray. *Photochemistry and Photobiology* 77: 226–233.
- Dong CH, Zolman BK, Bartel B, Lee BH, Stevenson B, et al. (2009) Disruption of Arabidopsis CHY1 reveals an important role of metabolic status in plant cold stress signaling. *Molecular plant* 2: 59–72.
- Huang X, Li Y, Zhang X, Zuo J, Yang S (2010) The Arabidopsis LSD1 gene plays an important role in the regulation of low temperature-dependent cell death. *The New phytologist* 187: 301–312.
- Nishiyama Y, Yamamoto H, Allakhverdiev SI, Inaba M, Yokota A, et al. (2001) Oxidative stress inhibits the repair of photodamage to the photosynthetic machinery. *The EMBO journal* 20: 5587–5594.
- Asada K (2006) Production and scavenging of reactive oxygen species in chloroplasts and their functions. *Plant physiology* 141: 391–396.
- Apel K, Hirt H (2004) Reactive oxygen species: metabolism, oxidative stress, and signal transduction. *Annual review of plant biology* 55: 373–399.
- Hernandez I, Alegre L, Van Breusegem F, Munne-Bosch S (2009) How relevant are flavonoids as antioxidants in plants? *Trends in plant science* 14: 125–132.
- Seyoum A, Asres K, El-Fiky FK (2006) Structure-radical scavenging activity relationships of flavonoids. *Phytochemistry* 67: 2058–2070.
- Bashandy T, Taconnat L, Renou JP, Meyer Y, Reichheld JP (2009) Accumulation of flavonoids in an ntra ntrb mutant leads to tolerance to UV-C. *Mol Plant* 2: 249–258.
- Kimura M, Yamamoto YY, Seki M, Sakurai T, Sato M, et al. (2003) Identification of Arabidopsis genes regulated by high light-stress using cDNA microarray. *Photochem Photobiol* 77: 226–233.
- Wu Z, Rodgers RP, Marshall AG (2004) Two- and three-dimensional van krevelen diagrams: a graphical analysis complementary to the kendrick mass plot for sorting elemental compositions of complex organic mixtures based on ultrahigh-resolution broadband fourier transform ion cyclotron resonance mass measurements. *Analytical chemistry* 76: 2511–2516.
- van Krevelen DW (1950) Graphical-statistical method for the study of structure and reaction processes of coal. *Fuel*: 269–284.
- Kai K, Takahashi H, Saga H, Ogawa T, Kanaya S, et al. (2011) Metabolomic characterization of the possible involvement of a Cytochrome P450, CYP81F4, in the biosynthesis of indolic glucosinolate in Arabidopsis. *Plant Biotechnology* 28: 379–385.
- Reemtsma T (2009) Determination of molecular formulas of natural organic matter molecules by (ultra-) high-resolution mass spectrometry: status and needs. *Journal of chromatography A* 1216: 3687–3701.
- Tohge T, Nishiyama Y, Hirai MY, Yano M, Nakajima J, et al. (2005) Functional genomics by integrated analysis of metabolome and transcriptome of Arabidopsis plants over-expressing an MYB transcription factor. *The Plant journal: for cell and molecular biology* 42: 218–235.
- Giallisco P, Hummel J, Lisek J, Inostroza AC, Catchpole G, et al. (2008) High-resolution direct infusion-based mass spectrometry in combination with whole ¹³C metabolome isotope labeling allows unambiguous assignment of chemical sum formulas. *Analytical chemistry* 80: 9417–9425.
- Rogers S, Scheltema RA, Girolami M, Breidling R (2009) Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics* 25: 512–518.
- Kind T, Fiehn O (2007) Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics* 8: 105.
- Karp PD, Riley M, Paley SM, Pellegrini-Toole A (2002) The MetaCyc Database. *Nucleic acids research* 30: 59–61.
- Kanehisa M (2002) The KEGG database. *Novartis Foundation symposium* 247: 91–101; discussion 101–103, 119–128, 244–152.
- Saito N, Tatsuzawa F, Nishiyama A, Yokoi M, Shigihara A, et al. (1995) Acylated cyanidin 3-sambubioside-5-glucosides in *Matthiola incana*. *Phytochemistry* 38: 1027–1032.
- Bloor SJ, Abrahams S (2002) The structure of the major anthocyanin in Arabidopsis thaliana. *Phytochemistry* 59: 343–346.
- Kasai HF, Saito N, Honda T (2011) Structural features of polyacylated anthocyanins using matrix-assisted laser desorption/ionization and electrospray ionization time-of-flight mass spectrometry. *Rapid communications in mass spectrometry: RCM* 25: 1051–1060.
- March RE, Miao X-S (2004) A fragmentation study of kaempferol using electrospray quadrupole time-of-flight mass spectrometry at high mass resolution. *International Journal of Mass Spectrometry* 231: 157–167.
- Abad-Garcia B, Garmon-Lobato S, Berrueta LA, Gallo B, Vicente F (2009) A fragmentation study of dihydroquercetin using triple quadrupole mass spectrometry and its application for identification of dihydroflavonols in Citrus juices. *Rapid communications in mass spectrometry: RCM* 23: 2785–2792.
- Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, et al. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic acids research* 42: D199–205.
- Junker BH, Klukas C, Schreiber F (2006) VANTED: a system for advanced data analysis and visualization in the context of biological networks. *BMC bioinformatics* 7: 109.
- Rohn H, Junker A, Hartmann A, Grafahrend-Belau E, Treutler H, et al. (2012) VANTED v2: a framework for systems biology applications. *BMC systems biology* 6: 139.

Acknowledgments

We would like to thank the gardeners for the excellent plant cultivation and the members of the Department of Ecogenomics and Systems Biology for the constructive discussions.

Author Contributions

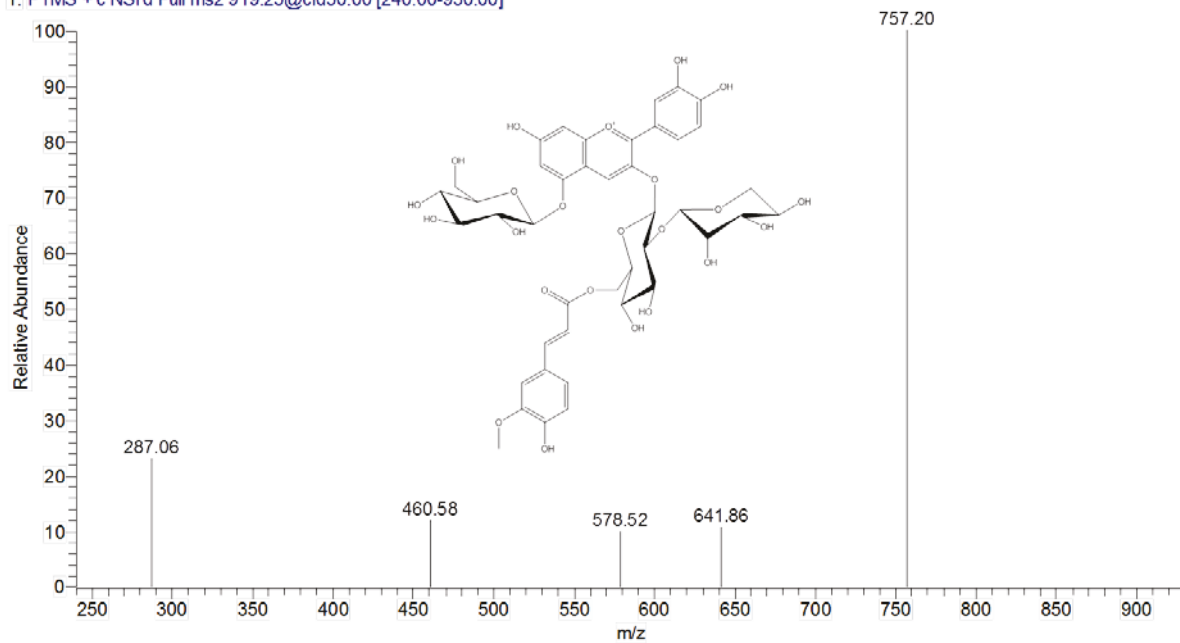
Conceived and designed the experiments: HD XS WW. Performed the experiments: HD XS. Analyzed the data: HD XS WW. Contributed reagents/materials/analysis tools: XS LW DE DL WW. Wrote the paper: HD XS LW DE DL WW.

Attachment

Structure proposals and MS² spectra of putatively new cyanidins

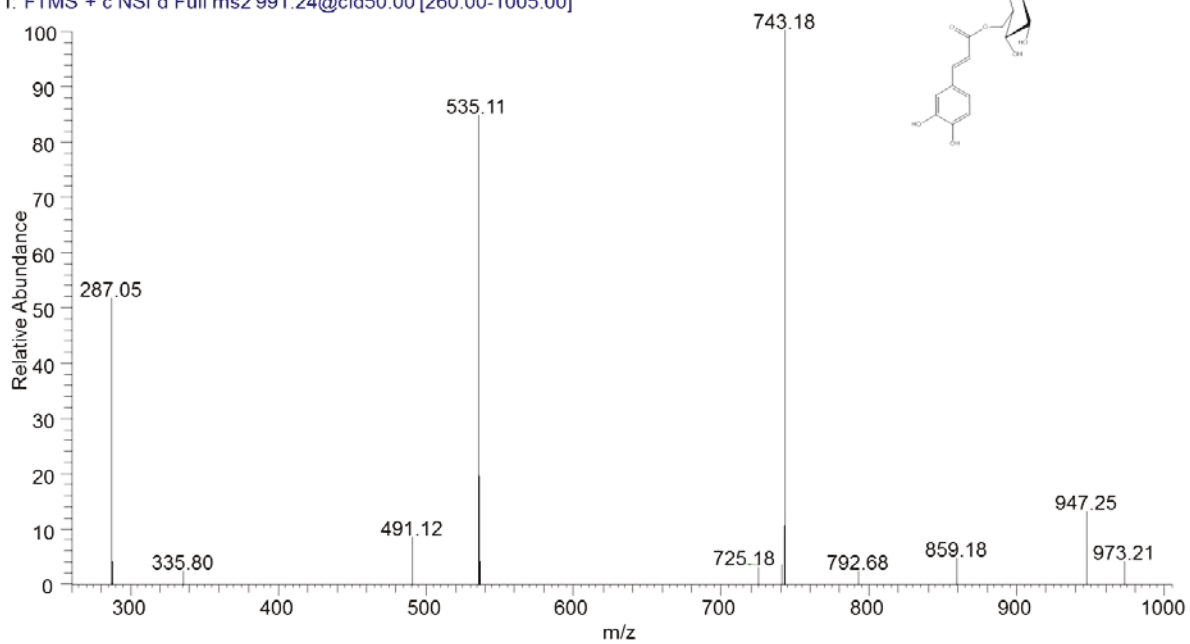
m/z 919

Ara4A#1390 RT: 31.61 AV: 1 NL: 8.34E3
T: FTMS + c NSI d Full ms2 919.25@cid50.00 [240.00-930.00]



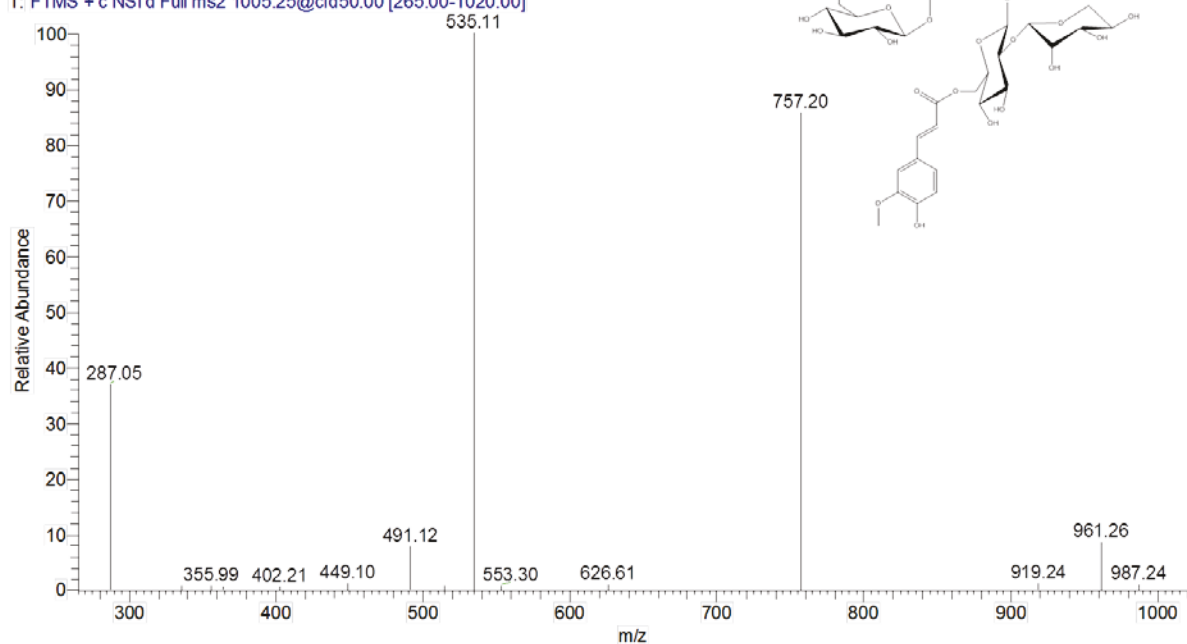
m/z 991

Ara4A#1574 RT: 35.82 AV: 1 NL: 3.90E4
T: FTMS + c NSI d Full ms2 991.24@cid50.00 [260.00-1005.00]



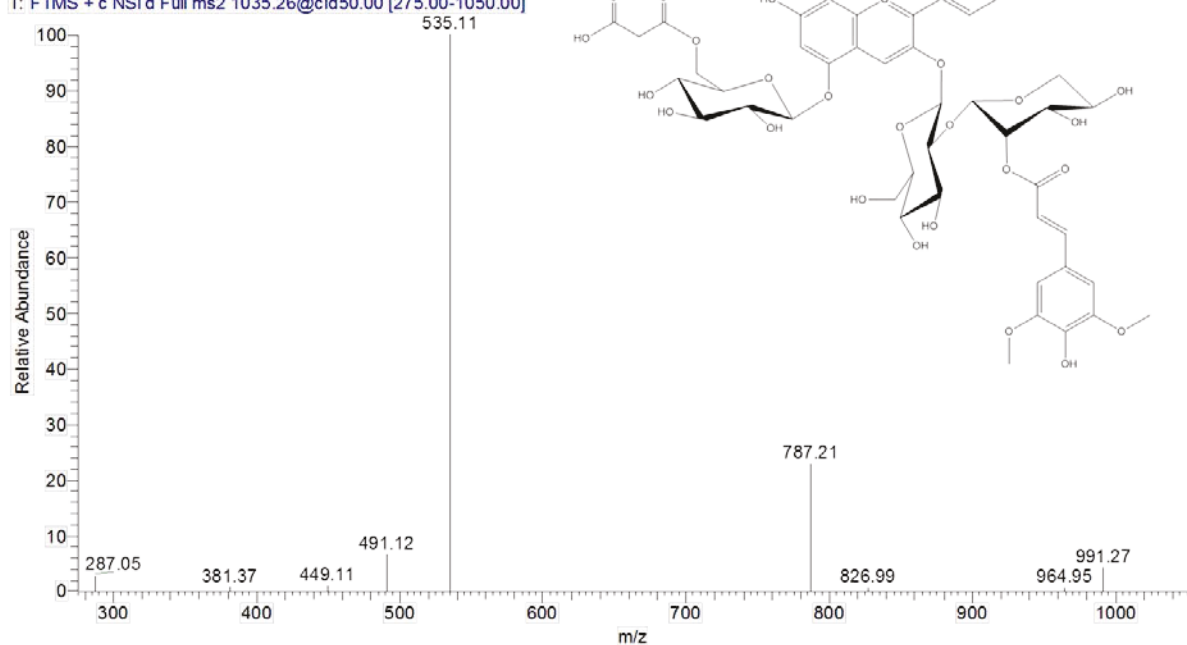
m/z 1005

Ara3A #1364 RT: 30.97 AV: 1 NL: 1.28E5
T: FTMS + c NSI d Full ms2 1005.25@cid50.00 [265.00-1020.00]



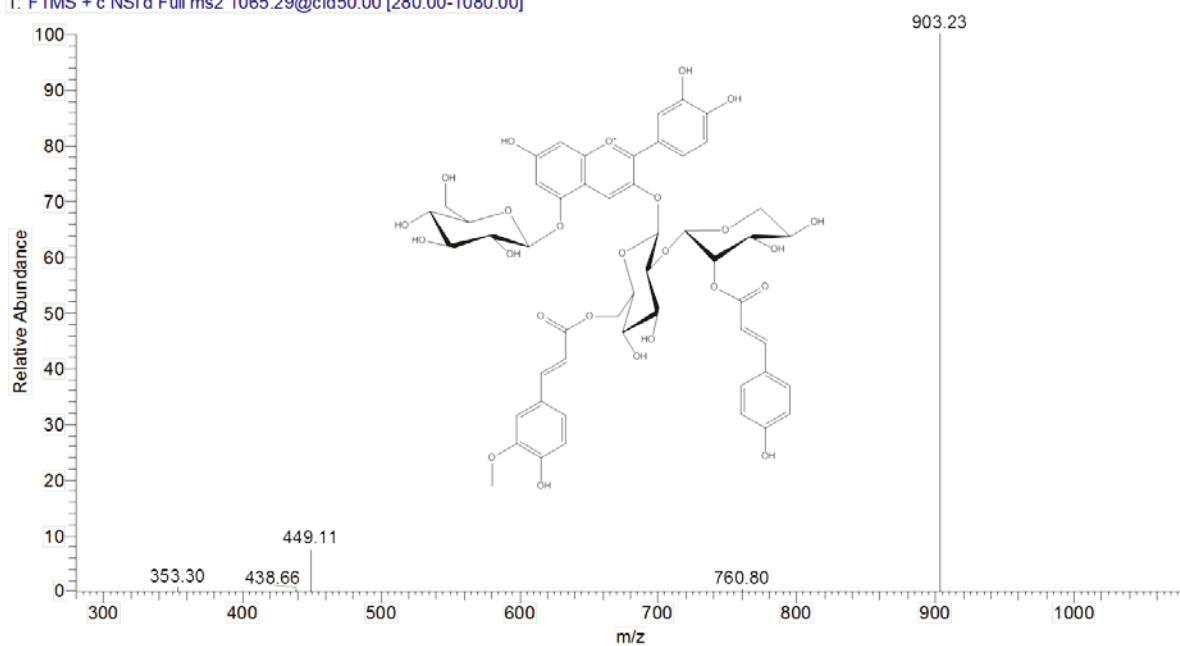
m/z 1035

Ara3A #994 RT: 22.59 AV: 1 NL: 1.31E5
T: FTMS + c NSI d Full ms2 1035.26@cid50.00 [275.00-1050.00]



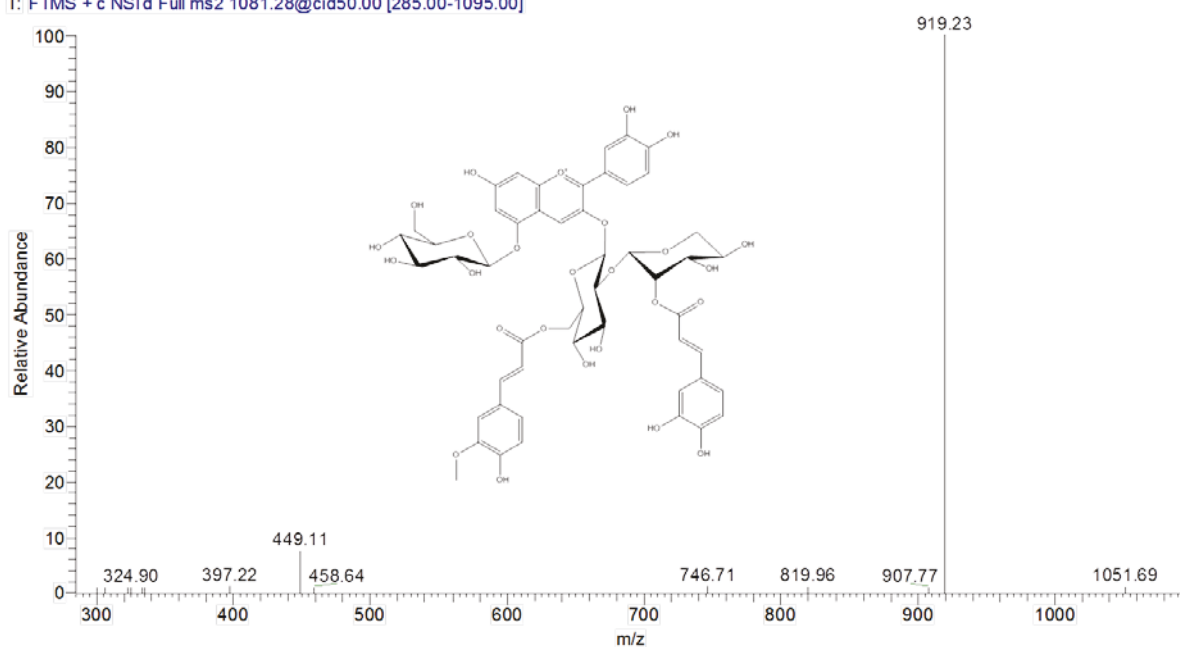
m/z 1065

Ara3A #1416 RT: 32.15 AV: 1 NL: 2.40E5
T: FTMS + c NSI d Full ms2 1065.29@cid50.00 [280.00-1080.00]



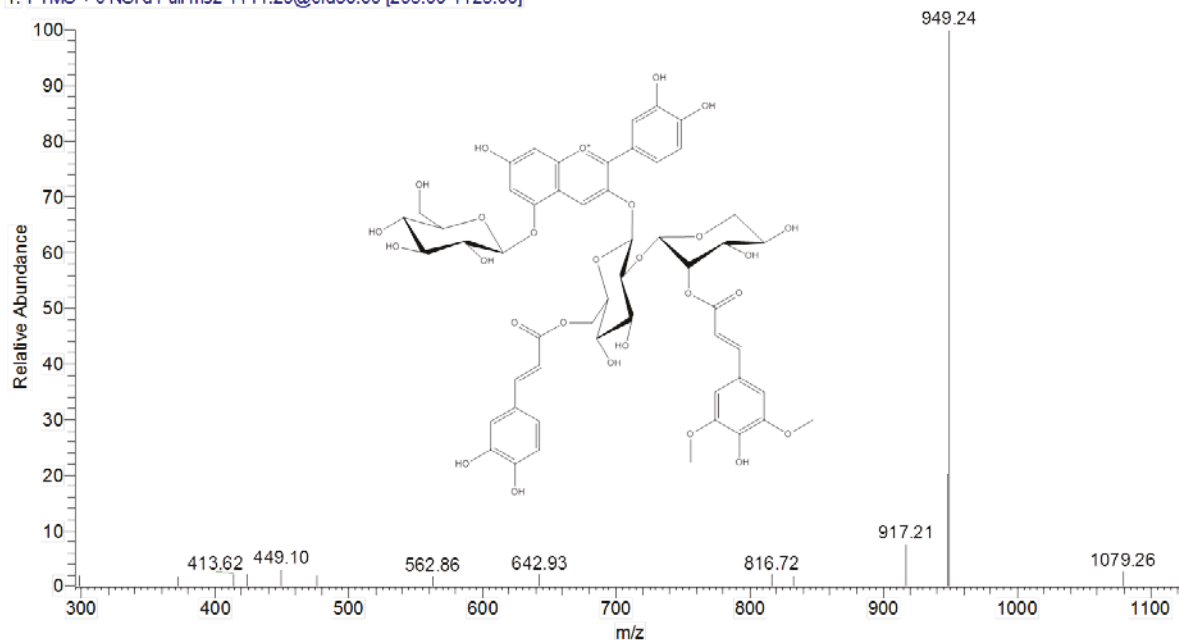
m/z 1081

Ara4C #1346 RT: 30.56 AV: 1 NL: 7.83E4
T: FTMS + c NSI d Full ms2 1081.28@cid50.00 [285.00-1095.00]



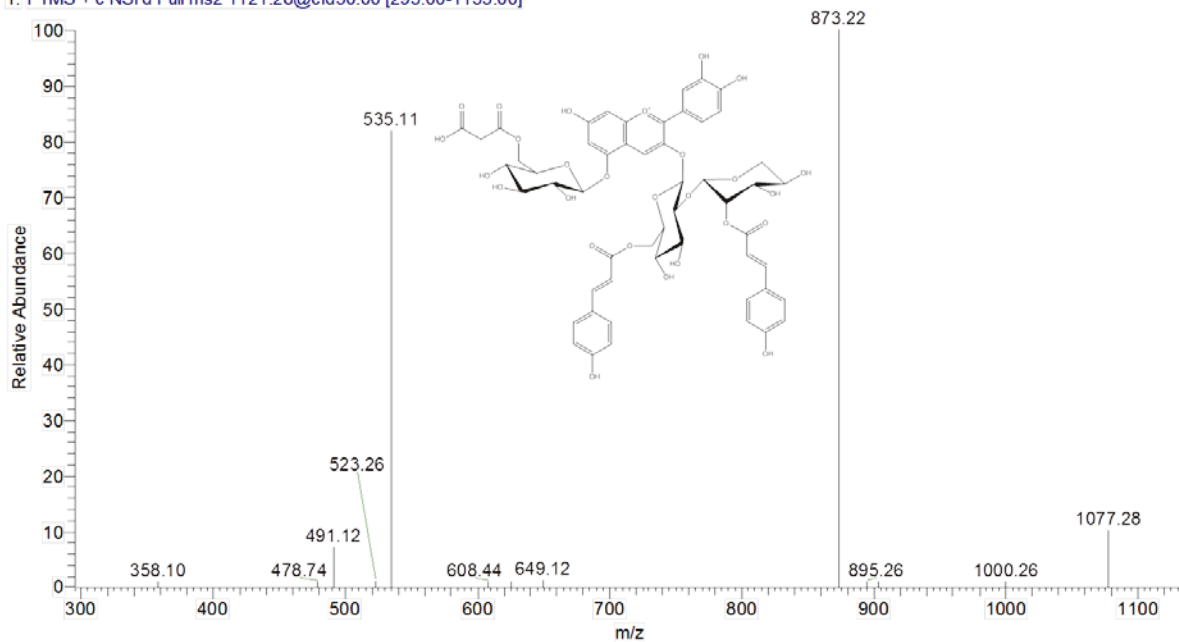
m/z 1111

Ara4A #1512 RT: 34.41 AV: 1 NL: 4.76E4
T: FTMS + c NSI d Full ms2 1111.29@cid50.00 [295.00-1125.00]



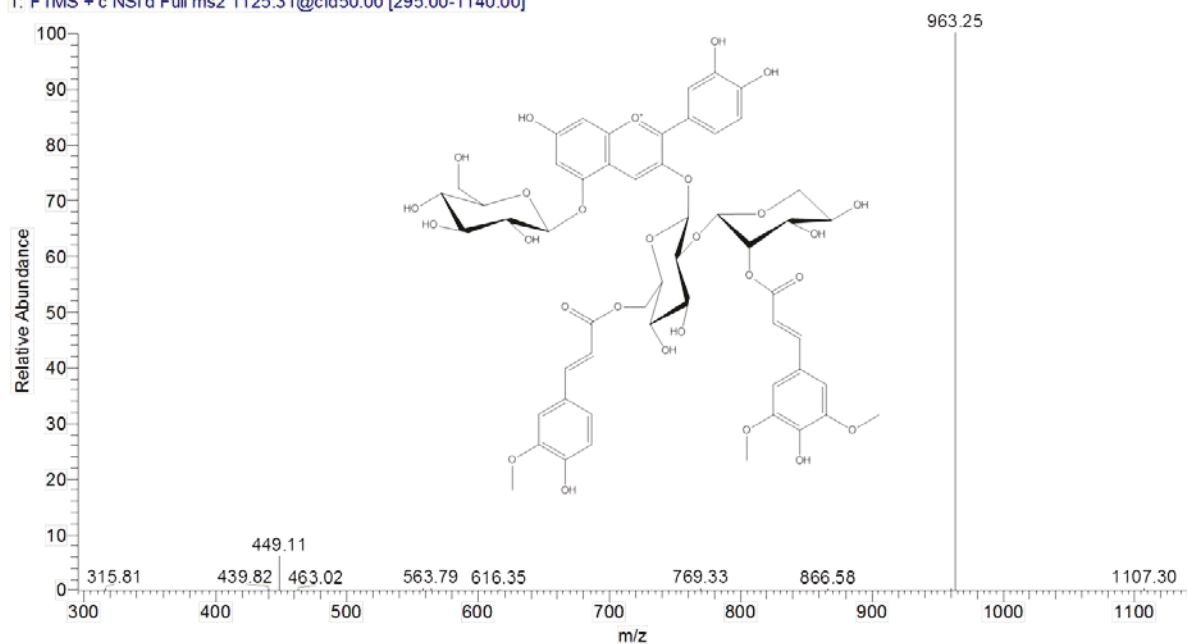
m/z 1121

Ara4A #1740 RT: 39.61 AV: 1 NL: 8.95E4
T: FTMS + c NSI d Full ms2 1121.28@cid50.00 [295.00-1135.00]



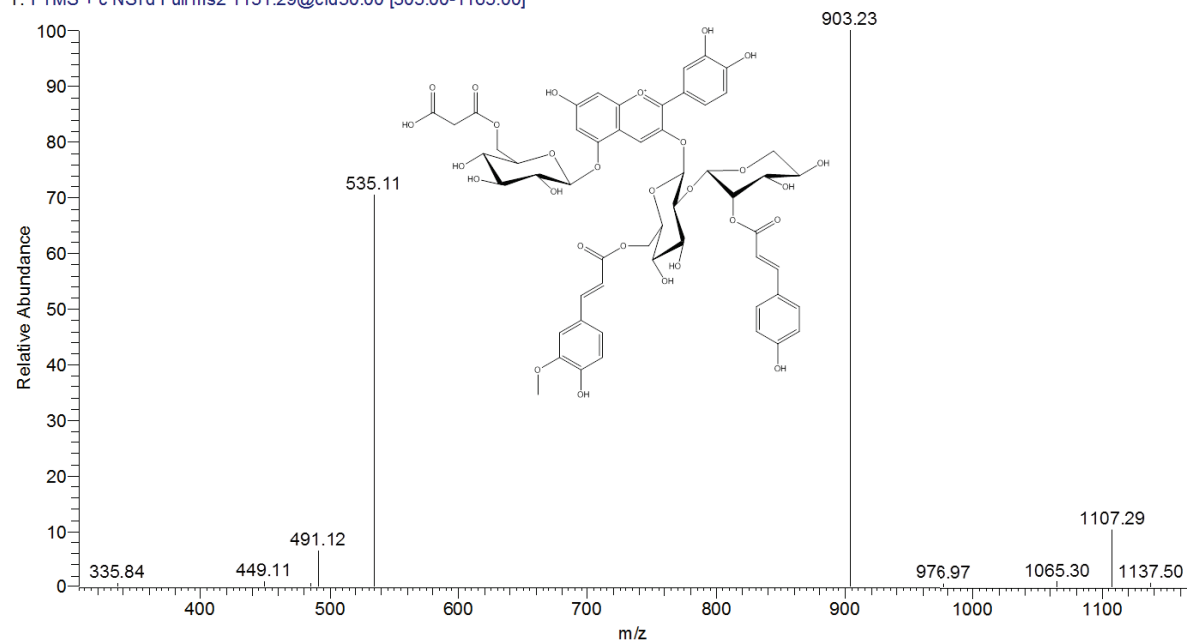
m/z 1125

Ara4A #1532 RT: 34.85 AV: 1 NL: 3.96E5
T: FTMS + c NSI d Full ms2 1125.31@cid50.00 [295.00-1140.00]



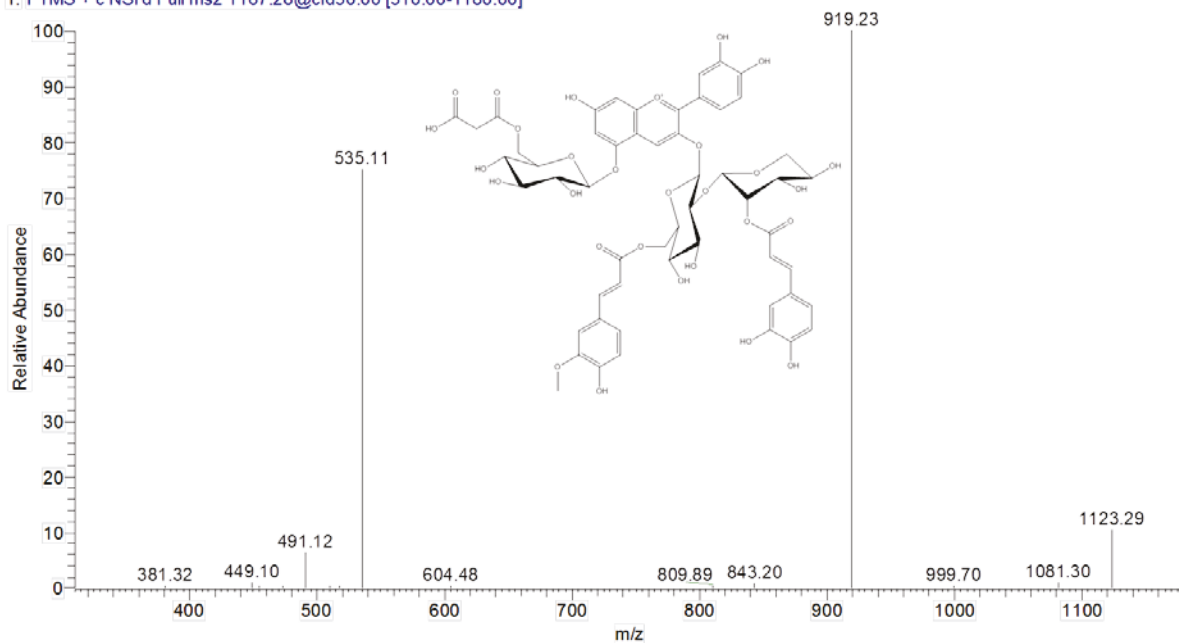
m/z 1151

Ara4A #1496 RT: 34.03 AV: 1 NL: 1.51E5
T: FTMS + c NSI d Full ms2 1151.29@cid50.00 [305.00-1165.00]



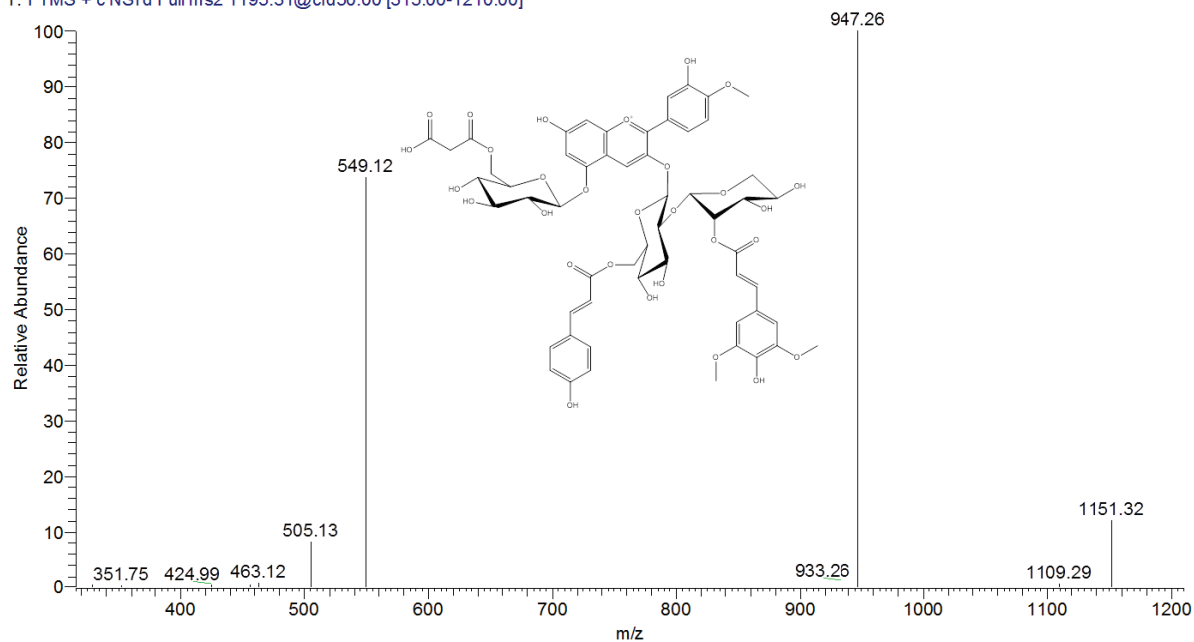
m/z 1167

Ara3A #1462 RT: 33.21 AV: 1 NL: 2.77E5
T: FTMS + c NSI d Full ms2 1167.28@cid50.00 [310.00-1180.00]



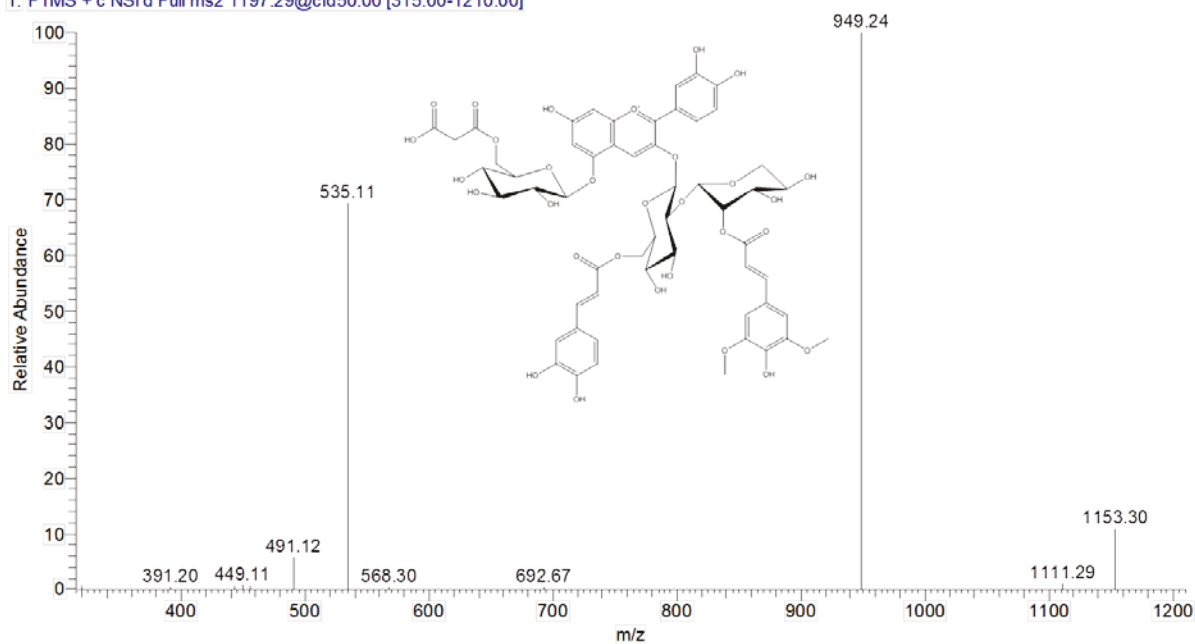
m/z 1195
position of methyl group on cyanidin core unclear

Ara3A #1620 RT: 36.82 AV: 1 NL: 2.58E5
T: FTMS + c NSI d Full ms2 1195.31@cid50.00 [315.00-1210.00]



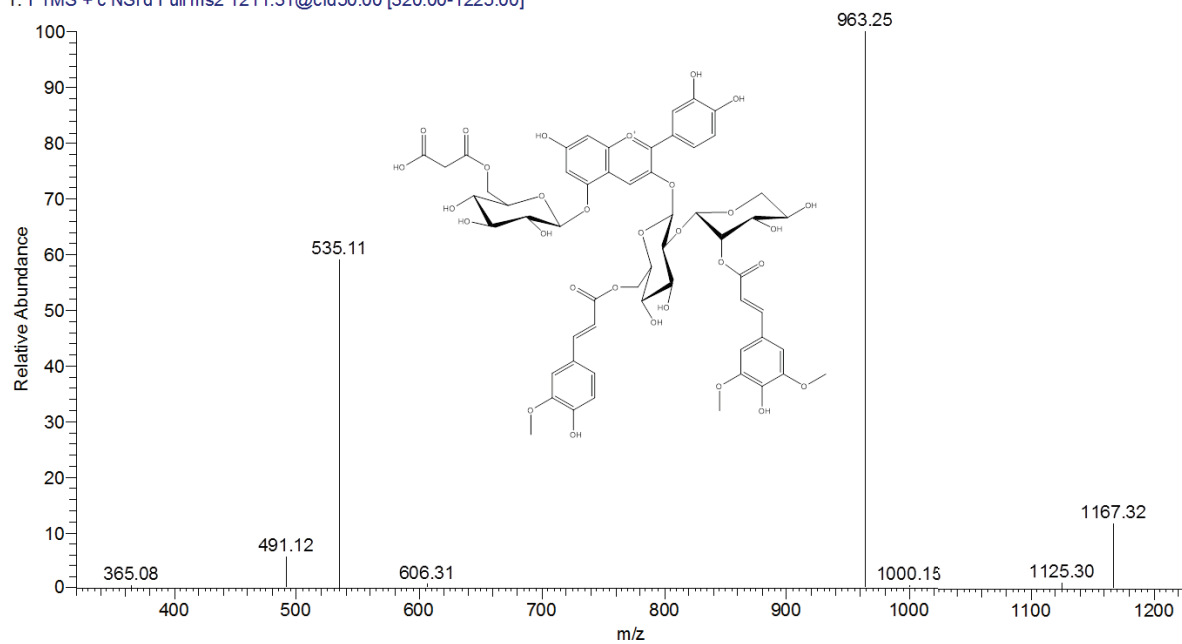
m/z 1197

Ara3A #1474 RT: 33.48 AV: 1 NL: 2.01E5
T: FTMS + c NSI d Full ms2 1197.29@cid50.00 [315.00-1210.00]



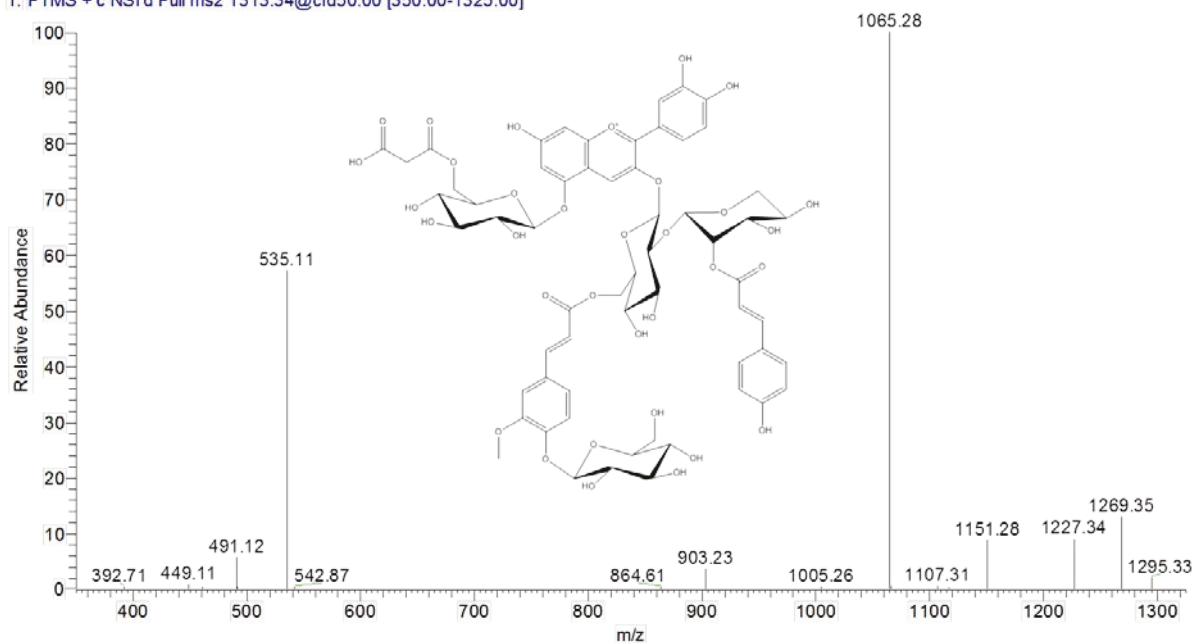
m/z 1211

Ara3A #1284 RT: 29.16 AV: 1 NL: 2.07E5
T: FTMS + c NSI d Full ms2 1211.31@cid50.00 [320.00-1225.00]



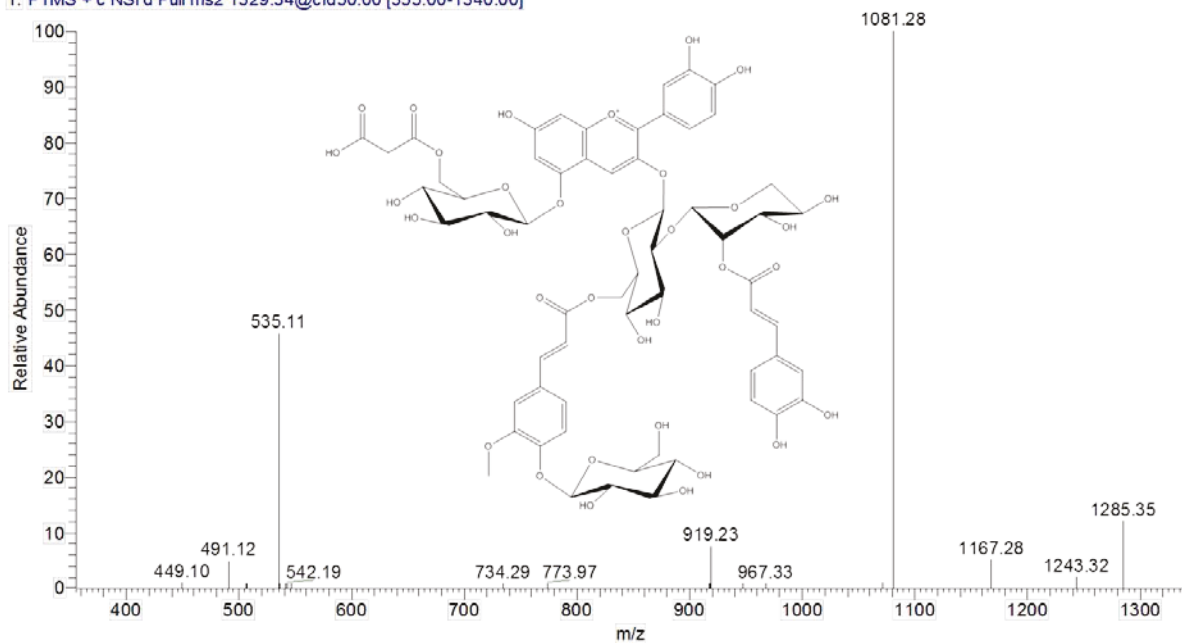
1313

Ara3B #1413 RT: 32.26 AV: 1 NL: 2.89E5
T: FTMS + c NSI d Full ms2 1313.34@cid50.00 [350.00-1325.00]



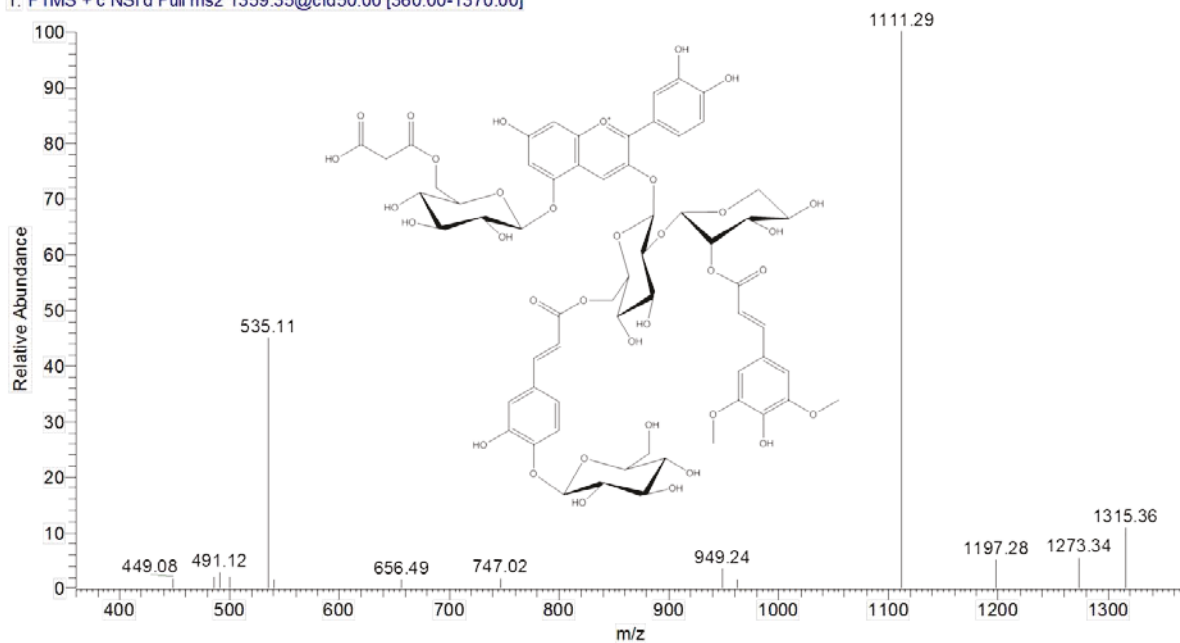
m/z 1329

Ara3A #1044 RT: 23.75 AV: 1 NL: 1.08E5
T: FTMS + c NSI d Full ms2 1329.34@cid50.00 [355.00-1340.00]



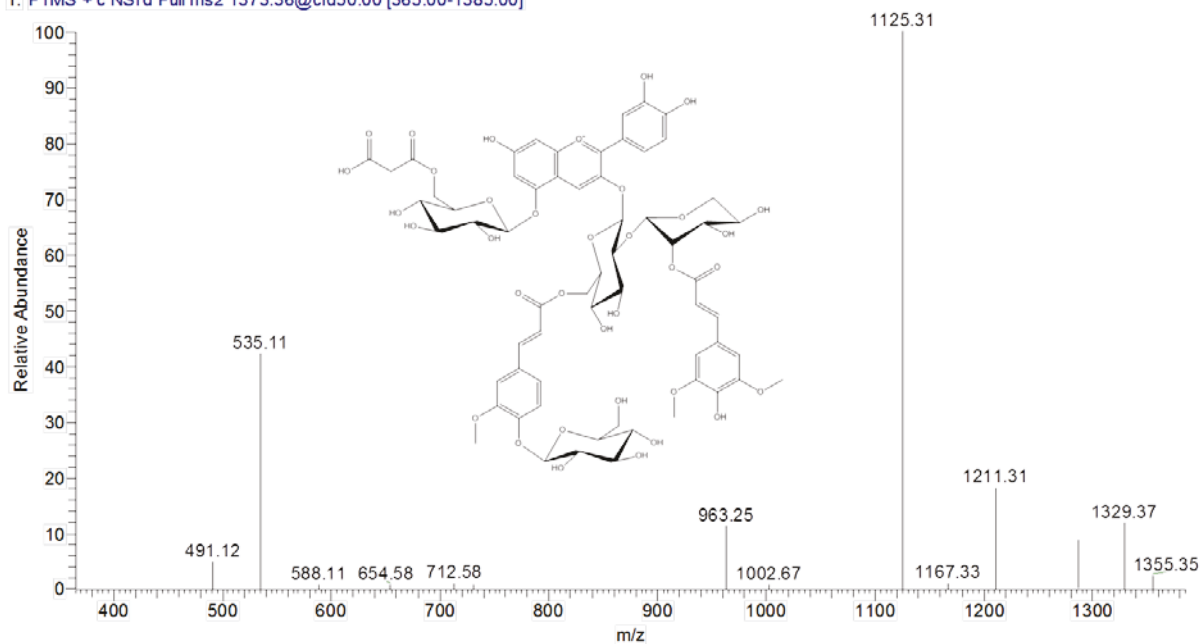
m/z 1359

Ara10C #1094 RT: 25.80 AV: 1 NL: 5.24E4
T: FTMS + c NSI d Full ms2 1359.35@cid50.00 [360.00-1370.00]



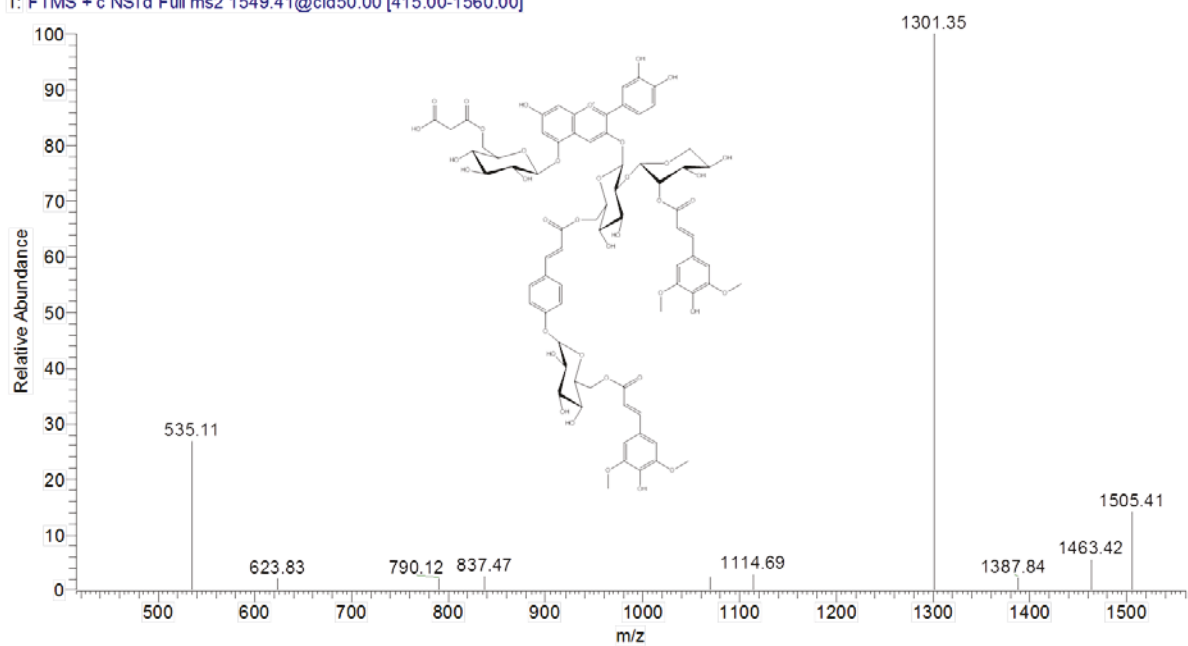
m/z 1373

Ara10C #1146 RT: 26.97 AV: 1 NL: 1.07E5
T: FTMS + c NSI d Full ms2 1373.36@cid50.00 [365.00-1385.00]



m/z 1549

Ara10A #1768 RT: 40.41 AV: 1 NL: 3.91E4
T: FTMS + c NSId Full ms2 1549.41@cid50.00 [415.00-1560.00]



Targeted MAPA for quantification of proteotypic peptides in complex shotgun proteomics samples – a case study on mature pollen of Tomato under heat stress

Palak Chaturvedi, Hannes Doerfler, M^a Angeles Castillejo, Stefanie Wienkoop, Volker Egelhofer and Wolfram Weckwerth*

Department of Molecular Systems Biology, Faculty of Sciences, University of Vienna, Althanstrasse 14, A-1090, Vienna, Austria

***Corresponding Author**

Prof. Dr Wolfram Weckwerth

Department of Molecular Systems Biology, Faculty of Sciences, University of Vienna, Althanstrasse 14, A-1090, Vienna, Austria

Phone: +43-6642067776550

Email: wolfram.weckwerth@univie.ac.at

Financial source:

Palak Chaturvedi is supported the by European Marie-Curie International training network “Solanaceae pollen thermotolerance SPOT-ITN”, grant agreement number 289220.

Abstract

Recently, we have developed a quantitative shotgun proteomics strategy called MAPA (mass accuracy precursor alignment)^{1, 2}. The MAPA algorithm uses high mass accuracy to bin mass-to-charge-ratios (m/z) of precursor ions from LC-MS analyses and to extract a quantitative sample versus m/z -ratio data alignment matrix from a multitude of samples. Here, we introduce a novel feature of this algorithm by allowing the extraction and alignment of proteotypic peptide precursor ions from complex shotgun proteomics data for accurate quantification of unique proteins. This strategy circumvents the problem of confusing the quantification of proteins due to undistinguishable protein isoforms by a typical shotgun proteomics approach^{3, 4}. We exemplified the approach by the comparison of control and heat-treated mature pollen. Pollen is a temperature-sensitive tissue involved in the reproductive cycle of plants and plays a major role for fruit setting and yield^{5, 6}. By applying the targeted MAPA data processing to a set of LC-MS analyses of control and heat-treated mature pollen samples 69 heat stress responsive proteins in mature pollen were identified using multivariate statistical analysis. Their potential function in heat stress protection is discussed.

43

44 **Introduction**

45 In the last decade shotgun proteomics has developed into one of the major technologies for high-
46 throughput proteomics^{7,8} and is a key discipline for functional genomics and systems biology⁹⁻¹¹. A
47 large problem for shotgun proteomics technology is the presence of protein isoforms with partially
48 identical primary amino acid sequences^{3,4,12,13}. If these regions have even identical tryptic cleavage
49 sites they cannot be distinguished by shotgun proteomics. Thus, unambiguous identification of
50 unique proteins is only guaranteed if so called proteotypic peptides are matched^{3,13-17}. Recently, we
51 developed a technique called MAPA (mass accuracy precursor alignment) which is able to extract and
52 align all precursor ions from a typical high mass accuracy shotgun proteomics analysis from a
53 multitude of samples^{2,18}. MAPA allows to extract all precursor ions independent on the
54 identification based on a genomic database and, thus, can also identify e.g. polymorphisms which are
55 not part of databases². The MAPA algorithm is freely available and can be downloaded¹. Recently
56 we introduced a novel feature: if a list of precursor ions is known they can be extracted from the
57 same dataset by using a targeted MAPA function¹. In the present study we apply this targeted MAPA
58 feature to extract all proteotypic peptides from a typical set of non-targeted shotgun proteomics
59 analyses. First, a list of proteotypic peptides is predicted from the genome database, and second, this
60 target list is uploaded to the MAPA software. Then, all LC-MS files are uploaded in mzXML format and
61 the proteotypic peptides are extracted and aligned in a data matrix. This process is very fast and can
62 be applied to a large set of samples as demonstrated in Hoehenwarter et al.^{2,13}. In the present

study, we have used a dataset of protein analysis in mature pollen isolated from plants grown under normal and heat stress conditions as an application of this novel MAPA feature.

Sexual reproduction in flowering plants is a very complex process and especially gametophytic development is highly sensitive to temperature fluctuations and other abiotic stresses such as drought, flooding and salinity, thus, controlling agricultural efficiency and production. We applied proteomic analysis to investigate developmental processes during pollen development⁵ and to understand how plants rapidly adapt to the fluctuating temperatures during their gametophytic developmental phase¹⁹. Gametophytic development takes place in male stamen and female pistil²⁰. Male gametophyte (or pollen grain) development is a well programmed process which includes elemental procedures such as cell polarity, cell cycle, regulation of gene expression and cell specification. This process can be divided into several distinct phases which lead to the formation of mature pollen. It takes place in the anther locules by the development of sporogenous tissue producing microsporocytes (or pollen mother cells) undergoing meiosis followed by asymmetric mitotic division (PM I) to produce bicellular pollen grain with two cells, a larger vegetative cell and smaller generative cell²¹. This whole developmental process occurs in a very short frame of time¹⁹. Mature pollen is an autonomous and highly simplified organism determined to disperse and transport male gametes to the female gametes where the fertilization process begins²².

Tomato (*Solanum lycopersicum* L.) is one of the most important food crop in the world and it shows highly negative response to extreme temperature having adverse effect on its reproductive and vegetative growth, which results into nearly 70% of loss in tomato production worldwide^{23, 24}. Nevertheless, heat stress has various effects which also depends upon genotypes, for example effect

of heat stress on heat-tolerant and heat sensitive tomato cultivars. Temperature stress most pronouncedly affects the development of the mature pollen grain, thereby reducing viability in heat sensitive cultivars²⁵.

Recent proteomic studies on tomato, Arabidopsis and rice reveal that mature pollen presynthesizes a large number of proteins which have predefined functions such as cell wall metabolism, energy metabolism, signaling, transport, cytoskeleton formation and others for the successful progress of fertilization^{5, 26-30}. Under harsh environmental conditions, there is simultaneous up- and downregulation of large numbers of genes/proteins³¹. Therefore it is necessary to acquire information about how these regulations are affected in mature pollen under stress conditions.

In the present study a highly sensitive shotgun proteomic approach (GEL-LTQ-Orbitrap MS³²) is employed to compare the proteome of tomato pollen under control and heat stress conditions. The quantification of the identified peptides was performed by applying targeted mass accuracy precursor alignment (MAPA) considering only “proteotypic peptides”. Subsequent multivariate statistical procedures were applied to identify putative marker proteins for stress responses. 69 proteins were assigned to heat stress conditions and their potential functions are discussed.

Experimental Procedure

Sampling and high temperature treatment

Tomato (ecotype *Red Setter*) plants were grown in the controlled greenhouse condition (12 h light and dark, 25-26°C during day time and 18-20°C at night). During the high temperature treatment the plants were transferred to the pre-warmed climatic chamber at 30°C/32°C. The chamber was then gradually heated until 38°C and plants were kept for an additional hour at 38°C before cooling down the chamber to 25°C. Harvesting of the flower buds was performed after 1h of the temperature stress treatment. Harvesting procedure and sampling of pollen were performed according to Chaturvedi et al.⁵. Flower buds were collected and anthers of individual buds were sampled in the 200 µl ice cold germination solution (2mM Boric acid, 2mM Calcium nitrate, 2mM Magnesium sulphate, 1mM Potassium nitrate) without sucrose. Anthers were squeezed open with gentle mechanical stress in the solution, vortexed and the supernatant containing pollen was collected in new tubes. Pollen was spun down at 100 x g for 1 min and washed twice with germination solution. Sampled pollen was analyzed under the microscope in order to confirm the maturity and to determine the quality of the preparation. Sufficient amount of pollen was used for protein extraction and further analysis. In addition, the reference dataset of controlled conditions published by Chaturvedi et al.⁵ is considered to compare the proteomes of mature pollen in two different conditions. Pollen for control conditions was sampled using 10% mannitol.

Protein Extraction, Prefractionation, Digestion, and LC-MS/MS

Protein extraction and analysis procedure is detailed by Chaturvedi et al.^{5, 11}. Proteins were extracted from freeze dried pollen pellets by grinding them for 2 min in a shaking mill using steel balls (~2 mm diameter). The homogenized pollen sample was disaggregated into 200 µl of protein extraction buffer (100 mM Tris- HCl, pH 8.0; 5% SDS, 10% glycerol; 10 mM DTT; 1% plant protease inhibitor cocktail (Sigma P9599)) and incubated for 5 min at room temperature followed by incubation for 2.5 min at 95°C and centrifugation at 21,000 x g for 5 min at room temperature. The supernatant was carefully transferred into a new tube. 200 µl of 1.4 M sucrose were added to the supernatant and proteins were extracted twice with 200 µl TE buffer equilibrated phenol followed by counter extraction with 200 µl of 1.4 M sucrose and 200 µl distilled water. Precipitation of proteins was carried out by adding 2.5 volumes of 0.1 M ammonium acetate in methanol to the combined phenol phases followed by incubation for 16h at -20°C. After incubation, samples were centrifuged for 5 min at 5000 x g. The pellet was washed twice with 0.1 M ammonium acetate, once with acetone, and air dried at room temperature. The pellet was re-dissolved in 6 M Urea and 5% SDS. Concentration of protein was determined by using the bicinchoninic acid assay.

Pre-fractionation of protein was carried out by SDS-PAGE. Approximately 40 µg of total protein were loaded onto a gel and run up to 1.5 cm. Gels were fixed and stained with in Methanol: Acetic Acid: Water: Coomassie Brilliant Blue R-250 (40:10:50:0.001). Gels were destained in methanol: water (40:60) and then each lane was divided into two fractions. Gel pieces were destained, equilibrated and digested with trypsin, desalted and concentrated. Prior to mass spectrometric measurement, the tryptic peptide pellets were dissolved in 4% (v/v) acetonitrile and 0.1% (v/v) formic acid. 10 µg of digested peptides were injected into a one-dimensional (1D) nano-flow LC-MS/MS system equipped

with a pre-column (Eksigent, Germany). Peptides were eluted using an Ascentis column (Ascentis Express, peptide ES-C18 HPLC column (SUPELCO Analytical, USA), dimension 15 cm x 100 µm, pore size 2.7 µm)) during an 80 min gradient from 5% to 50% (v/v) acetonitrile, 0.1% (v/v) formic acid. MS analysis was performed on an Orbitrap LTQ XL mass spectrometer (Thermo, Germany) with a controlled flow rate of 500 nL per minute. Specific tune settings for the MS were as follows: spray voltage was set to 1.8 kV; the temperature of the heated transfer capillary was set to 180 °C. Each full MS scan was followed by 10 MS/MS scans, in which the 10 most abundant peptide molecular ions were dynamically selected, with a dynamic exclusion window set to 90 s.

Peptide and Protein Identification

Raw data were searched with the SEQUEST algorithm present in Proteome Discoverer version 1.3 (Thermo, Germany) as described in Valledor & Weckwerth,³². Identification confidence was set to a 5% FDR and the variable modifications were set to: acetylation of N-terminus, oxidation of methionine and carbamidomethyl cysteine formation, with a mass tolerance of 10 ppm for the parent ion and 0.8 Da for the fragment ion.

The Tomato protein database was employed (ITAG 2.0 protein fasta, Sol genomic network). Peptides were matched against these databases plus decoys, considering a significant hit when the peptide confidence was high and an Xcorr threshold was established at 1 per charge (2 for +2 ions 3 for +3 ions etc.). All the spectra of the identified proteins and their meta-information such as spectral filtering, thresholds, cell-specificity are stored in the public plant proteomics databases PROMEX (<http://promex.pph.univie.ac.at/promex/>) with the searchable experiment ID Lyc escu003.

Targeted Mass Accuracy Precursor Alignment (MAPA) for quantification of proteotypic peptides

The Xcalibur raw files were converted to mzXML format with MassMatrix MS Data File Conversion v3.9 (<http://www.massmatrix.net/mm-cgi/downloads.py>). A target list of the m/z-ratios of all "proteotypic peptides" was prepared either from the full genome database ITAG 2.0 or from the output file of Proteome Discoverer (*File type: CSV format*). These m/z ratios were cut to the second decimal and uploaded within the ProtMax program. Preferences and settings were chosen according to Egelhofer et al., 2013¹, with slight modifications. The main employed ProtMax settings were: target list (method), intensity (quantification), cut after 2 decimals. Accepted charge states included +2, +3,+4,+5; "unite neighbors" option was active. This program is a windows forms application in the Common Language Runtime (CLR) environment. It can be downloaded from <http://www.univie.ac.at/mosys/software.html>¹.

Multivariate statistical and bioinformatic data analysis

Principal component analysis (PCA), box plots and ANOVA were performed using the statistical tool box COVAIN³³. For functional categorization of the identified proteins we exploited the Mapman mapping file *Solanum lycopersicum* (<http://mapman.gabipd.org/web/guest/mapmanstore>). The Venn diagram was produced by Venny (<http://bioinfogp.cnb.csic.es/tools/venny/index.html>). Subcellular locations were predicted using pSORT Wolf software (<http://wolfsort.seq.cbrc.jp/>),

183 TargetP (<http://www.cbs.dtu.dk/services/TargetP/>) and Predotar
184 (<http://urgi.versailles.inra.fr/predotar/predotar.html>).

185

Results and Discussion

Targeted MAPA for the quantification and subsequent multivariate statistics of proteotypic peptides in typical shotgun proteomics data

The workflow of the targeted MAPA strategy is shown in Figure 1. Extracted protein from pollen (See Experimental procedure) was digested with trypsin and peptides were analysed using nanoUPLC coupled to an Orbitrap LTQ XL mass spectrometer (Thermo, Germany). All the raw files obtained from control and heat stress samplings were searched with the SEQUEST algorithm of Proteome Discoverer version 1.3 (Thermo, Germany) as described by Valledor and Weckwerth³². In total 1633 proteins were identified (Supporting information 1).

Comparison and identification of potential heat stress related protein candidates was performed using targeted MAPA. A strategy was employed in which a target list of “proteotypic peptides” with their corresponding m/z ratio was prepared using the output file from proteome discoverer (*File type*: CSV format). It consisted of 4076 m/z precursors (Supporting information 2). This precursor ion list was uploaded to ProtMax (a software based on the concept of MAPA^{1,2}) along with the raw files converted into mzXML format. The output matrix in excel-format consisted of accurately measured 2896 m/z ratios as well as their corresponding retention times (Rt) and the scan-number which corresponded to the most intense MS signal from which an MS/MS event was triggered¹ (Supporting information 3). All the identified proteins, their respective peptide spectra and detailed information can be found online in the plant proteomics database PROMEX

(<http://promex.pph.univie.ac.at/promex/>) with the experiment ID Lyc escu003. Multivariate statistical analysis of the data matrix generated by ProtMax using principal component analysis (PCA) separated the two conditions by the first principal component PC1 (90.14%) which explains the strongest variation (Supporting information 4, Figure 2). Multivariate statistics (PCA) was performed to reduce the dimensionality of the data and to determine interesting m/z precursor ions. Further details of the basic principles of MAPA – extraction of precursor ions from raw-data and subsequent multivariate statistics to rank the precursor ion list according to their impact in sample separation – can be found in Hoehenwarter et al. 2008 ². After PCA, the highest positive and negative loadings from principal component 1 (PC1) were compared with the proteins identified in the total data set which is illustrated by a Venn diagram (Figure 3).

Comparison of targeted MAPA versus NSAF

Multivariate statistics were performed to reduce the dimensionality of the data and to rank m/z precursor masses extracted with ProtMAX according to Hoehenwarter et al ². Two hundred of the most significant positive and negative loadings from PC1 were then compared with the proteins identified with Proteome Discoverer® which is illustrated by a Venn diagram (Figure 3). Figure 3 shows the overlap of the identified proteins using the typical database search implemented in Proteome Discoverer® with the highest ranking proteins which were identified with the targeted MAPA approach. There is a large number of identified proteins which are different between control and heat stress treatment of mature pollen. 477 proteins are potential candidates for heat stress but they represent rather protein groups instead of individual proteins because identical tryptic peptides

in different protein isoforms cannot be distinguished¹³. By including all non-proteotypic peptides in the analysis ambiguous identification due to identical peptide sequences or repeated peptides due to highly homologous protein isoforms cannot be excluded^{3, 13}. Furthermore, the quantification approach can be problematic because summing up of these non-proteotypic peptides in total abundance scores of a protein can also lead to ambiguous quantification. Using the targeted MAPA approach, only proteotypic peptides are extracted from the complex raw-data. This strategy circumvents the problems of identification and quantification ambiguity as described above. In Figure 3, tryptic peptide precursor ions which were identified and quantified with targeted MAPA and showed a strong influence on the separation of control and heat stressed samples in PCA (Figure 2), are overlapped with the total number of identified proteins. 69 unique proteins are identified as potential heat stress marker. These 69 proteins are unambiguously identified by the targeted MAPA approach and further confirmed by matching to the total list of the 477 ambiguous protein identifications. A protein which was exclusively identified by this approach is e.g. a MAPKK protein (Solyc03g123800) (for a full list of the 69 proteins and their proteotypic peptides see supplementary information 1). This protein was identified by a specific unique proteotypic peptide by the MAPA approach as shown in figure 4.

Functional Categorization of the Heat stress proteome

Based on the comparative quantification approach using targeted MAPA, 69 unique proteins were identified which potentially could define protein marker for heat stress (Figure 3). These 69 proteins

were categorized into functional groups (Figure 5A) based on functional annotation by matching the identical proteins with Mapman files of *Solanum lycopersicum* (ITAG 2.3). More than 50% of the identified proteins represented three main groups: Protein synthesis (17%), Protein degradation (7%) and hypothetical proteins with unknown function (32%). The other groups included proteins involved in metabolism, stress, redox activity, signaling, transport, post translational modification and other functional groups (Supporting information 5; Supplementary information 1). Furthermore, for functional annotation, orthologs of these proteins were identified in *Arabidopsis thaliana* and rice (Table 1). We also considered the transcriptomic data of tomato flowers of *S. lycopersicum* (GEO Id: GSE33507) published in a study of Pietrella et al ³⁴ (see Table 1). The list of proteotypic peptides and subcellular location of these identified proteins is presented in Table 2. In Figure 5B a bi-clustering analysis of the functional categories is shown of the control versus the heat stress pollen samples. From the export of the targeted MAPA analysis, all proteins were assigned into functional groups and a bi-clustering analysis using the COVAIN toolbox for OMICS data integration and statistics was performed ³³. The heat-treated samples are separated from the controls. Several functional protein groups such as “stress abiotic, redox, DNA-repair, polyamine metabolism, ribosomal proteins, protein folding and cell organization” are highly upregulated relative to the control samples. This indicates that the response of pollen to heat is a multifactorial phenomenon and cannot be reduced to some few processes. Under heat stress conditions, regulation of proteins is determined by protein synthesis and protein degradation. At higher temperatures there is a decline in normal protein synthesis (see Figure 5B), together with an increase in the selective translation of mRNAs for characteristic sets of heat shock proteins (HSPs) ³⁵. In the present analysis, many proteins involved in translational activity have been identified, such as cytochrome b, 60S ribosomal protein (L4-B, L22-2),

271 eukaryotic translation initiation factor (eIF-4GB,eIF- 4G-1, eIF-3A), elongation factor Ts. Interestingly,
272 mature pollen of tomato do not synthesize new heat shock proteins under heat stress condition but
273 rather utilizes presynthesized HSPs ^{5, 6}. Therefore, these proteins might accelerate synthesis of many
274 proteins in mature pollen which can protect and help them to combat stress (Supporting information
275 5).

276 Protein degradative processes are often accelerated under heat stress condition leading to tissue
277 senescence or death ³⁶. Proteases, enzymes of ubiquitin/proteasome pathway and other proteolytic
278 enzymes mediate protein degradation ³⁷. Some of the proteins categorized in these function and
279 identified in heat stress mature pollen were proteasome subunit alpha type, ubiquitin-conjugating
280 enzyme E2 N and SKP-1 (Supporting information 5). S-phase kinase associated protein 1 (SKP1) is the
281 major unit in the SCF complex; this complex is one of the typical multimeric E3 (Ub Conjugation)
282 complexes which plays an important role in protein degradation ³⁸. The pollen-specific SKP1 genes
283 LSK1-LSK3 were also identified in *Lilium longiflorum* pollen and in *Antirrhinum hispanicum*
284 pollen (AhSSK1) specifically expressed in late pollen developmental stages and in pollen tube
285 elongation ^{39, 40}.

286 In our study we also identified some important heat stress marker protein candidates such as late
287 embryogenesis abundant protein (LEA) which takes part in stress defense mechanisms, by protecting
288 proper folding and conformation of both structural and functional proteins ⁴¹. Expression of this
289 protein is assumed to link with desiccation tolerance in seeds, pollen and anhydrobiotic plants ⁴².
290 Under desiccating conditions like heat and drought, this protein protects citrate synthase and
291 prevents protein aggregation⁴³. LEA proteins were identified in embryogenesis of cotton seeds and

292 maturation of Arabidopsis seeds ^{44, 45} (Supporting information 5). It is also reported that
293 overexpressed HVA 1, group 3 of LEA proteins from barley (*Hordeum vulgare* L.), conferred
294 dehydration tolerance in transgenic rice plant ⁴⁶.

295 Several proteins required to protect against oxidative stress were identified in our work (see also
296 Figure 5B). These were serine/threonine-protein phosphatase and superoxide dismutase (Cu/Zn-
297 SOD) which have been shown to be defensive in pollen metabolism against reactive oxygen species
298 (Supporting information 5). Another interesting candidate is a MAPKK (mitogen-activated protein
299 kinase kinase) protein (see Figure 4). MAPKK is an important protein involved in signal transduction
300 and the identified isoform might be involved in heat stress protection ⁴⁷. Many studies have been
301 carried out in Arabidopsis reporting the role of MAPKs in temperature stress ⁴⁸ (Supporting
302 information 5).

303 Quantitatively we identified some of the interesting protein candidates that showed increased levels
304 in heat stress condition compared to control (Reference data set). The regulation of these proteins
305 was represented by box plots, which were performed using the statistical tool box COVAIN
306 (Supporting information 6; Supplementary Figure 1; Figure 5).

307 Membrane fluidity has direct and immediate effect of heat stress which leads to the induction of
308 calcium influx and cytoskeletal rearrangement triggering downstream signaling cascades leading for
309 the upregulation of MAPK and CDPK ⁴⁹. We have identified a calcium binding protein Calnexin
310 (Figure 5 A) and Calreticulin2 (Figure 5 B) strongly increased in heat stress condition. Calreticulin and
311 calnexin are chaperons and Ca²⁺-binding proteins which are localized in the endoplasmic reticulum

312 acting in protein folding and control Ca^{2+} homeostasis⁵⁰. These proteins might help in rapid signaling
313 leading to production of many antioxidants to maintain thermostability in pollen.

314 Furthermore, as an important protein marker, late embryogenesis abundant protein (LEA) has shown
315 increased regulation under heat stress condition in mature pollen of tomato compared to control
316 conditions (Figure 5 C).

317 Plant hexokinase (HXK) has been shown to be involved in sugar sensing and signaling. This dual-
318 function enzyme has both catalytic and regulatory functions^{51, 52}. In this analysis it was observed that
319 hexokinase 1, which is categorized in carbohydrate metabolism, regulates differently in both the
320 conditions (Figure 5 D). Under heat stress condition it shows greater variability compared to control.
321 Hexokinase catalyzes two free hexoses to enter the pool of hexoses phosphate, which consists of
322 glucose-1-phosphate, glucose-6-phosphate and fructose-6-phosphate. The dynamics of the pool are
323 maintained by phosphoglucomutase and phosphoglucose isomerase, which are identified in mature
324 pollen of rice²⁹.

325 326 **Conclusion**

327 In this study, we present an approach to perform quantification of the identified proteins/ peptides
328 using targeted MAPA considering a target list of “proteotypic peptides” (4096 m/z ratios). In total,
329 2896 m/z features and their corresponding proteins were identified and quantified using multivariate
330 statistical analysis like PCA. Further, these identified proteins were compared with 1636 identified
331 proteins obtained from the SEQUEST algorithm.

332 This approach of quantification leads to the identification of 69 unique proteins which potentially
333 could define protein marker for heat stress. Functional analysis of these potential candidates reveals

334 their involvement in the process of protein synthesis and protein degradation which might protect
335 mature pollen and help them survive harsh temperature conditions. Some important heat stress
336 responsive proteins were identified, like LEA, MAPKK, ubiquitin, chaperone, and others, which might
337 have a protective function in pollen metabolism. Further proteins showing strong regulation in two
338 different temperature conditions (control and heat stress) were also determined. It was observed
339 that some proteins belonged to the calcium signaling pathways like calnexin and calreticulin 2. Other
340 proteins such as hexokinase 1 strongly accumulated under heat stress condition compared to
341 control, indicating the impairment of sugar metabolism in heat stressed mature pollen.

342 Altogether, this approach provides a first reference set and interesting protein candidates of tomato
343 mature pollen under heat stress condition. Peptide spectra of the identified proteins and their
344 detailed information can be reviewed online in the plant proteomics database PROMEX
345 (<http://promex.pph.univie.ac.at/promex/>). Future studies will focus on determining the proteome of
346 early stages of pollen development under heat stress conditions.

347

- 351 1. Egelhofer, V.; Hoehenwarter, W.; Lyon, D.; Weckwerth, W.; Wienkoop, S., Using ProtMAX to create
352 high-mass-accuracy precursor alignments from label-free quantitative mass spectrometry data generated in
353 shotgun proteomics experiments. *Nature protocols* **2013**, 8, (3), 595-601.
- 354 2. Hoehenwarter W, v. D. J., Wienkoop S, Steinfath M, Hummel J, Erban A, Sulpice R, Regierer B, Kopka J,
355 Geigenberger P, Weckwerth W., A rapid approach for phenotype-screening and database independent
356 detection of cSNP/protein polymorphism using mass accuracy precursor alignment. *Proteomics* **2008**, 8, 4214–
357 4225.
- 358 3. Hoehenwarter, W.; Chen, Y.; Recuenco-Munoz, L.; Wienkoop, S.; Weckwerth, W., Functional analysis
359 of proteins and protein species using shotgun proteomics and linear mathematics. *Amino Acids* **2011**, 41, (2),
360 329-341.
- 361 4. Jungblut, P. R., The proteomics quantification dilemma. *J Proteomics* **2014**, 107, 98-102.
- 362 5. Chaturvedi, P.; Ischebeck, T.; Egelhofer, V.; Lichtscheidl, I.; Weckwerth, W., Cell-specific analysis of the
363 tomato pollen proteome from pollen mother cell to mature pollen provides evidence for developmental
364 priming. *J Proteome Res* **2013**, 12, (11), 4892-4903.
- 365 6. Bokszczanin, K. L.; Network, S. P. T. I. T.; Consortium, S. F., Perspectives on deciphering mechanisms
366 underlying plant heat stress response and thermotolerance. *Frontiers in plant science* **2013**, 4.
- 367 7. Washburn, M. P.; Wolters, D.; Yates, J. R., Large-scale analysis of the yeast proteome by
368 multidimensional protein identification technology. *Nat Biotechnol* **2001**, 19, (3), 242-247.
- 369 8. Wolters, D. A.; Washburn, M. P.; Yates, J. R., An automated multidimensional protein identification
370 technology for shotgun proteomics. *Anal Chem* **2001**, 73, (23), 5683-5690.
- 371 9. Sabido, E.; Selevsek, N.; Aebersold, R., Mass spectrometry-based proteomics for systems biology. *Curr*
372 *Opin Biotechnol* **2012**, 23, (4), 591-7.
- 373 10. Weckwerth, W., Green systems biology - From single genomes, proteomes and metabolomes to
374 ecosystems research and biotechnology. *J Proteomics* **2011**, 75, (1), 284-305.
- 375 11. Cravatt, B. F.; Simon, G. M.; Yates, J. R., 3rd, The biological impact of mass-spectrometry-based
376 proteomics. *Nature* **2007**, 450, (7172), 991-1000.
- 377 12. Schlüter, H.; Jungblut, P. R., Special issue on protein species. *Amino Acids* **2011**, 41, (2), 217-218.
- 378 13. Hoehenwarter, W.; Larhlmi, A.; Hummel, J.; Egelhofer, V.; Selbig, J.; van Dongen, J. T.; Wienkoop, S.;
379 Weckwerth, W., MAPA distinguishes genotype-specific variability of highly similar regulatory protein isoforms
380 in potato tuber. *J Proteome Res* **2011**, 10, (7), 2979-2991.
- 381 14. Lehmann, U.; Wienkoop, S.; Tschoep, H.; Weckwerth, W., If the antibody fails—a mass Western
382 approach. *The Plant Journal* **2008**, 55, (6), 1039-1046.
- 383 15. Wienkoop, S.; Weckwerth, W., Relative and absolute quantitative shotgun proteomics: targeting low-
384 abundance proteins in *Arabidopsis thaliana*. *J Exp Bot* **2006**, 57, (7), 1529-1535.
- 385 16. Wienkoop, S.; Weiß, J.; May, P.; Kempa, S.; Irgang, S.; Recuenco-Munoz, L.; Pietzke, M.; Schwemmer,
386 T.; Rupprecht, J.; Egelhofer, V., Targeted proteomics for *Chlamydomonas reinhardtii* combined with rapid
387 subcellular protein fractionation, metabolomics and metabolic flux analyses. *Molecular BioSystems* **2010**, 6,
388 (6), 1018-1031.
- 389 17. Picotti, P.; Bodenmiller, B.; Mueller, L. N.; Domon, B.; Aebersold, R., Full dynamic range proteome
390 analysis of *S. cerevisiae* by targeted proteomics. *Cell* **2009**, 138, (4), 795-806.
- 391 18. Chen, Y.; Hoehenwarter, W.; Weckwerth, W., Comparative analysis of phytohormone-responsive
392 phosphoproteins in *Arabidopsis thaliana* using TiO₂-phosphopeptide enrichment and mass accuracy precursor
393 alignment. *The Plant Journal* **2010**, 63, (1), 1-17.

19. Kelly E. Zinn, M. T.-O., Jeffrey F. Harper, Temperature stress and plant sexual reproduction: uncovering the weakest links. *J Exp Bot* **2010**, 61 (7), 1959-1968.
20. RB, G., Plants: novel developmental processes. *Science* **1988**, 240(4858), 1460-1467.
21. Borg, M.; Brownfield, L.; Twell, D., Male gametophyte development: a molecular perspective. *J Exp Bot* **2009**, 60, (5), 1465-1478.
22. Monica A. Grobei, E. Q., Erich Brunner, Hubert Rehrauer, Runxuan Zhang, Bernd Roschitzki, Konrad Basler, Christian H. Ahrens, Ueli Grossniklaus, Deterministic protein inference for shotgun proteomics data provides new insights into Arabidopsis pollen development and function. *Genome Res* **2009**, 19, 1786-1800.
23. Pressman E, P. M., Pharr DM, The effect of heat stress on tomato pollen characteristics is associated with changes in carbohydrate concentration in developing anthers. *Ann Bot* **2002**, 90(5), 631-636.
24. Sato S, P. M., Gardner RG, Altered flower retention and development patterns in 9 tomato cultivars under elevated temperature. *Hort Sci* **2004**, 101, 95-101.
25. Levy A, R. H., Kedar N, Morphological and physiological characters affecting flower drop and fruit set of tomatoes at high temperature. *Euphytica* **1978**, 27(1), 211-218.
26. Sheoran, I. S.; Ross, A. R.; Olson, D. J.; Sawhney, V. K., Proteomic analysis of tomato (*Lycopersicon esculentum*) pollen. *J Exp Bot* **2007**, 58, (13), 3525-35.
27. Holmes-Davis R, T. C., Vensel WH, Hurkman WJ, McCormick S, Proteome mapping of mature pollen of *Arabidopsis thaliana*. *Proteomics* **2005**, 5, 4864-4884.
28. Noir S, B. u. A., Colby T, Schmidt J, Panstruga R., Anreference map of the Arabidopsis thaliana mature pollen proteome. *Biochem Bioph Res Co* **2005**, 337, 1257-1266.
29. Dai, S.; Li, L.; Chen, T.; Chong, K.; Xue, Y.; Wang, T., Proteomic analyses of Oryza sativa mature pollen reveal novel proteins associated with pollen germination and tube growth. *Proteomics* **2006**, 6, (8), 2504-29.
30. Ischebeck, T.; Valledor, L.; Lyon, D.; Gingl, S.; Nagler, M.; Meijón, M.; Egelhofer, V.; Weckwerth, W., Comprehensive cell-specific protein analysis in early and late pollen development from diploid microsporocytes to pollen tube growth. *Mol. Cell. Proteomics* **2014**, 13, (1), 295-310.
31. Ahsan N, D. T., Nouri M, Komatsu S, Tissue-specific defense and thermo adaptive mechanisms of soyabean seedlings under heat stress revealed by proteomic approach. *J Proteome Res* **2010**, 9, 4189-4204.
32. Valledor, L.; Weckwerth, W., An Improved Detergent-Compatible Gel-Fractionation LC-LTQ-Orbitrap workflow for plant and microbial proteomics. *Method in Molecular Biology* **2013**, in press.
33. Sun, X.; Weckwerth, W., COVAIN: a toolbox for uni- and multivariate statistics, time-series and correlation network analysis and inverse estimation of the differential Jacobian from metabolomics covariance data. *Metabolomics* **2012**, 8, (1), 81-93.
34. Pietrella M, G. G., Peano C, Fuligni F, De Bellis G, Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **2012**, 485(7400), 635-41.
35. Gurley WB, K. J., Transcriptional regulation of the heat-shock response: a plant perspective. *Biochemistry*. **1991**, 30(1), 1-12.
36. RC, H., Proteolytic activity during senescence of plants. *New Phytol.* **1990**, 116, 199-231.
37. A, S., A cut above the rest; the regulatory function of plant protease. *Planta* **2004**, 220, 183-197.
38. Pickart, C. M., Mechanisms underlying ubiquitination. *Annu. Rev. Biochem* **2001** 70, 503 - 533
39. Liang-Chi Chang, C.-L. G., Yun-Shan Lin, Hongyong Fu, Co-Shing Wang, Guang-Yuh Jauh, Pollen-Specific SKP1 like Proteins are Components of Functional SCF Complexes and Essential for Lily Pollen Tube Elongation. *Plant Cell Physiology* **2009**, 50, (8), 1558-1572.
40. Huang J, Z. L., Yang Q, Xue Y, AhSSK1, a novel SKP1 like protein that interacts with the S-locus F-box protein SLF. *Plant J* **2006**, 46, 780-793.
41. Basia Vinocur, A. A., Recent advances in engineering plant tolerance to abiotic stress: achievements and limitations

- Curr. Opin. Biotechnol.* **2005**, 16, 123-132.
42. Welin BV, O. A., Nylander M, Palva ET, Characterization and differential expression of dhv/lea/rab-like genes during cold acclimation and drought stress in *Arabidopsis thaliana*. *Plant Mol Biol* **1994**, 26, 131–144.
43. Goyal K, W. L., Tunnacliffe A, LEA proteins prevent protein aggregation due to water stress. *Biochem. J.* **2005**, 388, 51–157.
44. Dure L 3rd, G. S., Galau GA, Developmental biochemistry of cottonseed embryogenesis and germination: changing messenger ribonucleic acid populations as shown by in vitro and in vivo protein synthesis. *Biochemistry-Us* **1981**, 7 (14), 4162–4168.
45. Michel Delseny, N. B.-E., Christel Carles, Gillian Hull, Carlos Vicient, Monique Raynal, Françoise Grellet,; Aspart, L., Late Embryogenesis Abundant (LEA) protein gene regulation during *Arabidopsis* seed maturation. *J. Plant Physiol.* **2001**, 158, 419–427.
46. Chandrababu R, Z. J., Blum A, Hod THD, Wue R, Nguyen HT,, HAV1, a LEA gene from barley confers dehydration tolerance in transgenic rice (*Oryza sativa* L.) via cell membrane protection. *Plant sci* **2004**, 166, 85-862.
47. Alok Krishna Sinha, M. J., Badmi Raghuram, Narendra Tuteja, Mitogen-activated protein kinase signaling in plants under abiotic stress. *Plant Signaling & Behavior* **2011**, 6 (2) 196-203.
48. Kazuya Ichimura, T. M., Riichiro Yoshida, Takashi Yuasa, Kazuo Shinozaki, Various abiotic stresses rapidly activate *Arabidopsis* MAP kinases ATMPK4 and ATMPK6. *The Plant Journal* **2000**, 24(5), 655-665.
49. A. Wahid , S. G., M. Ashraf , M.R. Foolad, Heat tolerance in plants: An overview. *Environ. Exp. Bot.* **2007**, 61, 199–223.
50. LEV, D. B., The evolutionary history of calreticulin and calnexin genes in green plants. *Genetica* **2011**, 139 (2), 225–229.
51. Jang, J.-C. a. S., J. , Sugar sensing in higher plants. *Plant Cell* **1994**, 6, 1665–1679.
52. Jang, J.-C., León, P., Zhou, L. and Sheen, J., Hexokinase as a sugar sensor in higher plants. *Plant Cell* **1997**, 9, 5–19.

477

478

479 **Supporting information**

480 **Supporting information 1:** Details of pollen analysis, protein quantification by NSAF.

481 **Supporting information 2:** Details of identified proteotypic peptides

482 **Supporting information 3:** ProtMax Output

483 **Supporting information 4:** PCA loadings mature pollen in control and heat stress condition.

484 **Supporting information 5:** Details of functional analysis

485 **Supporting information 6:** Details of box plot analysis

486 **Supporting information 7:** Analysis of significantly different proteins

487 **Supplementary information 1:** Sequences of heat stress marker proteins identified in this analysis

488 **Supplementary Figure 1:** Box plots of interesting protein candidates that showed increased levels in

489 heat stress condition compared to control (Reference data set)

490

491

492

493

494

495

496

497

498 **Table 1** Functional annotation of heat stress responsive proteins in mature tomato pollen as well as
 499 their corresponding orthologs in *Arabidopsis thaliana* and rice. The data are also compared to
 500 transcriptomic data of tomato flower of *S. lycopersicum* (GEO Id: GSE33507) published in a study
 501 Pietrella et al ³⁴.

Solyc ID	Description	Transcriptom ic data of Tomato mature flower	Arabidopsis Id	Rice Id
Solyc08g075690	Acyl-CoA-binding protein	Yes	AT4G01400	LOC_Os08g06550
Solyc08g066820	Choline/ethanolamine kinase	Yes	AT2G26830,AT4G0 9760,AT1G71697	LOC_Os09g26700
Solyc04g081400	Hexokinase 1	Yes	AT1G47840	LOC_Os07g09890
Solyc06g007250	Leucine-rich repeat family protein	No	AT2G30105	LOC_Os10g31790
Solyc09g009180	Chaperone protein dnaJ	Yes	AT1G79940,AT4G2 1180	LOC_Os04g24180
Solyc03g025340	C2 domain-containing protein	Yes		No
Solyc08g079260	Serine/threonine-protein phosphatase (Fragment)	Yes	AT4G22670	LOC_Os02g01030
Solyc01g067740	Superoxide dismutase	Yes	AT1G08830, AT2G28190	LOC_Os07g46990
Solyc10g080890	3-oxoacyl-reductase	Yes	AT3G55290,AT3G4 6170,AT2G17845,A T3G55310,AT1G62 610, AT1G63380	LOC_Os08g39960
Solyc01g049680	Transcriptional activator protein Pur- alpha	Yes	AT2G32080	LOC_Os01g15600
Solyc05g005510	DNA-binding related protein (Fragment)	Yes		LOC_Os01g13700, LOC_Os01g55030
Solyc07g064510	Polyadenylate-binding protein 2	Yes	AT5G19350,AT1G1 1650,AT4G27000, AT5G54900	LOC_Os04g53440
Solyc01g090190	Nuclear RNA binding protein	Yes	AT4G16830	LOC_Os05g51180
Solyc02g021560	Heterogeneous nuclear ribonucleoprotein A3-like protein 2	Yes	AT3G15010	No
Solyc08g048500	Nucleic acid binding protein	Yes	AT5G46190,AT5G0 4430,AT4G18375,A T1G14170, AT2G22600	LOC_Os08g10030
Solyc10g009220	Polyadenylate-binding protein 4-like	Yes	AT4G00830	LOC_Os10g06130
Solyc06g082100	Related to ATP dependent RNA helicase	Yes	AT1G20960, AT2G42270	LOC_Os02g01740, LOC_Os03g53220, LOC_Os03g11470
Solyc01g103690	ATP-dependent RNA helicase A	Yes	AT1G20960, AT2G42270	LOC_Os04g35260
Solyc08g006030	Unknown Protein	Yes	AT4G31340,	LOC_Os09g32250

			AT2G24420	
Solyc03g044030	Pre-mRNA-splicing factor SYF1	No		No
Solyc00g047200	Cytochrome b	No		No
Solyc10g084350	60S ribosomal protein L4-B (L1)	Yes		No
Solyc01g099830	60S ribosomal protein L22-2	Yes	AT3G05560,AT5G27770,AT1G02830	LOC_Os03g22340
Solyc03g044000	Ribosomal protein L26	Yes	AT3G49910,AT5G67510	LOC_Os01g04730
Solyc01g103510	Ribosomal protein L3	Yes	AT1G61580,AT1G43170	LOC_Os12g07010, LOC_Os11g06750
Solyc07g009330	60S acidic ribosomal protein P3	Yes	AT5G57290,AT4G25890	LOC_Os01g13080
Solyc01g102590	Eukaryotic translation initiation factor 3 subunit 8-like protein	Yes		LOC_Os07g03230
Solyc00g007270	Translation initiation factor	Yes	AT3G55620	LOC_Os07g44620
Solyc01g111200	Eukaryotic translation initiation factor 4B	Yes	AT4G38710	LOC_Os02g24330
Solyc08g077700	Eukaryotic translation initiation factor 4 gamma 1	Yes		LOC_Os07g36940
Solyc09g010460	Eukaryotic translation initiation factor 3 subunit A	yes	AT4G11420	LOC_Os01g03070
Solyc11g062070	Outer envelope membrane protein	No	AT5G19620	LOC_Os02g10260
Solyc12g099820	Signal recognition particle protein	Yes	AT3G49100	LOC_Os06g14550
Solyc06g065170	Vacuolar protein sorting protein	Yes	AT2G05170	LOC_Os04g31390
Solyc03g123800	Mitogen-activated protein kinase	Yes	AT3G21220,AT5G40440, AT4G26070,AT5G56580,AT1G51660	LOC_Os06g09180
Solyc02g081700	Proteasome subunit alpha type	Yes	AT3G22110	LOC_Os06g07978, LOC_Os06g07140
Solyc11g042930	SKP1	Yes	AT1G75950,AT5G42190,AT2G03160,AT3G21850,AT2G03190,AT2G03170,AT2G25700,AT3G25650,AT3G60020,AT1G20140,AT3G60010,AT3G21830,AT4G34210,AT4G34470,AT1G10230,AT3G21860,AT2G20160	LOC_Os11g26910
Solyc11g008400	SKP1-like protein	No	AT1G75950,AT5G42190,AT2G03160,AT3G21850,AT2G03190,AT2G03170,AT2G25700,AT3G25650,AT3G60020,AT1G20140,AT3G60010,AT3G21830,AT4G34210,AT4G34470	No

			0,AT1G10230,AT3 G21860,AT2G2016 0	
Solyc08g077510	Poly(U)-specific endoribonuclease-A	Yes	AT4G17100	LOC_Os04g45910
Solyc06g073840	Peptidyl-prolyl cis-trans isomerase D	Yes	AT5G21990	LOC_Os07g06710
Solyc04g076140	Rho GTPase activating protein 12	Yes	AT4G24580	LOC_Os07g46450, LOC_Os03g24180
Solyc03g120270	Coatomer beta's subunit	Yes	AT1G79990,AT1G5 2360, AT3G15980	LOC_Os02g11840, LOC_Os02g11830
Solyc06g066610	TATA element modulatory factor 1	Yes	AT1G79830	No
Solyc08g082850	Uncharacterized ABC transporter ATP-binding protein	Yes	AT1G64550	LOC_Os02g58020
Solyc10g079200	Mitochondrial carrier protein	Yes	AT5G46800	LOC_Os10g42299
Solyc05g013130	Transmembrane emp24 domain- containing protein 10	Yes	AT1G26690,AT1G1 4010,AT1G69460,A T3G29070,AT2G03 290,AT2G03040	LOC_Os07g43890
Solyc02g082990	ABA-inducible protein-like protein	Yes	AT5G38760,AT5G5 3820	No
Solyc10g018060	Elicitor-responsive protein 3	Yes		No
Solyc09g061960	Late embryogenesis abundant protein	Yes	AT4G13230	No
Solyc02g089590	Pollen coat-like protein	Yes	AT5G38760,AT5G5 3820	No
Solyc02g067150	Pro-resilin	Yes	AT5G39570	No
Solyc02g091890	MRNA complete cds clone RAFL24- 22-E06	Yes	AT4G37820,AT2G2 2795	No
Solyc02g080960	TO62-3	No		No
Solyc06g065980	Nucleic acid binding protein	Yes		LOC_Os01g07810
Solyc01g096780	Sperm-associated SUN domain protein	No	AT5G04990,AT3G1 0730	LOC_Os05g18770
Solyc03g032130	High mobility group protein	Yes	AT5G23420,AT3G5 1880, AT1G20693,AT2G1 7560,AT1G20696,A T2G34450,AT4G35 570, AT1G20696	No
Solyc01g105980	Unknown Protein	No	AT4G33625	LOC_Os07g02710
Solyc01g111940	Coiled-coil domain-containing protein 124	Yes	AT1G16210	LOC_Os03g03980
Solyc08g081410	General vesicular transport factor p115	Yes	AT3G27530	LOC_Os12g35360
Solyc02g090090	TPR domain containing protein	Yes	AT4G37210	LOC_Os07g03070
Solyc11g070020	Unknown Protein	No	AT5G11810	LOC_Os02g04260
Solyc03g034080	Acidic leucine-rich nuclear phosphoprotein 32 family member A	Yes	AT3G50690	No
Solyc03g019970	F-box protein interaction domain containing protein	No	AT1G20795,AT1G2 0790, AT1G20800,AT1G2 0240,AT1G20360,A	No

			T1G57580,AT1G20940	
Solyc05g021100	PH domain containing protein	No	AT2G30880	LOC_Os03g46340, LOC_Os07g04550
Solyc07g056010	Conserved oligomeric Golgi complex subunit 4	Yes	AT4G01400	LOC_Os08g06550
Solyc12g011270	Elongation factor Ts	No	AT4G11120	LOC_Os08g25290
Solyc10g007260	Ubiquitin-conjugating enzyme E2 N	Yes	AT1G16890	LOC_Os01g48280
Solyc04g051720	Genomic DNA chromosome 5 P1 clone MXK3	Yes	AT5G64820	No
Solyc06g007610	Early tobacco anther 1	No	AT5G48480	No

502

503

504

505

506 **Table 2** List of proteotypic peptides, m/z-ratation and retention time as well as subcellular location of
 507 heat-stress responsive proteins

Heat stress marker protein	m/z	RT	Sequence	Subcellular prediction (pSort)
Solyc02g082990	459.23	13.7	AQDATNAVK	Mitochondria
Solyc10g018060	513.71	17.44	GGDEEGYSGR	Nucleus
Solyc09g061960	495.75	12.16	AEEKDAIK	Chloroplast
Solyc02g089590	732.34	41.36	ESVQETGQQMQAK	Chloroplast
Solyc02g067150	599.75	12.88	KNDDDEYGSR	Nucleus
Solyc02g091890	595.78	13.41	NKLDSDAGENK	Chloroplast
Solyc01g099830	593.3	65.67	ITVTSDSTFSK	Nucleus
Solyc08g079260	511.76	75.72	DYIENLGAK	Nucleus
Solyc00g047200	725.31	63.58	GIPNSYTDDETHT	Extracellular
Solyc10g084350	431.21	32.1	MANADLAR	Chloroplast
Solyc04g051720	437.73	15.62	SAAAAVEQK	Vacuoles
Solyc03g044000	642.29	103.77	FTAEVVMQTID	Cytoplasm
Solyc02g080960	501.25	29.93	AENALEAQR	Mitochondria
Solyc06g007610	525.25	55.76	NAFGAEEVGR	Cytoplasm
Solyc12g099820	572.77	53.46	GPDADISEVGGK	Nucleus
Solyc08g075690	501.75	88.26	QLLEEAASA	Mitochondria
Solyc01g049680	467.23	23.46	ISEVAGSDR	Cytoplasm
Solyc05g005510	498.7	23.24	NDQVDmMK	Chloroplast
Solyc06g065980	553.26	88.69	mEGITDGVNK	Cytoplasm
Solyc10g079200	476.24	52.72	YSGAIDAVR	Chloroplast
Solyc07g009330	468.7	27.38	GSGDEWSAK	Mitochondria
Solyc03g120270	573.8	55.03	DIAVVAQSESK	Chloroplast
Solyc10g007260	574.77	35.42	SNEAEAVETAK	Mitochondria
Solyc01g096780	548.27	51.61	VDYALSSGGAR	Cytoplasm
Solyc11g042930	632.79	23.11	TEDKASEDELK	Nucleus
Solyc04g076140	623.28	39.96	QAADVDDVEQR	Nucleus
Solyc01g067740	569.76	24.62	EHGAPEDVR	Cytoplasm
Solyc03g025340	458.7	27.88	AGGYEEK	Mitochondria
Solyc03g032130	441.71	40.56	LMDAYNR	Nucleus
Solyc11g008400	603.28	43.45	AVETGEEDDLK	Cytoplasm
Solyc01g103510	465.74	46.78	DINAQLEK	Cytoplasm
Solyc01g105980	631.8	37.81	DGEGIVSSQSQR	Chloroplast
Solyc01g102590	633.27	37.03	DGQDYAAAAGGGGR	Nucleus
Solyc01g111940	692.83	77.69	SPDNPLNQVADKS	Nucleus
Solyc08g081410	537.28	77.22	NATLAEDLAR	Chloroplast

Solyc08g066820	467.71	59.37	mEVTEGAR	Cytoplasm
Solyc02g090090	601.82	77.09	LATTDVDQIAR	Chloroplast
Solyc08g006030	600.29	39.81	ETLENHDLTK	Chloroplast
Solyc00g007270	592.32	73.57	NSLPDGVVVQR	Nucleus
Solyc05g013130	568.25	24.51	EAEmQELNR	Vacuoles
Solyc11g070020	715.3	73.97	ESDDSTFSIADSR	Nucleus
Solyc09g009180	466.25	91.01	SELNLDLK	Chloroplast
Solyc06g007250	707.31	34.7	VSSSGKFDEGADQS	Nucleus
Solyc02g081700	458.26	59.93	DGVVLVGEK	Chloroplast
Solyc03g034080	416.22	39.59	LNLSDNR	Nucleus
Solyc12g011270	661.29	52.64	EDVSSDALSNER	Chloroplast
Solyc01g090190	569.24	25.6	EGSNNENFAR	Nucleus
Solyc10g080890	514.23	36.16	mENHAEVR	Cytoplasm
Solyc06g073840	708.85	81.05	ADPDTLSALSGGNAK	Golgi
Solyc03g123800	514.26	40.07	DVDNPNVVR	Chloroplast
Solyc03g044030	488.72	88.03	DEGSNGEIR	Nucleus
Solyc02g021560	509.25	24.34	AVADADVSNR	Nucleus
Solyc03g019970	526.76	85.45	DGSVLAFTDK	Cytoplasm
Solyc07g064510	555.75	51.96	FSDESEQLR	Nucleus
Solyc04g081400	617.79	55.5	EGGNFELQQGR	Chloroplast
Solyc01g111200	524.74	26.03	TAEELDQSR	Nucleus
Solyc08g077510	548.25	47.32	TGYNIQDER	Nucleus
Solyc05g021100	466.22	17.58	VAEADAEAR	Chloroplast
Solyc08g048500	727.32	74.49	INDEDEDTVTFR	Nucleus
Solyc08g082850	699.78	79.57	MFDGMDEEEAPK	Cytoplasm
Solyc10g009220	481.72	46.98	DVSEEDLR	Nucleus
Solyc06g082100	498.73	47.65	TDYALEER	Cytoplasm
Solyc11g062070	556.25	30.8	GYEEGTVGSGR	Nucleus
Solyc07g056010	561.26	74.4	NLTDVGD MTR	Cytoplasm
Solyc08g077700	805.32	51.76	EEAEADSMEEDGAVK	Nucleus
Solyc06g066610	488.24	31.06	EAEALSEAR	Nucleus
Solyc06g065170	407.71	33.09	VPDDIAGK	Nucleus
Solyc09g010460	490.24	32.93	IAAEYEQR	Cytoplasm
Solyc01g103690	438.73	50.74	EAESLSLK	Nucleus

508

509

510

Figure legends

Figure 1: Targeted MAPA: Schematic representation of the approach for unbiased identification of protein marker based on proteotypic peptides analysed by shotgun proteomics.

Figure 2: Principal components analysis (PCA) of the data derived from the targeted MAPA approach depicted in figure 1. (C_M) mature pollen in control condition, (HS_M) mature pollen under heat stress condition.

Figure 3: Venn diagram, comparing the PCA-ranked proteotypic peptides by the targeted MAPA approach with the identified proteins in control and heat stress treated mature pollen samples. The comparison identifies 69 proteins which are potential heat stress markers.

Figure 4: Identification of low abundance potential protein marker by the targeted MAPA approach. (A) Chromatograms belongs to heat stress and control samples, showing MAPKK retention time (RT). (B) MS/MS spectrum from 514.26 precursor ion.

Figure 5: Functional categorization of heat stress responsive proteins identified by targeted MAPA **A** Functional categorization of 69 unique proteins which potentially could define protein marker for heat stress. **B** Biclustering of the quantitative targeted MAPA data and their corresponding functional protein categories was performed with the COVAIN toolbox³³ using average linkage of Euclidean distance between groups as the metric. The heat stress pollen samples are clearly separated from the control samples. Categories such as “stress abiotic” and “REDOX” are highly upregulated in the heat stress samples. For further discussion see text.

Figure 6: Box plot determining regulation of proteins. (A) Calcium binding protein Calnexin, (B) Calreticulin 2 calcium binding protein, (C) Late embryogenesis abundant protein D-34, (D) Hexokinase 1.

Figure 1

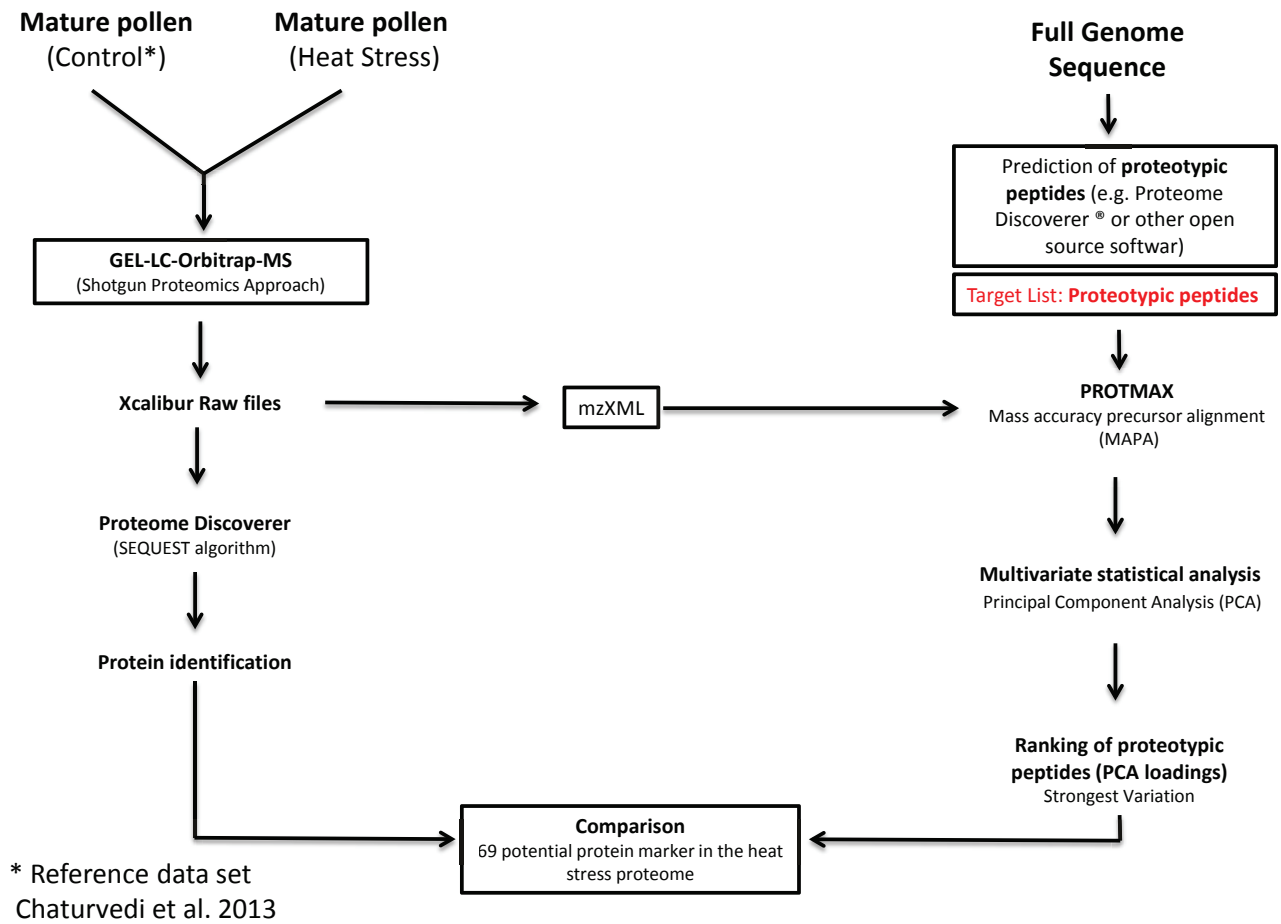
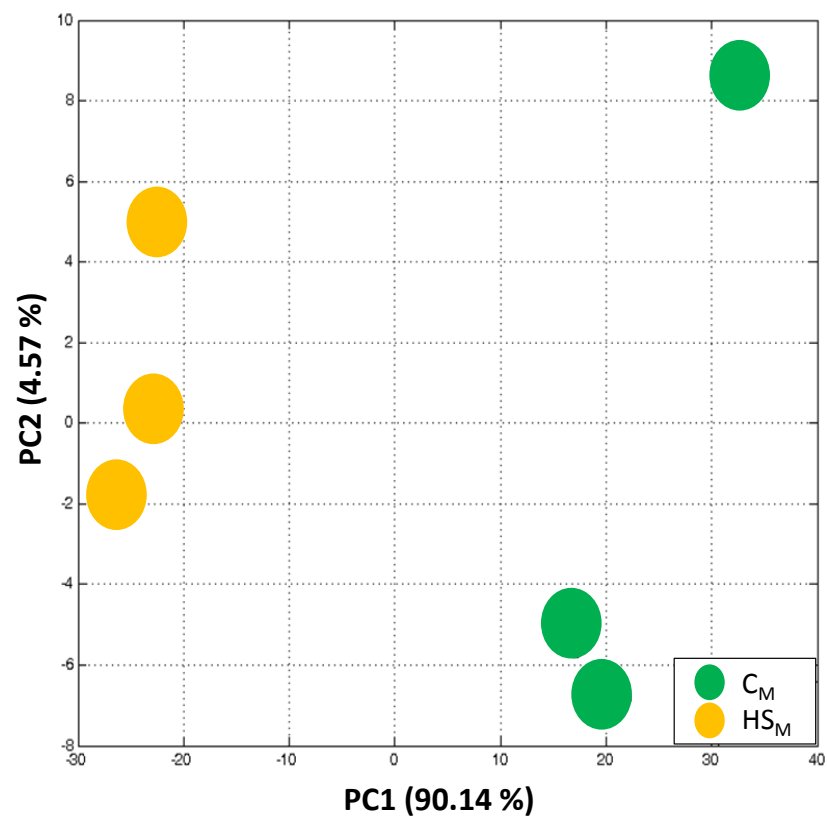


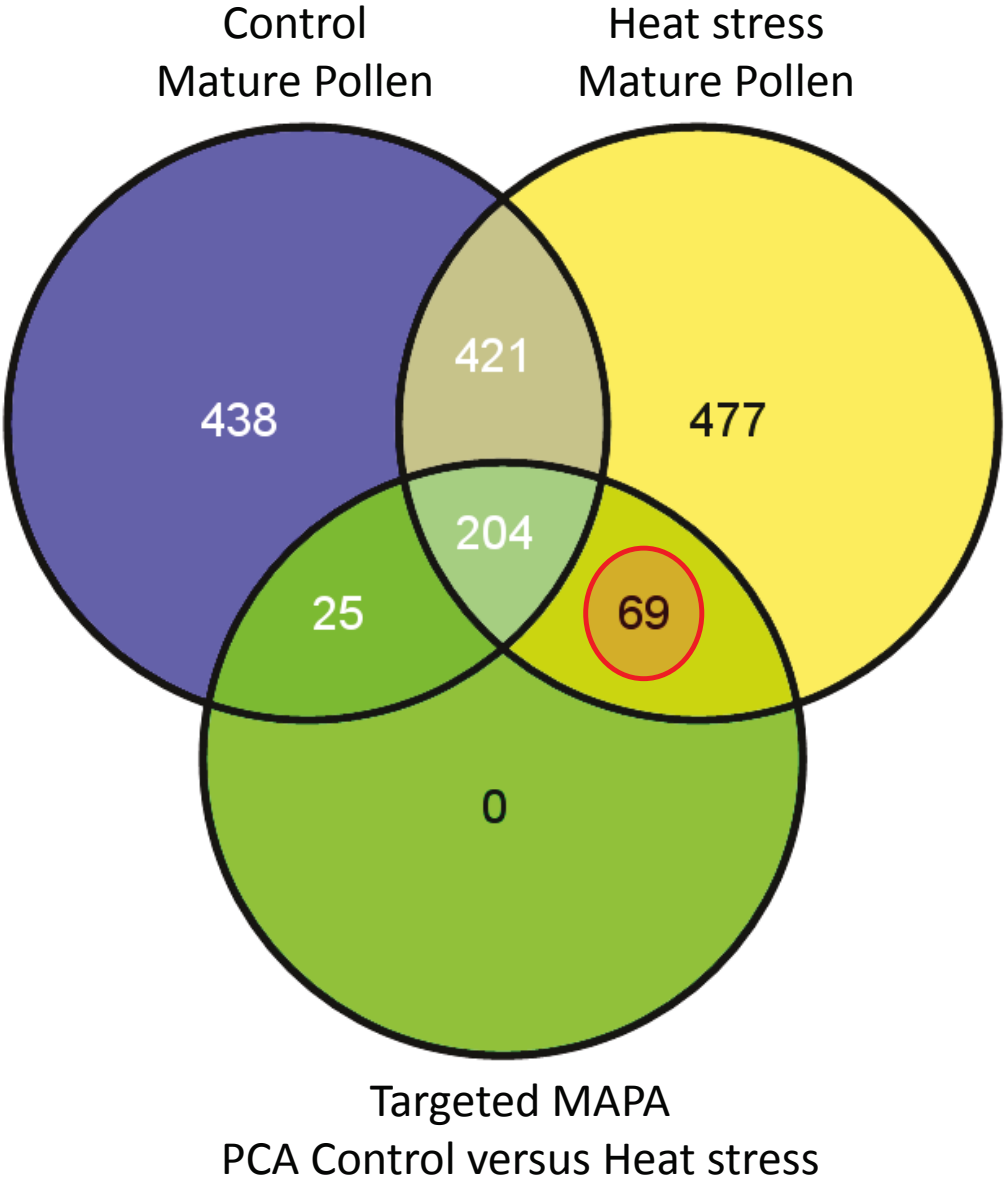
Figure 2



556

557 **Figure 3**

558



559

560

561

Figure 4

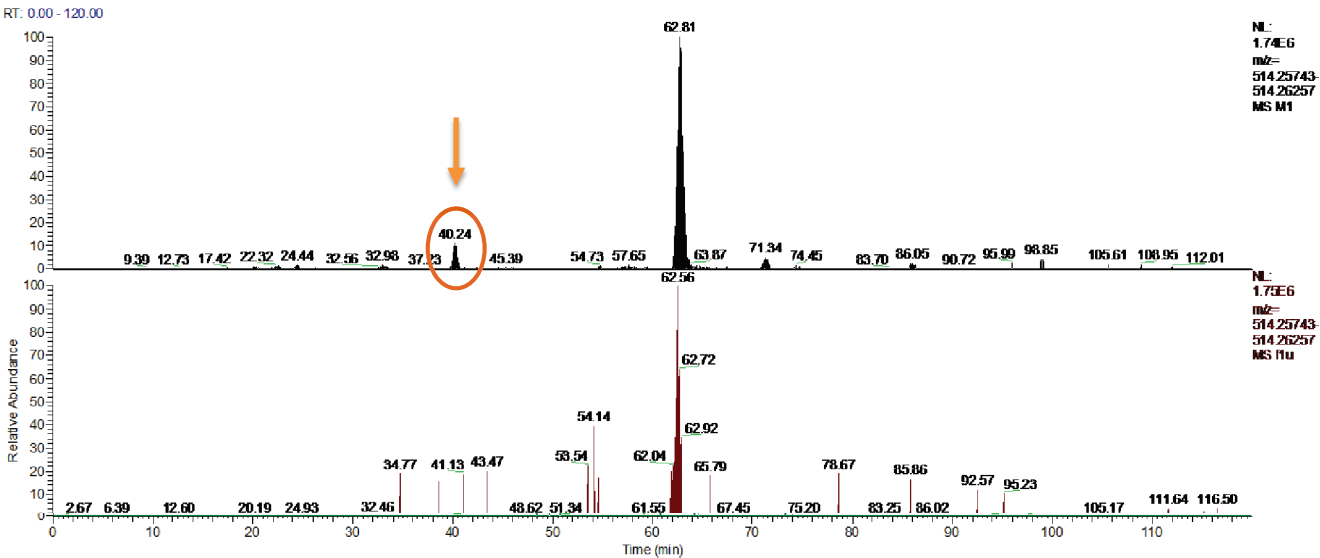


Figure 5A

A

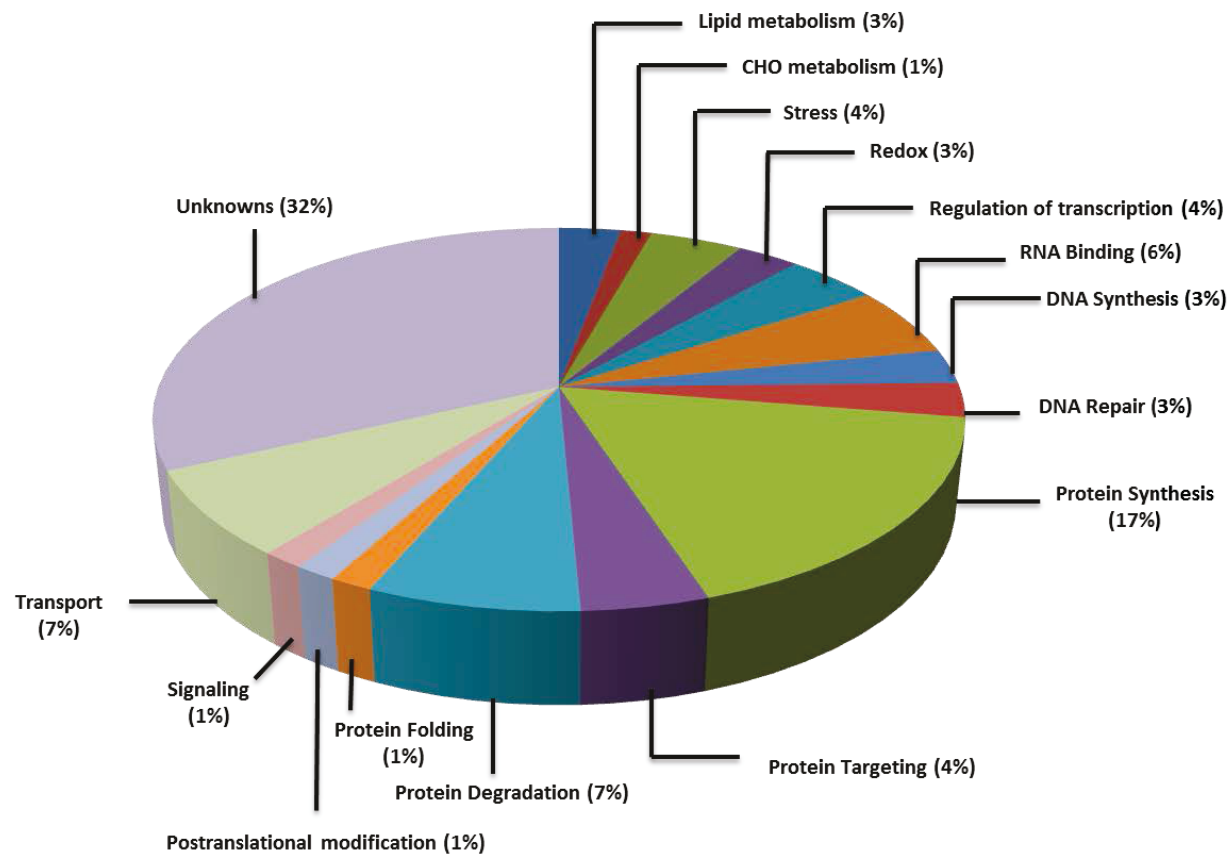
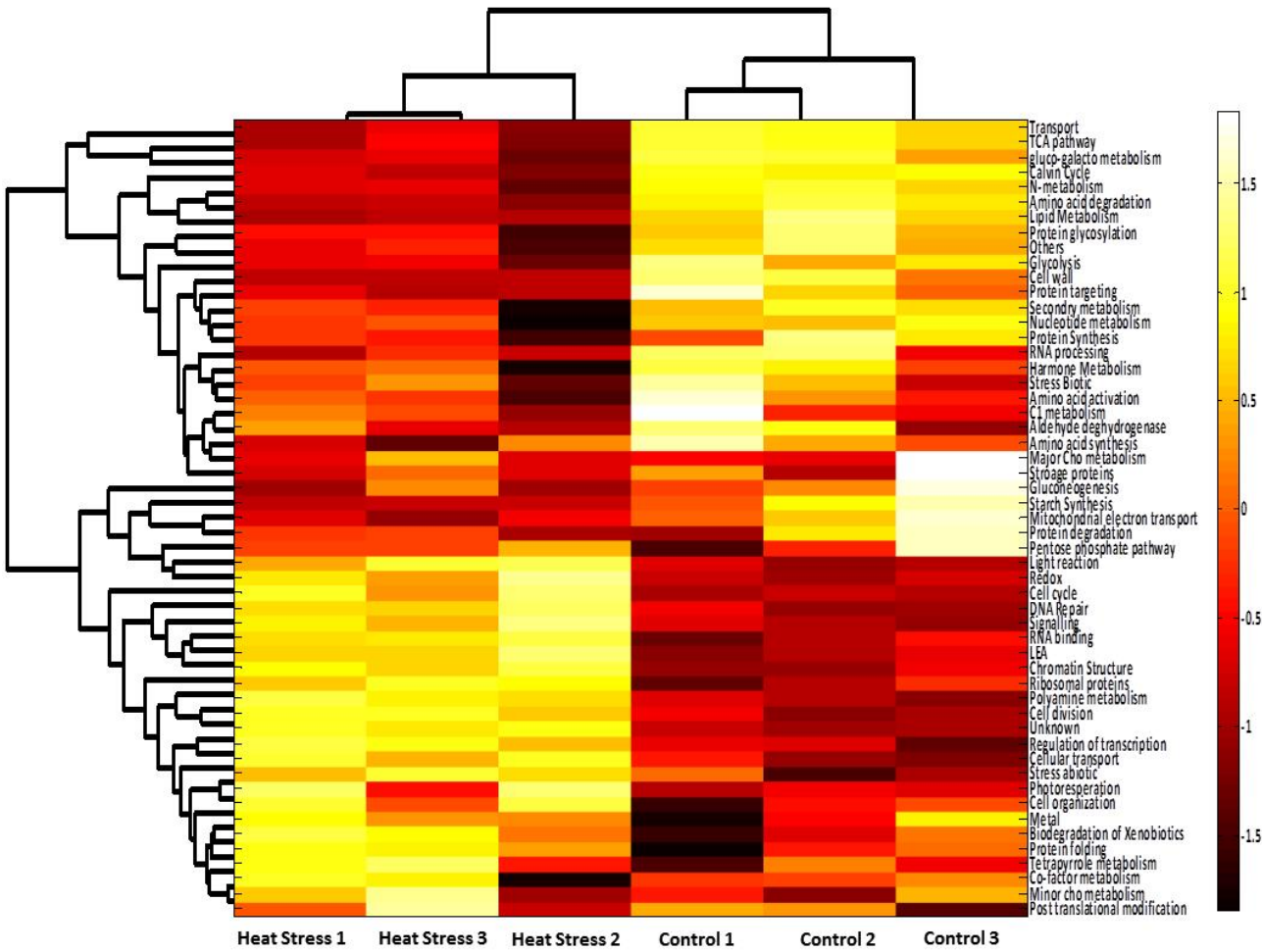


Figure 5B

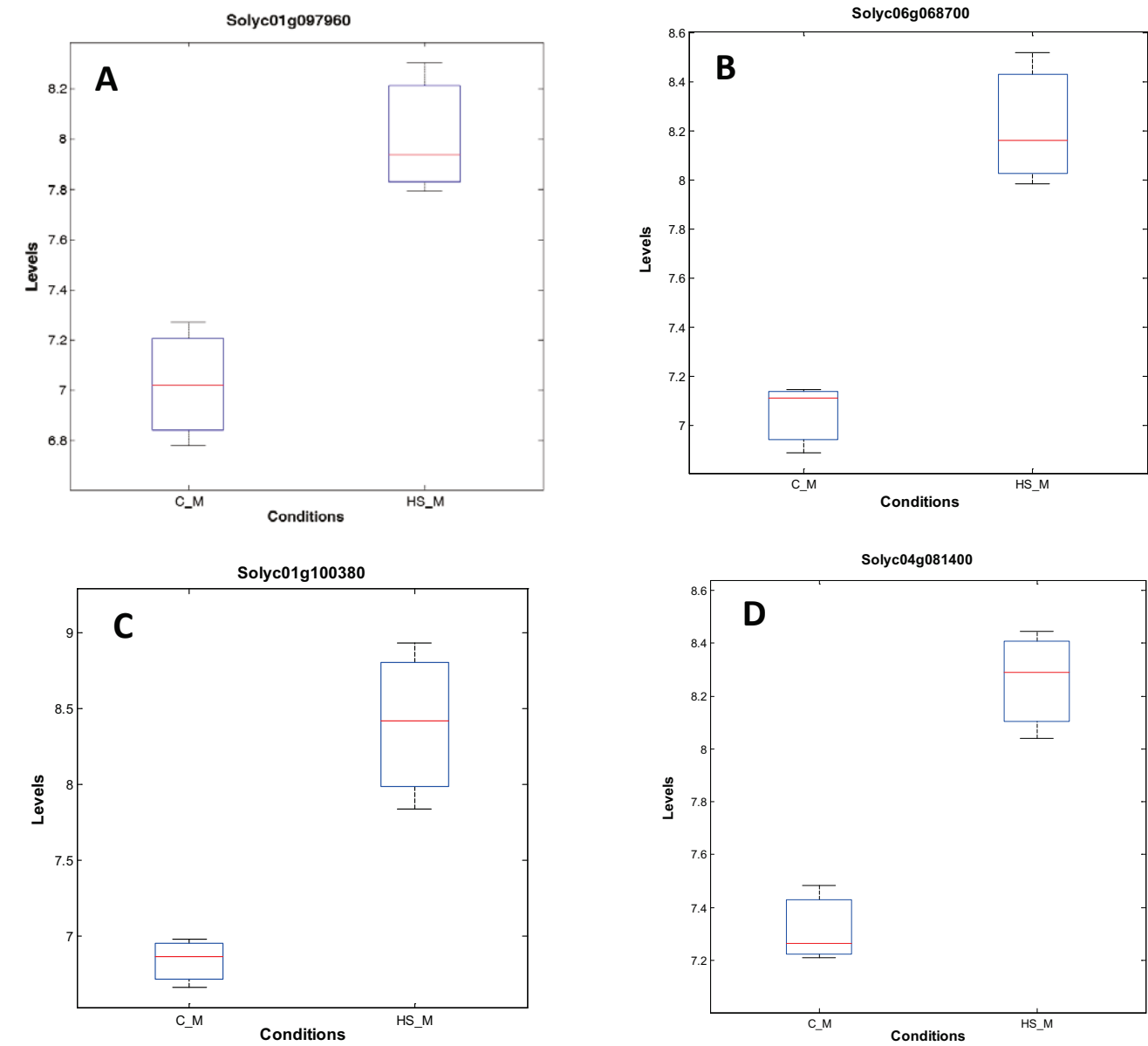
B



579

580 **Figure 6**

581



582

III. Concluding remarks

In the presented PhD thesis strategies were developed and established which might help push the metabolomics technique further concerning improved metabolite coverage, refined m/z feature picking and easier identification of unknown analytes in untargeted metabolomics-style experiments. Even though tremendous efforts have been done and are being executed by the metabolomics community, there are still various bottlenecks to overcome in the upcoming years, mainly in regards to absolute quantification and reliable identification of unknowns. Only when these problems are addressed, subsequent functional interpretation of “omics-type of data may proceed to the next level. For GC-MS side, it is largely the annotation of yet unidentified spectra as well as chromatographic problems like co-elution which have to be solved in the future, while on the LC-MS side the cornerstones are just about being laid with the beginning of the establishment of global databases and algorithms apt to handle the enormous amounts of data generated. Without question is the fact that highly accurate mass spectrometers in combination with high-performing liquid chromatography are a prerequisite for meaningful characterization of ultra-complex biological samples as only this setup can yield enough simultaneous information – like retention times, accurate masses, and product ion scans – for unfolding underlying biological relations. Especially ultra-high resolution mass spectrometry and improved deconvolution will give us new insights in the near future for the type of data that is being generated on a daily basis in a typical metabolomics lab. The current work presents methods for reproducible and meaningful untargeted analysis of typical LC-MS-applicable compounds like plant secondary metabolites, including methods of inferring mass spectral information about them. This, in return, opens up possibilities for establishing algorithms to characterize large datasets deriving from untargeted metabolomics experiments, as is being shown. While being a very potent technique, a lot of work concerning the LC-MS platform still will have to go into rendering databases comparable between labs, because compared to GC-MS, no standard specifications exist to date concerning the metabolic analysis of biological samples on the LC-MS setup. This is partly because of the vast array of possibilities that a state-of-the-art mass spectrometer offers for method development. Resolution, scan speed, fragmentation type, ionization type, etc. are only a few of the parameters the experimenter must consider on the mass spectrometric side for method development. Beside the need for working methods on the technical-analytical side, the development of new tools on the bioinformatic level is imperative in order to cope with the steadily increasing amount of data generated, as it is not feasible to handle “-omics”

data manually. However, even though algorithms for automated processing of high-resolution metabolite product ion scans are already being developed and applied, de-novo interpretation of metabolite fragmentation spectra is something which still demands manual work and good knowledge of organic chemistry.

Summing up, within the combination of highly sensitive analytical facilitation, bioinformatic data processing, and manual design of the experiment and data interpretation, metabolomics is able to yield new biological insights already, and will do even more so in the future.

IV. Abstract

While still a relatively young discipline, metabolomics – the identification and quantification of all metabolites present in a biological sample – has already proven to be a cornerstone technique in life sciences in recent years. Especially in view of ongoing technological advances in analytical chemistry, major scientific revelations can be expected in the near future by application of this technique to current biological questions. Even though the metabolomics field has progressed rapidly in recent years, a plethora of problems still has to be overcome to establish true metabolomic workflows for everyday analysis. The studies presented aim towards improving metabolomics in view of enhanced metabolic coverage by combination of different analytical platforms and an increased identification rate of unknowns in high-resolution MS-based metabolomics data.

The first study focused on improving the metabolic coverage from the same sample by combining the GC-TOF-MS with the LC-Orbitrap-MS platform, effectively increasing the range of detectable metabolites from 40 to 2000 m/z and enabling deeper functional interpretation of the data. To provide a proper experimental frame, a set of *Arabidopsis thaliana* specimen were subjected to high cold and light stress, in order to trigger global metabolic reprogramming of the plants' metabolism. Untargeted metabolomics analysis by GC-MS and LC-MS revealed highly altered metabolite profiles in the primary as well as the secondary plants metabolism compared to unstressed specimen. Subsequent data merging from both platforms revealed functional relations of metabolites deriving from the underlying metabolic network: application of the Granger causality test on the data set unfolded the presence of several typical secondary plant compounds ("flavonoids") in the cold-stressed plants deriving from activation of the phenylpropanoid biosynthetic pathway, correlating with the key precursor molecules shikimic acid and phenylalanine. While metabolite identification on the GC-MS platform is already developed and mostly a question of database availability, the LC-Orbitrap-MS analyses yielded thousands of possible metabolite signals and only a handful of them could be manually annotated. Thus, in the second study, a method for predicting novel metabolites in LC-MS data by exploiting the potential of high-resolution MS and accurate mass acquisition was established: A MATLAB script (*mzGroupAnalyzer*) was developed which automatically checks a list of possible sum formulas deriving from highly resolved precursor signals for a metabolic context. By application of the program to mass spectrometric data of the cold stressed plant profiles, together with structure validation by

product ion spectra, 16 putatively new substances in *Arabidopsis* could be predicted and complete biosynthetic pathways of secondary metabolites could be reconstructed. Additionally, for improved m/z feature extraction of “-omics”-type data, a method based on the successfully established MAPA (mass accuracy precursor alignment) algorithm was developed in the third study, which aims towards extraction and relative quantification of precursors in a targeted approach (compared to the untargeted method), which is shown to be superior compared to other quantification strategies (like NSAF – normalized spectral abundance factor) in several ways.

V. Zusammenfassung

Obschon noch eine relativ junge Technik, hat sich die Metabolomik – definiert als die Identifizierung und Quantifizierung aller Metaboliten in einer biologischen Probe – in den letzten Jahren als Schlüsseltechnik in den Lebenswissenschaften etabliert. Gerade in Hinblick auf die ständige Verfeinerung der Instrumentierung in der chemischen Analytik kann man in der nahen Zukunft durch Anwendung der Metabolomik auf biologische Fragestellungen neue Erkenntnisse und wissenschaftliche Durchbrüche erwarten. Obwohl die Metabolomik in den letzten Jahren eine rasche Weiterentwicklung durchlaufen hat gibt es noch einige Problembereiche, die gelöst werden müssen, damit metabolische Analysenverfahren als Standardmethoden in den Laborbereich übernommen werden können. Die in dieser Arbeit vorgestellten Publikationen zielen zum einen darauf ab, durch Kombination von analytischen Methoden einen größeren Bereich von Metaboliten abzudecken, und zum anderen, die Identifikationsrate von unbekannten Molekülen in Hochauflösungs-MS-basierten Datensätzen zu verbessern.

Die erste Studie demonstrierte die verbesserte Abdeckung der identifizierbaren Metabolite aus einem biologischen Sample durch Kombination der GC-TOF-MS und der LC-Orbitrap-MS – Technik, was zu einer Erhöhung der detektierbaren Moleküle in den Bereich von 40 – 2000 m/z führte, und überdies tiefergehende biologische Interpretation der Daten zuließ. Um einen geeigneten experimentellen Überbau zu gewährleisten, wurden mehrere *Arabidopsis thaliana* Exemplare hohem Licht- und Kältestress ausgesetzt, der dazu führen sollte, umfassendes metabolisches Reprogrammieren in der Pflanze auszulösen. Ungerichtete Metabolomanalysen durch parallelen Einsatz von GC-MS und LC-MS legten drastisch veränderte Metabolitprofile in den derart gestressten Pflanzen offen. Das darauffolgende Zusammenführen der Datensätze von den beiden unterschiedlichen analytischen Techniken ließ funktionelle Interpretation auf das darunterliegende metabolische Netzwerk zu: durch Anwendung des Granger-Kausalitätstestes konnte in den gestressten Pflanzen die Anwesenheit von mehreren typischen pflanzlichen Sekundärmetaboliten („Flavonoide“), ausgehend vom Phenylpropanoid-Metabolismus, festgestellt werden, welche mit Schlüsselsubstanzen des Primärmetabolismus, vor allem Shikimisäure und Phenylalanin, korrelierten. Während die Identifizierung von Metaboliten auf der GC-MS Seite bereits gut entwickelt und größtenteils eine Sache der Verfügbarkeit von Datenbanken ist, erbrachten die LC-Orbitrap-MS Analysen tausende von Metabolitsignalen, von denen wiederum nur einige

wenige manuell annotiert werden konnten. Im Zuge dessen wurde in der zweiten Studie eine Methode entwickelt, die die Vorhersage von neuen Verbindungen aus LC-MS-Daten ermöglicht, durch Ausnutzung der Hochauflösungs-MS und dementsprechend der Möglichkeit, exakte Massen für Datensignale zu definieren: Ein MATLAB Skript (*mzGroupAnalyzer*) wurde entwickelt, welches imstande ist, verschiedene Summenformel-Vorschläge aus hochaufgelösten Vorläuferionenmassen auf einen metabolischen Kontext hin zu überprüfen. Durch Anwendung dieses Programms auf die MS-Daten der gestressten Pflanzen, in Kombination mit manueller Interpretation der Produktionsspektren, konnten 16 bis dato unbeschriebene Verbindungen in *Arabidopsis thaliana* vorausgesagt und ein kompletter Biosyntheseweg aus der Cyanidin-Familie rekonstruiert werden.

Überdies wird in einer dritten Studie eine Technik zur verbesserten Extraktion von Vorläuferionen aus „omics“-Datensätzen vorgestellt: basierend auf dem MAPA (mass accuracy precursor alignment) – Algorithmus wurde eine Methode entwickelt, welche eine gezielte Extraktion und relative Quantifizierung von Vorläuferionen (im Gegensatz zu einer ungerichteten Herangehensweise) ermöglicht, welche sich im Vergleich zu anderen Quantifizierungsstrategien (etwa NSAF – normalized spectral abundance factor) in einigen Bereichen als überlegen erwies.

VI. Curriculum Vitae

Hannes Dörfler

Education

present	PhD-Student at the Department of Molecular Systems Biology, University of Vienna
28.11.2011	Conclusion of Diploma studies Biology and Chemistry (Mag. rer. nat.)
2010-2011	Diploma thesis at the Faculty of Life Sciences, Department of Molecular Systems Biology, Prof. Wolfram Weckwerth
2004-2010	Study of Chemistry and Biology at the University of Vienna Partial studies in Psychology and Philosophy
2003	Community service
June 2002	High School Diploma (Matura) at Stiftsgymnasium Melk

Scientific publications

- **Doerfler H**, Sun X, Wang L, Engelmeier D, Lyon D, Weckwerth W (2014) *mzGroupAnalyzer- Predicting Pathways and Novel Chemical Structures from Untargeted High-Throughput Metabolomics Data*. PLoS One 9: e96188
- **Doerfler H**, Lyon D, Nagele T, Sun X, Fagner L, Hadacek F, Egelhofer V, Weckwerth W (2013) *Granger causality in integrated GC-MS and LC-MS metabolomics data reveals the interface of primary and secondary metabolism*. Metabolomics 9: 564-574
- „Comparative metabolic studies in *Arabidopsis thaliana* via GC-MS and LC-MS“, diploma thesis

Qualifications

- Practical knowledge in instrumental chemistry, especially in the identification and quantification of small molecules via chromatography coupled to mass spectrometry („Metabolomics“ and „Lipidomics“); GC-MS (GC-QqQ-MS, GCxGC-TOF-MS) as well as HPLC-DAD and high-resolution LC-MS (UPLC-QqQ and UPLC-Orbitrap-MS);
- Software: Xcalibur, Chromeleon, LECO Pegasus
- Profound knowledge in the manual interpretation of MSⁿ-fragment spectra, structure elucidation of metabolites in high-throughput high-resolution MS-experiments and statistical methods
- Practical knowledge in electrochemical analysis of redox active organic molecules with the three-electrode-technique (cyclic, differential-pulse and square-wave voltammetry)
- Basic methods in environmental chemistry and ecology (elemental analysis, etc.)
- Basic knowledge in toxicology

- Profound knowledge in rhetoric and presentation skills
- Meticulous, proactive, team player, resilient

Teaching

- 2011-2012: Tutor for students' course „Metabolomics“

Conferences

- InMetA 2014 (2nd International Metabolomics Austria) 3.7.-4.7. 2014, Vienna University of Technology
Talk: *“mzGroupAnalyzer – Predicting Pathways and Novel Chemical Structures from Untargeted High-Throughput Metabolomics Data”*
- ICAR 2012 (International Conference of Arabidopsis Research), 3.7-7.7 2012, Vienna Hofburg
Poster presentation: *“Granger causality in integrated GC- and LC/MS metabolomics data reveals the interface-dynamics of central and secondary plant metabolism”*

Interests

- Chemistry, biology, pharmacy, toxicology; chemical analysis, „omics“-sciences; systems biology
- Psychology, philosophy, rhetoric, NLP
- Classical music, sports