universität wien

DISSERTATION

Titel der Dissertation

„Computationally assisted mining and processing of proteomics data of *Medicago truncatula* under drought stress.“

Verfasser

Mag. David Lyon

angestrebter akademischer Grad

Doctor of Philosophy (PhD)

Wien, 2014

Studienkennzahl lt. Studienblatt:	A 794 685 437
Dissertationsgebiet lt. Studienblatt:	Biologie
Betreuerin:	Dipl.-Biol. Dr. Stefanie Wienkoop, Privatdoz.

Declaration

I declare that the work presented in this thesis has been carried out at the Department of Ecogenomics and Molecular Systems Biology. It is, to the best of my knowledge, my original and own work. My contributions are stated at the beginning of each peer-reviewed publication. Furthermore this work, except for published manuscripts, has not been submitted, either in whole or in part, for publication, degree or any award at this or any other university.

Vienna, October 2014

Mag. David Lyon

“0100010101001100010101100100100101010110100110001001001010101100100010101001101001011”

Table of contents

Overview of the scientific field

Molecular Plant Systems Biology

Systems biology, a paradigm shift towards holistic rather than reductionist approaches to enable a systems understanding and study of complex interactions, has become a dominant trend in the biological sciences. Its foundation are the unbiased "omics" approaches, such as genomics, transcriptomics, proteomics, metabolomics, and the plant-specific phenomics, followed by data integration, statistics and computational modeling, trying to form as complete a picture as possible. Since a multitude of causes need to be observed, it is all about recognizing patterns and forming hypotheses after experimental design and data analysis, in contrast to measuring specific parameters due to a preformed hypothesis. This has become possible due to technological advances, specifically dramatic increases of computational power, next-generation sequencing, and advances in mass spectrometry (Aebersold and Mann 2003; Metzker 2010; Schwanhäusser, Busse, and Li 2011; Lommen, Gerssen, and Oosterink 2011; Wienkoop et al. 2004; Weckwerth, Wenzel, and Fiehn 2004; Hu et al. 2005).

In order to expand biological insights of important crop plants, well-known plant model organisms such as *Arabidopsis thaliana* L. (Brassicaceae) and *Medicago truncatula* Gaertn. (Fabaceae / Legumes), are used worldwide. The mixotrophic alga *Chlamydomonas reinhardtii* P. A. Dangeard (Chlamydomonadaceae) has been used in order to study the photosynthetic apparatus, and serves as a model organism for alga. The study of such model organisms enables fundamental research and strives to improve and to transfer knowledge on agronomics, nutrition, biotechnology and bioenergy. The availability of genomic data, as well as the critically important functional annotation is essential for systems biology, since genome sequences can be translated to protein sequences, transcripts can be mapped to genes, and metabolic pathways can be inferred from genes. Genetic modification such as knockdown and or knockout mutants are widely used to gain more insight into specific biological processes (**Staudinger et al. 2012**; Weckwerth 2011b; Valledor et al. 2013; Hebeler et al. 2008).

Mass Spectrometry-based concepts

Various methods exist in analytical chemistry to analyze complex biological samples, each with specific strengths and weaknesses. Mass Spectrometry (MS) is a very powerful post-genomic technique. Due to high sensitivity, it enables the measurement of analytes (e.g. proteins and metabolites) of about four to five orders of magnitude, depending on the instrument and sample. When coupling Chromatography with Mass Spectrometers, the complex mixture of a sample is separated in time as well as focused on the column. This enables the detection of low abundant compounds due to enrichment on the column, decreases the complexity of the sample at any given time, allows a more accurate quantification via chromatographic peak integration, and enables the Mass Spectrometer to perform various types of scans on specific mass to charge (m/z) regions (Weckwerth 2003). MS can only measure ions, charged molecules. Various ionization techniques exist. Two commonly used methods are Electro Spray Ionization (ESI), a soft ionization technique mostly generating intact ionized analytes, and the hard ionization Electron Impact (EI), which produces ionized fragments of analytes. The analysis of larger, more non-polar and less volatile compounds, e.g. peptides, is preferably performed with Liquid Chromatography (LC) coupled to Mass Spectrometry with ESI (nanoLC-ESI/MS). This is in contrast to smaller, more polar and more volatile compounds, such as amino acids and sugars, which are preferably analyzed by Gas Chromatography (GC) coupled to Mass Spectrometry with EI (GC-EI/MS) (Scherling et al. 2010). A focus of many shotgun-proteomics studies is the maximization of the sequence coverage of proteins (Michalski, Cox, and Mann 2011; Mann et al. 2011; Nagaraj et al. 2012). Improving the resolution of the chromatographic separation (by e.g. increasing the length of the LC-column or reducing the particle size of the stationary phase) is one factor, however, sensitivity is the crucial factor, since the dynamic range of complex samples spans several orders or magnitude and therefore only the most abundant proteins would be detected with conventional (4.6 mm inner diameter) chromatographic setups. Increasing the sensitivity can be achieved by concurrently reducing the inner diameter of LC-columns as well as the flow rate. Theoretically, reducing the inner diameter (ID) to a quarter of its original size enhances the sensitivity by a factor of sixteen (quadratic increase). Thus, for LC/MS-based shotgun proteomics, a general devel-

opment from micro- to nano-flow has occurred in most laboratories (Mitulovic and Mechtler 2006).

Metabolomics

Metabolomics, analogous to proteomics, is the study of the entirety of small-molecule fingerprints, the final products of cellular processes. Division into "primary" (necessary for growth, development, and reproduction) and "secondary" metabolites (not directly involved in the latter, but necessary for survivability, fecundity, etc.) can be made (D'Auria, Gershenzon, and Auria 2005; **Doerfler et al. 2013**; http://en.wikipedia.org/). Primary metabolites are well-characterized and experimentally validated, though naturally not detectable in every sample of plant origin. Plant secondary metabolites are extremely diverse, known for e.g. preformed and induced defense mechanisms, local and systemic localization, and complex biological interactions and activities. An inconceivable amount of novel, yet unknown metabolites (including derivatives) remain to be discovered. These theoretical numbers are dwindling, particularly due to species extinction in primary tropical forests. Typically, primary metabolites are measured with GC/MS and secondary metabolites using LC/MS (when applying Mass Spectrometry, otherwise HPLC/UV usually is a standard method) (Matsuda, Yonekura-Sakakibara, et al. 2009; Scherling et al. 2010). GC/MS metabolomics benefits from reproducible Retention times and Retention time indices as well as fragmentations (EI), while LC/MS metabolomics is not nearly as standardized and reproducible when comparing cross-laboratory methodology (Chromatography setup, Collision Energy settings, fragmentation type, etc.). The unpredictability of fragmentation patterns of metabolites in contrast to peptides constitutes a major difference between metabolomics and proteomics. Therefore, reference spectra are essential to identify/annotate compounds. Another major difference between GC and LC/MS is the superior availability of reference fragment spectra and Retention times of pure compounds in GC/MS (in the form of databases e.g. NIST-AMDIS, http://www.nist.gov/; GMD, http://gmd.mpimp-golm.mpg.de/). Acquiring high Mass Accuracy LC/MS metabolomics data, enables elemental composition determination (Kind and Fiehn 2006; Kind and Fiehn 2007). Many differences between GC and LC/MS metabolomics exist, which will not be indulged, since this is

not within the scope of this work. Targeted as well as unbiased approaches, using specialized instruments and data acquisition methods, in analogy to proteomics, exist for both GC as well as LC/MS (Matsuda, Yonekura-Sakakibara, et al. 2009; **Doerfler et al. 2013**; **Mari et al. 2013**; **Doerfler et al. 2014**).

Proteomics

The large-scale analysis of protein mixtures is termed proteomics. Apart from the targeted analysis of specific proteins/peptides, in systems biology, proteomics mostly aims to analyze the proteome, the entirety of all proteins of an organism of a given cell or tissue, at a given time, in a given state. Profiling proteomics delineates differently expressed abundance levels of proteins between samples. Functional proteomics inspects Post Translational Modification(s) (PTMs) of proteins, the interaction of proteins with substrates and small molecules, aiming at a functional characterization (Mitulovic and Mechtler 2006; Baerenfaller et al. 2008).

Two-dimensional Sodium-Dodecyl-Sulfate Poly-Acrylamide-Gel-Electrophoresis (2D SDS-PAGE) is a well-known technique for the separation of complex protein mixtures. It employs IsoElectric-Focussing (IEF), the separation of proteins based on their isoelectric point, in the first dimension and subsequently the separation by mass in an electric field. Due to the combination of two orthogonal separation techniques, a much higher resolution can be achieved, compared to using either technique by itself. After the separation processes, proteins can be made visible with various staining solutions, relative quantification based on spot intensity can be performed, the spots can be excised for further analysis, or transferred to other matrices, e.g. for Western Blots (Görg et al. 2002; Towbin, Staehelin, and Gordon 1979). Established laboratory protocols exist for the application of 2D SDS-PAGE. It is a robust and inexpensive method, but has inherent limitations. Namely, resolution limits owing to the hydrophobicity, isoelectric point, and molecular weight range of proteins resolvable and to protein-loading capacity; low-abundant proteins lying below the Limit Of Detection (LOD), PTMs cause proteins coded by the same gene to migrate to different locations on the gel (Glinski and Weckwerth 2006). A complementary non-gel-based approach called MudPIT (Multidimensional Protein Identification Technology) demonstrated the rapid and

large-scale analysis of the yeast proteome, by using a biphasic stationary phase composed of Strong Cation eXchange (SCX) and Reversed Phase (RP) materials, coupled to Mass Spectrometry (Washburn, Wolters, and Yates 2001).

To date 2D SDS-PAGE and Western Blots are commonly used, though a trend towards the high-throughput nanoLC-ESI/MS shotgun proteomics techniques, as well as absolute quantification based on stable isotope-labeled internal standards can be observed (Lehmann et al. 2008; Wienkoop et al. 2008; Aebersold, Burlingame, and Bradshaw 2013; **Recuenco-Munoz et al. 2014** *accepted*).

Unbiased approach

In „bottom up"-**shotgun proteomics**, all of the proteins (of a complex sample) are digested with a protease (e.g. Trypsin), subsequently the resulting peptides are measured. This protein-profiling strategy enables a very high sample throughput due to a relatively short sample preparation time. Shotgun proteomics is capable of resolving complex samples and can provide a high protein identification rate. Typically, the complex peptide mixture is separated by Reversed Phase Liquid Chromatography (RPLC), ionized by Electro Spray Ionization (ESI), and the precursors (entire peptides) are measured by MS and their corresponding fragments by MS/MS (Aebersold and Mann 2003). This data can be used to search against genomic databases and Expressed Sequence Tag (EST) libraries for peptide identification (as well as reference databases of previously identified spectra, such as "ProMEX", or could even be "de-novo-sequenced") (Hummel et al. 2007; Wienkoop et al. 2012). Label-free relative quantitation can also be performed either by „intensity based" approaches (e.g. integrating the area under the chromatographic peak) or by „spectral counting" (enumerating the occurrence of MS/MS scan events of a given precursor ion) (Hoehenwarter and Wienkoop 2010). Label-based relative quantification, such as Stable Isotope Labeling by Amino acids in Cell culture (SILAC), has often been applied to mammalian samples, e.g. mouse (Ong et al. 2002). Since mammalian cells cannot synthesize a number of "essential" amino acids, these are provided in the medium to support cell growth. Amino acids labeled with heavy stable isotopes are introduced into the medium, which are incorporated into proteins. Subsequent shotgun-proteomics analysis of mixed samples, grown on natural and heavy media, result in the detection of light and heavy peptide

signals, enabling accurate relative quantification. In contrast, photoautotrophic plants can synthesize all amino acids, therefore Stable Isotope Labeling In Planta (SILIP), which uses ^{15}N enriched nitrogen sources, was developed (Schaff et al. 2008). Depending on the labeling-time and the growth rate of the organism, partial as well as full metabolic labeling can thus be achieved (see Specific background and objectives of the FWF project) (Kline and Sussman 2010).

Sequenced organisms offer a major advantage, since the known nucleotide sequence can be translated into an amino acid sequence, which can be digested in silico (providing e.g. tryptic peptides). The experimental MS/MS spectra and their corresponding precursor masses are matched against the in silico database using e.g. commercially available algorithms/programs like „Sequest" and „Mascot" to identify peptides and infer proteins (Eng et al. 1994; Pappin, Hojrup, and Bleasby 1993). The experimental data yields even more stringent results when working with instruments capable of high Mass Accuracy and resolving power (e.g. „LTQ-Orbitrap-MS"). With the rapid generation of genome sequence information in conjunction with advances in Mass Spectrometry technology as well as bioinformatics, shotgun proteomics has emerged as a promising field of protein research and a routine in many laboratories. Two general categories can be distinguished: **Unbiased (untargeted) analyses** of samples using Data Independent (e.g. SWATH, MS^E) and Data Dependent Analysis (e.g. dynamic MS/MS triggering of the LTQ-Orbitrap) for protein identification as well as PTM analysis thereof (qualitative analysis); and **biased (targeted) analyses** using specialized Mass Spectrometers such as triple-quadrupoles (QqQ) (quantitative analysis) (Gillet et al. 2012; Collins et al. 2013; Moran et al. 2014). Biological interpretation necessitates quantitative information in addition to identifications. Some of the new generation Mass Spectrometers aim to comply with both demands and novel data-processing methods improve quantitation (e.g. Orbitrap Fusion). Thus, a strict separation of qualitative and quantitative methods is not always possible. Rather, the goal would most often, if not always, be to get as much qualitative and quantitative information as possible, which depending on the technical platform as well as time and cost, is not always possible. Naturally, this leads to specialized methods and instruments (as previously mentioned) (Schulze and Usadel 2010).

Targeted approach

The "Mass Western" is a targeted proteomics technique used to quantify specific proteins of interest. The term was coined in analogy to the well-known "Western Blot". A number of selected proteotypic peptides (specific only to the protein of interest) are synthesized, each containing one amino acid labeled with ^{13}C and ^{15}N, and can thus serve as internal standards. The synthetic peptides are tuned on the mass spectrometer in order to achieve maximum sensitivity. A dynamic range of about four orders of magnitude can be measured. The internal standard is introduced to the sample as early as possible and the proteins are digested. Subsequent LC/MS measurements are typically performed on a Triple Quadrupole type instrument. The synthetic peptides coelute with their natural counterparts on the LC system, but can be differentiated on the MS level, because of a mass shift due to the introduction of the "heavy amino acid". This technique allows absolute quantification of low-abundant proteins in complex samples (Lehmann et al. 2008; **Lyon, Weckwerth, and Wienkoop 2014**; **Recuenco-Munoz et al. 2014** *accepted*). The benefits of absolute in contrast to relative quantification are the comparability of data, independent of the sample, experiment, tissue or organism, as well as the increase in accuracy of the measurement. The drawbacks are an increased cost of reagents (internal standards), time needed to establish the specific methods (finding proper SRM transitions, tuning of the Mass Spectrometer and measuring calibration curves), and that only specific analytes will be detected regardless of the contents of the sample. With absolute quantification it is possible to distinguish isoforms or gene families all in one LC/MS run. In order to reach a selection of proteotypic peptides, two different approaches can be employed. The **experimental approach** consists of the LC/MS analyses of the tryptic digest of a heterogeneous protein mixture. Offline pre-fractionation and/or enrichment and online two-dimensional LC can be used to increase the sensitivity/dynamic range of the method as well as the number of detected peptides. The resulting data is matched against a database in order to identify which of the measured peptides belong to the protein of interest. From these peptides, a number of suitable signature peptides (proteotypic peptides) are chosen to be used as stable isotope-labeled synthetic peptide standards. The **theoretical approach** uses the sequence of the protein of interest, which is digested in

silico, yielding tryptic peptides. A number of proteotypic peptides (specific only for the protein of interest) are chosen, taking size, amino acid composition, and tryptic efficiency into account. The experimental setup has to be validated and potential additional iterations concerning the choice of peptides have to be done (Lehmann et al. 2008; **Lyon, Weckwerth, and Wienkoop 2014**; Schulze and Usadel 2010; Wienkoop et al. 2008; Lange et al. 2008).

Databases

The **UniProt** Knowledgebase (UniProtKB) is not only a mere repository for amino acid sequences, but contains functional information on proteins, with accurate, consistent and rich annotation, as well as cross-references to experimental data, amongst others. UniProtKB is divided into two sections, Swiss-Prot, a labor-intensive, high-quality, manually annotated and non-redundant protein-sequence database, as well as TrEMBL, computer-annotated translated nucleotide sequences. UniProt Reference Clusters (UniRef) clusters sets of sequences, from UniProtKB and UniParc, to gain complete coverage of the sequence space, while merging redundant sequences and/or fragments. "The UniRef100 database combines identical sequences and sub-fragments from any organism into a single entry" (http://www.uniprot.org/; Suzek et al. 2007).

A large collection of non-redundant protein sequences of several sources can be found at the National Center for Biotechnology Information (**NCBI**) at the following web-page http://www.ncbi.nlm.nih.gov/protein/. This resource includes protein sequences translated from nucleotide sequences of coding regions of DNA (GenBank), annotated protein-sequence records derived from data in public sequence archives and from computation, curation and collaboration (RefSeq), experimental and inferential, manually annotated sequences (SwissProt), functional information (PIR), and structural information (PDB) (K. Pruitt et al. 2002; K. D. Pruitt et al. 2014; Tatusova et al. 2014; http://www.ncbi.nlm.nih.gov/).

More and more genome sequencing projects are currently running (incomplete projects 21461) and are being completed (completed projects 6646) (https://gold.jgi-psf.org/index, as of October 2014). One specific project aims to sequence and annotate the genome of *M. truncatula* which is still not complete (http://www.jcvi.org/). The umbrella association International Medicago Genome

Annotation Group (**IMGAG**) is a combined effort to re-sequence various inbred *M. truncatula* lines, to characterize Single Nucleotide Polymorphisms (SNP), Insertions/Deletions (INDELs) and Copy Number Variants (CNV), in order to describe the population structure and identify haplotypes. Thus, a long-term, community-accessible Genome-Wide Association (GWA) mapping resource has been created (http://www.medicagohapmap.org; http://www.jcvi.org/).

The generation of Expressed Sequence Tags (**ESTs**) is (or in former times was) faster and cheaper than the generation of entire genomes (Adams et al. 1991). These nucleotide sequences can be translated to protein sequences via Six-Frame-Translation (using e.g. EMBOSS) (Rice, Longden, and Bleasby 2000). Such a collection of protein sequences can serve as a database for shotgun proteomics identification (Larrainzar et al. 2007; Weckwerth 2011a).

ProMEX is a public mass-spectral library of experimental data. The database consists of mass spectra of tryptic peptides derived from *A. thaliana*, *C. reinhardtii*, *M. truncatula* and *Solanum tuberosum* L. Since it is independent of genomic data, it is especially suitable to search for PTMs (Post Translational Modifications) such as phosphorylation sites. New gene models can be derived from proteomics data using this tool. The database includes data from subcellular fractionation, metal oxide affinity chromatography (MOAC) as a phosphoprotein-enrichment strategy, from neutral loss scanning for p-site quantification, drought-stress protein markers, novel protein allergens and pollen development, amongst others (Hummel et al. 2007; Wienkoop et al. 2012; Hoehenwarter et al. 2008; Larrainzar et al. 2009; May et al. 2008; Wienkoop et al. 2008; Lehmann et al. 2008; Reumann et al. 2007; Kierszniowska, Walther, and Schulze 2009; Wienkoop et al. 2010; **Ischebeck et al. 2014**).

Functional annotation

Reasons to use BLAST

Even though the genome of an organism may be fully sequenced, this neither immediately results in the proper assignment of all coding regions, nor the unambiguous functional annotation thereof. Additionally, splice variants, Single Nucleotide Polymorphisms (SNPs) and Insertions/Deletions (INDELs) exist, complicating the picture. Genetically highly diverse species, such as *Pisum sativum* L., and plant polyploidy (e.g. Magnolia) additionally increase the

complexity of the data and its interpretation (Bourgeois et al. 2011; Parris et al. 2010). Similar sequences do not necessarily have the same function and differing sequences have shown to possess the same function. Nevertheless, the curious mind of a biologist can utilize well-known tools such as BLAST in the hope of gaining insights in homology, function, and localization, amongst others (Altschul et al. 1990; http://blast.ncbi.nlm.nih.gov/Blast.cgi).

Applying BLAST

Basic Local Alignment Search Tool (BLAST) is an umbrella term for tools enabling biological sequence comparison to assess homology. Multiple specialized variants (e.g. blastp, blastn, blastx, PSI-blast, DELTA-blast, etc.) exist, dependent on the sequence and problem at hand. Taking the sequence length and the evolutionary distance between the query sequence and the database into account, various substitution matrices exist (e.g. BLOSUM 62, PAM30, etc.), producing different scores that evaluate the search result. Usually, the E-value (expectancy value) is used to evaluate the data. The latter is produced by calculating the fraction of the search space (database size in bits) by the adjusted score (in bits). Thus, low E-values indicate that the rarity of the score overcomes the search space. In other words, a low E-value shows that the hit would not be expected by chance alone, but is due to sequence identity and similarity. Due to the fact that the E-value depends on the query size as well as the database size, there is not a single ubiquitous threshold cutoff value to be used. Within the publications presented in this work, a cutoff value of 1 * e-3 was used. This particular value was chosen in order to produce a sufficient amount of query hits, while trying to retain the stringency to get meaningful results, since only 1 * e-3 hits would be expected to be seen by chance alone.

A. thaliana has enjoyed most attention in the world of molecular plant biology, therefore the genome annotation is more advanced compared to other model plants. Subsequently, homology-based information transfer is often performed against *A. thaliana* (i.e. TAIR10 at the present time) (**Ischebeck et al. 2014**; http://blast.ncbi.nlm.nih.gov/Blast.cgi).

MapMan and GO

Mercator is an automated pipeline, an efficient annotation tool for functional genomic and proteomic sequences. It utilizes the MapMan bin ontologies for functional annotation of plant omics data. Controlled vocabularies and func-

tional ontologies simplify the exchange of information and enable computational approaches. 36 major MapMan bin categories exist (e.g. Photosynthesis, Tri Carboxylic Acid cycle (TCA), etc.), each with various specialized sub-categories, modeled as a tree (Thimm et al. 2004; Lohse et al. 2014; Usadel et al. 2009). A typical use case would be the visualization of gene expression and/or integrated proteomics and metabolomics data, comparing a control to a treatment group. By selecting different so-called "pathways", versatile data exploration can be performed, since a general overview as well as particular functional categories can be visualized and thus inspected for patterns of differential expression in abundance levels (**Staudinger et al. 2012**; Tellström et al. 2007; **Ischebeck et al. 2014**). MapMan was originally designed for plants, specifically *A. thaliana*, in contrast to the species-unspecific Gene Ontology (GO). There are three GO ontologies that conform to independent categories of gene function: molecular function (GO-MF), biological processes (GO-BP), and cellular component (GO-CC). The dependencies are realized in a directed acyclic graph (Ashburner et al. 2000; Klie and Nikoloski 2012).

Another important resource is the Protein Ontology (PO, formerly PRO), providing an ontological representation of protein-related entities and showing relationships between them. Genes, taxon-neutral to species specific protein sequences, protein complexes, and PTMs are represented (if available). PO complements databases such as UniProtKB and interoperates with other ontologies such as GO (Natale et al. 2014; http://proconsortium.org).

Data Integration and Interpretation – Bioinformatics

The goal of many systems biology approaches is to recognize differential patterns of expression as a consequence of specific treatments and/or time courses. Observing that transcripts/proteins/metabolites associated with particular pathways are enriched/up- or down-regulated (Valledor et al. 2013; Hoehenwarter et al. 2008; Jurgen Cox et al. 2014; Wienkoop and Weckwerth 2006). In situ, analytes differ by several orders of magnitude in abundance, thus posing difficulties for data acquisition and therefore also when evaluating their relative expression patterns. Additionally, technical and batch-to-batch

variability have to be accounted for, to delineate the latter from biological variability and differential expression due to sample treatments. In order to draw as holistic a picture as possible and thus maximize the information content, all MS-measured analytes and other (e.g. physiological) parameters can be merged into a single data matrix (**Doerfler et al. 2013**; **Mari et al. 2013**). Since these values often consist of a wild mixture of SI (The International System of Units) as well as arbitrary units, and can differ by several orders of magnitude, data normalization/transformation/standardization is needed. To illustrate: High-abundance proteins have higher values (peak area, spectral count) compared to low-abundance or small proteins; using multivariate statistics this may cause high ranking for high-abundance proteins. Data transformation is necessary to rescale data, enabling small, but potentially statistically significant changes to emerge from the data set. Thus, log (e.g. base 2 or base 10) and/or "z-transformation"/"standardization" (or other methods) should be applied (**Staudinger et al. 2012**; Weckwerth 2007). In shotgun proteomics, when e.g. using spectral counting, the protein size will greatly influence the magnitude of the measured abundance value. One method of resolution is the Normalized Spectral Abundance Factor (NSAF), another the exponentially modified Protein Abundance Index (emPAI) (Schulze and Usadel 2010; **Ischebeck et al. 2014**). Novel methods are constantly being developed to increase the quantitative accuracy of MS data. One, for instance, is the "proteomic ruler", which normalizes protein abundance to histone content (Winiewski et al. 2014). In metabolomics, internal standards are often used for sample normalization. Technical variability (e.g. ESI-quality, sample injection volume, skimmer or S-lens pollution, etc.) can dramatically influence the signal intensity of MS measurements. In order to cope with such problems, normalization can be applied in an intra-experimental/batch-fashion. This can be illustrated with a simple example of building the sum of all signals for each measurement (sum_m), as well as the overall sum, the sum of all these sums (sum_all). Subsequently, each individual value is divided by the sum_m of its corresponding measurement and multiplied by the fixed sum_all value. Thereby, the varying signal intensities of individual measurements (e.g. one LC/MS "run") are represented as the fraction of the total intensity of the measurement, and then all values are scaled to the same "baseline-intensity", while retaining the biologically relevant variabil-

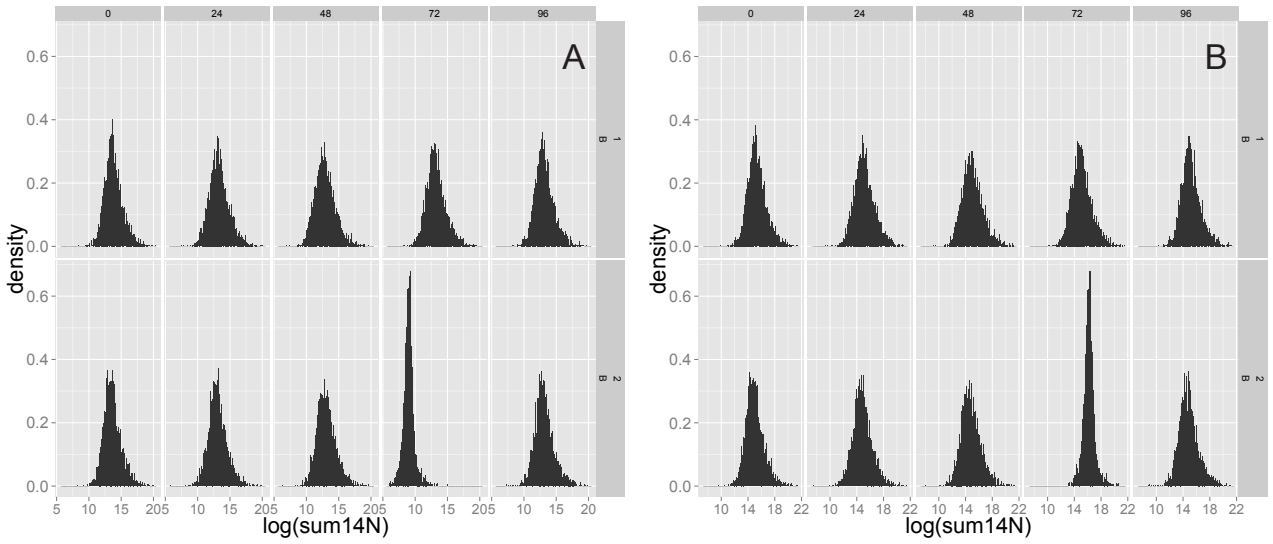


Figure 1 legend:
The histograms display the intensity of all peptide signals (log base 10) versus the density of the occurrence. The sub-plots are divided on the abscissa by five distinct TimePoints (0h, 24h, 48h, 72h, 96h) and the ordinate is divided by two technical replicates (both of the biological replicate "B"). Plot A shows the original data, and plot B the transformed data.

ity (see Figure 1). Figure1.Sub-plot A.sample72h.B2 stands out in comparison to the histograms of the other samples, due to its shift on the x-axis (smaller values) as well as its more positive kurtosis. Sub-plot B.sample72h.B2 shows a clear shift of the values to the right, to elevated numbers, conforming with the other sub-plots. However, this shift does not influence the distribution of values and therefore the positive kurtosis remains unchanged. Naturally, such a data transformation does not resolve any type of difficulty concerning outliers or even missing values. Numeric and visual inspection of the data and testing various transformation strategies is an important aspect of data analysis. This data is from an unpublished data set.

Data to stats: getting data ready for analysis

Preparing data for statistical analysis is often anything but trivial. Careful considerations have to be taken concerning missing values and data normalization/transformation/standardization (Gromski et al. 2014). Another difficulty is the expected input format of software capable of statistical analysis. E.g. "R", "Matlab-COVAIN", or commercial programs such as "SIMCA" and "Statgraphics" (to name only a few) all expect a specific input, which has to be

accounted for (Sun and Weckwerth 2012; R Development Core Team 2008). I have used scripting languages such as “Python” or “R” to apply established functions such as “melt” or “pivot”, as well as implemented other data transformations to perform exploratory data analysis (**Lyon and Castillejo et al. 2014** *in preparation*).

Software

As technological aspects of Mass Spectrometry are rapidly developing, experimental design towards a better understanding of biology adapts and evolves (Bantscheff et al. 2007). Since research doesn’t end with the mere acquisition of experimental data, but necessitates the computational analysis of such biological information, adaptation of existing as well as creation of new software tools and programs is essential (Kohlbacher et al. 2007; Pluskal et al. 2010; Deutsch et al. 2010; Lommen 2009; Lommen, Gerssen, and Oosterink 2011). This includes adaptation to novel community standards (e.g. file formats such as “mzXML” to “mzML” and novel standards such as “qcML” and “mzTAB”), data processing parallelization to increase speed, connecting available individual resources by creating pipelines, automation and inventing novel algorithms, to generally enable high throughput analysis of omics data (this is not meant to be a comprehensive list of bioinformatics tasks) (Martens et al. 2011; Walzer et al. 2014; Bald et al. 2012; Goloborodko et al. 2013; Choi et al. 2014; Griss et al. 2014). Last but definitely not least, the visualization of data serves to get an overview, to zoom into specific aspects (see Figure 1), it can highlight mistakes or technical problems and aids in the explanation of interpretations when conveying information in the form of publications or lectures (Gehlenborg et al. 2010). Manual data analysis is not feasible for high-throughput MS data, simply due to the excess of information, as well as potentially justifiable, but irreproducible decisions taken by the data analyst. Therefore, the automated identification and quantification of high-throughput MS data is a necessity (Jürgen Cox and Mann 2008; Jürgen Cox et al. 2011; Lommen 2009; Lommen, Gerssen, and Oosterink 2011; Deutsch et al. 2010; Kohlbacher et al. 2007; Pluskal et al. 2010; Röst et al. 2014; Egelhofer et al. 2013; Smith et al. 2006; McIlwain et al. 2014). The difficulty lies in the translation from human pattern recognition combined with specialized knowledge and intuitive decision-making to stringent math-

ematical descriptions, operations and heuristics that can be implemented as an algorithm in software. This type of data processing can be done with various software solutions. Two prominent and adequate programming/scripting languages are “Python” and “R”, which I have used extensively throughout the present work. “Python” is known for its ease of use, flexibility and rapid development times, and “R” for statistical calculations as well as its powerful graphics engine (Rossum and Drake 2001; R Development Core Team 2008). I have extensively used “IPython” (web-based interactive computation environment for Python) as well as the R package “ggplot2” (Grammar of Graphics) (Perez and Granger 2007; Wickham 2009). For high performance computing, other programing languages such as (the low-level programming language) C++ are preferably used.

Application to plant research

The following section outlines the specific background and objectives of the PhD thesis. The major part of this work was conducted during the FWF project (Fonds zur wissenschaftlichen Förderung/Austrian Research Fund), which will be described in more detail as follows (FWF project number P23441-B20, http://homepage.univie.ac.at/stefanie.wienkoop/fwf23441_en.html).

LC/MS method development and data analysis

Establishing a robust and reproducible LC/MS metabolomics platform on an Orbitrap Mass Spectrometer, combined with subsequent data analysis, is not a straightforward task, if samples of various species, organs and fractions are to be considered. A single methodology is not capable of coping with all possible analytical tasks, and if a nearly all-encompassing method would be established, it might be too impractical, due to extended laboratory work and measurement time. Therefore, depending on the analytes to be analyzed, choices have to be made concerning the Mass Spectrometry setup: measuring in positive or negative ion mode, scan range, Mass Resolving Power, LockMass,

number of MS/MS scan events, micro scans, fragmentation mode and collision energy settings, dynamic exclusion/inclusion lists, charge state recognition, intensity thresholds, etc. (**Doerfler et al. 2013**; **Doerfler et al. 2014**; **Mari et al. 2013**; Koulman et al. 2009; Xu et al. 2010; Vincent et al. 2013). Additionally, due to the Ion Trap, interesting MS^n scan events can be created. Many derivatives of known Flavonoids exist, whose MS/MS spectra cannot be annotated, since only the neutral loss of sugars or organic acids from the precursor is visible in the spectrum. MS^n spectral trees can result in the fragmentation of e.g. the Agylcon backbone of a glycosylated-Flavonoid. Such a spectrum can be compared to spectra of known reference compounds. The Aglycon compound can therefore be identified and the sugars/organic acids inferred (Matsuda, Yonekura-Sakakibara, et al. 2009; **Doerfler et al. 2013**; **Doerfler et al. 2014**). Such MS^n trees can be performed in a "flat" fashion (e.g. the five most abundant ions of the MS/MS scan event will be used for MS/MS/MS fragmentation, i.e. 1 x MS, 1 x MS/MS, and 5 x MS/MS/MS), or a "deep" fashion (e.g. the most abundant ion from the MS/MS scan event will be used for MS/MS/MS fragmentation, subsequently the most abundant ion will be subjected to MS/MS/MS/MS fragmentation and so on and so forth, i.e. 1 x MS, 1 x MS/MS, 1 x MS/MS/MS, 1 x MS/MS/MS/MS, and 1 x MS/MS/MS/MS/MS). Due to limited scan event numbers (inherent limitation of the software), considering a sufficient number of MS scans for peak integration (duty cycle time), lack of sufficient ions to acquire sufficient Signal to Noise ratios, etc., these scan events have to be restricted (even with modern and fast Mass Spectrometers). I have developed several methods for MS/MS as well as MS^n fragmentation trees, which resulted in an improved reproducibility of spectra and an increase in the Signal to Noise ratio, while retaining sufficient MS scans to perform quantitation (methods and data largely unpublished, though some aspects were applied in the following publications) (**Doerfler et al. 2013**; **Mari et al. 2013**; **Doerfler et al. 2014**). Furthermore, I've suggested the use of spiking peptides into the metabolite sample as internal reference standards, in order to use the subsequent signals to build a Retention time index in analogy to GC/MS metabolomics.

Many of these settings additionally depend on the chromatographic setup

and the acquisition speed of the Mass Spectrometer, since peak widths and the duty cycle time of the entire MS method should be considered when setting the dynamic exclusion duration of precursors as well as the number of scan events and types. Sample injection volumes of 5 µl have a great influence on the chromatographic performance of 100 µm I.D. columns, in contrast to more conventional 4.6 mm I.D. columns. This needs to be considered when choosing sample solvents for injection, as well as the initial gradient conditions. Naturally, not all types of analytes can be separated with commonly used Reversed Phased Liquid Chromatography (RPLC) or are highly soluble at initial gradient conditions. Therefore, other modes such as Hydrophilic Interaction Liquid Chromatography (HILIC) have to be employed, MS compatible mobile phases used, and during the subsequent MS data analysis, attention paid to potentially differing adduct formations, dependent on the mobile phase and the concentration of the analyte in question. Comparing Retention times of analytes measured with orthogonal chromatographic techniques, in order to gain additional information concerning the annotation, is difficult to impossible (Lu et al. 2010; Koulman et al. 2009; Xu et al. 2010; Vuckovic and Pawliszyn 2011; Matsuda, Yonekura-Sakakibara, et al. 2009; Matsuda, Shinbo, et al. 2009; Goodacre et al. 2004; Lee et al. 2011). Apart from these technical MS aspects, there is a very limited to non-existent choice of software solutions available to cope with such data.

Following data acquisition, data analysis is the next challenging task. A multitude of methods exist, the choice of which depends on the data acquisition. These techniques will not be described, since this would go beyond the scope of this work. Rather, a brief overview of the data analysis used for the publications pertinent to this thesis will follow. Depending on the data acquisition method and the samples, the Mass Spectrometer chooses which precursors to select for MS/MS fragmentation (data dependent MS/MS triggering). Therein lies a lot of information which was used by ProtMAX to align and extract data from multiple LC/MS measurements, building a data matrix of unidentified m/z-features and their intensities (Egelhofer et al. 2013). The raw LC/MS data was manually analyzed, in order to find proper settings for ProtMAX data extraction. The LC/MS data matrix was filtered for reproducible signals, merged with

the analogous GC/MS data matrix, subsequently transformed and subjected to statistical analysis and modelling (see "Data Integration and Interpretation – Bioinformatics" and (**Doerfler et al. 2013**; **Mari et al. 2013**; Egelhofer et al. 2013)).

Many of the aforementioned technical considerations apply to Proteomics MS analysis as well. For more accurate quantification and a wider dynamic range, Triple Quadrupole Mass Spectrometers (QqQ) were used (**Lyon, Weckwerth, and Wienkoop 2014**).

Specific background and objectives of the FWF project

The production of nitrogen fertilizers is an energy-demanding process, the use of such fertilizers is economically expensive and by leaching and eutrophication can negatively influence biodiversity and ecological processes. The capability of mutualistic symbiosis of legumes with rhizobia is significant for sustainable agricultural systems worldwide, specifically due to the metabolic exchange of nitrogen-fixing bacteria with their host plant. In this relationship, legume plants can form so-called root-nodules (bulbous growths interspersed throughout the length of the plant root-system) encapsulating specialized bacteria, specific to the host plant. The bacteria can assimilate the abundant atmospheric nitrogen (~ 78%), transform it to ammonia, and subsequently provide the plant with this important and potentially growth-limiting nutrient. In exchange, the plant feeds the bacteria with sugars and other nutrients. This process is called **Symbiotic Nitrogen Fixation** (**SNF**). When legumes grow without these endosymbionts, they need to be fertilized with nitrogen (amongst other nutrients). Accordingly, the opposing terms "**N-fix**", for SNF, and "**N-fed**", for fertilized plants, can be distinguished. SNF is susceptible to environmental stresses, particularly drought, which inhibits SNF and subsequently reduces crop yields. *M. truncatula*, a sequenced model organism of legumes, serves to study physiological, metabolic as well as proteomic responses to drought stress within the FWF project. The molecular mechanisms regulating the differential control of water relations during drought is of fundamental importance to plant physiol-

ogy. Elucidated mechanisms could hopefully be applied to closely related crop species such as *P. sativum* and *Medicago sativa* L. (Larrainzar et al. 2009; Larrainzar et al. 2007; **Staudinger et al. 2012**).

M. truncatula is a diploid, annual, forage plant that serves as a model organism for the large Fabaceae family (Legumes) of which most species can perform SNF. Other prominent species of the family are *Lens culinaris* Medik., *Phaseolus vulgaris* L., and *Glycine max* (L.) Merr. Apart from the Poaceae, Fabaceae are the most import plant family to humans, due their use as grains, pasture, and in agroforestry. Due to SNF, they can colonize low-nitrogen environments, thereby improving the soil and thus play a critical role in natural ecosystems, agriculture, and agroforestry (http://medicago.vbi.vt.edu/; Graham and Vance 2003; Trevaskis et al. 2002). Next-generation sequencing technologies enable time and cost efficient acquisition of entire genomes, thus also paving the way for database dependent shotgun proteomics experiments (Metzker 2010). Though *M. truncatula* is almost completely sequenced, the genome annotation is not complete. Proteomic databases specific for Medicago do not reflect the "entirety" of potential proteins (e.g. UniProt entries specific for *M. truncatula*, (http://www.uniprot.org)) or are imperfectly annotated (e.g. International Medicago Genome Annotation Group (IMGAG), (http://medicago.org/genome/IMGAG/)). Consequently, many proteins of *M. truncatula* are regarded as putative proteins, particularly protein databases created from nucleotide sequences via automated gene prediction. In an effort to increase the number of protein identifications while concurrently reducing computational cost, I have created non-redundant fused databases (for *M. truncatula* and other species) using various sources, with the intent of providing data to ameliorate genome annotation (e.g. ProMEX data is cross-referenced in UniProt). Hence, such a single database can be used to analyze shotgun-proteomics LC/MS data, in contrast to evaluating the same data with multiple databases (**Staudinger et al. 2012**; **Lyon et al. 2014**). Larger non-redundant databases will automatically lead to an increase in Peptide Spectrum Matches (PSMs) for the target as well as the decoy database. Discord exists about the use of such fused databases, largely due to inconsistent functional annotation of proteins. Reassigning the identification results to the original databases poses another difficulty. With different versions of genome annotations (of the same species, utilizing the exact same

genomic sequence), some confidently and repeatedly experimentally validated proteins can vanish (Valledor et al. 2012). It is apparent that functional annotation of *M. truncatula* is ongoing and that unbiased, reproducible shotgun proteomics data is a valuable resource for proteogenomics, amongst others.

In order to gain more functional information, I have performed homology searches, using BLAST, against the well-characterized species *A. thaliana* (**Ischebeck et al. 2014**). I have used the Mercator pipeline for the functional annotation of the aforementioned fused protein sequence databases specific to species such as *M. truncatula*. Due to size restrictions and ambiguous difficulties with protein sequence recognition (largely due to interspersed asterisks, "X" and other single letter amino acid codes not describing the most simplistic 20 proteinogenic amino acids for humans, short repeats, sequence length, etc., mostly derived from six-frame-translation of nucleotide sequences), semi-automated filtering had to be performed preceding the upload to the Mercator pipeline. This task was achieved using in-house unpublished Python scripts as well as manual investigation, which I performed. The functional categorization served not only for visualization, but also for clustering the data (in analogy to GO-term enrichment). The resulting data were used within the following publications (**Staudinger et al. 2012**; **Lyon et al. 2014**; **Ischebeck et al. 2014**).

Progress in unbiased as well as targeted analyses of proteins and metabolites is tightly linked to technological advances especially in the field of Mass Spectrometry (MS) (Glinski and Weckwerth 2006; Hu et al. 2005; Olsen et al. 2005). Shotgun proteomics using Liquid Chromatography coupled to Mass Spectrometry (LC/MS and LC/MS/MS) generates tremendous amounts of data (expressed protein sequence information). Within this project, proteins identified with high confidence, through computer assisted database dependent identification, are stored in the publicly accessible spectral reference database PROMEX (http://promex.pph.univie.ac.at/promex/) (Hummel et al. 2007; Wienkoop et al. 2012). This reference database is cross-linked to UniProt (Apweiler and Consortium 2012; http://www.uniprot.org). Thus, experimentally validated protein sequences characterized as putative can ameliorate genome annotation. Furthermore, information for better functional characterization can potentially be supplied through sample information (e.g. sample of roots

and nodules vs. leaves, subcellular fractionation, enrichment, etc.), experimental design (comparing sample groups), and sequence homology searches. Thus, the research objectives are defined as follows:

1. Functional characterization and visualization of proteins for legumes

Irrespective of database-dependent identification, computationally assisted data interpretation and visualization is necessary to cope with the given data. Mapping functional annotation to protein sequences and therefore to unique identifiers such as Accession Numbers, enables biologically meaningful data interpretation, functional clustering, and additional visualization methods. As previously mentioned, one task consisted of merging protein FASTA files (FASTA-format-files will henceforth be referred to as FASTA files) of various sources, creating a non-redundant protein database specific for the given organism. These protein FASTA files were subjected to the aforementioned Mercator pipeline to utilize the resulting MapMan functional categories. In order to find homologs and therefore gain additional information, protein FASTA databases were subjected to BLAST against e.g. TAIR10 or other better characterized databases of closely related species. This type of bioinformatics work was performed not only for *M. truncatula*, but also for *Nicotiana tabacum*, *G. max*, *Phaseolus vulgaris*, *Lotus japonicus*, *Oriza sativa*, *Physcomitrella patens*, *P. sativum*, and *Triticum aestivum* (**Staudinger et al. 2012**; **Ischebeck et al. 2014**; **Gil-Quintana 2014** *submitted*; **Meisrimler et al. 2014** *in preparation*; as well as unpublished work).

2. ^{15}N labeling, development of an automated program for protein turnover calculations

In shotgun proteomics, the comparison of the abundance of peptides corresponding to specific proteins yields the Fold Changes of Proteins (FCP) (Hoehenwarter and Wienkoop 2010; L. Li et al. 2012). The FCP is widely used to compare protein levels of various samples, but neither resolves the dynamics of the proteome in the different biological states that are being compared nor the mechanisms whereby the system changes from one state to the other (Pratt et al. 2002). The measurement of a protein in steady-state conditions will be the

result of the change in its synthesis rate compared with the change in its degradation rate (Pratt et al. 2002). Metabolic (in situ) labeling is characterized by the incorporation of stable isotopes into the proteins of organisms via growth on media or food (Kline and Sussman 2010). The advantage of this technique is that the tags are very subtle with insignificant impact on cellular processes and allow the fully functional proteins to be produced and distributed within cells in a normal context (L. Li et al. 2012). Since only fully labeled or unlabeled peptides can be identified automatically with current software, the SELPEX (Selective Peptide Extraction) approach was used for robust MS data extraction (Castillejo et al., 2012). Mass spectrometric measurement of the ratio between light (naturally occurring isotopic distribution) and heavy (enriched with ^{15}N) isotopic peaks and their respective degrees of enrichment provide a means to measure the synthesis and degradation rates of individual proteins (Pratt et al. 2002; Cargile et al. 2004). Such approaches have been shown for cell cultures, and most approaches rely on previously identified sequences (Martin et al. 2012; L. Li et al. 2012). Software tools coping with partial metabolic labeling data in an automated fashion have recently been published. Nevertheless, novel algorithms to approach such complex data are needed (**Lyon et al. 2014**).

The main goal of the proposed PhD project was the mining and processing of LC/MS proteomics data from partial metabolic labeling experiments; focusing on developing a robust automated system to extract and align peptide isotopic envelopes in order to determine the ratio between light (naturally occurring isotopic distribution) and heavy (enriched with ^{15}N) peptide spectra. The major objectives and challenges are the allocation of individual peptide signals within multiple measurements, the link between database dependent identification and ^{15}N data analysis, and database independent ^{15}N data analysis. A peer-reviewed article showing the successful implementation of a robust automated program determining peptide label ratios from LC/MS analyses was published (**Lyon et al. 2014**).

The relation of rhizobia with their host plants, and thus Symbiotic Nitrogen Fixation, its importance to the nitrogen cycle and the influence of drought on the latter, is the central theme of the FWF project (FWF project number

P23441-B20, http://homepage.univie.ac.at/stefanie.wienkoop/fwf23441_en.html). Protein turnover information from *M. truncatula* is unavailable to date. Studying drought stress of plants is of fundamental importance for sustainable human agriculture worldwide. In order to study the differential regulation of protein degradation and synthesis during the recovery phase of drought stress, an experiment was conducted, comprising *M. truncatula* growing under controlled conditions. Six-week-old plants were subjected to drought stress and re-watered with ordinary as well as ^{15}N enriched inorganic fertilizer. The incorporation of heavy nitrogen sources (NH_4NO_3) into amino acids and subsequently into proteins results in complex composite isotopic envelopes, changing in time as the incorporation progresses (**Lyon and Castillejo et al. 2014** *in preparation*). This elegant experimental design enables the study of the recovery phase of drought stress by analyzing protein turnover information in conjunction with more conventional proteomics data. This was done to gain novel insights and broaden our understanding of the molecular dynamics under such conditions and in the hope of transferring this knowledge to crop plants (see Outlook).

Publications

A brief overview of all published manuscripts (8) I have contributed to ensues, followed by more detailed descriptions and the publications themselves.

The following two publications are an integral part of the previously described FWF project:

Lyon, David, Maria Angeles Castillejo, Christiana Staudinger, Wolfram Weckwerth, Stefanie Wienkoop, and Volker Egelhofer. 2014. "Automated Protein Turnover Calculations from ^{15}N Partial Metabolic Labeling LC/MS Shotgun Proteomics Data." PloS One 9 (4): e94692. doi:10.1371/journal.pone.0094692.

Staudinger, Christiana, Vlora Mehmeti, Reinhard Turetschek, **David Lyon**, Volker Egelhofer, and Stefanie Wienkoop. 2012. "Possible Role of Nutritional Priming for Early Salt and Drought Stress Responses in Medicago Truncatula." Frontiers in Plant Science 3 (December): 285. doi:10.3389/fpls.2012.00285.

The following two publications, as well as the last-mentioned publication, are first-author publications, and therefore an integral part of this thesis:

Lyon, David, Wolfram Weckwerth, and Stefanie Wienkoop. 2014. „ Mass Western for absolute quantification of target proteins and considerations about the instrument of choice.", Plant Proteomics. Edited by Jesus V. Jorrin-Novo, Setsuko Komatsu, Wolfram Weckwerth, and Stefanie Wienkoop. Vol. 1072. Methods in Molecular Biology. Totowa, NJ: Humana Press. doi:10.1007/978-1-62703-631-3. http://link.springer.com/10.1007/978-1-62703-631-3.

Doerfler, Hannes, **David Lyon,** Thomas Nägele, Xiaoliang Sun, Lena Fragner, Franz Hadacek, Volker Egelhofer, and Wolfram Weckwerth. 2012. "Granger Causality in Integrated GC–MS and LC–MS Metabolomics Data Reveals the Interface of Primary and Secondary Metabolism." Metabolomics, October. doi:10.1007/s11306-012-0470-0. (Hannes Doerfler, **David Lyon**, Thomas Naegele and Xiaoliang Sun **contributed equally to this work**.)

Finally, the following publications were published during this PhD thesis and are pertinent to the latter in form and content. The first publication of the following list was a contribution related to the development of the FWF project, transferred to other plant species:

Ischebeck, Till, Luis Valledor, **David Lyon**, Stephanie Gingl, Matthias Nagler, Mónica Meijón, Volker Egelhofer, and Wolfram Weckwerth. 2014. "Comprehensive Cell-Specific Protein Analysis in Early and Late Pollen Development from Diploid Microsporocytes

to Pollen Tube Growth." Molecular & Cellular Proteomics : MCP 13 (1): 295–310. doi:10.1074/mcp.M113.028100.

Doerfler, Hannes, Xiaoliang Sun, Lei Wang, Doris Engelmeier, **David Lyon**, and Wolfram Weckwerth. 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.

Egelhofer, Volker, Wolfgang Hoehenwarter, **David Lyon**, Wolfram Weckwerth, and Stefanie Wienkoop. 2013. "Using ProtMAX to Create High-Mass-Accuracy Precursor Alignments from Label-Free Quantitative Mass Spectrometry Data Generated in Shotgun Proteomics Experiments." Nature Protocols 8 (3): 595–601. doi:10.1038/nprot.2013.013.

Mari, Angela, **David Lyon**, Lena Fragner, Paola Montoro, Sonia Piacente, Stefanie Wienkoop, Volker Egelhofer, and Wolfram Weckwerth. 2013. "Phytochemical Composition of Potentilla Anserina L. Analyzed by an Integrative GC-MS and LC-MS Metabolomics Platform." Metabolomics : Official Journal of the Metabolomic Society 9 (3): 599–607. doi:10.1007/s11306-012-0473-x.

Automated Protein Turnover Calculations from ^{15}N Partial Metabolic Labeling LC/MS Shotgun Proteomics Data

Protein turnover reflects the combination of protein synthesis and degradation, independent of the absolute or relative abundance of proteins. Systems biology approaches demand high throughput measurements, which in turn generate copious amounts of data, not only in shotgun proteomics. Novel algorithms and their implementation into user-friendly computer programs, aiming to cope with such large amounts of data, are needed, since manual data analysis is not feasible.

Declaration of authorship

The results of this chapter are presented in the form of a manuscript published in the journal „PLOS One“. The work presented in the following manuscript is largely my own. I have developed the **ProtOver** algorithm, programmed the software, contributed to the LC/MS method and measurement, and written the manuscript.

Published manuscript

OPEN ACCESS Freely available online

PLOS ONE

Automated Protein Turnover Calculations from ^{15}N Partial Metabolic Labeling LC/MS Shotgun Proteomics Data

David Lyon, Maria Angeles Castillejo, Christiana Staudinger, Wolfram Weckwerth, Stefanie Wienkoop, Volker Egelhofer*

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

Abstract

Protein turnover is a well-controlled process in which polypeptides are constantly being degraded and subsequently replaced with newly synthesized copies. Extraction of composite spectral envelopes from complex LC/MS shotgun proteomics data can be a challenging task, due to the inherent complexity of biological samples. With partial metabolic labeling experiments this complexity increases as a result of the emergence of additional isotopic peaks. Automated spectral extraction and subsequent protein turnover calculations enable the analysis of gigabytes of data within minutes, a prerequisite for systems biology high throughput studies. Here we present a fully automated method for protein turnover calculations from shotgun proteomics data. The approach enables the analysis of complex shotgun LC/MS ^{15}N partial metabolic labeling experiments. Spectral envelopes of 1419 peptides can be extracted within an hour. The method quantifies turnover by calculating the Relative Isotope Abundance (RIA), which is defined as the ratio between the intensity sum of all heavy (^{15}N) to the intensity sum of all light (^{14}N) and heavy peaks. To facilitate this process, we have developed a computer program based on our method, which is freely available to download at http://promex.pph.univie.ac.at/protover.

Citation: Lyon D, Castillejo MA, Staudinger C, Weckwerth W, Wienkoop S, et al. (2014) Automated Protein Turnover Calculations from ^{15}N Partial Metabolic Labeling LC/MS Shotgun Proteomics Data. PLoS ONE 9(4): e94692. doi:10.1371/journal.pone.0094692

Editor: Lennart Martens, UGent/VIB, Belgium

Received January 30, 2014; **Accepted** March 18, 2014; **Published** April 15, 2014

Funding: This work was funded by FWF, Project Number: P23441-B20 (Austrian Science Fund), and Spanish Ministry of Education, through the Mobility Program R-D + I 2008-2011, code 2010-0236. 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: volker.egelhofer@univie.ac.at

Introduction

In shotgun proteomics the FCP (Fold Change in Protein) is widely used to compare protein levels of various samples, but neither resolves the dynamics of the proteome in the different biological states that are being compared, nor the mechanisms whereby the system changes from one state to the other [1–3]. The elevated abundance of a protein could be the result of an increased synthesis or a decreased degradation rate or a combination of the latter. In recent years, numerous publications employed protein turnover to gain more insight into the regulation of protein abundance [4–13]. SILAC-based experimental data can be analyzed with the freely available MaxQuant software for identification and quantification purposes [14]. However, a user-friendly, fully automated, and freely available tool is needed, enabling the extraction of complex partial metabolic labeling data for high throughput studies.

Since plants are capable of synthesizing their own amino acids, supplying them with an inorganic nitrogen source enriched with ^{15}N leads to the incorporation of ^{15}N into amino acids and subsequently into fully functional proteins. The higher the degree of ^{15}N incorporation, the higher the mass shift of the resulting mass spectrum. Full incorporation of ^{15}N results in a mass shift of all isotopic peaks compared to the ^{14}N form of the peptide (see purple spectrum in Figure 1). In the latter spectrum, there are still isotopic peaks present, mainly due to the contribution of ^{13}C. A vast number of combinatorial possibilities of isotopomers and isotopologues range from the light ^{14}N to the pure ^{15}N form, known as partially labeled peptides. The resulting mass spectra of individual proteolytic peptides are a composite of all peptide species of variable ^{15}N incorporation (see also example Figure 1). This adds to the inherent complexity of biological shotgun-proteomics samples, due to the increased isotopic envelope of individual spectra. Therefore, the main objective of this work was to develop an efficient algorithm for fully automated protein turnover calculations, which can be applied to any kind of sample data arising from partial metabolic ^{15}N labeling experiments, no matter the type of organism or tissue.

Software tools coping with partial metabolic labeling data in an automated fashion already exist. Commercial in conjunction with freely available software were used to analyze mammalian pulse chase LC/MS data [15–17]. The latter method relies on a combination of ^{14}N and ^{15}N spectral counts with MS1 information, and requires every peptide quantitation event to have an associated ^{15}N MS2 peptide identification [15]. Thus, fully ^{15}N labeled peptide species are essential, in contrast to the method presented within this manuscript, which aims to analyze partially ^{15}N labeled peptides.

The software "ProTurnyzer" introduced by [18] is available upon request. It accepts pep.xml files in conjunction with RAW data files from Thermo Scientific. Each RAW file (LC/MS file) depends on one corresponding pep.xml file containing the peptide

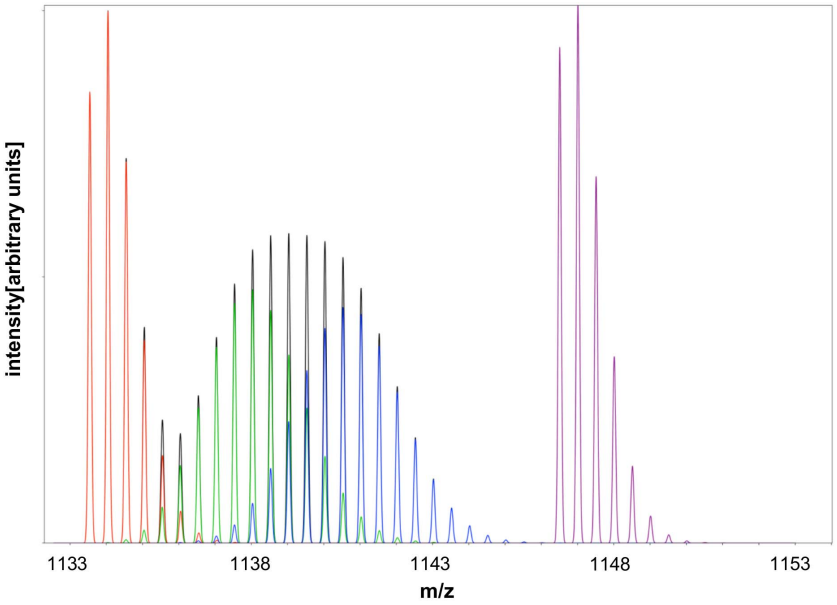


Figure 1. Simulated spectrum of isotopic distribution of the peptide sequence ''MPSAVGYQPTLGTEMGTLQER'' (charge state 2). The spectrum consists of a peptide species with natural isotopic distribution (red), a peptide with 30% ^{15}N incorporation (green), a peptide with 50% ^{15}N incorporation (blue), and a peptide with 100% ^{15}N incorporation (purple). The sum of all composite spectra is displayed in black.
doi:10.1371/journal.pone.0094692.g001

sequence and retention time information necessary to extract data from the RAW files. This means that every RAW file is used for peptide identification purposes as well as for protein turnover analysis. Accordingly, each LC/MS measurement has to be subjected to database dependent identification. Since the vast majority of shotgun proteomics search engines rely on the MS2 spectra of monoisotopic precursors for identification purposes, this approach is only applicable for very low partial metabolic labeling rates.

Another software processing ^{15}N partial metabolic labeling data, named "TurnStile" [9], is available upon request. The program uses centroided mzXML and MS Excel files (providing peptide sequence, charge state, and retention time start and end information) to extract spectral envelopes. Subsequently, multiple spectra are averaged and fitted in order to derive the ^{15}N incorporation percentage and intensities of the light and heavy isotopic envelopes. Retention time values can be adapted for each file individually.

Both "ProTurnyzer" and "TurnStile" process each LC/MS file individually, average over multiple spectra (producing a single averaged spectrum for the extraction of the spectral envelope of each peptide) and subsequently fit experimental data to theoretical values. One of the main differences between "TurnStile" and "ProTurnyzer" is how they calculate the averaged spectra. "TurnStile" averages over all scans within a given retention time start and end point, and have applied a 3 min window for their data [9]. In contrast, "ProTurnyzer" extracts peak intensities from RAW MS1 scans within an elution time window of 60 s before and after the corresponding MS2 scan, by summing up all intensities bound by local minima surrounding the maximum within 20 ppm [18].

The presented method is fundamentally different to the previously mentioned methods, since all LC/MS files of a time series experiment are processed together, in reverse chronological order (from the maximally to the minimally labeled state). The basic idea behind this approach is the assumption that the spectral envelope of the maximally labeled Time Point will always have the maximum number and intensity of isotopic peaks, given that the monoisotopic precursor is still present. This leads to the best signal to noise ratio for the isotopic peaks. The peak picking of every Time Point depends on the previous one. Thus, an interdependency of Time Points is established, that reduces picking of noise. The application expects centroid or profile mode mzML files in conjunction with a text file containing peptide identification information. This algorithm has been implemented in a program, written in Python, which is freely available to the scientific community at http://promex.pph.univie.ac.at/protover.

Methods

Since we cannot assume that every protein will be present in the sample at any given time (or present in a detectable quantity), the question remains for which proteins/peptides to look in a partial metabolic labeling LC/MS shotgun proteomics data set, if this data cannot be used for peptide identification. We have circumvented this problem with the experimental design of our study. Parallel to a ^{15}N labeled sample group, we have grown a set of ^{14}N control plants. The LC/MS data generated from samples of the latter group was used for peptide identification (for details see Document S1). Seven-week-old *M. truncatula* plants were split into two groups a control (non-labeled) and a treatment (fertilized with ^{15}N enriched ammonium nitrate) group. Samples were taken for five consecutive days. After protein extraction and digestion, the samples were analyzed by LC/MS. Since the incorporation of ^{15}N leads to fully functional proteins, we assumed a very similar protein composition for the control and the treated sample groups. Subsequently, the control group was used for peptide identification, generating a list of peptide sequences, their corresponding charge state and retention time as well as the accession number of the inferred protein [19]. This list together with the samples of the treatment served as the input for the program at hand (for a more detailed description see Document S1).

The current software does not require unlabeled data/samples in any way. We've employed an experimental design which includes control samples that are unlabeled in order to use these unlabeled samples for peptide/protein identification, which in turn serves as the input for the program. If previously identified peptide sequences, charge states and retention times are known, this unlabeled control is unnecessary. Any partially labeled ^{15}N LC/MS shotgun proteomics data could be evaluated with this software.

Method Outline

The following steps were used in our algorithm:

- Sort the input files in reverse chronological order
- Calculate the isotopic peaks (isotopic envelope) for a given peptide sequence and charge
- Pick peaks according to template
- Filter out co-eluted picked peaks
- Choose best scan within retention time-range
- Set new template from experimental data for the next file
- Filter noise at TP_0 (first Time Point)
- Calculate the RIA (Relative Isotope Abundance)
- Post processing filter
- Data export
- Compatibility

Sort Input Files in Reverse Chronological Order

Since partial metabolic labeling experiments consist of time course measurements, regardless of pulse-chase or other experimental designs, the chronological order of the measurements can be taken into consideration. In our approach we search the files in reverse chronological order (from the maximally labeled to the minimally/non-labeled). The number of peaks of the isotopic envelope increase with time, as more ^{15}N is incorporated, therefore decreasing when reversing the order. The first file (TP_{MAX}) is searched with a template of theoretically calculated m/z values for a given peptide, producing the picked peaks of the measured spectrum. The template for the next Time Point (TP) consists of only those peaks that could be picked for the previous TP. The extracted (experimental) spectrum serves as the template (replacing the number and position of the peaks in the template, but not their theoretically calculated value). Thus the m/z values do not change, but the number of values in the template changes dependent on how many of them were found in the previous TP. This leads to the next extracted spectrum which is again used to extract the spectrum of the next file (TP_{MAX}-2). This approach enables the algorithm to never pick more peaks than in the previous time point, which in turn reflects the biology of the underlying data.

Calculate the Isotopic Peaks (Isotopic Envelope) for a Given Peptide Sequence and Charge

The possible isotopic envelope and thus the peaks for the theoretical template are calculated as follows: For each peptide sequence, the sum of its individual C, H, O, N, S atoms is built and multiplied with the mass of its most abundant isotope. This produces the monoisotopic peak. All subsequent peaks are calculated by exchanging the mass of a ^{14}N by a ^{15}N atom. The largest isotopic peak is the ^{15}N monoisotopic peak. Thus, the template consists of as many peaks as there are nitrogen atoms plus one (n+1). Finally, the mass values are converted to m/z values by the addition of as many protons as charges, divided by the number of charges.

Pick Peaks According to Template

For any given peptide, the mzML file is searched within a user defined retention time window, allowing for common retention time deviations occurring in Liquid Chromatography (LC). Every Full Scan within this window is processed as follows:

- The most abundant m/z value is picked within a user-defined range (e.g. +/−10 ppm) of the monoisotopic peak.
- The algorithm only searches for subsequent peaks if the first peak (the monoisotopic peak) was found.
- All subsequent peaks are picked analogously (since the mass accuracy decreases with decreasing intensity, this value can also be adjusted separately by the user dependent on the given data).

Filter Out Co-eluted Picked Peaks

In order to remove overlapping peaks belonging to another peptide, the following filter was implemented. If the ratio of the current peak is 3 times higher to the preceding peak (empirically found value), the current peak is removed from the raw data. Subsequently, the appropriate peak is picked again. This routine of removal and re-picking is iterated either until no more peaks are removed from the raw data, or no more peaks remain to be picked from the raw data (see Figure 2.A and 2.B).

Furthermore, the application of the co-eluted picked peaks filter in conjunction with the penalty of the total score addresses the issue of complex overlapping envelopes.

Choose Best Scan within Retention Time-range

A single scan is used for peak picking, and not a scan as a result of averaging over multiple scans. The latter could potentially lead to an increase in noise and or elevate the complexity of the spectrum, since analytes eluting with similar retention times are prone to produce overlapping isotopic envelopes, especially for partial metabolic labeling data. In order to choose the best retention time (scan) from within the given retention time range and to evaluate the quality of the selected data points for a given peptide, a total score (TS) is calculated for each scan. The maximum score is selected and the corresponding data points saved

$$TS = I_{MIP0} - W_{ppm} + C_{TP-n} - P \qquad \text{(I)}$$

The total score is composed of the following components:

• $\mathbf{I_{MIP0}}$: logarithm to the base 10 of the intensity of the Monoisotopic Precursor (MIP0) in arbitrary units.

• $\mathbf{W_{ppm}}$**:** Weighted sum of ppm deviations of a given peptide spectrum.

$$\textbf{(II)}\ W_{ppm} = \sum_{i=1}^{n} \left(\frac{I_k}{\sum_{i=1}^{n} I_i} \right) * |\Delta ppm|, \text{ with}$$

- I_k: Intensity of a peak in the given peptide spectrum (in arbitrary units).
- $|\Delta ppm|$: The absolute value of the ppm deviation of the m/z value of compared to the theoretically calculated m/z value.
- $\sum_{i=1}^{n} I_i$: Sum of all peak intensities in the given peptide spectrum (in arbitrary units).

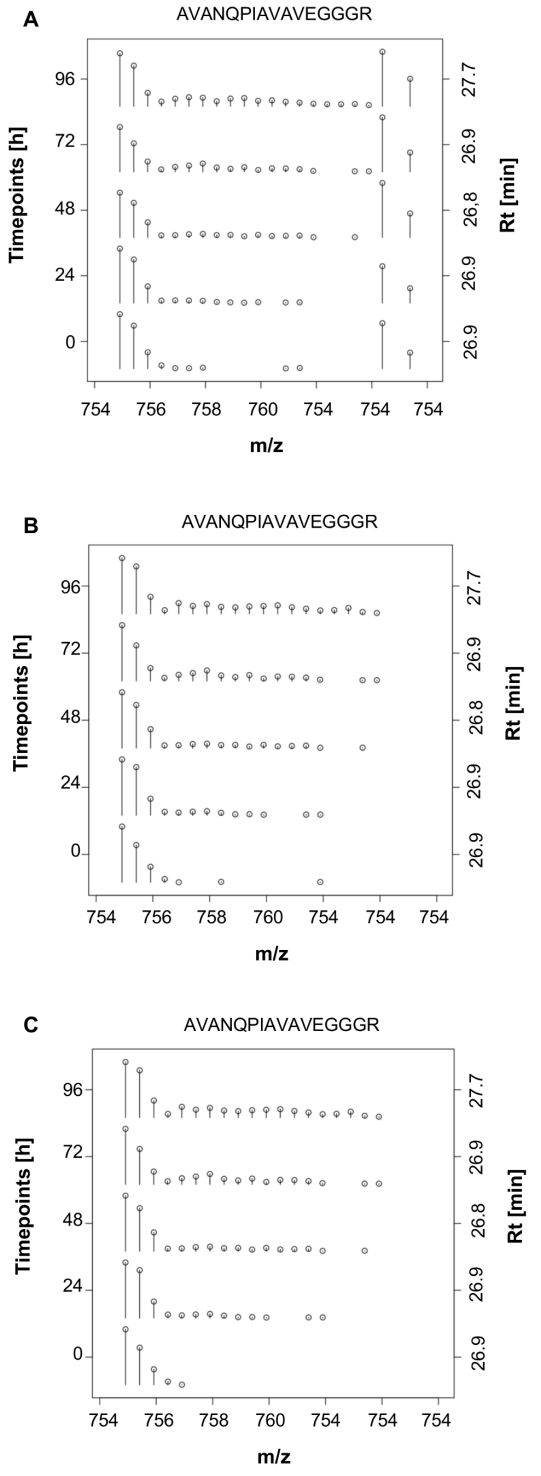


Figure 2. Picked peaks of the peptide sequence "AVANQPIAVAVEGGGR" at all Time Points (TP). The abscissa indicates the mass to charge ratio. Left ordinate indicates Time Points (corresponding to the user-given number in the "Experiment file"), right ordinate indicates the retention time (in minutes) of the scan used to pick the peaks. The individual spectra are normalized to the base peak of the given spectrum. A: Without the application of any filters. B: Filter out co-eluted picked peaks. C: Filter out co-eluted picked peaks and Filter noise at TP_0.
doi:10.1371/journal.pone.0094692.g002

• **Coverage: (III)** $C_{TP-n} = \frac{N_{TP-n}}{N_{TP}}$

– C_{TP-n}: Coverage of a given Time Point (TP) with n from Zero to the maximum number of Time Points, in reverse chronological order.
– N_{TP-n}: Number of picked peaks of the experimental spectrum of the previous Time Point, respectively number of theoretically calculated peaks of a given peptide sequence for last Time Point (TP_{MAX}).
– N_{TP}: Number of picked peaks of the experimental spectrum of the given Time Point.

• **Penalty: (IV)**

$$\left\{ \begin{array}{c} 3, \left(I_{[m/z]_s} > 0\right) \wedge \left(I_{[m/z]_s} < 2 * I_{[m/z]_k}\right) \wedge \left(I_{[m/z]_s} > 0.5 * I_{[m/z]_k}\right) \\ 0, \left(I_{[m/z]_s} = 0\right) \end{array} \right\}$$

with: $[m/z]_s = m/z_k - (m/z_{k+1} - m/z_k)$.

If the first peak of the experimental spectrum is an isotopic peak of another peptide a penalty is applied. First, the mass difference $(m/z_{k+1} - m/z_k)$of the first to the second peak is calculated. This value is deducted from the first peak. Within a range of +/− 10 ppm of m/z the peak with the highest intensity is selected. If the intensity of this peak is higher than half and less than twice of the intensity of $\mathbf{I}_{MIP0}$, then a penalty of 3 is applied.

Set New Template from Experimental Data for the Next File

After the processing of the initial file (TP_{MAX}), the template changed from all theoretically calculated peaks to only those peaks that could be picked for the previous TP. If not a single peak could be picked in the "previous round", all theoretically calculated values remain as the template.

Filter Noise at TP_0 (First Time Point)

The spectrum of the minimally labeled measurement is used to determine which peaks represent the ^{14}N peptide (natural abundance of N). In order to remove data points that are rather considered noise than low abundant peaks, all data points following a missing peak are removed from the spectrum (from low to high m/z-values, see Figures 2.B and 2.C).

Calculate the RIA (Relative Isotope Abundance)

The Relative Isotope Abundance (RIA) is defined as the ratio of the ^{15}N to all isotopic peaks [1,20]. Since no ^{15}N incorporation has taken place at the very first measurement, all peaks present in TP_0 are considered part of the ^{14}N peptide species. For each individual experimental spectrum the RIA is calculated as follows:

$$RIA = \frac{A_{15}}{A_{14} + A_{15}} \quad (V)$$

with

– A_{14}: Sum of intensities of all ^{14}N peaks (natural abundance).
– A_{15}: Sum of intensities of all ^{15}N peaks (isotopically labeled).

In order to differentiate the natural abundance from the enriched part of an overlapping isotopic peak (see Figure 1 red and green species overlapping at e.g. the 5^{th} isotopic peak), the relative intensity values at TP_0 are taken into account when calculating

A_{14} and A_{15} for all other TPs. (for more details please see Document S1).

Post-processing Filter

Due to the incorporation of ^{15}N, novel synthesis of a given protein will produce an increase of the A_{15} term when measuring its proteolytic peptides. We assume that the RIA for a given protein will stay constant or increase with time due to the following reasons. A fictitious protein without novel synthesis, but with degradation, would produce a constant A_{15} and a decreasing A_{14} term (see formula (V)), and thus a constant numerator and a decreasing denominator with time, leading to an increase of the RIA over time. A fictitious protein with novel synthesis, but without degradation, would produce an increasing A_{15} and a constant A_{14} term, also leading to an increase of the RIA over time. Furthermore, a protein with novel synthesis and degradation will always produce an increasing RIA over time. Therefore, a post-processing filter was devised, removing all peptides whose RIA decreased over time (see Figure 3.A and 3.B as well as Figure 4). If data for one Time Point of a peptide is missing, the RIA for that Time Point is not calculated. Subsequently, this peptide will not pass the post processing filter even if all other Time Points produced a positive result.

This very stringent filter reduces the data to the most stable signals that can be traced throughout the entire data set (see Figure 4.D). Please see Document S1 for configuration options (Post-processing configuration options).

Data Export

In order to save the output of the data analysis, the extracted spectra and useful additional information is saved to tab delimited txt files for easy import into Excel (see Document S1). Additionally, two-dimensional plots of peptide spectra for all the Time Points can be plotted as pdf files (see example Figure 2.C). The RIA for each protein can be plotted as well (see example Figure 3.A and 3.B without the regression line).

Compatibility

Furthermore, the presented program runs on all commonly used operating systems (Windows, OSX, and Linux), is independent of the tissue being analyzed, and is not restricted to any specific type mass spectrometric data.

Results and Discussion

Protein turnover experiments are most often performed using cell cultures of human or plant cell lines. The uniformity of the given cell type and the possibility to quickly exchange the growth medium enable full incorporation of heavy labels within hours or at maximum a couple of days [1,3,8,13]. The experimental design of the present study is inherently different, due to the fact that entire plants were grown in pots to their fully functioning potential, closely resembling the phenotype of the species in the wilderness of nature. This results in dramatically reduced measurable turnover rates due to the following reasons: The exchange of the light by the heavy amino acid pool cannot be performed by simple plating (as in cell cultures), but by supplying an inorganic nitrogen source that has to be taken up by the roots and incorporated into amino acids and subsequently into proteins, in contrast to SILAC experiments [21] where fully labeled amino acids are provided in excess, and plant cell cultures, where a labeled nitrogen source replaces the unlabeled form immediately. The degradation of existing light or marginally labeled proteins feeds the light amino acid pool, thereby counteracting the relative increase of the heavy amino acids. In order to ensure full labeling, plants were grown with ^{15}N medium for over 12 weeks [22]. Therefore, the RIA values of the data set utilized within this study are generally low, but much closer to in situ-growth conditions. After 5 days of labeling, the mean of all RIAs is still below 50% (data not shown). The higher the intensity of the signal, the higher the mass accuracy and vice versa. Manual inspection of the extracted spectra and comparison with the raw data showed that highly abundant peptides lead to fewer missing peaks as well as to congruence of the resulting RIAs, while low abundant peptides showed higher variability, since true positive peaks might not fall within the calculated mass range, but random noise could. The three biological replicates of the test data set showed that over 800 peptides of the 1419 identified peptides passed all three previously described filters, with a variability that can be expected of independent biological replicates. Naturally, the quality of the extracted spectra and thus the output strongly depends on the quality of the input data. Measuring the LC/MS data with a high mass resolution is beneficial, since overlapping peaks are more likely to be resolved and thus enable the algorithm to pick the proper peaks. Instability or poor ESI-spray quality can lead to missing or noisy spectra and reduce mass accuracy. Peptides with missing spectra at any given Time Point will eventually fail to pass the filters. The major steps of the algorithm will be discussed as follows.

Performance of the Applied Filter

The effect of removing co-eluted picked peaks filter as described in Methods becomes apparent when comparing Figure 2.A to 2.B, as the two peaks with the highest m/z-values were excluded from the spectrum. The effect of filter out noise at the first Time Point (TP_0) is visible when inspecting the extracted spectrum of Figure 2.C compared to 2.A or 2.B, as all peaks following an empty position (missing peak) are removed from the spectrum. The ameliorated peak picking of spectral envelopes of peptides, due to the incorporation of the latter two filters, not only affects the extracted spectra, but also the resulting RIA of the associated proteins. The post-processing filter, described in the Methods part, removes the peptide sequence "NAVFGDSSALAPGGVR" (hollow circle as symbol) (see Figure 3.A and B), due to the lack of an increasing RIA over time. Linear regression of mean RIA values per Time Point, yielded an increase in the regression coefficient from 0.978 (Figure 3.A) to 0.997 (Figure 3.B). Only the application of the co-eluting picked peaks filter affects the total score (lowers the coverage term) and thus potentially alters which scan is chosen for spectral extraction.

The variability of the calculated RIAs for a protein decreases when applying the previously described filters. The overall effects of the various filters are illustrated in Figure 4.A to D. For each protein, all associated peptide RIAs were averaged for each Time Point and a linear regression calculated. The density distribution of regression coefficients (R^2) of all 422 proteins with and without the application of the previously described filters are shown in Figure 4. The fraction of high R^2 values increases with the application of the filters. Since the post-processing filter removes peptides, all subsequently removed protein R^2 values were set to Zero (see Figure 4.D). The fraction of proteins with a regression coefficient between 0.95 and 1.0 starts at 59%, without the application of any filters, increases to 64%, with the application of the co-eluted picked peaks filter, increases further to 66% with the additional application of the filter noise at TP_0 filter, and finally reaches 89% with the additional application of the post-processing filter (due to the removal of values). The increase in precision of the RIA values after application of the filters is corroborated by the change in the regression coefficients.

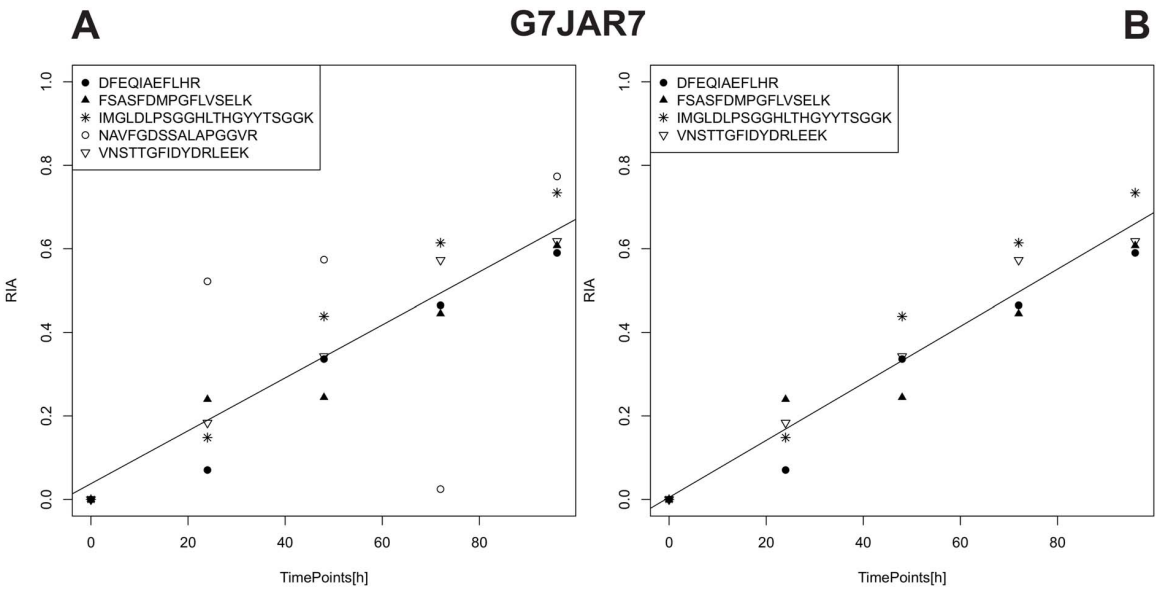


Figure 3. Relative Isotope Abundance (RIA) plots. The abscissa represents the Time Points (provided by the user in the Experiment File) and the ordinate the RIA ratio at the given Time Point. The titles of the plots indicate the Accession Number for the given data. The legend shows all peptide sequences that could be attributed to the given protein. A: illustrates the RIA plot for G7JAR7 without the application of any filters. B: RIA plot for G7JAR7 with the application of post-processing filter.
doi:10.1371/journal.pone.0094692.g003

Each histogram shown in Figure 5 displays the frequency as a function of the coverage of peptides at a given Time Point. Generally, when comparing the histograms in a reverse chronological order (from maximum time point (TP_{MAX}) to minimum time point (TP_{MIN}), 96 h to 0 h) and thus from the maximally labeled to the minimally labeled state, a trend from a negative skew to a positive skew with intermediate stages can be observed (see Figure 5. A to E). The distribution at TP_{MAX} clearly shows a high coverage for most of the peptides and as the coverage decreases so does the number of peptides. This reflects the underlying biology of the experimental setup showing the partial labeling state of the proteins and thus of their proteolytic peptides. Due to the varying turnover rates, the coverage cannot be constant for all 1419 cases (peptides) at any given time (it should however be constant for all peptides associated with a protein at any given time). As described in the Methods (see "Set new template from experimental data for the next file"), the algorithm was trained to produce a decreasing coverage over time. Figure 5 illustrates the results of the implementation of this desired functionality.

Performance of the developed strategy is demonstrated by the automated protein turnover calculations of 1419 peptides from five Time Points (n = 3, three biological replicates).

Many studies express protein turnover as % turnover per hour (log ratio of heavy to light per hour for SILAC experiments show linear correlation). We are dealing with an entire organism, not a specific cell type, thus we would not expect the synthesis and degradation rates to be constant over time, but rather showing distinct biologically relevant and interesting dynamic kinetics.

Biological Applicability

The amount of information that can be generated with the presented automated method, is very high, specifically due to the coupling of partial metabolic labeling with high throughput shotgun proteomics, in contrast to the excision of proteins from gel spots [22,23]. Within the given dataset, the Glycine-rich RNA binding protein (Uniprot accession number: G7JG67) showed a high turnover rate (RIA) in all biological replicates (0.716 mean +/−0.01 standard deviation of 3 biological replicates at TP_{MAX}). The protein plays a functional role in processing, transport, localization, translation and stability of mRNAs and the high turnover rates are in accordance to previous plant protein turnover measurements [23,24]. In contrast, a low protein turnover rate (RIA) was observed for the Harpin binding protein containing a conserved fibrillin domain (Uniprot accession number: G7I4U4) (0.398 mean +/−0.021 standard deviation of 3 biological replicates at TP_{MAX}). Plant fibrillins expression increases during acclimation to various biotic and abiotic stresses (reviewed by [25]). The observed low RIA after five days of ^{15}N metabolic labeling is in line with the assumption of low stresses during the experimental period.

Comparison to Other Approaches

The presented algorithm is based on data analysis in reverse chronological order (a unique and novel feature), and doesn't subsequently fit data to theoretical relative isotope abundances, but uses experimentally derived intensity values for subsequent RIA calculations. Assuming that the monoisotopic precursor is still present, the spectral envelope of the maximally labeled Time Point will always have the maximum number and intensity of isotopic peaks, leading to the best signal to noise ratio. An interdependency of Time Points is established that reduces picking of noise, since the peak picking of every Time Point depends on the previous one.

The presented algorithm is trained to pick the best possible scan within the user-given retention time range, enabling large retention time deviations that can occur in high throughput studies. Switching (renewing) liquid chromatography columns (sometimes done between batches of samples), often leads to retention time shifts. A major strength of our approach is that it can cope very well with these shifts. The user has to simply set a higher retention time range in the "experiment-file", which will

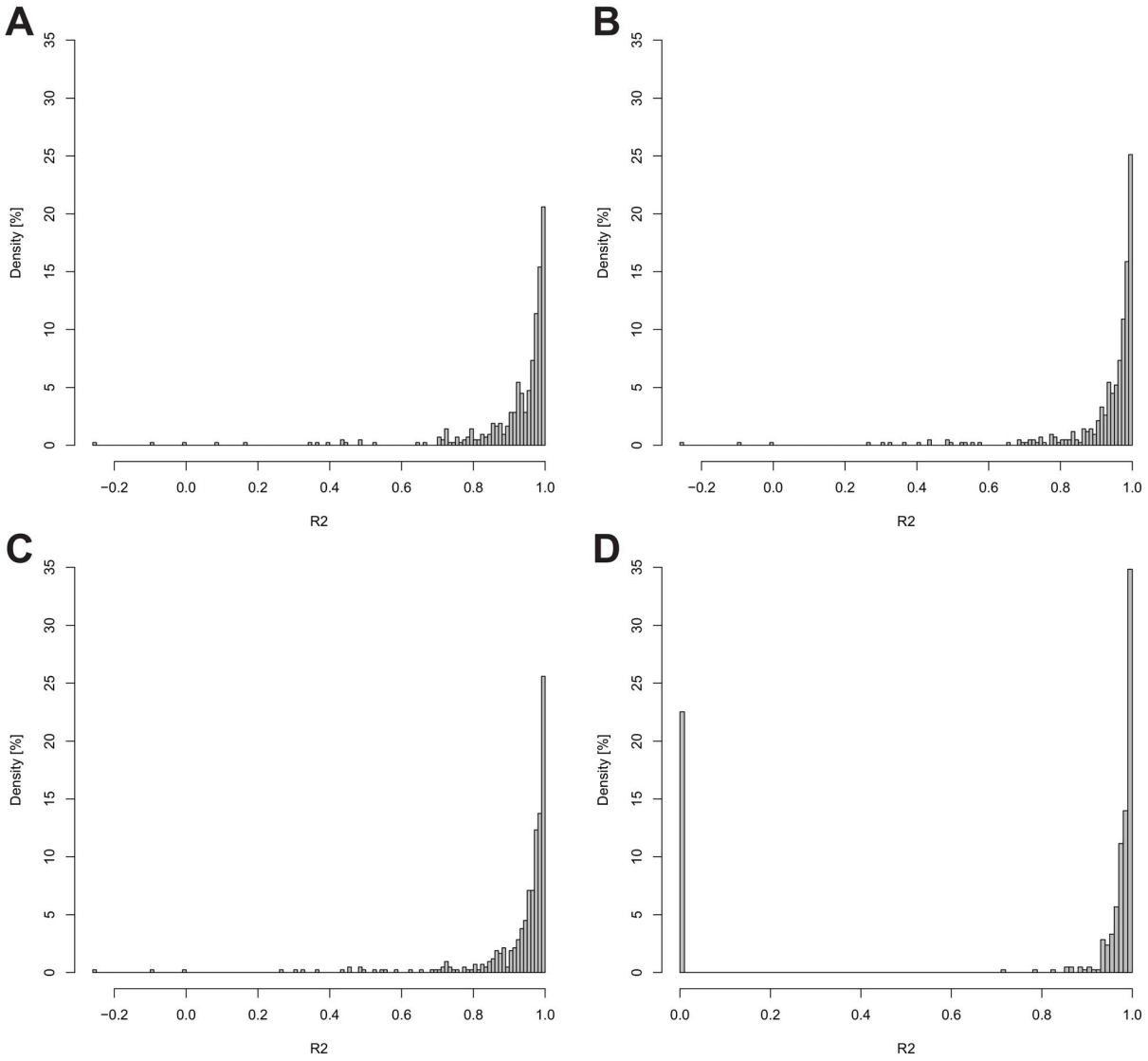


Figure 4. Histograms of the regression coefficient versus the density of proteins. Histograms of the regression coefficient versus the density of proteins, comparing no filter (A), co-eluted picked peaks filter (B), co-eluted picked peaks and filter noise at TP_0 (C), and all filters combined (D). (A-D) with 422 cases each. For each Time Point, all peptide RIA values (associated with an Accession Number) were averaged. Subsequently the linear regression was calculated, and thereof Histograms produced. (D) includes (the 94 of the 422) proteins that were removed by the post-processing filter.
doi:10.1371/journal.pone.0094692.g004

lead to a prolonged runtime (more data has to be processed). The algorithm will still pick the correct scan, making high-throughput studies feasible. In contrast, "TurnStile" [9] averages over a user-defined retention time range, thus potentially averaging over isobaric or isomeric peaks or even noise not belonging to the target. In order to circumvent this behavior, the user would need to set a very narrow retention time range and potentially adapt this setting for each individual file, leading to an enormous work-load contradicting the computational automation of the workflow and impeding high-throughput data analysis (see Figure S1).

Comparing the calculated RIA values, the protein with the Uniprot accession number G7IF28, with 9 associated peptides, displays a low protein turnover when applying our approach (RIA values ranging from 0.0 to 0.35, and linear regression coefficient (R^2) of mean RIA values is 0.998, see Figure S2.A and S2.B). In contrast, the output generated with "TurnStile" displayed a spread of data, the resulting RIA values reach from 0.23 to 0.99, encompassing a large part of the range of possible values (with $R^2 = -0.2213$). The protein with the accession number G7JG67, with 5 associated peptides, displays a high protein turnover when applying our approach, with a linear regression coefficient of $R^2 = 0.992$ (RIA values ranging from 0 to 0.76). Except for the last two Time Points the RIA values of the peptides at a given Time Point derived from "TurnStile" analysis are neither similar nor do they indicate a trend towards an increase in RIA over time (with $R^2 = -0.568$, and the spread of the data reaches from 0.09 to 0.85 for the RIA) (see Figure S2.C and S2.D). For further comparison see Document S1.

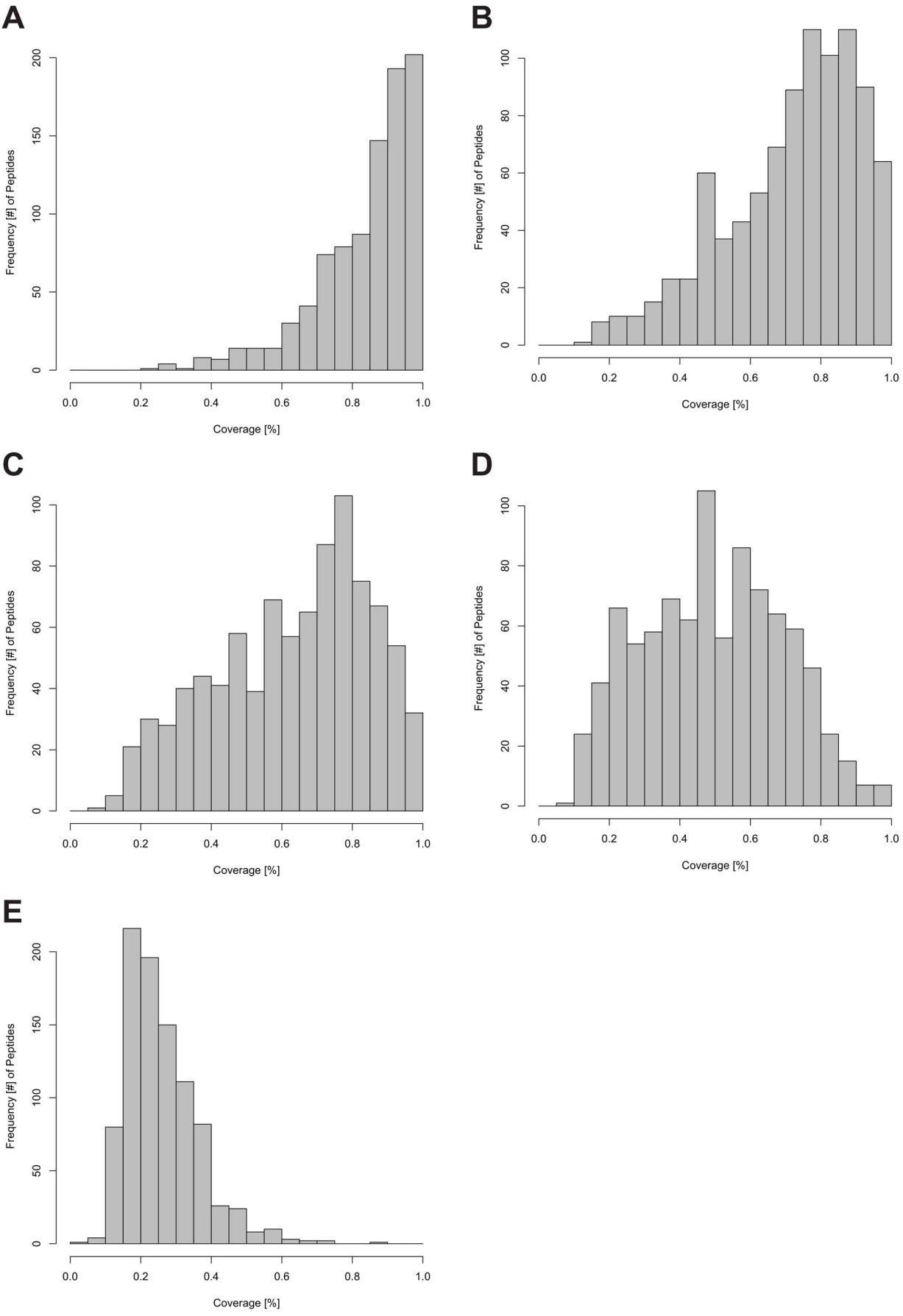

A
Frequency [#] of Peptides
0
50
100
150
200
0.0
0.2
0.4
0.6
0.8
1.0
Coverage [%]
B
Frequency [#] of Peptides
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0
Coverage [%]
C
Frequency [#] of Peptides
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0
Coverage [%]
D
Frequency [#] of Peptides
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0
Coverage [%]
E
Frequency [#] of Peptides
0
50
100
150
200
0.0
0.2
0.4
0.6
0.8
1.0
Coverage [%]

Figure 5. Histogram plots of the coverage versus frequency of peptides. The abscissa represents the coverage in percent. The coverage is calculated analogous to formula (III) the sole difference being the constant denominator N_{MAX}. Frequency of peptides indicates the number of peptide sequences with a given coverage (number of cases 1419). All five Time Points (from A to E, of one biological replicate) from 96 h (A) to 0 h (E), in 24 h intervals of ^{15}N incorporation are shown.
doi:10.1371/journal.pone.0094692.g005

Computation Rate

The computation rate depends on the amount of LC/MS data, the number of identified peptides, and the user-defined retention time range. The runtime increases linearly as a function of the retention time range and/or the number of identified peptides. E.g. given five mzML-files with about 5 GB of data, 1419 identified peptides, and 2 min retention time range, the runtime was about 40 min. Principally, many files can be processed with the given program (it was tested with about 60 GB of data). One strength of the algorithm is to pick the proper scan despite isobaric peptides in the chromatographic domain. Therefore, using a high retention time range is recommended despite the extended runtime.

Outlook

- Implementation of a Graphical User Interface (GUI).
- Post-Translational Modification (PTM) support.
- Differential data analysis of treatment groups with repect to biological and technical replicates.

Supporting Information

Figure S1 TurnStile output strongly depends on Rt range. The abscissa represents the Time Points and the ordinate the RIA ratio at the given Time Point. The titles of the plots indicate the Accession Number for the given data as well as the retention time window used for data analysis. From Ai to Aii to Aiii (note: the legend for these sub-plots shown at the bottom) and from Bi to Bii to Biii (note: the legend for these sub-plots shown at the bottom) the retention time window decreases from 10 min to 90 s to individually adapted values for every peptide for every file (in the range of 15 to 45 seconds).
(TIF)

Figure S2 Qualitative comparison of TurnStile vs. Protover. The abscissa represents the Time Points and the ordinate the RIA ratio at the given Time Point. The titles of the plots indicate the Accession Number for the given data. The legends show all peptide sequences that could be attributed to the given protein. A and B show the RIA plots for G7IF28 (note: the legend for both sub-plots only shown in the right sub-plot). C and D show the RIA for G7JG67 (note: the legend for both sub-plots only shown in the right sub-plot). The data illustrated in A and D were processed using TurnStile with a 90 s retention time window (the recommended setting). B and D were processed using Protover with a 10 min retention time window (+/−5 min) (the recommended setting).
(TIF)

Document S1 Supplements. Detailed description of experimental procedures and methods, data analysis and comparison with other algorithm. Calculation of relative RIA, discussion of averaging scans as well as comparison of Retention Time settings.
(DOC)

Acknowledgments

The authors would like to give special thanks to Stefan Karner and to Thomas Naegele for valuable discussions.

Author Contributions

Performed the experiments: MAC. Contributed reagents/materials/analysis tools: WW. Wrote the paper: DL VE. Contributed to the experiments: CS. Conceived the experimental part: SW. Supervised MS data analyses: SW. Developed the algorithm: DL. Programmed the software: DL. Contributed to the LC/MS method and measurement: DL. Conceived the algorithm and supervised the development: VE. Responsible for project coordination: VE.

References

1. Pratt JM, Petty J, Riba-Garcia I, Robertson DHL, Gaskell SJ, et al. (2002) Dynamics of Protein Turnover, a Missing Dimension in Proteomics. Mol Cell Proteomics 1: 579–591. doi:10.1074/mcp.M200046-MCP200.
2. Hoehenwarter W, Wienkoop S (2010) Spectral counting robust on high mass accuracy mass spectrometers. Rapid Commun Mass Spectrom: 3609–3614. doi:10.1002/rcm.
3. Li L, Nelson CJ, Solheim C, Whelan J, Millar AH (2012) Determining Degradation and Synthe-sis Rates of Arabidopsis Proteins Using the Kinetics of Progressive 15N Labeling of Two-dimensional Gel-separated Protein Spots. Mol Cell Proteomics 11: M111.010025. doi:10.1074/mcp.M111.010025.
4. Belle A, Tanay A, Bitincka L, Shamir R, Shea EKO, et al. (2006) Quantification of protein half-lives in the budding yeast proteome. Proc Natl Acad Sci U S A 103: 13004–13009. doi:10.1073/pnas.0605420103.
5. Yen H-CS, Xu Q, Chou DM, Zhao Z, Elledge SJ (2008) Global protein stability profiling in mammalian cells. Science 322: 918–923. doi:10.1126/science.1160489.
6. Selbach M, Schwanhäusser B, Thierfelder N, Fang Z, Khanin R, et al. (2008) Widespread changes in protein synthesis induced by microRNAs. Nature 455: 58–63. doi:10.1038/nature07228.
7. Maier T, Schmidt A, Güell M, Kühner S, Gavin A-C, et al. (2011) Quantification of mRNA and protein and integration with protein turnover in a bacterium. Mol Syst Biol 7: 511. doi:10.1038/msb.2011.38.
8. Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, et al. (2013) Corrigendum: Global quantification of mammalian gene expression control. Nature 495: 126–127. doi:10.1038/nature11848.
9. Martin SF, Munagapati VS, Salvo-Chirnside E, Kerr LE, Le Bihan T (2012) Proteome turnover in the green alga Ostreococcus tauri by time course 15N metabolic labeling mass spectrometry. J Proteome Res 11: 476–486. doi:10.1021/pr2009302.
10. Price JC, Guan S, Burlingame A, Prusiner SB, Ghaemmaghami S (2010) Analysis of proteome dynamics in the mouse brain. PNAS 2010. doi:10.1073/pnas.1006551107/−/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1006551107.
11. Zhang Y, Reckow S, Webhofer C, Boehme M, Gormanns P, et al. (2011) Proteome Scale Turno-ver Analysis in Live Animals Using Stable Isotope Metabolic Labeling. Anal Chem: 1665–1672. doi:10.1021/ac102755n.
12. Toyama BH, Savas JN, Park SK, Harris MS, Ingolia NT, et al. (2013) Identification of long-lived proteins reveals exceptional stability of essential cellular structures. Cell 154: 971–982. doi:10.1016/j.cell.2013.07.037.
13. Schwanhäusser B, Busse D, Li N (2011) Global quantification of mammalian gene expression control. Nature: 337–342. doi:10.1038/nature10098.
14. Cox J, Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 26: 1367–1372. doi:10.1038/nbt.1511.
15. Savas JN, Toyama BH, Xu T, Yates JR, Hetzer MW (2012) Extremely long-lived nuclear pore proteins in the rat brain. Science 335: 942. doi:10.1126/science.1217421.
16. Toyama BH, Savas JN, Park SK, Harris MS, Ingolia NT, et al. (2013) Identification of long-lived proteins reveals exceptional stability of essential cellular structures. Cell 154: 971–982. doi:10.1016/j.cell.2013.07.037.
17. Park SK, Venable JD, Xu T, Iii JRY (2008) A quantitative analysis software tool for mass spec-trometry–based proteomics. 5. doi:10.1038/NMETH.1195.
18. Zhang Y, Reckow S, Webhofer C, Boehme M, Gormanns P, et al. (2011) Proteome Scale Turno-ver Analysis in Live Animals Using Stable Isotope Metabolic Labeling. Anal Chem: 1665–1672.
19. Castillejo MA, Staudinger C, Egelhofer V, Wienkoop S (2014) Medicago truncatula proteomics for systems biology: novel rapid shotgun LC-MS

approach for relative quantification based on Full-Scan Selective Peptide Extraction (Selpex). Plant Proteomics Methods Mol Biol 1072: 303–313.
20. Gustavsson N, Greber B, Kreitler T, Himmelbauer H, Lehrach H, et al. (2005) A proteomic method for the analysis of changes in protein concentrations in response to systemic perturba-tions using metabolic incorporation of stable isotopes and mass spectrometry. Proteomics 5: 3563–3570. doi:10.1002/pmic.200401193.
21. Ong S, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, et al. (2002) Stable Isotope Label-ing by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics *: 376–386. doi:10.1074/mcp.M200025-MCP200.
22. Hebeler R, Oeljeklaus S, Reidegeld KA, Eisenacher M, Stephan C, et al. (2008) Study of early leaf senescence in Arabidopsis thaliana by quantitative proteomics using reciprocal 14N/15N la-beling and difference gel electrophoresis. Mol Cell Proteomics 7: 108–120. doi:10.1074/mcp.M700340-MCP200.
23. Li L, Nelson CJ, Solheim C, Whelan J, Millar AH (2012) Determining degradation and synthesis rates of arabidopsis proteins using the kinetics of progressive 15N labeling of two-dimensional gel-separated protein spots. Mol Cell Proteomics 11: M111.010025. doi:10.1074/mcp.M111.010025.
24. Singh U, Deb D, Singh A, Grover A (2011) Glycine-rich RNA binding protein of Oryza sativa inhibits growth of M15 E. coli cells. BMC Res Notes 4: 18. doi:10.1186/1756–0500–4-18.
25. Singh DK, McNellis TW (2011) Fibrillin protein function: the tip of the iceberg? Trends Plant Sci 16: 432–441. doi:10.1016/j.tplants.2011.03.014.

Supplementary Material

Plant Growth

The seeds of barrel medic (*M. truncatula* A17 cv. Jemalong) were surface sterilized and sown in pots containing a mixture of perlite: vermiculite 2:5 (v:v). Plants were grown under controlled conditions in a growth chamber (14h day and 10h night; 300 µmol m^{-2} s^{-1} photosynthetic photon flux density; 22°C day and 16°C night temperatures; 50–60% relative humidity). During the first week of growth, plants were watered with nutrient solution (Evans, 1981) containing 0.5mM ammonium nitrate. The following 6 weeks a nutrient solution with ammonium nitrate concentration of 2.5 mM was used for watering in order to enhance biomass accumulation and to keep plant growth performance identical during the initial developmental stage.

^{15}N labeling experiment

Seven week old plants were randomly separated into two subsets: 1) control plants were further watered with ^{14}N- ammonium nitrate fertilizer (2.5 mM), while 2) another set of plants was transferred to ^{15}N- labeled ammonium nitrate (^{15}N nitrate, 98% ^{15}N; Sigma-Aldrich) containing growth medium. Plants were washed two times with water before medium application. Growth medium was supplied daily to pot capacity for 5 days. *M. truncatula* shoots were collected each day from the first day of ^{15}N application, frozen in liquid nitrogen and stored to -80ºC until further processing.

Protein Extraction

Three biological replicates were used for protein extraction. Two hundred mg of liquid nitrogen frozen material (fresh weight) were homogenized in 1 ml of urea buffer containing 50mM HEPES, pH 7.8 and 8 M Urea using a glass homogenizer. After centrifugation (10000g, 10min, 4°C), the urea-soluble proteins in the supernatant were precipitated overnight in five volumes of -20°C cold acetone containing 0.5% β-mercaptoethanol. The precipitate was pelleted at 4000g, 4°C for 15 min. The resulting pellet was washed with -20°C cold methanol containing 0.5% β-mercaptoethanol and again centrifuged (4000g, 4°C, 10 min). Air-dried protein pellets were dissolved in 800 µl of urea buffer (described above). Protein concentration was determined by Bradford assay using BSA as a standard.

Protein Digestion

One hundred µg of protein was initially digested using 0.1 µg (1:1000, w/w) of endoproteinase LysC (Roche, Mannheim, Germany) during 5 h at 30°C. For the second digestion step, samples were diluted with trypsin buffer (10% Acetonitrile, 50mM Ammonium bicarbonate, 2 mM $CaCl_2$) to a final concentration of 2 M Urea, and incubated overnight at 37°C with Poroszyme immobilized trypsin beads (3.3:100, v/w; Applied Biosystems, Darmstadt, Germany). The digest was desalted with C18-SPEC96-well plates (Varian, Darmstadt, Germany) according to the manufacturer's instructions. The eluted peptides were vacuum-dried.

nanoESI LC-MSMS

Peptide digests (2.5 µg each) were randomly applied to a RP column (Supelco Ascentis® Express Peptide ES-C18, 150x0.1mm) separated during a 90 min gradient ranging from 98% solvent A (0.1% FA in water, 2% ACN) to 80% solvent B (90% acetonitrile, 0.1% FA in water). For each treatment tree biological and two technical replicates were randomly analyzed to discriminate technical from biological variation. MS analyses were performed on a LTQ-Orbitrap XL (Thermo Fisher Scientific, Bremen, Germany), applying a top seven method. A full-scan range from 350 to 1,600 m/z was used. The resolution was set to 60,000. Dynamic exclusion settings were as described in [1]. Briefly, repeat count was set to one, repeat duration 20 s, exclusion list size 500, exclusion duration 60 s and exclusion mass width 10 ppm. Charge state screening was enabled with rejection of unassigned and 1+ charge states. Minimum signal threshold counts were set to 50,000.

Peptide identification and generation of input list

The control (^{14}N) as well as the treated (^{15}N) samples resulted in a total of 15 files each (three biological replicates x five time points). Only the raw data files from non-labeled (^{14}N) samples were processed and quantified using MaxQuant (v 1.4.0.3) with a combined protein fasta database (see 5.1, for MaxQuant details see 5.2).

Medicago truncatula composite fasta generation

A composite protein-fasta-file was created by fusing the following three databases:

1.) Uniprot UniRef100 Medicago, origin: www.uniprot.org, Uniprot advanced-search Medicago truncatula (3880), UniRef100. The search was performed on May 7th 2013. 54246 entries.

2.) IMGA, origin: http://medicago.org/genome/IMGAG/. 64123 entries.

3.) Database of Plant Ubiquitin Proteoasome System, origin: http://bioinformatics.cau.edu.cn/plantsUPS/. 1010 entries

4.) DCFI/MTGI/TC, origin: http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=medicago. Originally 412908 entries reduced to 59598 entries (after picking the longest continuous amino acid sequence).

The six-frame-translation was performed with an in-house tool. (The six-frame-translated longest-open-reading-frame protein-fasta-file contains six entries per accession number.) A new protein-fasta-file was created by picking only the longest continuous amino acid sequence per accession number.

The three fasta files described above were combined, producing a new fasta containing 131338 entries, which will be referred to as MT-fasta henceforth. Protein sequences 100% identical in sequence and in length were combined by subsequently adding one header after the other, separating them by the following characters " __***__ " (no matter if the redundancies originated from one or multiple fasta-files). All other entries were simply added to the new file. The first accession number of the header was repeatedly written to the very beginning of the header line, separated by a " | " (pipe character), in order to consistently view and parse the accession numbers.

Peptide identification settings using MaxQuant

Trypsin (trypsin/P) was selected as the digestion enzyme, a maximum of two missed cleavages and no static or dynamic modifications were selected. Default settings were used except for the following modifications: “Global Parameters / Identification / Min. Peptides” was set to 2 and the “Parameters / Match between runs” box was selected. Within the “Group-specific parameters” the “Multiplicity” of 1 was selected, no variable modifications were selected, and for “Label-free quantification” LFQ was selected and “Fast LFQ” was deselected. Within the “Global parameters” no fixed modifications were selected. The default decoy database settings containing reversed sequences were used to estimate the false discovery rate (FDR) and a PSM as well as protein identification cutoff of FDR ≤ 1%.

Generation of input list

We used a similar strategy to the Selective Peptide Extraction (SelPEx) [2] that allows for the targeted quantification of ^{15}N-labeled peptides. The input list for *Medicago truncatula*

shoots was generated from the MaxQuant data matrix derived from the ^{14}N control samples (the "evidence" file). This list was filtered for target peptides according to the following criteria:

a) All peptide sequences were assigned a "+" in the "Reverse" or "Contaminant" column were disregarded.

b) Only peptide sequences with an "Intensity" value, a "Score" and a "PEP" score were retained.

This resulted in three separate lists of peptide identifications (A: 1055, B: 1118, and C: 916) which were combined and the duplicates removed. Finally, this resulted in a list of 1419 peptides. The peptide sequence, the dominant charge state (mean value of charge states rounded to the closest integer), as well as the recalibrated retention time (calculated by MaxQuant) were used to build the input list.

Generation of mzML files

The proprietary "raw" file format (from Thermo Scientific™) is only legible under windows operating systems with the use of proprietary software. Thus the cross platform compatible, open standard mzML was used. All "raw" files were converted to "mzML" files using "msconvert" [3].

Generation of LC/MS test data

To enable a rapid download and quick execution of the test data, the raw files were converted (analogous to 5.4 but) with the following restrictions. Limited to only MS1 scans, mz-range restricted to 402 to 866 mz, and retention time-range restricted from 28.85 to 32.3 min.

Calculation of A_{14} and A_{15}

Assuming that no enrichment of ^{15}N has occurred at TP_0 (first Time Point), the ratio of the monoisotopic peak to the intensity values of all other isotopic peaks of the spectral envelope of a peptide, reflect the experimentally derived intensity distribution of the natural isotopic composition. In order to differentiate the natural abundance from the enriched part (light and heavy) of an overlapping isotopic peak (see Figure 1 in Manuscript red and green species overlapping at e.g. the 5^{th} isotopic peak), the relative intensity values at TP_0 are taken into account when calculating A_{14} and A_{15}. At TP_0 the sum of all intensity values

produces A_{14}, and A_{15} is set to Zero. For all other TPs, all intensity values $I_{[m/z]_{iTPn}}$ of the isotopic peaks *i* occurring at TP_0 are split and one part assigned to A_{14} and the rest to A_{15}, depending on the intensity ratio $\frac{I_{[m/z]_{iTPn}}}{I_{[m/z]_{mTPn}}}$ of the analogue isotopic peak *i* at TP0 to the monoisotope *m*. The calculation is as follows:

$$A_{14} = \sum_{i=1}^{n_{TP0}} \frac{I_{[m/z]_{iTPn}} * I_{[m/z]_{iTP0}}}{I_{[m/z]_{mTP0}}},$$

$$A_{15} = \sum_{i=1}^{n_{TP0}} \left(I_{[m/z]_{iTPn}} - \frac{I_{[m/z]_{mTPn}} * I_{[m/z]_{iTP0}}}{I_{[m/z]_{mTP0}}} \right)$$

given that for each peak: $\frac{I_{[m/z]_{iTPn}} * I_{[m/z]_{iTP0}}}{I_{[m/z]_{mTP0}}} \quad I_{[m/z]_{iTPn}}$

the fractional part assigned to A_{14} is not greater than the intensity of the peak itself. Otherwise, A_{14} is simply the sum of all intensity values and no corresponding intensity is assigned to A_{15}.

$$A_{14} = \sum_{i=1}^{n_{TP0}} I_{[m/z]_{iTPn}}$$

$$A_{15} = 0$$

For all TPs, all intensity values $I_{[m/z]_{iTPn}}$ of isotopic peaks NOT occurring at TP_0 are assigned to A_{15}.

$$A_{15} = \sum_{i=nTP0+1}^{n_{TP0}} I_{[m/z]_{iTPi}}$$

Post processing configuration options

The possibility of a negative effect of this filter strongly depends on the type of data used for analysis. If the experimenter realizes that this filter is too stringent for the data at hand, he/she could simply change a single line of code (from „RIAincreasing = True" to

„RIAincreasing = False" in „run_experimentfile.py"), thus stopping the application of this filter.

Averaging over multiple scans vs. one single scan

Chromatographic peak widths are not constant for all analytes, can vary between replicates and are dependent on analyte concentration and chromatographic conditions. Thus "fronting" and or "tailing", very narrow as well as broad peaks, and peaks with multiple shoulders are commonly observed in liquid chromatography of highly complex samples. Determining the beginning and end of an eluting substance is a challenging task by itself. Averaging over multiple scans in order to produce one single spectrum necessitates the additional step of binning in the mass (m/z) dimension (a simple arithmetic mean will certainly not suffice, since the intensities of almost identical m/z values would not be summed). Also, the fact that the mass accuracy and relative isotope abundances increase with the signal intensity needs to be considered for such an endeavor. A reliable and fast averaging method could lead to smoother spectral envelopes and a better signal to noise ratio in comparison to single scans, but could also raise the complexity of the spectrum or lead to overlaps of isotopic peaks from different analytes that would not occur in a single scan. In a nutshell, there are advantages and disadvantages for both approaches. The utilization of a single scan is straightforward, but averaging over multiple scans isn't. The presented algorithm is trained to pick the best possible scan within the user-given retention time range, coping with large retention time deviations that can occur in high throughput studies. Additionally, the functionality of the presented program could easily be extended with a proper algorithm (if available) that averages over multiple spectra.

Comparison of Retention Time settings

Figure S1.Ai, Aii, and Aiii illustrate how the individual RIA values of peptides attributed to the same relatively high turnover protein are drawn closer together when applying smaller Retention time settings and that the best results are achieved with individually optimized Retention time window settings for each peptide for each file (for the data presented in Figure S1.Aiii and Biii this means 140 manual data entries for the beginning and the end of the retention time (14 peptides * 5 LC/MS measurements * 2 beginning/ending Rt). The analogue processing steps were performed for a relatively low turnover protein, see Figure S1.Bi, Bii, and Biii. In the latter plots, the values are less stringent, probably due to the fact

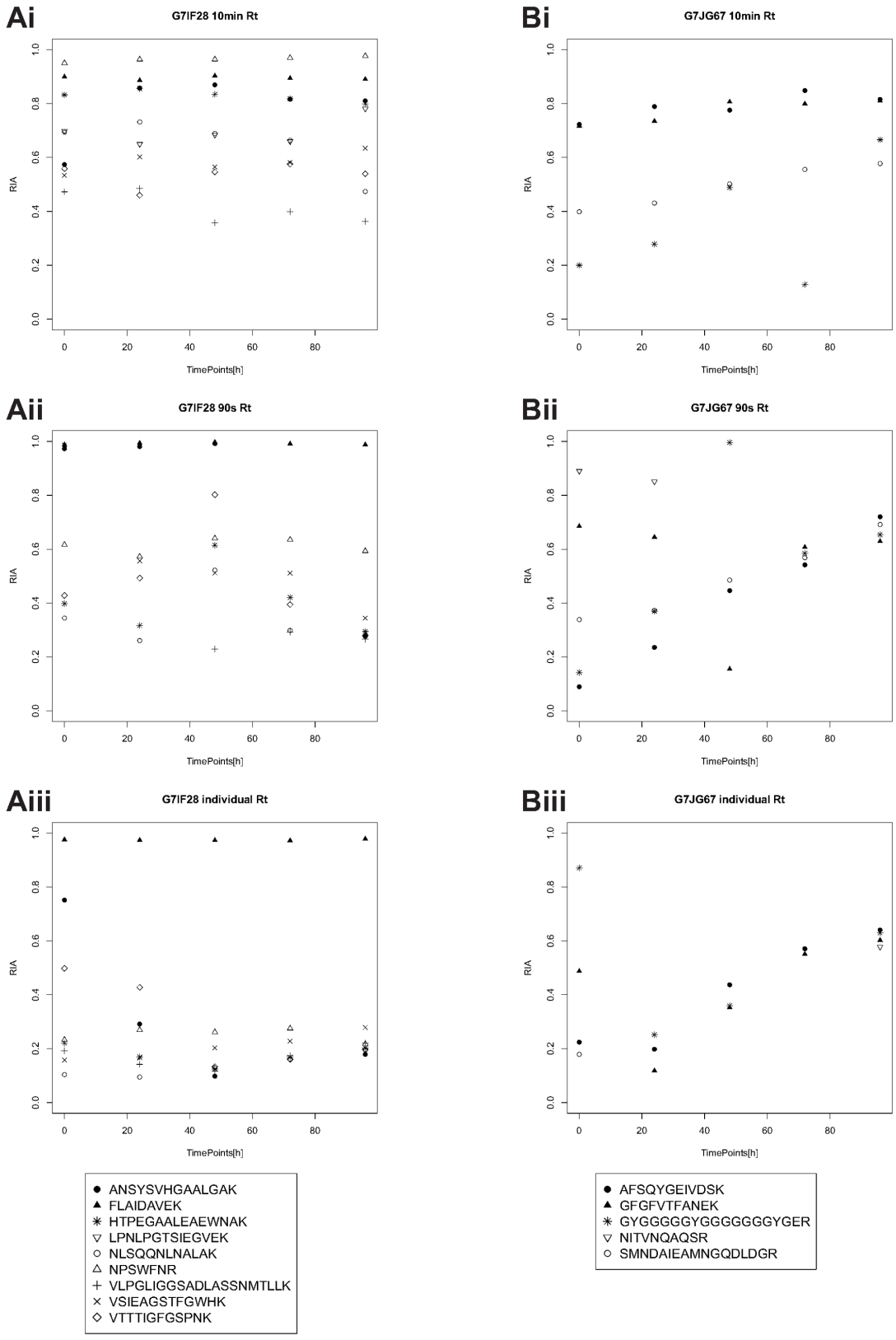
that low protein turnover produces low intensity heavy isotope envelopes, which are hard to differentiate from noise when averaging over multiple scans.

References

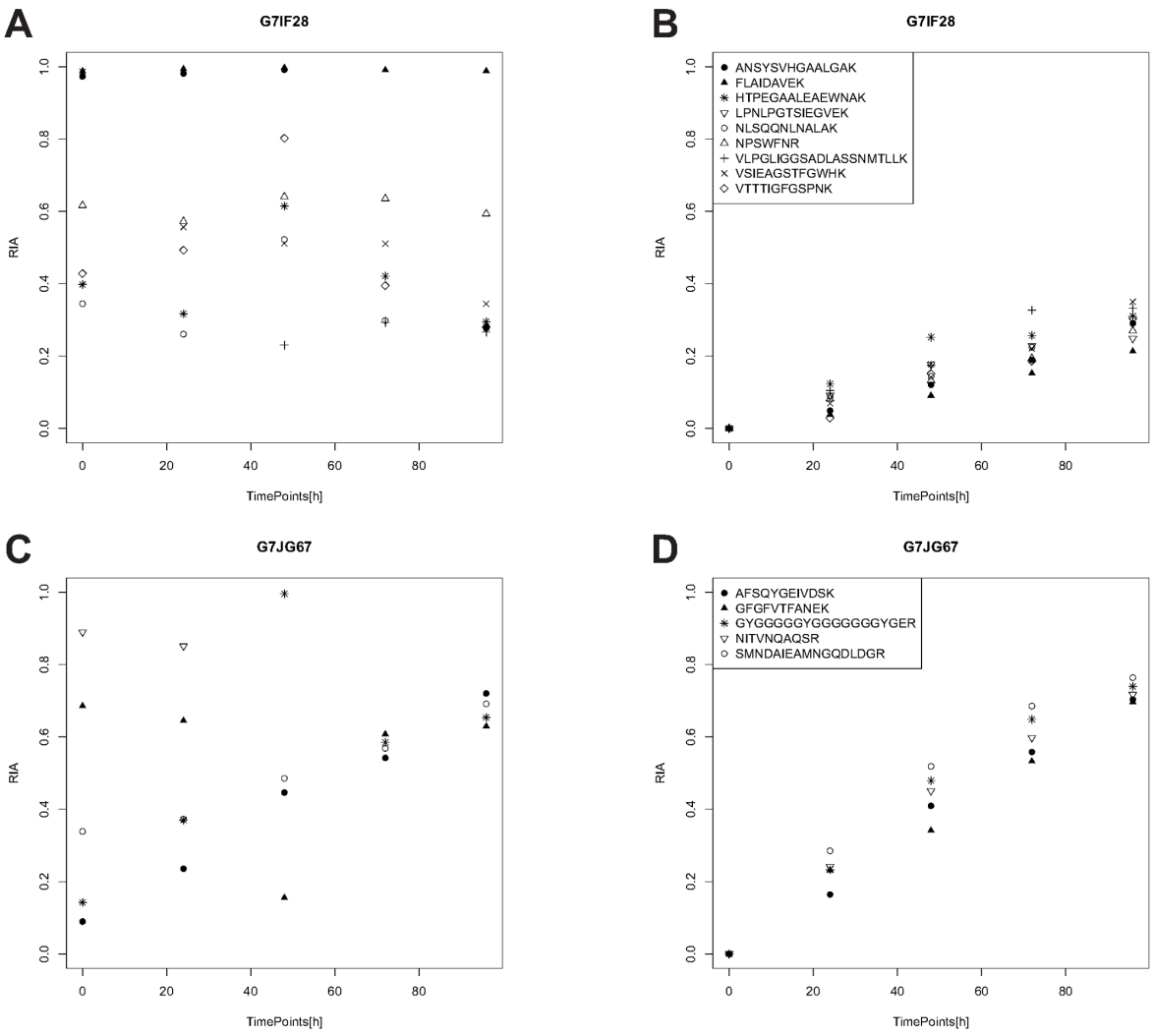
1. Hoehenwarter W, Wienkoop S (2010) Spectral counting robust on high mass accuracy mass spectrometers. Rapid Commun Mass Spectrom: 3609–3614. doi:10.1002/rcm.

2. Castillejo MA, Staudinger C, Egelhofer V, Wienkoop S (2014) Medicago truncatula proteomics for systems biology: novel rapid shotgun LC-MS approach for relative quantification based on Full-Scan Selective Peptide Extraction (Selpex). Plant Proteomics Methods Mol Biol 1072: 303–313.

3. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, et al. (2012) A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol 30: 918–920. doi:10.1038/nbt.2377.



Figure_S1 TurnStile output strongly depends on Rt range. The abscissa represents the Time Points and the ordinate the RIA ratio at the given Time Point. The titles of the plots indicate the Accession Number for the given data as well as the retention time window used for data analysis. From Ai to Aii to Aiii (note: the legend for these sub-plots shown at the bottom) and from Bi to Bii to Biii (note: the legend for these sub-plots shown at the bottom) the retention time window decreases from 10 min to 90 s to individually adapted values for every peptide for every file (in the range of 15 to 45 seconds).



Figure_S2 Qualitative comparison of TurnStile vs. Protover. The abscissa represents the Time Points and the ordinate the RIA ratio at the given Time Point. The titles of the plots indicate the Accession Number for the given data. The legends show all peptide sequences that could be attributed to the given protein. A and B show the RIA plots for G7IF28 (note: the legend for both sub-plots only shown in the right sub-plot). C and D show the RIA for G7JG67 (note: the legend for both sub-plots only shown in the right sub-plot). The data illustrated in A and D were processed using TurnStile with a 90 s retention time window (the recommended setting). B and D were processed using Protover with a 10 min retention time window (+/−5 min) (the recommended setting).

Additional remarks

Setting the Mass Resolving Power to a higher value raises the likelihood of baseline separation of overlapping peaks in the MS spectra. However, this comes at the cost of a loss of scan time (and an increase in file size). Thus, less MS/MS spectra were measured in order to increase the duty cycle time, which results in more MS spectra. This is necessary to obtain a sufficient amount of MS spectra to clearly delineate peaks in the chromatographic domain. I have provided the MS method for the analysis of the data.

Two complementary programs, named Isodist and Envelope, were developed for the analysis of isotopic distributions of MS data (Sykes and Williamson 2008; Sperling et al. 2008). Envelope enables the modeling of complex isotopic distributions of composite peptide species (e.g. a given amino acid sequence with natural isotopic distribution plus the same sequence enriched with 20% ^{15}N, in a ratio of 1 to 3). Isodist provides functionality to calculate the isotopic distribution of complex labeling patterns, computing the fractional degree of labeling as well as the ratio of the peptide species (e.g. given the latter example of a complex spectrum consisting of 2 merged peptide species, the fractional degree of labeling, i.e. 20% ^{15}N, as well as the ratio of 1 to 3, can be calculated) (Sperling et al. 2008; Sykes and Williamson 2008). I have tested the coupling of ProtOver to Isodist, since the fractional degree of labeling of a composite peptide spectrum is of interest for calculations, such as the estimation of the degree of labeling of the total precursor pool or even individual amino acids, as well as protein synthesis and degradation rates (Nelson, Li, and Millar 2014; Hinkson and Elias 2011; Claydon and Beynon 2012; Q. Li 2010). Envelope was used to generate simulated spectra of varying degrees of isotopic labeling. Specifically, three amino acid sequences “ANSYSVHGAALGAK”, “FLAIDAVEK”, and “VLPGLIGGSADLASSNMTLLK” (attributed to the Accession Number “G7IF28”), of charge state two, were used (unpublished data). For each sequence four individual spectra were generated. 1. Natural isotopic distribution; 2. Natural isotopic distribution plus 20% ^{15}N incorporation, in a ratio of 1 to 1; 3. Natural isotopic distribution plus 40% ^{15}N incorporation, in a ratio of 1 to 1; 4. Natural isotopic distribution plus 60% ^{15}N incorporation, in a ratio of 1 to 1 (see Figure 2: A1, B1, and C1). The resulting spectra were converted to an “mgf-style” format, subsequently converted to mzML-files using

msconvert from ProteoWizard, and the “ms level” changed from 2 to 1 (Martens et al. 2011; Chambers et al. 2012). These mzML-files served as the input for the ProtOver software, generating a data matrix of extracted peaks (analogous to centroids of profile mode peaks, see Figure 2: A1, B1, and C1), which was subsequently converted into individual spectra for each peptide for each degree of labeling, in order to serve as the input for Isodist. Isodist processing was performed using a two-population model, one consisting of the natural isotopic distribution and the other fully enriched with ^{15}N, otherwise default settings were applied. The Isodist output, consisting of the amplitude of the light and the heavy species, as well as the fractional degree of ^{15}N enrichment, was entered into Envelope in order to produce simulated spectra (see Figure 2: A2, B2, and C2). All intermediate steps of the above data preparation that could be automated were performed using an unpublished Python script. Figure 2: A1, B1, and C1 show that ProtOver picked the apex of each profile mode peak, as expected. A2, B2, and C2 show that only the first two sub-graphs (from bottom to top) were correctly simulated, while the remaining sub-graphs strongly deviate from the desired values, which should exactly resemble the corresponding sub-graphs A1, B1, and C1. Since several parameters can be modified for Isodist calculations, an adaptation of the latter seems likely to produce the desired results. A potential and beneficial approach could be to use existing RIA values to automatically adjust these parameters. This work remains to be done and published in the future.

In naturally occurring biological organisms, two distinct populations of vastly differing labeling degree (such as the simulated data described above) would not occur using a partial metabolic labeling approach. Rather many intermediate populations of varying degrees of labeling exist. The latter data is not feasible to be computed with the described Isodist program, since an unknown amount of populations entails many permutations to consider, such functionality is not given, and the computational cost would be exuberant. Nevertheless, a simplified two- or even three-species population model could be employed to estimate the fractional degree of labeling. The ProtOver algorithm extracts data from a single MS scan and the mass spectrometer used (LTQ-Orbitrap-XL) produces spectra of very high Mass Accuracy, but not of highly accurate relative intensities (Xu et al. 2010). Therefore, a future outlook would be to average over

Figure 2 Legend:

A1 and A2 show the amino acid sequence “ANSYSVHGAALGAK”; B1 and B2 “FLAIDAVEK”; C1 and C2 “VLPGLIGGSADLASSNMTLLK”. A1, B1, and C1 display the output of the Envelope simulation in blue and the ProtOver extracted data points in red. A2, B2, and C2 display the output of Isodist in blue and the same ProtOver extracted data points as in A1, B1, and C1 in blue, as a visual aid and for reference. Each sub-graph shows an increased degree of labeling, from bottom to top: 1. Natural isotopic distribution; 2. Natural isotopic distribution plus 20% ^{15}N incorporation, in a ratio of 1 to 1; 3. Natural isotopic distribution plus 40% ^{15}N incorporation, in a ratio of 1 to 1; 4. Natural isotopic distribution plus 60% ^{15}N incorporation, in a ratio of 1 to 1.

multiple MS scans and thus produce a “smoothed” averaged spectrum. Such “smoothed” spectra would be much more likely to yield successful results when coupling ProtOver to Isodist, but could also result in erroneous spectra due to co-elution. Furthermore, a Graphical User Interface (GUI) is planned, as well as parallelization of computation in order to increase speed and thus throughput, as well as potentially coupling/integrating the Isodist software (see Outlook).

Possible role of nutritional priming for early salt and drought stress responses in *Medicago truncatula*

Abiotic stress, specifically salt and drought stress, are becoming even more important with the decline of freshwater availability. It is therefore critical to aspire a more profound understanding of regulatory mechanisms of plants on a molecular level. The utilization and integration of unbiased (untargeted) GC/MS, LC/MS, as well as physiological data facilitates a systems biology approach to study salt and drought stress of the model plant *M. truncatula*.

I have adapted a so-called "mapping-file" (provided by the MapMan-store http://mapman.gabipd.org/web/guest/mapmanstore) of the publicly accessible tool MapMan (Thimm et al. 2004; Lohse et al. 2014) to operate with multiple specific protein databases, used for protein identification. This aided in the visualization and thus interpretation of highly complex shotgun proteomics data. Furthermore, I adapted a "metabolomics-mapping", a functional categorization of metabolites (also provided by the MapMan-store) and incorporated it into the "proteomics-mapping". I have achieved the latter tasks by programming (unpublished in-house Python script) using both string comparison techniques and standalone BLAST from http://blast.ncbi.nlm.nih.gov for sequence similarity search. Functional categories, so-called "bins", within this "mapping-file" were (and still are) manually adapted, and are freely available to download at the departments' website (http://www.univie.ac.at/mosys/databases.html) for the scientific community.

Declaration of authorship

The results of this chapter are presented in the form of a manuscript published in the journal „Frontiers in Plant Science". I have provided a critical contribution to the following publication, though the largest part the work was performed by the coauthors.

I have contributed by creating a non-redundant protein FASTA database from various sources, which was used for identification of shotgun proteomics data, using unpublished Python scripts. Furthermore, using online resources (see publication) as well as unpublished Python scripts, I've created a so-

called „mapping“ file. The latter links protein identifiers (Accession Numbers) to multiple functional categories. Thus aiding and enabling the interpretation of complex data. The „mapping file“ was used for visualization (see Figure 1 of the following publication) amongst other things. Further unpublished Python scripts which I created were used to retrieve sub-selections of various data. Finally, I’ve participated in pertinent discussions, proof-read the manuscript, and written the respective part of the manuscript.

Additional remarks

Due to manual refinement of protein and metabolite functional categories, addition of complementary databases, BLAST comparisons (homology searches), etc., the „mapping file“ is constantly amended and used for various tasks.

Published manuscript

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 21 December 2012
doi: 10.3389/fpls.2012.00285

Possible role of nutritional priming for early salt and drought stress responses in *Medicago truncatula*

Christiana Staudinger†, Vlora Mehmeti†, Reinhard Turetschek, David Lyon, Volker Egelhofer and Stefanie Wienkoop*

Department of Molecular Systems Biology, University of Vienna, Vienna, Austria

Edited by:
Dominique Job, Centre National de la Recherche Scientifique, France

Reviewed by:
Roque Bru-Martinez, Universidad de Alicante, Spain
Jean-Michel Ané, University of Wisconsin–Madison, USA

***Correspondence:**
Stefanie Wienkoop, Department of Molecular Systems Biology, University of Vienna, Althanstrasse 14, 1090 Vienna, Austria.
e-mail: stefanie.wienkoop@univie.ac.at

†Christiana Staudinger and Vlora Mehmeti have contributed equally to this work.

Most legume species establish a symbiotic association with soil bacteria. The plant accommodates the differentiated rhizobia in specialized organs, the root nodules. In this environment, the microsymbiont reduces atmospheric nitrogen (N) making it available for plant metabolism. Symbiotic N-fixation is driven by the respiration of the host photosynthates and thus constitutes an additional carbon sink for the plant. Molecular phenotypes of symbiotic and non-symbiotic *Medicago truncatula* are identified. The implication of nodule symbiosis on plant abiotic stress response mechanisms is not well understood. In this study, we exposed nodulated and non-symbiotic N-fertilized plants to salt and drought conditions. We assessed the stress effects with proteomic and metabolomic methods and found a nutritionally regulated phenotypic plasticity pivotal for a differential stress adjustment strategy.

Keywords: salt stress, plant-microbe interactions, drought stress, *Medicago truncatula*, mapman mapping

INTRODUCTION

Reduced water availability will dramatically impact agricultural productivity in the next 40 years. According to demographic and climate change models, the human population will double by 2050 and the variability in rainfalls will increase (IPCC, 2007). Therefore, we need a profound understanding of plant physiology and metabolism under water limiting conditions.

Drought and salinity are environmental constraints accounting for substantial yield losses. Both decrease the amount of water available to plants, leading to reduced growth, and photosynthesis (Chaves et al., 2009). Thus, it has been proposed that early acclimatory responses to both stresses share strong commonalities (Munns, 2002).

Legumes play an important role in increasing the sustainability of agricultural land use. Amongst several studies on drought and salt stress effects in model legumes, many have been conducted with *Medicago* spp. recently (Lopez et al., 2008; Bianco and Defez, 2009; Salah et al., 2009; Aranjuelo et al., 2011; Filippou et al., 2011; Kang et al., 2011). Noticeably, the symbiotic status amongst the studies is very diverse. The stress response of N-fixation in root nodules was extensively studied (Larrainzar et al., 2007, 2009; Naya et al., 2007; Lopez et al., 2008; Salah et al., 2009). However, various publications have been conducted with non-symbiotic (not inoculated with rhizobia) legumes (Sanchez et al., 2008a; Noreen and Ashraf, 2009; Diaz et al., 2010). Interestingly, a positive impact of rhizobial symbiotic interaction to stress has been proposed (Frechilla et al., 2000; Miransari and Smith, 2009). However, the influence of symbiotic interactions on abiotic stress acclimatory mechanisms is still in its infancy.

During their life-cycle plants acclimate to environmental constraints by a wide range of mechanisms that are conceptually classified as avoidance or tolerance strategies (Levitt, 1980). In case of lowered water availability in the environment, stress avoidance essentially aims at maintaining the initial plant water status and lowering the rate of stress imposed at the tissue or cellular level. Tolerance strategies aim at preventing damage and maintaining metabolism, once water deficit has been established. Avoidance and tolerance mechanisms are neither mutually exclusive nor active in a temporal sequence. Their distinction is conceptual, but useful when investigating plant stress responses (Verslues et al., 2006).

Plant acclimatory responses are complex exhibiting multigenic and interrelated properties. In addition, comparability with previous work is known to be hampered, due to heterogeneities in factors influencing stress responses such as plant age, growth conditions, diurnal changes, and the experimental treatment, such as severity, duration, and method of stress imposition (Aguirrezabal et al., 2006). Consequently, robust parameters for a specific definition of stress are still missing. Due to the complexity of plant stress response and its interlinked mechanisms and influencing factors, it becomes necessary to extend research to multilevel analyses (Jogaiah et al., 2012). Using systems biology approaches the integration

Abbreviations: C, control; D, drought; DW, dry weight; F', chlorophyll fluorescence in the light-adapted state; FDR, false discovery rate; $F_{m'}$, chlorophyll fluorescence when PSII centers are maximally closed in the light-adapted state; $F_{q'}$, difference between $F'$and $F_{m'}$; FW, fresh weight; g_s, stomatal conductance; IS, internal standard; N, nitrogen; N-fed, nitrogen fertilization; N-fix, nitrogen fixation; PS, photosynthesis; PSII, photosystem II; S, salt; SE, standard error; WC, water content; Ψ_{leaf}, leaf water potential.

of -omics data such as metabolomics and proteomics may also compensate method specific limitations.

To date, data of proteomic studies are still behind in numbers of identifications that of transcript data. Nevertheless, the informative value on the protein level seems high for several reasons. For instance, the direct translation of transcript abundance to protein abundance in terms of one point abundance and changes over time is still under controversial debate. Especially, in the context of changes in time- and stress dependent manner it has been shown that transcript and protein data do not correlate significantly (Hajduch et al., 2010). As a possible reason they suggest for instance regulation via post-translational protein modification. A temporal lag that causes, e.g., a delay in adjustment of enzyme abundance when transcript levels have already changed, have extensively been discussed by Gibon et al. (2004, 2006).

So far, most studies focused on genetic engineering using, e.g., Quantitative Trait Loci (QTL) mapping have shown only limited success (Rispail et al., 2010). Thus, knowledge transfer from transcript and genome data complemented with postgenomic metabolite and proteome data will enhance the success for smart breeding in future.

In the present study, early stress response mechanisms to salt and drought stress have been investigated. The aim of this work was to (i) unravel robust and easily detectable putative stress response markers on a physiological, metabolite as well as protein level and (ii) to find novel insights for a regulatory relevant role of the nutritional priming comparing shoots of N-fixing with fertilized *M. truncatula* plants.

MATERIALS AND METHODS

PLANT GROWTH AND SAMPLING CONDITIONS

The seeds of barrel medic (*M. truncatula* A17 cv. Jemalong) were surface sterilized and sown in pots containing a mixture of perlite:vermiculite 2:5 (v:v). The experimental setup was based on the protocol used by (Larrainzar et al., 2009). Plants were grown under controlled conditions in a growth chamber (14-h day and 10-h night; 270 µmol m^{-2} s^{-1} photosynthetic photon flux density; 22°C day and 16°C night temperatures; 50–60% relative humidity). During the first week of growth, plants were watered with nutrient solution (Evans, 1981) containing 0.5 mM ammonium nitrate. The following 2 weeks a nutrient solution with ammonium nitrate concentration of 2.5 mM was used for watering in order to enhance biomass accumulation and to keep plant growth performance identical during the initial developmental stage. After 3 weeks, half of the plants were randomly selected and inoculated with *S. meliloti* 2011. Furthermore, for inoculated plants nutrient solution was N free while the other subset was fertilized with 2.5 mM ammonium nitrate. After 7 weeks plants were randomly separated into sub-sets: control and drought or salt stressed, respectively. Control plants were supplied daily with nutrient solution to pot capacity whereas abiotic stress was applied to the other groups as follows. Drought stress was imposed by withholding water and nutrients; after flushing pots with deionized water, nutrient solutions containing 200 mM NaCl were applied every day to salt stressed plants. After 6 days of stress, plants were harvested 6 h after the onset of light. *M. truncatula* shoot and root tissue was separated, flash-frozen in liquid nitrogen, and stored at −80°C until further processing. Analysis was carried out as previously described using 3 biological replicates for each condition: N-fertilized and inoculated plants [*n* the following referred to as N-fed and nitrogen fixing (N-fix)] exposed to salt stress or water deprivation as well as control without stress treatment.

PHYSIOLOGICAL PARAMETERS

Stomatal conductance (g_s) was measured 3 h after onset of the photoperiod with a steady-state porometer (PMR-4, PP Systems, Hitchin, UK) connected to the EGM-4 gas monitor serving as data logger. About 0.5 cm^2 of terminal leaflets of fully expanded leaves were placed into a cuvette. Records were taken after ∼20 s, when equilibrium was established. The inlet air flow rate was kept constant at 75 ml/min. The porometer then measured the air humidity of inlet and outlet air flow, air temperature and the PPFD reaching the leaf. From these parameters g_s was calculated. The water content (WC) of the leaves and roots was calculated as (FW−DW)*FW^{-1}(FW = fresh weight; DW = dry weight). Leaf water potential was measured 3 h after the onset of the photoperiod with a Scholander pressure bomb. Primary chlorophyll fluorescence parameters ($F_{m'}$, F') were assessed employing a saturation pulse method, using the MINI-head version of the IMAGING-PAM chlorophyll fluorometer M-series (Heinz Walz GmbH, Effeltrich, Germany). The PSII operating efficiency was calculated by $F_q'/F_{m'} = (F_m' - F')/F_m'$ (Baker, 2008). Analysis was carried out on three biological replicates for each of the previously described conditions (**Table 1**).

EXTRACTION AND DERIVATIZATION OF METABOLITES

Medicago truncatula roots and shoots were ground to a fine powder under liquid nitrogen and subsequently lyophilized. About 10 and 30 mg of the powdered shoots and roots were used for the extraction with 1 ml of freshly prepared and pre-cooled extraction buffer ($MeOH:CHCl_3:H_2O$, 2.5:1:0.5), respectively. In order to avoid any degradation or modification of metabolites the samples were kept on ice for 8 min. During this time the samples were vortexed

Table 1 | Effect of drought and salt treatments on plant water status and physiological parameters in N-fed (A) and N-fix (B) *M. truncatula*.

Parameter	Control	Drought	Salt
(A) N-FED			
WC_{shoot}	0.82 ± 0.06 a	0.78 ± 0.01 b	0.82 ± 0.01 a
Ψ_{leaf} (MPa)	−0.68 ± 0.07 a	−1.06 ± 0.08 b	−0.69 ± 0.18 a
g_s [mmol m^{-2} s^{-1}]	381.52 ± 139.02 a	121.52 ± 30.32 b	37.01 ± 5.46 c
F'_q/F'_m	0.55 ± 0.05 a	0.44 ± 0.07 b	0.58 ± 0.03 a
(B) N-FIX			
WC_{shoot}	0.89 ± 0.01 a	0.89 ± 0.01 a	0.90 ± 0.01 a
Ψ_{leaf} (MPa)	−0.73 ± 0.10 a	−0.98 ± 0.09 b	−0.75 ± 0.17 a
g_s [mmol m^{-2} s^{-1}]	425.95 ± 156.23 a	165.71 ± 36.15 b	36.14 ± 6.40 c
F'_q/F'_m	0.57 ± 0.02 a	0.56 ± 0.01 a	0.58 ± 0.02 a

Values represent the mean ± SE (n = 3). The letters a, b, and c indicate significant differences between control and stress treatments (Student's t test $p < 0.05$). WC, water content; Ψ_{leaf}, leaf water potential; g_s, stomatal conductance; F'_q/F'_m, PSII operating efficiency.

regularly and afterward centrifuged for 4 min at 14,000 *g*/min, at 4˚C. The supernatant was added to another tube which contained 500 µl of ultrapure water and shaken thoroughly. After the phase separation by centrifugation (4 min, 14,000 *g*/min), the upper polar phase was split into two aliquots. Internal standard (IS) was added (10 µl of 0.1 g/l $^{13}C_6$-Sorbitol) and the samples were dried out using a vacuum concentrator at room temperature. For metabolite derivatization, 20 µl of the freshly prepared methoxymation mixture (40 g/l methoxyamine hydrochloride CH_3ONH_2*HCL in pyridine) were added to the dried samples and shaken for 90 min at 30˚C. After adding 80 µl of the silylation mixture: 1 ml of MSTFA (*N*-methyl-*N*-trimethylsilyl trifluoroacetamide) spiked with 30 µl of the alkane standard mixture (C10-C40, each 50 mg/l) as retention index (RI) marker, the samples were incubated for 30 min with shaking at 37˚C and then centrifuged (14,000 *g*/min) for 2 min to remove any insoluble material. The supernatant was carefully taken and transferred into glass vials with micro inserts. One microliter of the derivatized sample was injected. Six replicates per treatment (three biological, two technical) were randomly injected to discriminate technical from biological variation.

GC-TSQ-MS SETTINGS

For metabolite profiling GC-MS is mostly the method of choice. Here we used GC hyphenated to triple quadrupole (Thermo Scientific TSQ Quantum GC™, Bremen, Germany). In order to identify a large number of metabolites, a profiling analysis in full-scan mode was performed with a scan range of m/z 40–600 and a scan time of 200 ms. The metabolite separation was performed on a HP-5MS capillary column (30 m × 0.25 mm × 0.25 µm; Agilent Technologies, Santa Clara, CA, USA), at a constant flow 1 ml/min helium. The split less injection of 1 µl of the sample was done by the TriPlus auto sampler (Thermo Scientific, Bremen, Germany). The temperature of the injector was 230˚C. Compound elution settings were 1 min at 70˚C isotherm, ramp to 76˚C at 1˚C per min heating rate, then to 350˚C at a 6˚C per min rate and hold for 1 min. Post run temperature was set to 325˚C for 10 min. The transfer line temperature was set to 340˚C and ion source temperature was 250˚C. Electron Impact (EI) ionization was set to 70 eV.

METABOLITE DETECTION, IDENTIFICATION, AND RELATIVE QUANTIFICATION

The criteria used for identification were fragmentation patterns that are characteristic for the particular compound, the retention time (RT) and RI. Combining these criteria, it is possible to unambiguously identify metabolites and distinguish between the components even if they are chemically very similar. The identification of each analyte was achieved by matching the MS-spectra and RT against (a) an in-house library (modified gmd database)[1]; (b) AMDIS (calculation of retention indices and comparison with RI of compounds in the mass spectral library); and (c) matching against the in-house measured standards. Calculation of retention indices was performed using the RT of the detected compound and the RT of the RT-index marker (alkane mixture), calculated with AMDIS for representative samples of different treatments. Due to derivatization, in some cases more than one peak was detected for one metabolite. These peaks were initially analyzed separately and summed up for further analysis or data mining. About 15% of the detected analytes were identified as unknown compounds. Calculation of the peak areas was performed using LC-Quan for the GC-TSQ-MS data, which is suitable to calculate the peak area for all compounds in all samples according to given parameters. Here the determined RT as well as the quant mass for each component was used to automatically extract data from all sample replicates. An initial data matrix of the calculated peak area for each detected compound was obtained separately. The list of detected components and calculated areas was exported to an Excel file. We used an in-house Matlab tool to produce a complete data matrix automatically. The data matrix was normalized to the sample DW and the IS for relative quantification.

PROTEIN EXTRACTION

The same three biological replicates as those taken for metabolite analysis have been used for protein extraction. Two hundred milligrams of liquid nitrogen frozen shoot material were cryoground using a Retsch MM400 ball mill and homogenized in 1 ml of urea buffer containing 50 mM HEPES, pH 7.8, 5 mM PMSF, and 8 M Urea. After centrifugation (10,000 *g*, 10 min, 4˚C) the urea soluble proteins in the supernatant were precipitated overnight in five volumes of −20˚C cold acetone containing 0.5% β-mercaptoethanol. The precipitate was pelleted at 4,000 *g*, 4˚C for 15 min. The resulting pellet was washed with −20˚C cold methanol and again centrifuged (4,000 *g*, 4˚C, 10 min).

PROTEIN DIGESTION

Air-dried protein pellets were dissolved in 500 µl urea buffer the protein concentration was determined by Bradford assay, using BSA as a standard. 100 µg of protein was initially digested using endoproteinase LysC (1: 100 vol/vol, 5 h, 30˚C, Roche, Mannheim, Germany). For the second digestion step, samples were diluted with trypsin buffer (10% ACN, 50 mM AmBic, 2 mM $CaCl_2$) to a final concentration of 2 M Urea and incubated overnight at 37˚C with Porosyzme immobilized trypsin beads (1:10, vol/vol; Applied Biosystems, Darmstadt, Germany). The digest was desalted with C18-SPEC 96- well plates (Varian, Darmstadt, Germany) according to the manufacturer's instructions. The eluted peptides were vacuum-dried.

nanoESI LC-MS/MS

Peptide digests (0.5 µg each) were randomly applied to a RP monolithic capillary column (50 µm internal diameter, 15 cm length, Merck, Darmstadt, Germany) separated during a 120 min gradient ranging from 90% solvent A (0.1% FA in water) to 80% solvent B (80% acetonitrile, 0.1% FA in water). For each treatment tree biological and three technical replicates were randomly analyzed. MS analyses were performed on a LTQ-Orbitrap XL (Thermo Fisher Scientific, Bremen, Germany). For the database dependent spectral count analysis (Wienkoop, 2011), a top five MS analysis setting was used with the full-scan range from 350 to 1,800 m/z. Dynamic exclusion settings were as described in Hoehenwarter and Wienkoop (2010). Briefly, repeat count was set to

[1] http://gmd.mpimp-golm.mpg.de/download/

one, repeat duration 20 s, exclusion list size 500, exclusion duration 60 s and exclusion mass width 10 ppm. Charge state screening was enabled with rejection of unassigned and 1+ charge states. Minimum signal threshold counts were set to 1,000.

PROTEIN IDENTIFICATION AND RELATIVE QUANTIFICATION

We used the SEQUEST algorithm and the Proteome Discoverer (v 1.3, Thermo Scientific) to search MS data against a fasta file we created from a *Medicago* spp. and *Sinorhizobium* spp. subset of UniProt Knowledgebase[2] containing 63,688 sequences as of April 2012. *In silico* peptide lists were generated with the following settings: trypsin as the digestion enzyme, a maximum of three missed cleavages and methionine oxidation as dynamic modification. Mass tolerance was set to 5 ppm for precursor ions and 0.8 Da for fragment ions. Additionally, a decoy database containing reversed sequences was used to estimate the false discovery rate (FDR) Only high confidence (FDR $\leq$ 0.01%) peptide identifications with a minimum XCorr of 2.2 and proteins with at least two distinct peptides were considered. Peptide spectra are stored in the ProMEX library (Wienkoop et al., 2012) and can be checked under its ID "Med trun001." Protein relative quantification is based on database dependent spectral counting as described previously (Larrainzar et al., 2009). Six replicates per treatment (three biological, two technical) were randomly injected to discriminate technical from biological variation.

STATISTICAL ANALYSIS

Detailed analysis of the physiology, as well as metabolite and protein data was performed by calculating the ratios between control and treated samples. Significant differences between these were determined using Student's *t* test at $p < 0.05$ and fold change ≥ 2 (**Tables 2** and **3**).

[2]http://www.uniprot.org/

MAPMAN MAPPING FILE FOR *M. TRUNCATULA* PROTEINS AND METABOLITES

A new Mapman mapping file was created on the basis of the mapping file "Mt_Mt3.5_0411" and "MappingMetabolites" acquired from http://mapman.gabipd.org/web/guest/mapmanstore. This mapping file corresponds to MTGI release "Mt3.5v3 RELEASE 20100825" ("Mt3.5_GenesProteinSeq_20100825.fasta" subsequently called MTGI-fasta-DB) which can be found at http://www.jcvi.org/. Shotgun proteomics experimental data were evaluated with the Uniprot database fasta file (see Protein Identification and Relative Quantification).

The "Identifier" and "Description" categories of entries from the "Mt_Mt3.5_0411" mapping file correspond to accession numbers and header information of the MTGI-fasta-DB. The mapping file "Med trun_mappingformapman_Mosys_v1_20120913.txt" was created by comparing the protein sequences of the Uniprot-fasta-DB (MT only) to the MTGI-fasta-DB. Comparison was performed using string comparison (unpublished Python script) as well as standalone BLAST from http://blast.ncbi.nlm.nih.gov. Mapping file entries corresponding to completely identical sequences were replaced. The "Bincode" and "Name" remained unchanged, but the "Identifier" and the "Description" were replaced by the corresponding Uniprot accession number and header, furthermore the "Type" was set to "*P.*" All Uniprot entries not 100% identical in sequence and length to an entry in MTGI were blasted against a database created from the entire MTGI fasta file. Uniprot entry hits with an *e*-value equal to or lower than10^{-3} replaced mapping file entries as previously described. Uniprot entry hits with *e*-values higher than 10^{-3} were added to the bincode "35.2.1" with the name "not assigned.unknown.evalhigh" and entry hits resulting in no query hit at all were assigned to the bincode "35.2.2" with the description "not assigned.unknown.blastwithouthits." The pertinent information of the metabolite mapping file from the MapMan Store was incorporated into the current mapping file by simply adding the respective entries (at the proper bin location). Certain entries were manually curated and shifted from "not assigned.unknown" bins to appropriate categories. Six metabolites

Table 2 | Ratios of stress responsive root metabolites (stressed/control).

	SN-fed vs.CN-fed	SN-fix vs.CN-fix	DN-fed vs. CN-fed	DN-fix vs.CN-fix
GABA	3.1 (0.019)	0.5 (0.046)	ns	2.0 (0.043)
Aspartate	3.7 (0.011)	3.3 (0.048)	2.5 (0.049)	2.8 (0.008)
Leucine	2.0 (0.007)	ns	2.5 (0.005)	ns
Threonate	3.0 (0.046)	2.9 (0.048)	ns	2.8 (0.001)
Glutamate	4.9 (0.024)	ns	2.0 (0.001)	ns
Proline	ns	ns	10.5 (0.001)	12.1 (0.005)
Fumarate	3.5 (0.025)	2.3 (0.004)	3.1 (0.001)	3.3 (0.009)
Galactonate	3.8 (0.001)	2.8 (0.013)	ns	2.0 (0.003)
Sucrose	2.7 (0.003)	4.4 (0.006)	2.8 (0.001)	2.1 (0.003)
Myo-Inositol	4.0 (0.003)	ns	2.3 (0.002)	2.0 (0.001)
Ononitol	3.0 (0.035)	2.0 (0.003)	ns	2.0 (0.003)
Pinitol	2.6 (0.048)	ns	3.0 (0.018)	ns

Fold change $\geq$ 2 and student's t test $p < 0.05$ in brackets (n = 6). CN-fed, mean of controls of N-fertilized plant roots; SN-fed, mean of salt stressed, N-fertilized plant roots; CN-fix, mean of controls of N-fixing plant roots; DN-fix, mean of salt stressed, N-fixing plant roots; ns, not significantly changed.

Table 3 | Stress responsive shoot proteins and metabolites of six replicates as fold change.

	Stressed Drought		Stressed Salt		Non-stressed Controls
	DN-fed/CN-fed	**DN-fix/CN-fix**	**SN-fed/CN-fed**	**SN-fix/CN-fix**	**CN-fed/CN-fix**
PROTEINS					
1. Photosystem (PS)					
1.1 PS.lightreaction					
G7IJ45 photosystem II 10 kDa polypeptid	ns	ns	ns	3.8 (0.002)	ns
G7JH56 photosystem II CP47 chlorophyll apoprotein	ns	ns	2.2 (0.005)	ns	ns
G7JE46 thylokoid luminal 16.5 kDa protein	ns	ns	ns	0.4 (0.009)	ns
G7JAX6 photosystem I reaction center subunit N	ns	0.3 (0.029)	ns	ns	2.9 (0.043)
G7JQA7 apocytochrom *f*	0.3 (0.0029)	ns	ns	ns	ns
B7FIU4 ATP synthase gamma chain	ns	2.8 (0.003)	ns	ns	0.5 (0.010)
B7FIR4 ATP synthase gamma chain	ns	0.5 (0.009)	ns	ns	ns
G7JAI2 ATP synthase	ns	ns	3.5 (0.0030)	ns	ns
1.2 PS.photorespiration					
G7JAR7 serin hydroxymethyltransferase	ns	0.5 (0.019)	ns	ns	ns
1.3 PS.calvin cycle					
G7J252 ribulose bisphosphate carboxylase small chain	ns	ns	3.3 (0.020)	ns	3.2 (0.005)
2.1.2 Major CHO metabolism.synthesis.starch.transporter					
G7LDP4 ADP, ATP carrier protein	ns	ns	2.1 (0.014)	ns	ns
3.4 Minor CHO metabolism.myo-inositol					
G7J4B5 l-myo-inositol-1 phosphate synthase	5.3 (0.0001)	ns	ns	ns	4.0 (0.033)
G7LAD5 l-myo-inositol-1 phosphate synthase	2.0 (0.0204)	ns	ns	ns	ns
6.3 Gluconeogenesis.Malate DH					
G7JTZ0 Malate dehydrogenase	ns	ns	ns	0.4 (0.000)	ns
7.1 OPP.oxidative PP.6-phosphogluconate dehydrogenase					
Q2HVD9 6-phosphogluconate dehydrogenase	ns	0.5 (0.001)	ns	ns	2.0 (0.001)
9.9 Mitochondrial electron transport/ATP synthesis.F1-ATPase					
G7LCJ4 ATP synthase delta subunit	2.5 (0.0001)	ns	ns	ns	2.2 (0.048)
10.1 Cell wall.precursor synthesis					
G7L571 UDP-glucose 6-dehydrogenase	ns	0.3 (0.001)	ns	ns	ns
11.1 Lipid metabolism.FA synthesis and FA elongation					
G7LIV6 biotin carboxylase	ns	0.4 (0.032)	ns	ns	2.6 (0.023)
G7JNN1 Acyl-[acyl-carrier-protein] desaturase	ns	ns	3.6 (0.014)	ns	ns
11.6 Lipid metabolism.lipid-transfer proteins					
G7JID0 non-specific lipid-transfer protein	2.5 (0.002)	ns	ns	ns	ns
12.2 *N*-metabolism.ammonia metabolism.glutamate synthase					
Q2HW53 ferredoxin-dependent glutamate synthase	ns	0.5 (0.000)	ns	ns	ns
P04078 glutamine synthetase cytosolic isozyme	ns	ns	ns	0.5 (0.004)	3.8 (0.004)
13.1 Amino acid metabolism.synthesis					
Q6J9 × 6 SAMS	2.2 (0.0079)	ns	ns	ns	ns
A4ULF8 SAMS	2.4 (0.0007)	ns	ns	ns	ns
A4PU48 SAMS	ns	0.5 (0.009)	ns	ns	ns
G7L3W1 SAMS	ns	0.5 (0.002)	ns	ns	ns
G7JTY4 LL-diaminopimelate aminotransferase	ns	ns	0.4 (0.021)	ns	ns
G7J013 alanine glyoxylate aminotransferase	ns	ns	ns	2.4 (0.005)	ns
15.2 Metal handling.binding, chelation, and storage					
G7K283 ferritin	ns	ns	4.0 (0.018)	ns	ns
G7JLS7 ferritin	11.4 (0.005)	ns	10.0 (0.0004)	ns	ns
16.2 Secondary metabolism.phenylpropanoids					
G7JTH6 caffeic acid 3-*O*-methyltransferase	6.5 (0.0000)	ns	4.0 (0.0023)	ns	ns

(Continued)

Table 3 | Continued

	Stressed Drought		Stressed Salt		Non-stressed Controls
	DN-fed/CN-fed	**DN-fix/CN-fix**	**SN-fed/CN-fed**	**SN-fix/CN-fix**	**CN-fed/CN-fix**
19.10 Tetrapyrrole synthesis					
G7IK85 Mg-chelatase subunit chlI	0.3 (0.0002)	0.3 (0.0005)	ns	0.5 (0.004)	ns
20.1 Stress.biotic					
B0RZH7 putative thaumatin-like protein	ns	0.4 (0.000)	ns	ns	2.2 (0.001)
G7IYL0 receptor-like protein kinase	ns	0.5 (0.002)	ns	ns	ns
20.2 Stress.abiotic					
Q2HT97 heat shock protein Hsp70	ns	ns	ns	0.5 (0.034)	ns
G7JGC6 low-temperature inducible	2.3 (0.0001)	ns	ns	ns	2.3 (0.001)
G7JGC9 low-temperature inducible	ns	ns	ns	0.3 (0.021)	ns
21.5 Redox.peroxiredoxin					
G7JS60 peroxiredoxin Q	2.6 (0.0000)	ns	ns	ns	ns
23.4 Nucleotide metabolism.phosphotransfer and pyrophosphatases					
G7JMM2 nucleoside diphosphate kinase	ns	ns	2.0 (0.040)	ns	ns
B7FIM7 soluble inorganic pyrophosphatase	ns	03 (0.006)	ns	ns	ns
26.20 Misc.ferredoxin-like					
G7KWY5 ferredoxin	ns	0.3 (0.011)	ns	ns	2.1 (0.037)
26.4 Misc.beta-1,3 glucan hydrolases					
G7JQL4 endo-beta-1,3-glucanase	ns	ns	0.5 (0.014)	ns	ns
27.1 RNA.processing					
G7JK09 Poly(A)-binding protein	ns	0.4 (0.000)	ns	ns	ns
27.4 RNA.RNA binding					
G7JG67 glycerine-rich RNA binding protein	0.5 (0.0059)	ns	ns	ns	ns
29.2 Protein.synthesis					
Q945F4 eukaryotic translation initiation factor 5A-2	ns	0.4 (0.003)	ns	ns	2.5(0.003)
G7IH13 elongation factor EF-2	ns	ns	ns	2.5 (0.000)	ns
29.5 Protein.degradation					
G7LIT0 ATP-dependent Clp protease	0.4 (0.0226)	ns	ns	ns	ns
G7ZVC0 presequence protease	ns	0.5 (0.010)	ns	ns	ns
G7K8J5 bi-ubiquitin	ns	ns	0.3 (0.024)	ns	ns
G7LB82 proteasome subunit alpha type	ns	ns	2.1 (0.019)	ns	ns
31.1 'Cell.organization					
G7IAN2 tubulin ßchain	ns	ns	5.4 (0.0006)	ns	ns
G7L5V0 tubulin ßchain	ns	ns	3.0 (0.0205)	0.4 (0.046)	ns
G7KB73 annexin	2.0 (0.0394)	ns	ns	ns	0.5 (0.000)
G7JAX5 actin	ns	ns	3.8 (0.0001)	ns	ns
34.1 Transport. p- and v-ATPases					
A6Y950 Vacuolar H + -ATPase B subunit	ns	0.5 (0.001)	ns	ns	ns
"PUTATIVE" UNCHARACTERIZED PROTEINS					
B7FJY9 similar 94.0% Q9SQL2, CB24_PEA, chlorophyll a-b binding protein P4, chloropl., *Pisum sativum* (garden pea), $e = 1.0 \times 10^{-178}$	3.0 (0.0001)	ns	ns	ns	ns
B7FMC4 similar 73.0% Q03666, GSTX4_TOBAC, probable glutathione *S*-transferase, *Nicotiana tabacum* (common tobacco), $e = 1.0 \times 10^{-121}$	2.1 (0.007)	ns	ns	ns	2.6 (0.036)
B7FJR8 similar 83.0% Q9LZG0, ADK2_ARATH, adenosine kinase 2, *Arabidopsis thaliana* (mouse-ear cress), $e = 0$	ns	0.4 (0.000)	ns	ns	2.1 (0.015)
B7FM78 similar 97.0% P81406, GAPN_PEA, NADP-dependent glyceraldehyde-3-phosphate dehydrogenase, *Pisum sativum* (garden pea), $e = 0$	ns	0.5 (0.023)	ns	ns	ns

(Continued)

Table 3 | Continued

	Stressed Drought		Stressed Salt		Non-stressed Controls
	DN-fed/CN-fed	**DN-fix/CN-fix**	**SN-fed/CN-fed**	**SN-fix/CN-fix**	**CN-fed/CN-fix**
B7FKR5 similar 99.0% O24076, GBLP_MEDSA, guanine nucleotide-binding protein subunit beta, *Medicago sativa* (alfalfa), $e = 0$	ns	0.4 (0.005)	ns	ns	2.6 (0.001)
B7FI14 similar 64.0% Q9LEH3, PER15_IPOBA, peroxidase 15, *Ipomea batatas* (sweet potato) (*Convolvulus batatas*), $e = 1.0 \times 10^{-132}$	ns	0.4 (0.000)	ns	ns	2.2 (0.001)
B7FL15 similar 85.0% P13443, DHGY_CUCSA, glycerate dehydrogenase, *Cucumis sativus* (cucumber), $e = 2.0 \times 10^{-71}$	ns	0.3 (0.000)	ns	ns	2.7 (0.000)
G7I4F9 uncharacterized protein	ns	0.4 (0.021)	ns	ns	2.1 (0.035)
B7FHX0 similar 98.0% P29500, TBB1_PEA, tubulin beta-1 chain, *Pisum sativum* (garden pea), $e = 0$	ns	ns	4.4 (0.0035)	ns	2.2 (0.002)
B7ZWQ5 similar 90.0% Q40977, MDAR_PEA, monodehydroascorbate reductase, *Pisum sativum* (garden pea), $e = 0$	ns	ns	2.0 (0.035)	ns	ns
B7FL16 similar 84.0% P13443, DHGY_CUCSA, glycerate dehydrogenase, *Cucumis sativus* (Cucumber), $e = 2.0 \times 10^{-88}$	ns	ns	ns	0.5 (0.043)	ns
B7FI41 similar 52.0% Q41160, LCB3_ROBPS, putative bark agglutinin LECRPA3, *Robinia pseudoacacia* (BLAQCK locust), $e = 5.0 \times 10^{-87}$	ns	ns	ns	0.4 (0.023)	ns
G7KAG7 similar 71.0% Q9THX6, TL29_SOLLC, thylakoid lumenal 29 kDa protein, chloroplast, *Solanum lycopersicum* (tomato; *Lycopersicon esculentum*), $e = 1.0 \times 10^{-172}$	ns	ns	ns	0.5 (0.032)	ns
B7FNH1 similar 79.0% O23755, EF2_BETVU, elongation factor 2, Beta vulgaris (sugar beet), $e = 2.0 \times 10^{-67}$	ns	ns	ns	0.3 (0.004)	3.6 (0.004)
METABOLITES					
Major CHO metabolism					
Glucose	ns	10 (0.034)	ns	0.3 (0.009)	0.5 (0.014)
Glucose-1-p	ns	ns	ns	5.1 (0.000)	0.5 (0.014)
Maltose	ns	ns	ns	2.3 (0.003)	ns
Ribitol	3.2 (0.010)	ns	ns	ns	ns
Amino acid metabolism					
Glutamate	ns	ns	2.1 (0.010)	ns	ns
Leucine	ns	6.1 (0.000)	2.7 (0.006)	2.2 (0.012)	ns
Proline	0.5 (0.049)	ns	2.6 (0.006)	ns	ns
Valine	ns	ns	2.4 (0.000)	2.4 (0.012)	ns
Aspartate	ns	0.3 (0.021)	ns	ns	ns
TCA					
2-Oxoglutarate	ns	ns	ns	0.3 (0.040)	ns
Citrate	ns	ns	0.5 (0.008)	0.3 (0.000)	ns
Succinate	ns	ns	ns	0.5 (0.008)	ns
Malate	ns	ns	ns	0.2 (0.001)	2.0 (0.012)
Malonate	ns	0.5 (0.029)	0.4 (0.000)	0.1 (0.001)	2.1 (0.001)
Others					
Phosphate	ns	ns	0.5 (0.006)	0.2 (0.000)	ns

(Student's t test $p < 0.05$ in brackets and fold change ≥ 2; $n = 6$) with significantly altered abundances in response to spectral counts of stress proteins and peak area of metabolites (IS and DW normalized). Protein category headers including binning numbers of the MapMan mapping file. CN-fed, control, N-fertilized; CN-fix, control, N-fixation; DN-fed, drought, N-fertilized; DN-fix, drought, N-fixation; SN-fed, salt, N-fertilized; SN-fix, salt, N-fixation; numbers 1 – 6 indicate replicates. ns, not significant; SAMS, S-adenosylmethionone synthetase.

not previously contained in the mapping were added. Separate bins were created for *S. meliloti* and *S. medicae*. Identified and pertinent protein accessions of these two endosymbionts were manually classified and thus put into sub-bins. The mapping file can be downloaded at http://www.univie.ac.at/mosys/databases.html. It will be updated in accordance to novel identifications/insights.

FUNCTIONAL CHARACTERIZATION OF STRESS RESPONSIVE PUTATIVE UNCHARACTERIZED PROTEINS

For a functional characterization of the stress responsive, so far putative proteins of unknown function in our analysis, we have used BLAST to find entries in phylogenetically related organisms by sequence similarity (see also **Table 3**).

RESULTS

PHYSIOLOGICAL RESPONSES TO SALT AND DROUGHT IN *M. TRUNCATULA*

Medicago truncatula was chosen in order to study the early stress acclimation under two N-nutritional conditions combined with two different environmental perturbations (four different stress treatments). The effect of reduced water availability on plant performance was analyzed in order to assess the degree of stress as alterations in water status in both nutritional phenotypes in *M. truncatula* (N-fed and N-fix; **Table 1**). The effect of drought stress was significant for most of the analyzed parameters depending on nutritional status. Water potential was significantly reduced during drought (potential dropped to −0.98 MPa and −1.06 MPa for N-fix and N-fed plants respectively), but not during salt stress. The PSII operating efficiency in terms of ($F'q/F'm$) was significantly decreased only in the leaves of drought stressed N-fed plants. Stomatal conductance was significantly reduced upon perturbation. In order to get a more holistic insight into the extent of plant acclimatory responses, significantly changing root metabolites were assessed (**Table 2**). Most of the significantly changed metabolites in roots did not change significantly in the leaves and vice versa. However, the degree of stress in terms of the fold change was more significant in the roots. Most of the responsive metabolites increased during stress. However, especially organic acids and a few amino acids of the leaves showed a decline in response to stress (**Table 3**).

DESCRIPTIVE ANALYSIS OF THE DETECTED PROTEINS AND METABOLITES

All identified proteins and metabolites were functionally categorized and visualized with Mapman (Thimm et al., 2004) using a new *M. truncatula* mapping file we created for UniProt data (**Figure 1**, see Mapman Mapping File for *M. truncatula* Proteins and Metabolites). Upon all identified proteins (643), "protein regulation" (20%), and "PS" (13%) are the dominant functional categories. In addition, the proteins assigned to the PS show highest relative abundance (spectral count per protein weight). Other categories are "redox," "amino acid," and "cell," each accounting for 5% of all identified proteins. Stress and signaling together reach 7% of all protein identifications followed by several other small categories (**Figure 1**). For the metabolites we found the major categories of the primary metabolism including amino acids "AA metabolism," the "TCA" cycle (organic acids), sugars "major COH metabolism," and "others."

QUANTITATIVE DATA MINING FOR SALT AND DROUGHT RESPONSIVE METABOLITES AND PROTEINS OF NUTRITIONAL *M. TRUNCATULA* PHENOTYPES

About 11% of all identified proteins (69 of 643) and 33% of all identified metabolites (15 of 45) changed significantly upon early stress acclimation ($p \leq 0.05$ and fold change ≥ 2; $n = 6$). GC-MS based metabolite profiling generally results in the identification of metabolites associated with the primary metabolism. Here, we found that most metabolites responding to stress were corresponding to the major sugar and amino acid metabolism and the TCA cycle. The protein categories with the highest percentage involved in stress response are: "PS," "amino acid," and "cell" with 12% each (**Table 3**). A small overlap of responsive compounds across the two stress treatments was observed (7 of 98, **Figure 2**). However, no analyzed compound was responsive during stress acclimation across all treatments. The Mg-chelatase subunit chlI (G7IK85), leucine, and malonate have been found to respond to three of the four different treatments. Of all the significantly altered levels of proteins and metabolites, only a particular subset responded to a specific treatment. Approaching the data from a different perspective, **Figure 2A** shows that more responsive compounds are shared between the salt than between the drought treated phenotypes. In contrast, a few specific response features were observed when dissecting the nutritional phenotypes (**Figure 2B**). Altogether, we found that the majority of significantly changed compounds of the nitrogen fertilized (N-fed) plants increased while the majority of significantly changed compounds of the N-fix plants decreased independent of the stress type (**Figures 3** and **4**).

We then compared the control levels of the proteins of the nutritional phenotypes with the response levels of perturbation (**Figure 4**). Interestingly, for the drought stressed plants, an approximation in protein levels between the two phenotypes has been observed. Thirteen responsive proteins of the N-fix plants show a higher control level compared to the N-fed controls. At the analyzed time of drought acclimation, those proteins decreased significantly, reaching the level of the N-fed plants (which have not changed during drought stress). Vice versa, control levels of six responsive proteins of the N-fed plants increased during drought, reaching unchanged control levels of the N-fix plants. This mechanism is less distinct for salt stress (**Figure 4**).

DISCUSSION

DEFINITION OF THE DEGREE OF STRESS AND THE CHALLENGE OF COMPARING DIFFERENT CONSTRAINTS

Salt and drought, two major environmental constraints have been compared. A moderate stress level was applied in order to study the early acclimation responses of *M. truncatula* growing under two different nutritional conditions. A biphasic growth inhibition model by saline conditions has been proposed earlier (Munns, 2002). During the first phase, growth inhibition is mainly governed by the decreased water availability due to higher solute concentrations in the soil solution, lowering soil water potential. If salt stress is prolonged ion toxicity effects gain importance in constraining plant metabolism and survival, described as the second phase, the salt stress specific phase (Sanchez et al., 2008a). To obtain similar early stress response levels for both stress types, keeping morphological parameters comparable, plants were harvested at the same

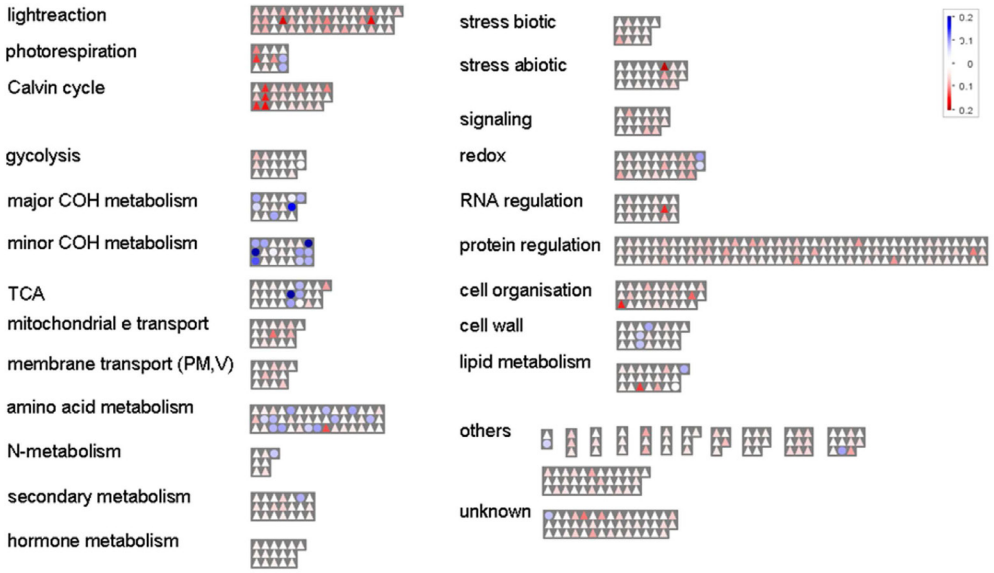


FIGURE 1 | Mapman overview. Functional distribution and relative abundance of the 643 shoot proteins (Log10 of the spectral count normalized by protein weight; FDR ≤ 0.01 %) and 45 metabolites (Log10 of peak area normalized by DW and IS) identified. Triangles = proteins; circles = metabolites. The strength of the color indicates the abundance of the compounds.

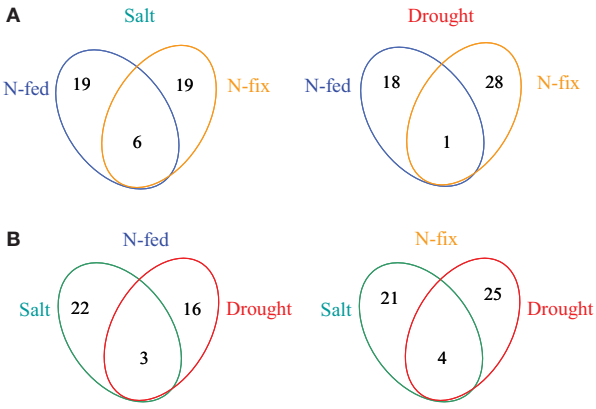


FIGURE 2 | Venn diagrams of the number of stress responsive proteins and metabolites. (A) overview of N-source depended changes for drought and salt separately; **(B)** overview of stress dependent overlap for N-fed (nitrogen fertilization) and N-fix (nitrogen fixation) separately.

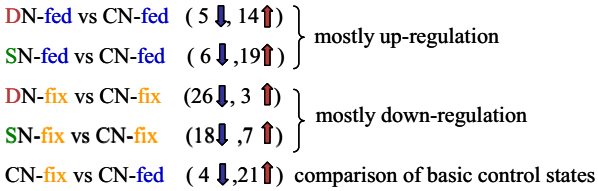


FIGURE 3 | Summary of the number of up- or down-regulated shoot compounds (proteins and metabolites) of the different treatments. Compounds, mostly up-regulated in N-fed and mostly down-regulated in N-fix plants during stress (independent on stress condition). D, drought; C, control; S, salt; N-fed, nitrogen fertilization; N-fix, nitrogen fixation.

age (duration-of-stress was 6 days). We compared the response of *M. truncatula* to water deficit resulting from a progressive mild drought treatment and a high initial 200 mM salt treatment. After 6 days of treatment, water-withholding and salt stress treatment resulted in stress responses. In order to assess the degree of stress at the plant level, several physiological parameters showing typical responses to decreasing water availability were analyzed as direct and indirect measures of plant water status (**Table 1**). All four stress treatments elicited acclimatory responses, as evidenced by significant decreases in stomatal conductance. Furthermore, our data also indicate that stress treatments had a low effect on photosynthesis. The PSII operating efficiency was neither affected by salt nor severely by drought (**Table 1**). This supports the onset of an early phase of stress acclimation. Drought experiments of soybean have shown that rates of photosynthesis were inhibited when leaf water potential dropped below −1.1 MPa (Boyer, 1970). This is consistent with our data; since photosynthesis was only affected in drought treated N-fed plants, when leaf water potential reached threshold. Salt and drought constraints are initially encountered at the root part of the plants. This might also contribute to the fact that in legumes, N-fixation is impaired in response to water deficit, before a decrease in photosynthetic rate can be observed (Durand et al., 1987; Djekoun and Planchon, 1991). As expected, when testing for some significant changes of metabolites in roots compared to the shoots (**Table 2**), the extent of stress-induced was more

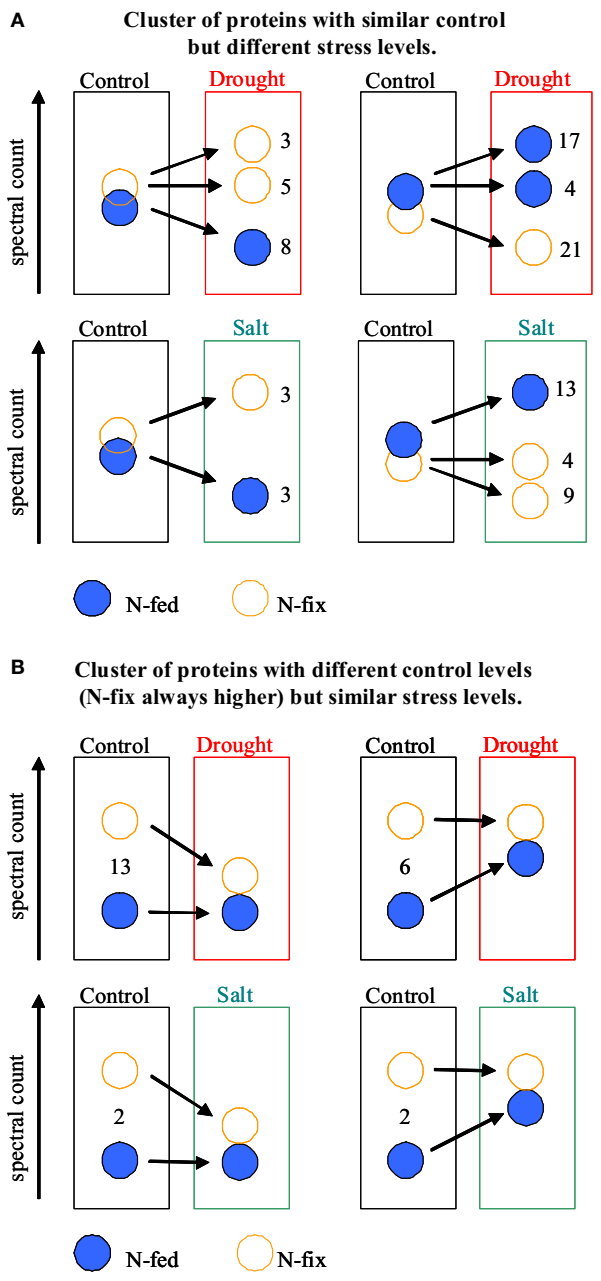


FIGURE 4 | Schematic overview of differentially up or down-regulated protein clusters depending on control level and nutritional status. (A) Similar control levels but different stress levels. **(B)** Different control levels but similar stress response levels. N-fed, nitrogen fertilization; N-fix, nitrogen fixation.

important in roots than in shoots. Some typical stress response marker such as proline, GABA and the polyols pinitol, ononitol, and myo-inositol (Vernon and Bohnert, 1992) were partially found to solely or more distinctly accumulate in the roots. Surprisingly, proline was only significantly increased in roots (∼10-fold) exposed to drought and shoots (∼twofold) exposed to salt. In leaves of N-fed plants it was found to even decrease. Since proline has been reported to increase during drought (Delauney and Verma, 1993) and other abiotic effects (Szabados and Savouré, 2010), the data suggest a moderate stress response where proline accumulation has not been fully established. This observation could result from the more pronounced stomatal closure in salt stressed than in drought stressed plants. As the water loss through stomata is lower, tissue WC, and water potential remain constant. Thus the degree of stress at the plant tissue level might not yet induce a substantial accumulation of osmoprotectants such as proline. The results for drought are in agreement with the data of Filippou et al. (2011), where, e.g., proline accumulation in *M. truncatula* leaves occurred only after 9 days of water-withholding, whereas in roots already after 3 days. The biological role of proline accumulation during stress is under extensive discussion (Verbruggen and Hermans, 2008). Drought stress experiments in *Lotus japonicus* strengthened the hypothesis that proline is necessary for the rehydration ability of the plants (Diaz et al., 2010). In agreement with our data, they showed that it does not reduce the rate of water loss. Interestingly, the counter-correlation of salt stressed plants showing no changes in leaf water potential suggests that this might be due to the more significant decrease in stomatal conductance, regulated by an increased ABA level. It was shown that the stomatal conductance was controlled by the root water potential when the ABA level of the xylem sap was increased (Tardieu et al., 1991). Thus our data demonstrate that salt and drought have impact on stomatal conductance but to a different degree, indicating higher stress response to salt than to drought. In contrast, water potential decreased significantly only during drought and more severely in N-fed compared to N-fix plants leading to the conclusion that drought stress has a stronger and thus earlier impact on water availability than salt. However, effects are still in a moderate range revealing an early stress response for both constraints. Altogether the physiological results lead to the following conclusions: (a) Indifferent to stress treatment and nutritional status stomatal conductance is an early stress response parameter; (b) Proline and the other observed, typical stress responsive metabolites as well as photosynthetic efficiency seem to be robust markers only for severe stress in leaves; (c) the root is the first place adjusting and controlling acclimation of stress; (d) all physiological parameters showing significant differences when comparing control to stressed groups, interestingly also showing significant differences between the two stress treatments; (e) in order to establish the highest possible similarity in plant water status between the two constraints, numerous salt concentrations and time points need to be assessed (and possibly additional parameters measured). However, an identical response seems very unlikely.

MOLECULAR STRESS ADJUSTMENTS DEPENDING ON THE NUTRITIONAL PHENOTYPE – CHAOS WITH SYSTEM?

Numerous studies on salt and/or drought stress in plants have been summarized recently (Pinheiro and Chaves, 2011; Krasensky and Jonak, 2012). Drought and salinity reduce soil water availability and induce common stress avoidance strategies such as shoot growth inhibition and lower stomatal conductance. However, there is not much overlap between the molecular data sets

published so far. This is probably due to the fact that experimental setups and application of stresses are very different and an appropriate definition of the degree of stress (in terms of experimental conditions as well as plant water status) for a better comparison is difficult and often missing (Jones, 2007). Another reason may be the differential steady-state of the plants such as growth state (Chaves et al., 2009) and nutritional status prior to stress exposure (Frechilla et al., 2000). The molecular data presented here shows that salt and drought stress share few common features in terms of changes in compound abundance (**Table 3** and **Figures 2A,B**). First of all, the significantly responding compounds appear randomly distributed across treatments and most functional categories of the metabolic network. This result is not surprising since stress effects seem not severe and plant metabolism has not yet been fully adjusted at the time of analyses. However, in agreement with other data (Sanchez et al., 2008a,b), we found a down-regulation of organic acids and an up-regulation of amino acids that seem typical for salt stress (**Table 3**). The results suggest that the TCA cycle is almost exclusively responding to early salt stress but not to drought. Within the N-fix phenotype of the salt stress group all five of the responsive metabolites of the TCA cycle were down-regulated.

Amino acids most significantly change in salt stressed and N-fed plants while most of the responsive sugars significantly changed in N-fix plants. The protein levels of the functional categories of amino acid and N-metabolism decreased, while the amino acids accumulated in response to stress. This trend could be observed within all stress treatments, except for drought stressed N-fed plants where this trend was inverted (**Table 3**). Possibly, increased amino acid levels are the cause for the down-regulation of proteins involved in amino acid synthesis and/or the consequence of protein degradation. Interestingly, this correlation has also previously been observed in root nodules of drought stressed *M. truncatula* (Larrainzar et al., 2009). They also found some glutamine sythetase isoforms decreasing during drought. However, while amino acid synthetases and asparagine aminotransferases seemed to play an important role during drought stress acclimation in nodules, *S*-adenosyl-L-methionine synthases (SAMS) seem to be more specifically involved in leaves. In addition, the SAMS isoforms seem only involved in early response to drought but not to salt stress. Furthermore, the four identified SAMS isoforms respond differently to drought. SAMS is a key enzyme, catalyzing the biosynthesis of SAM using methionine and ATP. It has been described that some of the SAMS genes were expressed constitutively, whereas others seemed specifically regulated by developmental and/or environmental factors depending on the requirement for SAM (Boerjan et al., 1994; Gómez-Gómez and Carrasco, 1998). SAM is a methyl donor, involved in many regulatory relevant processes on the transcript and protein level (Gómez-Gómez and Carrasco, 1998). However, further studies need to be conducted to unravel the regulatory function of the different SAMS isoforms during plant responses to water deficits.

Sugars are usually described to increase during osmotic stress adjustment (Clifford et al., 1998; Hummel et al., 2010). Surprisingly, glucose was decreasing in salt stressed plants. However, under drought stress glucose increased and other carbon metabolites increased as well. Interestingly, on the protein level, cell organization seemed most responsive in salt stressed N-fed plants. Distinctively, the two tubulin β chains (G7IAN2 and G7L5V0) and actin (G7JAX5) were found to be up-regulated. These components are involved in the dynamics of the cytoskeleton. Several studies in Arabidopsis have shown a relationship between the plant cytoskeleton and salt stress tolerance by the induction of actin filament assembly and bundle formation (Wang et al., 2010, 2011). This result may indicate a more specific response of salt stressed plants that are N-fertilized.

Besides malonate (down-regulation) and leucine (up-regulation), the metabolites found to respond in three out of the four treatments, Mg-chelatase subunit chlI (G7IK85) was also significantly changed (down-regulated) in both drought phenotypes as well as the salt stressed N-fix plants. The Mg-chelatase, composed of three different subunits, is the first enzyme involved in chlorophyll biosynthesis. It has been described to be involved in several stress-induced alterations. Dalal and Tripathy (2012) summarized the stress response of enzyme activity and on the protein and transcript level. They showed that Mg-chelatase protein abundance and gene expression are generally down-regulated during drought, salt, cold, and heat stress. A study on pea revealed that the Mg-chelatase chlI activity is redox regulated by chloroplast thioredoxins (Luo et al., 2012). Intriguingly, there are controversial discussions dealing with the Mg-chelatase subunit chlH. Initially it has been reported to act as an ABA receptor (Shen et al., 2006). However, Müller and Hansson (2009) reported that ABA had no effect on subunit chlH. Recently, Tsuzuki et al. (2011) presented evidence for the chlH subunit affecting ABA signaling of stomata guard cells but not acting as ABA receptor. These data strongly support that the Mg-chelatase is an important key player of chlorophyll degradation already during early stress response. The role of subunit chlI, however, needs to be studied in more detail.

Most other stress responsive compounds found, appear to be selectively distributed. However, we found interesting response patterns that might be explained by regulatory important mechanism: noticeably, the ratio between up and down-regulated compounds is grouping the nutritional phenotypes (**Figure 3**). The different molecular control levels of the two nutritional traits are leading to these response patterns. Starting with the comparison of the phenotypes, we found 25 of the stress responsive compounds also significantly distinguish N-fix from N-fed plants under control condition (**Table 3**). Here in general, protein and metabolite levels are higher in the control steady states of the N-fix plants compared to the N-fed plants. Furthermore, the ratio of up- vs. down-regulated proteins and metabolites during early stress response is generally higher in N-fed plants and vice versa the ratio of down-regulation higher in N-fed plants. Several distinct proteins seemed to change randomly coming from the same control state (**Figure 4A**). However, when analyzing the phenotypes after early stress adjustment, the proteomic data revealed a process of approximation to a similar molecular stress-steady-state (**Figure 4B**). Especially the protein response-pattern to drought aligned the way that proteins of the N-fix shoots of higher control level decreased to the level of N-fed shoots and vice versa. Taking these data together, there is evidence that the N-fed plants invest more energy in stress adjustment of protein levels than the N-fixing plants, where down-regulation of proteins is dominating the

process of acclimation. Interestingly, there is an overlap of six for the salt- compared to one stress responsive protein of the drought treatment (**Figure 2A**). Thus, salt stress response seems less dependent on the nutritional status than drought. Thus, we propose that (a) the initial molecular steady-state of the plants in terms of nutritional status seems pivotal for the downstream stress adjustment strategy; (b) during stress-acclimation-phase plants try to adjust their metabolic network to an approximate level (more significantly during the drought stress response); and (c) N-fix plants may need less energy for the stress adjustment than N-fed *M. truncatula* plants.

CONCLUSION

In the case of *M. truncatula*, our results suggest the following.

- Our drought stress treatment, led to a more pronounced water deficiency at the plant level than the salt stress treatment. This finding points to stress type specific acclimation strategies, especially stress avoidance mechanisms such as stomatal conductance. Either way, physiological, metabolomic, and proteomic data revealed significant differences in the degree and strategy of early drought, as compared to salt stress response, under identical growth conditions.
- Mg-chelatase subunit chlI, leucin, and malonate were significantly affected in three out of four stress treatments (two stress types, two nutritional conditions). Thus, they are likely robust early stress response markers. Further evaluation studies are necessary for confirmation.
- Proteomic adjustment seems low cost for N-fixing, as compared to N-fertilized plants, suggesting a potentially increased tolerance to stress. Whether this can be explained by symbiotic interaction itself or a more general kind of nutritional priming remains to be investigated further. However our results underline that the N-nutritional condition seems of crucial importance for plant stress acclimation.

ACKNOWLEDGMENTS

Vlora Mehmeti, Christiana Staudinger, and David Lyon were funded by the Austrian Fonds zur Förderung der wissenschaftlichen Forschung“ (FWF), P23441-B20.

REFERENCES

Aguirrezabal, L., Bouchier-Combaud, S., Radziejwoski, A., Dauzat, M., Cookson, S. J., and Granier, C. (2006). Plasticity to soil water deficit in *Arabidopsis thaliana*: dissection of leaf development into underlying growth dynamic and cellular variables reveals invisible phenotypes. *Plant Cell Environ.* 29, 2216–2227.

Aranjuelo, I., Molero, G., Erice, G., Christophe Avice, J., and Nogues, S. (2011). Plant physiology and proteomics reveals the leaf response to drought in alfalfa (*Medicago sativa* L.). *J. Exp. Bot.* 62, 111–123.

Baker, N. R. (2008). Chlorophyll fluorescence: a probe of photosynthesis in vivo. *Annu. Rev. Plant Biol.* 59, 89–113.

Bianco, C., and Defez, R. (2009). *Medicago truncatula* improves salt tolerance when nodulated by an indole-3-acetic acid-overproducing *Sinorhizobium meliloti* strain. *J. Exp. Bot.* 60, 3097–3107.

Boerjan, W., Bauw, G., Van Montagu, M., and Inzé, D. (1994). Distinct phenotypes generated by overexpression and suppression of S-adenosyl-L-methionine synthetase reveal developmental patterns of gene silencing in tobacco. *Plant Cell* 6, 1401–1414.

Boyer, J. S. (1970). Differing sensitivity of photosynthesis to low leaf water potentials in corn and soybean. *Plant Physiol.* 46, 236–239.

Chaves, M. M., Flexas, J., and Pinheiro, C. (2009). Photosynthesis under drought and salt stress: regulation mechanisms from whole plant to cell. *Ann. Bot.* 103, 551–560.

Clifford, S., Arndt, S., Corlett, J., Joshi, S., Sankhla, N., Popp, M., et al. (1998). The role of solute accumulation, osmotic adjustment and changes in cell wall elasticity in drought tolerance in *Ziziphus mauritania* (Lamk.). *J. Exp. Bot.* 49, 967–977.

Dalal, V. K., and Tripathy, B. C. (2012). Modulation of chlorophyll biosynthesis by water stress in rice seedlings during chloroplast biogenesis∗. *Plant Cell Environ.* 35, 1685–1703.

Delauney, A. J., and Verma, D. P. S. (1993). Proline biosynthesis and osmoregulation in plants. *Plant J.* 4, 215–223.

Diaz, P., Betti, M., Sanchez, D. H., Udvardi, M. K., Monza, J., and Marquez, A. J. (2010). Deficiency in plastidic glutamine synthetase alters proline metabolism and transcriptomic response in *Lotus japonicus* under drought stress. *New Phytol.* 188, 1001–1013.

Djekoun, C., and Planchon, A. (1991). Water status effect on dinitrogen fixation and photosynthesis in soybean. *Agron. J.* 83, 316–322.

Durand, J. L., Sheehy, E. J., and Minchin, F. R. (1987). Nitrogenase activity, photosynthesis and nodule water potential in soyabean plants experiencing water deprivation. *J. Exp. Bot.* 38, 311–321.

Evans, H. J. (1981). “Symbiotic nitrogen fixation in legume nodules,” in *Research Experiences in Plant Physiology*, ed. T. C. Moore (New York: Springer-Verlag), 294–310.

Filippou, P., Antoniou, C., and Fotopoulos, V. (2011). Effect of drought and rewatering on the cellular status and antioxidant response of *Medicago truncatula* plants. *Plant Signal. Behav.* 6, 270–277.

Frechilla, S., Gonzalez, E. M., Royuela, M., Minchin, F. R., Apaicio-Tejo, P. M., and Arrese-Igor, C. (2000). Source of nitrogen nutrition (nirogen fixation or nitrate assimilation) is a major factor involved in pea response to moderate water stress. *J. Plant Physiol.* 157, 609–617.

Gibon, Y., Blasing, O. E., Palacios-Rojas, N., Pankovic, D., Hendriks, J. H., Fisahn, J., et al. (2004). Adjustment of diurnal starch turnover to short days: depletion of sugar during the night leads to a temporary inhibition of carbohydrate utilization, accumulation of sugars and post-translational activation of ADP-glucose pyrophosphorylase in the following the light period. *Plant J.* 39, 847–862.

Gibon, Y., Usadel, B., Blaesing, O. E., Kamlage, B., Hoehne, M., Trethewey, R., et al. (2006). Integration of metabolite with transcript and enzyme activity profiling during diurnal cycles in Arabidopsis rosettes. *Plant Cell Environ.* 32, 859–874.

Gómez-Gómez, L., and Carrasco, P. (1998). Differential expression of the S-adenosyl-L-methionine synthase genes during pea development. *Plant Physiol.* 117, 397–405.

Hajduch, M., Hearne, L. B., Miernyk, J. A., Casteel, J. E., Joshi, T., Agrawal, G. K., et al. (2010). Systems analysis of seed filling in *Arabidopsis*: using general linear modeling to assess concordance of transcript and protein expression. *Plant Physiol.* 152, 2078–2087.

Hoehenwarter, W., and Wienkoop, S. (2010). Spectral counting robust on high mass accuracy mass spectrometers. *Rapid Commun. Mass Spectrom.* 24, 3609–3614.

Hummel, I., Pantin, F., Sulpice, R., Piques, M., Rolland, G., Dauzat, M., et al. (2010). *Arabidopsis* plants acclimate to water deficit at low cost through changes of carbon usage: an integrated perspective using growth, metabolite, enzyme, and gene expression analysis. *Plant Physiol.* 154, 357–372.

IPCC. (2007). *Fourth Asessment Report: Synthesis Report.* Available at: http://ipcc.ch/publications_and_data/ar4/syr/en/contents.html.

Jogaiah, S., Govind, S. R., and Tran, L.-S. P. (2012). Systems biology-based approaches toward understanding drought tolerance in food crops. *Crit. Rev. Biotechnol.* 1–17. doi:10.3109/07388551.2012.659174

Jones, H. G. (2007). Monitoring plant and soil water status: established and novel methods revisited and their relevance to studies of drought tolerance. *J. Exp. Bot.* 58, 119–130.

Kang, Y., Han, Y., Torres-Jerez, I., Wang, M., Tang, Y., Monteros, M., et al. (2011). System responses to long-term drought and re-watering of two contrasting alfalfa varieties. *Plant J.* 68, 871–889.

Krasensky, J., and Jonak, C. (2012). Drought, salt, and temperature stress-induced metabolic rearrangements and regulatory networks. *J. Exp. Bot.* 63, 1593–608.

Larrainzar, E., Wienkoop, S., Scherling, C., Kempa, S., Ladrera, R., Arrese-Igor, C., et al. (2009). Carbon metabolism and bacteroid functioning are involved in the regulation of nitrogen fixation in *Medicago truncatula* under drought and recovery. *Mol. Plant Microbe Interact.* 22, 1565–1576.

Larrainzar, E., Wienkoop, S., Weckwerth, W., Ladrera, R., Arrese-Igor, C., and Gonzalez, E. M. (2007). Medicago truncatula root nodule proteome analysis reveals differential plant and bacteroid responses to drought stress. *Plant Physiol.* 144, 1495–1507.

Levitt, J. (1980). "Responses of plants to environmental stresses," in *Water, Radiation, Salt and Others Stresses.* Vol. 2. New York: Academic Press.

Lopez, M., Herrera-Cervera, J. A., Iribarne, C., Tejera, N. A., and Lluch, C. (2008). Growth and nitrogen fixation in *Lotus japonicus* and *Medicago truncatula* under NaCl stress: nodule carbon metabolism. *J. Plant Physiol.* 165, 641–650.

Luo, T., Fan, T., Liu, Y., Rothbart, M., Yu, J., Zhou, S., et al. (2012). Thioredoxin redox regulates ATPase activity of Mg-chelatase chlI subunit and modulates redox-mediated signaling in tetrapyrrole biosynthesis and homeostasis of reactive oxygen species in pea plants. *Plant Physiol.* 159, 118–130.

Miransari, M., and Smith, D. L. (2009). Alleviating salt stress on soybean (*Glycine max* (L.) Merr.) – *Bradyrhizobium japonicum* symbiosis, using signal molecule genistein. *Eur. J. Soil Biol.* 45, 146–152.

Müller, A. H., and Hansson, M. (2009). The barley magnesium chelatase 150-kd subunit is not an abscisic acid receptor. *Plant Physiol.* 150, 157–166.

Munns, R. (2002). Comparative physiology of salt and water stress. *Plant Cell Environ.* 25, 239–250.

Naya, L., Ladrera, R., Ramos, J., González, E. M., Arrese-Igor, C., Minchin, F. R., et al. (2007). The response of carbon metabolism and antioxidant defenses of alfalfa nodules to drought stress and to the subsequent recovery of plants. *Plant Physiol.* 144, 1104–1114.

Noreen, Z., and Ashraf, M. (2009). Assessment of variation in antioxidative defense system in salt-treated pea (*Pisum sativum*) cultivars and its putative use as salinity tolerance markers. *J. Plant Physiol.* 166, 1764–1774.

Pinheiro, C., and Chaves, M. M. (2011). Photosynthesis and drought: can we make metabolic connections from available data? *J. Exp. Bot.* 62, 869–882.

Rispail, N., Kaló, P., Kiss, G. B., Ellis, T. H. N., Gallardo, K., Thompson, R. D., et al. (2010). Model legumes contribute to faba bean breeding. *Field Crops Res.* 115, 253–269.

Salah, I., Albacete, A., Martínez Andújar, C., Haouala, R., Labidi, N., Zribi, F., et al. (2009). Response of nitrogen fixation in relation to nodule carbohydrate metabolism in Medicago ciliaris lines subjected to salt stress. *J. Plant Physiol.* 166, 477–488.

Sanchez, D. H., Lippold, F., Redestig, H., Hannah, M. A., Erban, A., Krämer, U., et al. (2008a). Integrative functional genomics of salt acclimatization in the model legume *Lotus japonicus*. *Plant J.* 53, 973–987.

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

Shen, Y.-Y., Wang, X.-F., Wu, F.-Q., Du, S.-Y., Cao, Z., Shang, Y., et al. (2006). The Mg-chelatase H subunit is an abscisic acid receptor. *Nature* 443, 823–826.

Szabados, L., and Savouré, A. (2010). Proline: a multifunctional amino acid. *Trends Plant Sci.* 15, 89–97.

Tardieu, F., Katerji, N., Bethenod, O., Zhang, J., and Davies, W. J. (1991). Maize stomatal conductance in the field: its relationship with soil and plant water potentials, mechanical constraints and ABA concentrations in the xylem sap. *Plant Cell Environ.* 53, 205–214.

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., et al. (2004). Mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* 37, 914–939.

Tsuzuki, T., Takahashi, K., Inoue, S., Okigaki, Y., Tomiyama, M., Hossain, M. A., et al. (2011). Mg-chelatase H subunit affects ABA signaling in stomatal guard cells, but is not an ABA receptor in Arabidopsis thaliana. *J. Plant Res.* 124, 527–538.

Verbruggen, N., and Hermans, C. (2008). Proline accumulation in plants: a review. *Amino Acids* 35, 753–759.

Vernon, D. M., and Bohnert, H. J. (1992). A novel methyl transferase induced by osmotic stress in the facultative halophyte Mesembryanthemum crystallinum. *EMBO J.* 11, 2077–2085.

Verslues, P. E., Agarwal, M., Katiyar-Agarwal, S., Zhu, J., and Zhu, J.-K. (2006). Methods and concepts in quantifying resistance to drought, salt and freezing, abiotic stresses that affect plant water status. *Plant J.* 45, 523–539.

Wang, C., Zhang, L., Yuan, M., Ge, Y., Liu, Y., Fan, J., et al. (2010). The microfilament cytoskeleton plays a vital role in salt and osmotic stress tolerance in Arabidopsis. *Plant Biol. (Stuttg.)* 12, 70–78.

Wang, C., Zhang, L.-J., and Huang, R.-D. (2011). Cytoskeleton and plant salt stress tolerance. *Plant Signal. Behav.* 6, 29–31.

Wienkoop, S. (2011). "Proteomics and metabolomics for systems biology in legumes," in *Cool Season Grain Legumes,* eds M. Perez de la Vega, A. M. Torres, J. I. Cubero, and C. Kole (New Hampshire: Science Publishers), 303–314.

Wienkoop, S., Staudinger, C., Hoehenwarter, W., Weckwerth, W., and Egelhofer, V. (2012). ProMEX – a mass spectral reference database for plant proteomics. *Front. Plant Sci.* 3:125. doi:10.3389/fpls.2012.00125

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 11 October 2012; paper pending published: 18 October 2012; accepted: 30 November 2012; published online: 21 December 2012.

Citation: Staudinger C, Mehmeti V, Turetschek R, Lyon D, Egelhofer V and Wienkoop S (2012) Possible role of nutritional priming for early salt and drought stress responses in Medicago truncatula. Front. Plant Sci. ***3****:285. doi: 10.3389/fpls.2012.00285*

This article was submitted to Frontiers in Plant Proteomics, a specialty of Frontiers in Plant Science.

Mass Western for Absolute Quantification of Target Proteins and Considerations About the Instrument of Choice

The Western Blot is a well-known and established technique for the specific detection and quantification of target proteins. Despite numerous successful applications, this antibody-based method is potentially prone to unspecific binding and cross-reactivity. In analogy to the famous Western Blot, the Mass Western, an alternative Mass-Spectrometry-based technique for the absolute quantification of target proteins, was developed (Lehmann et al. 2008). This method relies on external and internal synthetic standard peptides, labeled with stable isotopes. Due to the coupling of liquid chromatography to mass spectrometry, the Mass Western additionally enables multiplexing. The study highlights the importance of using specialized instruments, e.g. Triple Quadrupoles, and methods, e.g. Selective Reaction Monitoring (SMR), in order to maximize quantitative accuracy and the Linear Dynamic Range.

Declaration of authorship

The results of this chapter are presented in the form of a manuscript published as a chapter in the book series „Methods in Molecular Biology“. The work presented in the following manuscript is largely my own. I have conducted all experiments, analyzed the data and written the manuscript.

Additional remarks

In order to increase the accuracy of stoichiometry, a modified approach of the “MassWestern” was developed in analogy to Holzmann et al. 2009. When determining the exact stoichiometry of e.g. two proteins, it is imperative to spike the sample with equal amounts of respective internal heavy peptide standards. Therefore, cross-concatenated standard peptides were designed. A concatenated peptide consists of a (e.g. tryptic) peptide from each of the two proteins which could be cleaved using a protease, thereby ensuring identical amounts of standard peptides for both proteins (**Recuenco-Munoz et al. 2014** *accepted*). I assisted in the design of this method.

Published manuscript

Chapter 15

Mass Western for Absolute Quantification of Target Proteins and Considerations About the Instrument of Choice

David Lyon, Wolfram Weckwerth, and Stefanie Wienkoop

Abstract

The Mass Western describes the absolute quantification of proteins based on stable isotope labeled integral standard peptides and liquid chromatography coupled selective reaction monitoring triple quadrupole mass spectrometry (LC-SRM/MS). Here, we present a detailed workflow including tips and we discuss advantages and disadvantages of using different types of MS for absolute quantification.

Key words Mass Western, SRM, Heavy peptide internal standard, Absolute quantification, Triple quadrupole, Orbitrap

Abbreviations

SID	Stable isotope dilution
SRM	Selective reaction monitoring
PTM	Posttranslational modification
QqQ	Triple quadrupole
LC-SRM/MS	Liquid chromatography coupled selective reaction monitoring triple quadrupole mass spectrometry
HP	Heavy peptide ($^{15}N^{13}C$ labeled synthetic standard)
LP	Light peptide (native peptide, no labeling)
HCD	High energy collision induced dissociation
CE	Collision energy
LDR	Linear dynamic range
LOD	Limit of detection
LOQ	Limit of quantification

1 Introduction

Absolute protein quantification using stable isotope dilution (SID) in conjunction with liquid chromatography (LC) and selective reaction monitoring (SRM) triple quadrupole (QqQ) mass

Jesus V. Jorrin-Novo et al. (eds.), *Plant Proteomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1072,
DOI 10.1007/978-1-62703-631-3_15,

spectrometry (MS) has been successfully applied to highly complex crude proteomics samples [1, 2]. In contrast to relative quantification, absolute quantitative data result in absolute concentration levels. Thus besides comparison of different experimental treatments, absolute quantification enables the analysis of protein stoichiometry within a sample, the differentiation of isoforms as well as the comparison of inter-experimental conditions, such as different species. Additionally, it leads to highly verifiable data. In analogy to the well-known Western Blot, Lehmann et al. [3] first coined the suitable term Mass Western. While for Western Blot analysis synthetic peptides can be used for the synthesis of antibodies, they can be directly applied for the sensitive and targeted detection and absolute quantification using the Mass Western. It is the integration of stable isotope labeled synthetic peptides in combination with gel based or gel-free LC-SRM/MS. A theoretical and an experimental approach to set up the Mass Western can be distinguished (Fig. 1). Both approaches start out by defining at least one protein of interest. The theoretical approach continues with in silico digestion of proteins and prediction of proteotypic peptides (understood as a unique amino acid sequence of a peptide, unambiguously identifying a specific protein of interest within a given proteome) for the target proteins (in reference to the proteome of the sample). Subsequently, the

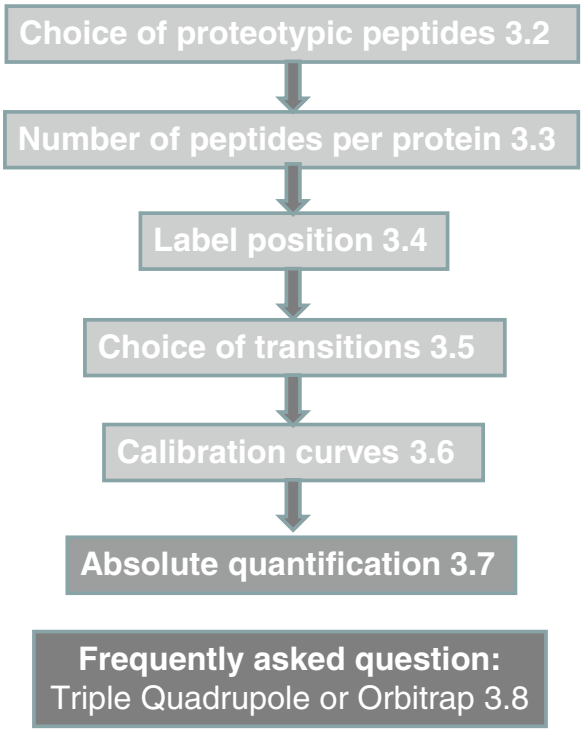


Fig. 1 Workflow diagram of the steps for Mass Western as described in the text

chosen peptides are synthesized, including (at least) one stable isotope labeled amino acid (being highly enriched with ^{15}N and ^{13}C) per peptide. In contrast to the latter, within the experimental approach, proteotypic peptides are chosen through data mining of shotgun-proteomics experiments, mass spectral reference databases such as the plant proteomic spectral library ProMEX [4], as well as protein fractionation, enrichment, and subsequent mass spectral analysis. The ensuing synthesis of stable isotope labeled peptides is identical to the theoretical approach, as well as the rest of the procedure. LC-SRM/MS method development follows and completes the workflow of the Mass Western approach. The detailed workflow is presented in the specific context of quantification capacity of a QqQ, here a TSQ (Thermo Triple Stage Quadrupole Vantage), compared to a Linear Trap Quadrupole-Orbitrap (LTQ-Orbitrap).

2 Materials

2.1 Software

We recommend the use of Skyline ([5], https://brendanx-uw1.gs.washington.edu/) for method development and data analysis. Proteolytic cleavage probability, hydrophobicity calculation, and PTM prediction are part of many useful features available at Expasy Tools (http://ca.expasy.org/). Skyline in conjunction with SRM collider can be used for the predition of proteotypic peptides and to find interferences in a given background proteome [6].

2.2 Mass Spectrometer

TSQ (Thermo Triple Stage Quadrupole, Vantage).

LTQ-Orbitrap (Thermo Linear Trap Quadrupole-Orbitrap XL).

3 Methods

3.1 General Overview

The heavy synthetic standard peptide (containing one amino acid labeled with ^{15}N and ^{13}C; HP) and its native counterpart, the light peptide (with no artificially introduced label; LP), share identical physicochemical properties. Retention time, ionization efficiency, and fragmentation properties of a HP-LP pair are assumed to be identical. The HP is used to tune the mass spectrometer and to create an external calibration curve. The peak area of the LP is set into the calibration curve to calculate the absolute quantity. When measuring samples for quantification, equal amounts of HP are spiked into each sample, as internal standards. The internal standard serves as a quality control. The retention time, peak shape, and relative intensities of individual transitions (*see* **Notes 1** and **2**) of the LP are compared to the HP. This gives the experimenter great confidence concerning the accuracy of the data used for

quantification. The incorporation of a HP internal standard is the only approach that enables absolute quantification. The peak area of the HP can be used to normalize data across multiple samples (by dividing each individual peak area of a sample by the peak area of the HP of that particular sample). Since a known amount of HP is spiked into the sample, the peak area of the HP can be compared to the mean and standard deviation of its calibration curve at the given concentration. Sample handling and or technical aberrations can thereby be detected. Subsequently, the measurement can be repeated or a correction factor applied. The discrete steps to set up a Mass Western experiment are described as follows.

3.2 Choice of Proteotypic Peptides

When setting up a Mass Western using the experimental approach, all identified peptides of the target protein are considered as putative candidates. The amino acid sequences of each peptide can be BLASTed against the proteome of the organism, in order to distinguish proteotypic peptides (other tools such as Skylineare available as well). Sequences containing missed cleavage sites should be avoided, due to the potential occurrence of proteolytic peptides consisting of partial sequences of the targeted peptide. Subsequently, the total amount of the targeted peptide will decrease and thus the accuracy and sensitivity of the experiment. Methionine should be avoided, since it is prone to oxidation. Cysteine residues tend to build disulfide bridges and can result in dimers or cyclization, though alkylation can counteract this problem. Generally, reproducibly identifiable proteotypic peptides are very promising candidates for a Mass Western experiment. Peptides with the highest ionization efficiency (and thus the most intense signal), and best chromatographic peak shape should be chosen.

When approaching a Mass Western experiment theoretically, the putative list of proteolytic peptides, resulting from in silico digestion of the protein of interest, needs to be reduced to proteotypic candidates. This can be performed by BLAST, as previously mentioned. Further selection criteria are as follows: Peptide length should preferably be between eight and twenty amino acids. Proteolytic efficiency should be as high as possible and can be estimated by tools such as Peptide Cutter (http://web.expasy.org/peptide_cutter/). N-terminal Glutamine has a tendency to form cyclic peptides. Low coupling efficiency due to the hydrophobic and steric characteristics of Tryptophan can pose problems when synthesizing peptides. Generally hydrophobicity and thus solubility of the peptides needs to be considered not only concerning synthesis, but also concerning extraction, digestion and resuspension of proteins/peptides. Sequences containing Proline can produce multiple chromatographic peaks due to enantiomers, but are also known to easily and prominently fragment in MS/MS experiments. Furthermore, potential Post Translational Modifications (PTMs) should be considered. To the best of the author's

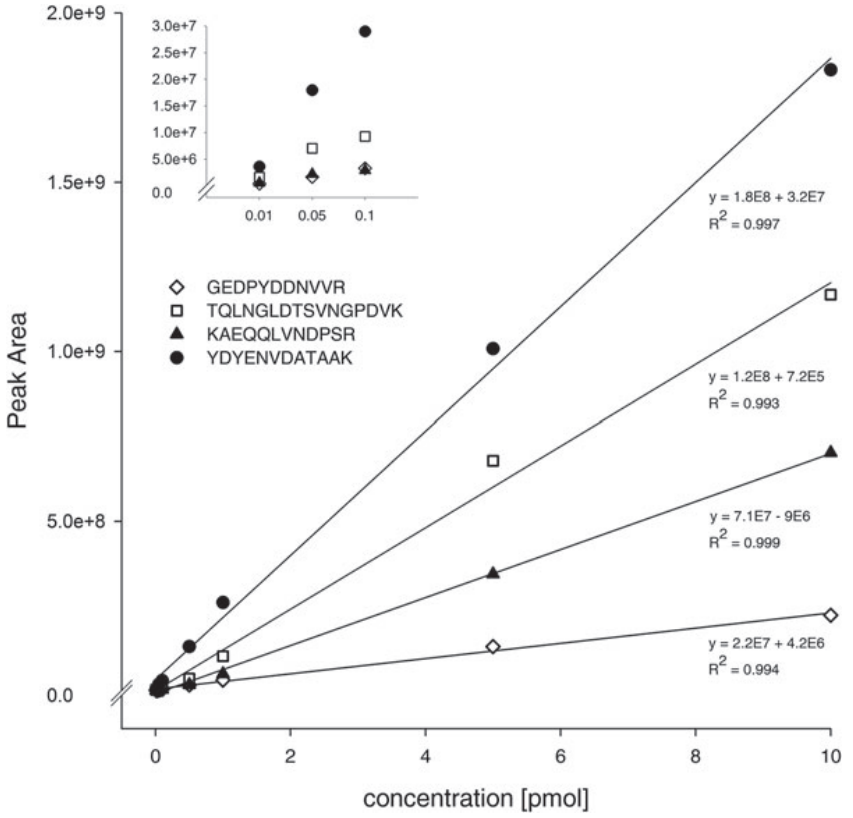


Fig. 2 Concentration versus Peak Area. A dilution series of four HPs was measured on a Thermo TSQ Vantage QqQ-MS. The ionization efficiency, and thus the signal response, is dependent on the amino acid sequence, as can be seen by the differing slopes of the linear regression lines

knowledge, no robust method exists to predict the signal intensity (ionization efficiency) of a given peptide sequence (Fig. 2). This can potentially be a major drawback, since the ionization efficiency can only be determined experimentally. Subsequently, an otherwise suitable peptide can exhibit very low signal intensities, and thus pose analytical difficulties. Digestion efficiency, peptide solubility, ionization efficiency, and matrix effects are accounted for within the experimental approach, in contrast to the theoretical approach.

3.3 Number of Peptides per Protein

Peptides should be chosen from different regions (e.g., the middle of and close to the C-terminus) of the protein. The more peptides per protein measured, leading to identical results, the more certainty about the quality of the results can be assumed. Principally, one peptide per protein should be sufficient for quantification. Nonetheless, at least two peptides per protein should be used for the Mass Western, if possible, due to variability in proteolytic efficiency and differing recovery rates. Necessary validation can thus be performed by comparing individual peptide results for a given protein.

3.4 Label Position

No matter the label position, the precursor *m/z* values of a HP-LP pair (isotopologues) will differ, dependent on the labeled amino acid. In principal, it doesn't matter which amino acid is chosen, if ^{15}N and ^{13}C labeling is applied, since delta M (the mass shift) will be at least 4 Da (for Alanine) which can easily be separated by modern mass spectrometers. However, it is recommended choosing the C-terminal amino acid of the synthetic peptide to be labeled with ^{15}N and ^{13}C. Y-ions are predominantly used for SRM transitions in QqQ, due to their abundance and intensity but b-ions may be chosen as well. Subsequently, the product *m/z* values of any Y-ion will also differ for a HP-LP pair, resulting in highly selective transition pairs. In theory, the selection of one transition is enough. However, the more transitions per peptide selected, the higher the certainty. More transitions may also increase signal intensity while improving sensitivity, but increase the duty cycle at the same time.

3.5 Choice of Transitions

It is imperative to use identical types of transitions for a HP and its LP counterpart (The charge state of the precursor of the HP needs to be the same as for the LP. If a singly charged Y_3-type-ion is chosen for the HP, the same needs to be chosen for the LP.) A similar ion fragmentation intensity pattern has been described comparing HCD collision with QqQ fragmentation [7]. However, tuning of the mass spectrometers collision energy (CE) by direct infusion (flow injection) of the synthetic peptides enables maximum sensitivity for specific transitions. Atleast for some mass spectrometers a semiautomatic ramping of CE and selection of the most intense transitions is possible [8]. The occurrence of other parameters such as Declustering Potential, S-lens, Collision Exit Potential, Ion Transfer Capillary Offset Voltage, etc. is vendor specific. Fine tuning of all possible parameters guarantees the highest possible sensitivity of the experiment. Y-ions N-terminal to Proline frequently result in high signal intensities, and are therefore preferably chosen. Selecting the most abundant fragments aims at maximum sensitivity of the assay (*see* **Note 1**). High selectivity can usually be achieved by choosing transitions whose product *m/z* values are higher than their precursor *m/z* value (possible due to for example precursor charge state 2, product charge state 1). A combination of sensitive and selective transitions often results in the most effective experimental setup. In general two transitions per peptide are sufficient for sensitivity and selectivity since the HP internal standard includes the retention time as additional confidence identification parameter (*see* **Note 2**).

3.6 Calibration Curves

Comparing HP and LP peak areas to deduce quantitative results (single point calibration) is not recommended due to the well known fact that peptide ionization efficiency varies significantly (Fig. 2). Except for nearly identical peak areas, quantitation will not be as

accurate as with external calibration curves. An external calibration curve is recorded as follows. A dilution series of the synthetic peptides is measured using optimized settings for the instrument given. The linear dynamic range (LDR), the limit of detection (LOD), and the limit of quantification (LOQ) can thereby be estimated. If the standard peptides are spiked into a complex matrix (reflecting the nature of the sample to be analyzed), the LDR, LOD, and LOQ in matrix can be assessed more precisely concerning the performance of the assay. Accurate quantification can only be achieved within the linear dynamic range. For each HP a linear regression is calculated for data within its linear dynamic range.

3.7 Absolute Quantification

When measuring samples for quantification, the amount of HP added to the individual samples should be kept constant. The heavy standard peptide (HP) is introduced to the sample at the earliest possible point of time. The signal intensity of the HP standard should be as close as possible to that of the native target. The peak area integral of the HP can be used to normalize data across multiple samples. The LP peak area can thus be adjusted and set into the regression equation of the external calibration curve to calculate a quantitative value. Best results are expected if the values are in the middle of the linear regression.

3.8 Frequently Asked Question: Triple Quadrupole or Orbitrap

The SRM approach can generally be executed on different types of MS. At present, for absolute quantification triple quadrupole mass spectrometers (QqQ) are routinely being used due to a wide linear dynamic range, excellent sensitivity and selectivity, as well as acquisition speed. In contrast, due to the very long duty cycles and incomparable sensitivity, quantification based on SRM on the hybrid MS instrument LTQ-Orbitrap-XL is not recommended. The LTQ-Orbitrap-XL MS features high mass accuracy and resolution, and is an excellent tool for discovery experiments (such as shotgun proteomics), and relative quantification [7, 8]. The latter MS instrument has also successfully been used for LC-MS based quantification with the drawback of a reduced linear dynamic range compared to QqQ LC-SRM/MS. Recent articles state that the Orbitrap-Exactive is "equal or better" than QqQ if full scan acquisition based quantification is applied [9, 10] (*see* **Note 3**). However, at present the use of QqQ holds advantages compared to other types of instruments such as a Linear Trap Quadrupole-Orbitrap (LTQ-Orbitrap). Figure 3 illustrates the differing linear dynamic range of the aforementioned instruments for three distinct standard HP. The two MS instruments are compared for applicability of absolute quantification (methodological details can be found at http://www.univie.ac.at/mosys/publications.html). The figure shows that even though LTQ-Orbitrap MS signal intensities in data dependent mode appears higher, the linear dynamic range is bigger for the QqQ Triple Stage Quadrupole (TSQ) MS (*see* **Note 4**).

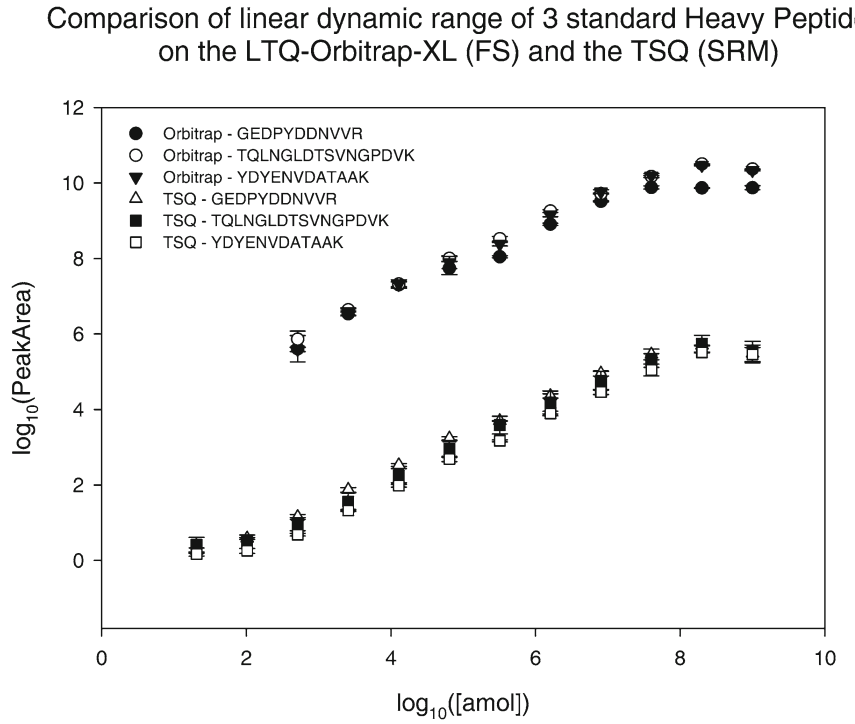


Fig. 3 Concentration versus Peak Area. Three Stable isotope labeled synthetic peptides was subjected to a dilution series, and measured with the LTQ-Orbitrap-XL, in data-dependent acquisition mode, and the TSQ-Vantage, in SRM-mode. Data indistinguishable from noise or simply not present are not depicted, and thus the "missing" data points in the Orbitrap function at lower concentration levels. The linear and nonlinear parts of the two functions are distinguished by the flattening of the respective functions. In order to attain an optimal coefficient of determination (R^2) similar for both instruments, the linear regression was calculated with a subset of the data points. The resulting regression lines are depicted in the graph. The upper but especially the lower limit of the TSQ regression line clearly extends to lower concentration levels, indicating heightened sensitivity over a wider linear dynamic range

This is due to the fact that low concentration signals are detectable in TSQ while the signal-to-noise ratio is already too low in the LTQ-Orbitrap. Thus, when aiming to quantify very low abundant targets in complex matrices, QqQs are the instruments of choice. Recent developments in the field of mass spectrometry include optimized Ion Optics (enabling larger amounts of ions to be focused in a shorter amount of time), data acquisition speed, data acquisition range, sensitivity [10–13] as well as the emergence of novel DIA (Data Independent Acquisition) instruments, methods and software [14]. Thus high resolution and high mass accuracy instruments, such as the latest generation of the Orbitrap, are increasingly competitive to QqQs. Additionally data produced by these instruments is not restricted to quantification as with SRM. These and other studies show that at present the use of an LTQ-Orbitrap MS may be an alternative to QqQ if the sensitivity is not limited [10]. For improved sensitivity of low abundant proteins using FullScan high resolution high mass accuracy MS some instrument adjustments may be useful (*see* **Note 5**).

4 Notes

1. When multiple SRM transitions for a given substance are chosen, the individual signal intensities are added to a single signal. This should also be considered when comparing high mass accuracy and resolution FullScan methods with SRM/MRM methods as well as the generation of the mass spectrometer, instrument parameters, and complexity of the underlying sample matrix.
2. False positive signals can be disregarded and do not necessarily distort quantification, since true signals can reliably be identified by comparing the retention time, peak shape, and relative transition order of a HP-LP-pair.
3. Since metabolites most frequently are detected with a single charge, the precursor is larger than its product m/z value, and SRM transitions whose precursor is smaller than its product m/z value can't be selected.
4. The systematic shift of Peak Area values (arbitrary units), between the two instruments does not imply any qualitative or quantitative difference.
5. Reducing the scan range (Full Scan or Selected Ion Monitoring) elevates the signal to noise ratio and improves sensitivity, however, useful information may thus become unavailable [15]. Elevating target Automatic Gain Control (AGC) values can improve the dynamic range, due to elevated sensitivity for low concentrations, but can lead to negative space charge effects at high concentrations [12].

Acknowledgments

We thank Thomas Naegele for useful comments and discussions.

References

1. Desiderio DM, Kai M (1983) Preparation of stable isotope-incorporated peptide internal standards for field desorption mass spectrometry quantification of peptides in biologic tissue. Biomed Mass Spectrom 10:471–479
2. Wienkoop S, Weckwerth W (2006) Relative and absolute quantitative shotgun proteomics: targeting low-abundance proteins in *Arabidopsis thaliana*. J Exp Bot 57:1529–1535
3. Lehmann U, Wienkoop S, Tschoep H et al (2008) If the antibody fails—a mass western approach. Plant J 55:1039–1046
4. Wienkoop S, Staudinger C, Hoehenwarter W et al (2012) ProMEX—a mass spectral reference database for plant proteomics. Front Plant Sci 3:125
5. MacLean B, Tomazela DM, Shulman N et al (2010) Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics 26:966–978
6. Röst H, Malmström L, Aebersold R (2012) A Computational Tool to Detect and Avoid Redundancy in Selected Reaction Monitoring. Mol Cell Proteomics : MCP 11 (8) (August): 540–549. doi:10.1074/mcp.M111.013045. http://www.ncbi.nlm.nih.gov/pubmed/22535207

7. Makarov A, Denisov E, Kholomeev A et al (2006) Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. Anal Chem 78:2113–2120
8. Olsen JV, Godoy LMFD, Li G et al (2005) Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. Mol Cell Proteomics 4:2010–2021
9. Gallien S, Duriez E, Crone C et al (2012) Targeted proteomic quantification on quadrupole-orbitrap mass spectrometer. Mol Cell Proteomics mcp.O112.019802
10. Henry H, Sobhi HR, Scheibner O et al (2012) Comparison between a high-resolution single-stage Orbitrap and a triple quadrupole mass spectrometer for quantitative analyses of drugs. Rapid Comm Mass Spectrom 26:499–509
11. Bateman KP, Kellmann M, Muenster H et al (2009) Quantitative-qualitative data acquisition using a benchtop Orbitrap mass spectrometer. J Am Soc Mass Spectrom 20:1441–1450
12. Lu W, Clasquin MF, Melamud E et al (2010) Metabolomic analysis via reversed-phase ion-pairing liquid chromatography coupled to a stand alone orbitrap mass spectrometer. Anal Chem 82:3212–3221
13. Seppälä U, Dauly C, Robinson S et al (2011) Absolute quantification of allergens from complex mixtures: a new sensitive tool for standardization of allergen extracts for specific immunotherapy. J Proteome Res 10:2113–2122
14. Gillet LC, Navarro P, Tate S et al (2012) Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: a New Concept for Consistent and Accurate Proteome Analysis.Mol Cell Proteomics: MCP 11 (6) (June): O111.016717. doi: 10.1074/mcp.O111.016717. http://www.ncbi.nlm.nih.gov/pubmed/22261725.
15. Venable JD, Dong M, Wohlschlegel J et al (2004) Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. Nat Methods 1:1–7

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

To date, no single analytical platform can cope with the detection of the immense structural diversity of all plant metabolites (not even from a single species). Variations in molecular weight, polarity, hydrophobicity, volatility, as well as other physicochemical properties imply the necessity for specialized analytical techniques such as GC/MS and LC/MS. The comprehensive acquisition of analytes strives towards a more holistic understanding of biological processes, by e.g. mapping primary and secondary metabolites to pathways. Data integration of primary metabolites, measured with GC/MS, and secondary metabolites measured with LC/MS, followed by a combined data analysis, enables a more comprehensive interpretation of metabolic responses to environmental stress. Since the analysis of a sample can be considered as a metabolic snapshot in time, the acquisition of time-series data aims to understand developmental processes over time, and can be used for modeling efforts.

Declaration of authorship

The results of this chapter are presented in the form of a manuscript published in the journal „Metabolomics“. Hannes Doerfler, David Lyon, Thomas Naegele and Xiaoliang Sun contributed equally to this work.

I have contributed by creating the LC/MS/MS method, performing the LC/MS data acquisition, carrying out the subsequent data extraction, transformation, analysis, and visualization of the LC/MS data (e.g. PCA not shown in the publication), as well as assisted in the data integration of GC/MS and LC/MS data. Furthermore, I've written the respective methodological parts of the manuscript.

Additional remarks

A similar, slightly adapted LC/MS data acquisition as well as data analysis method was employed for the publication of **Mari et al. 2013** (see below).

Published manuscript

Metabolomics
DOI 10.1007/s11306-012-0470-0

ORIGINAL ARTICLE

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

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

Published online: 25 October 2012

(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):

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

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). $^{13}C_6$-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 µmol m^{-2} 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 $^{13}C_6$-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×*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×*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*-trimethylsilyltrifluoroacetamid (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/H_2O 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×*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/H_2O 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 ChromaTOF® 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 H_2O 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 MS^2 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 MS^2 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 ($^{13}C_6$-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.

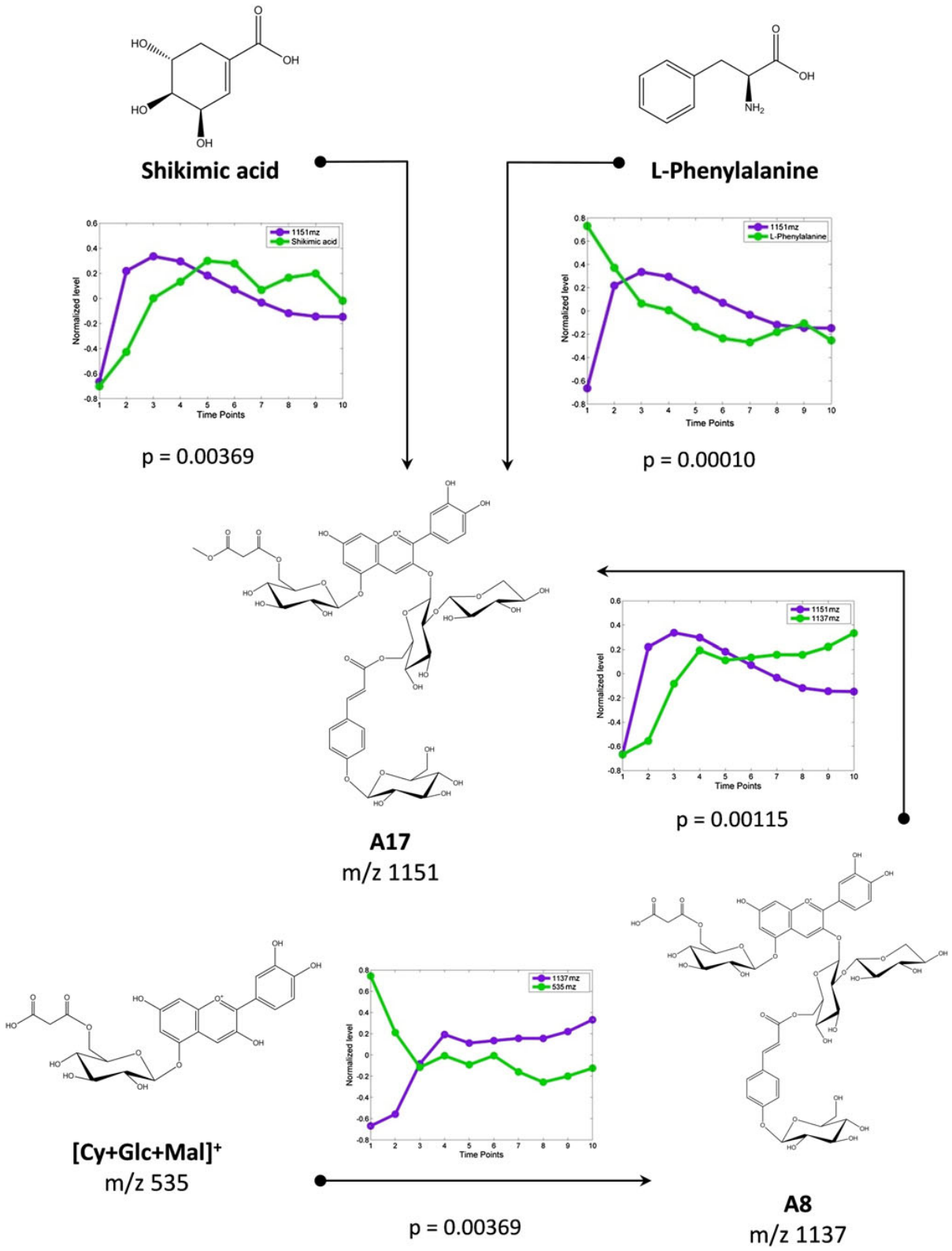
3.2 Analysis of the interaction between primary and secondary metabolism

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 $[Cy + Glc + Mal]^+$ (*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
L-Phenylalanine
1151mz
Shikimic acid
1151mz
L-Phenylalanine
Normalized level
Time Points
p = 0.00369
p = 0.00010
A17
m/z 1151
1151mz
1137mz
p = 0.00115
[Cy+Glc+Mal]+
m/z 535
1137mz
535mz
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 (*m/z*)	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-(*p*-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-(*p*-O-(glucosyl)-*p*-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-(*p*-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-(*p*-O-(glucosyl)*p*-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-(*p*-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-(*p*-O-(glucosyl)*p*-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-(*p*-O-(glucosyl)-*p*-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-(*p*-O-(glucosyl)*p*-coumaroyl)glucoside] 5-O-[6‴-O-(methyl-malonyl)glucoside]	Shi and Xie 2010	0.7817	n.a.

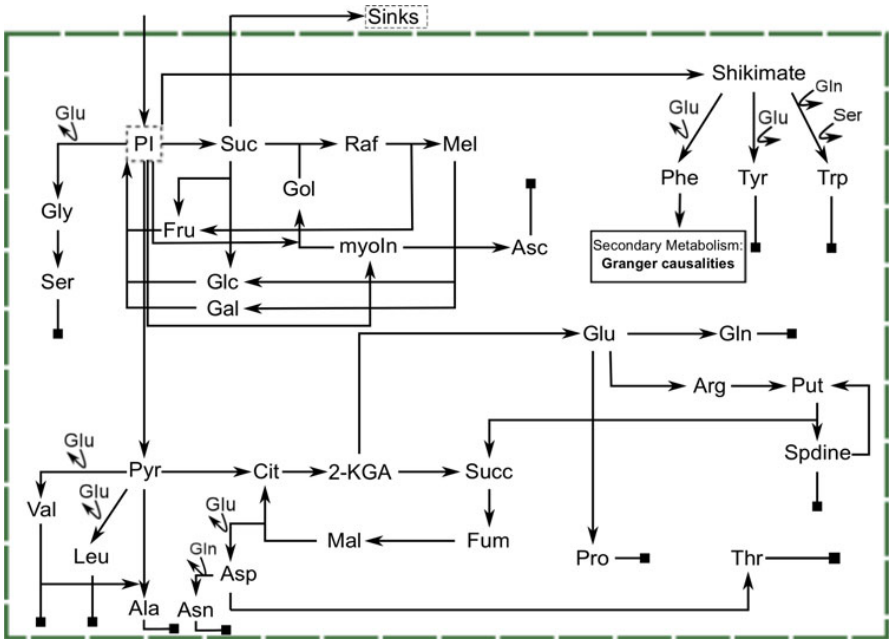


Fig. 2 Schematic representation of the primary metabolism in leaf cells of *A. thaliana*. Secondary metabolites identified by Granger causalities are exemplarily integrated derived from phenylalanine. *PI* phosphorylated intermediates, *Glu* glutamate, *Gln* glutamine, *Gly* Glycine, *Ser* serine, *Suc* 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, *Spdine* 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 µmol m^{-2} 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 coworkers 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

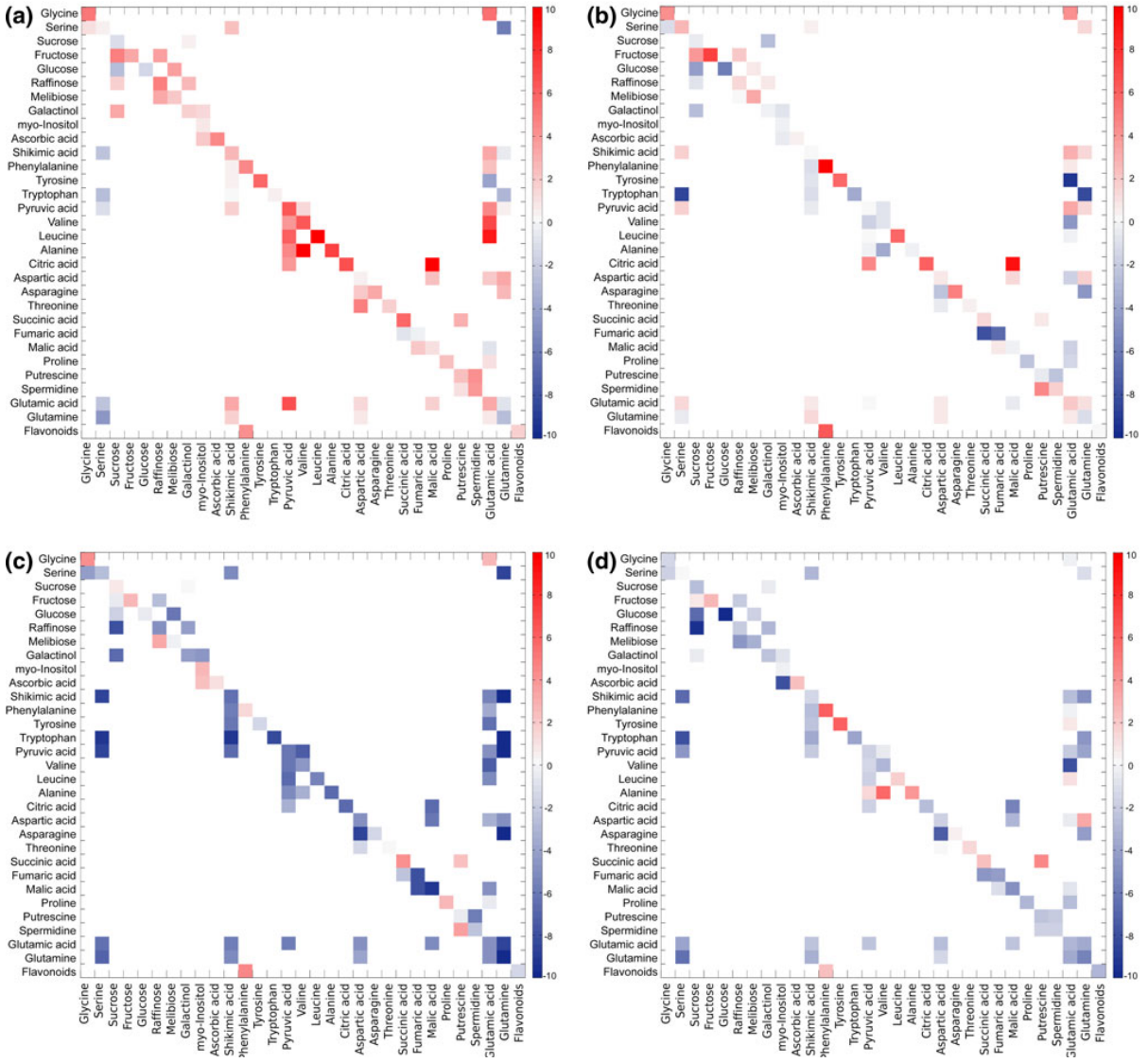


Fig. 3 Entries of the differential Jacobian after 2d(**a**), 8d(**b**), 14d(**c**) and 18d(**d**) of cold exposure relative to the non-acclimated condition are visualized by the heat map. Red colours indicate an increase of putative metabolic interaction while blue colours indicate a decrease relative to the non-acclimated plants. Entries of the differential Jacobian matrices represent mean values of 30 calculations

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 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

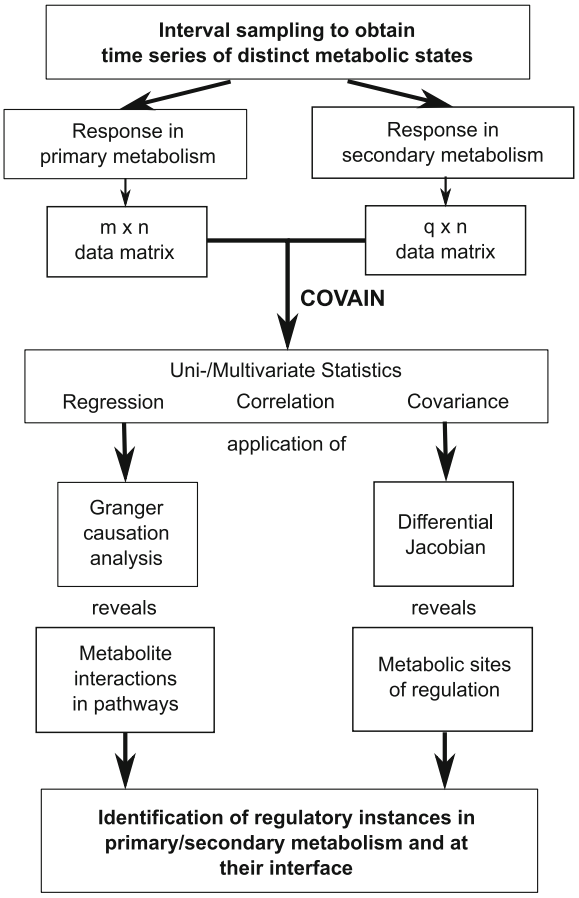


Fig. 4 Granger causality analysis and the differential Jacobian for comprehensive analysis of GC–MS and LC–MS data sets to evaluate pathway interactions and regulatory instances in metabolism

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.

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 *tt* 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., Bergmuller, 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.1365-313X.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., & Smirnoff, 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-Estepa, 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., Bläsing, 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.

Comprehensive Cell-specific Protein Analysis in Early and Late Pollen Development from Diploid Microsporocytes to Pollen Tube Growth

A thorough understanding of pollen development on a molecular level is necessary, since plant reproduction and productivity is dependent on the latter. The analysis of the differential expression of protein patterns dependent on developmental stages of pollen growth and the public access to such data is of interest to the scientific community. *Nicotiana tabacum* produces relatively large flowers compared to other important crop species of the Solanaceae family such as *Solanum lycopersicum*, *Solanum tuberosum*, and *Solanum melongena* and was therefore well-suited for the study.

Declaration of authorship

The results of this chapter are presented in the form of a manuscript published in the journal „Molecular & Cellular Proteomics“. I have provided a critical contribution to the following publication, though the largest part the work was performed by the coauthors.

Since tobacco is not as well-characterized as e.g. *A. thaliana*, three distinct protein FASTA files (databases) were used for identification as well as all downstream analyses. Using unpublished Python scripts, I have fused the FASTA files into a single non-redundant database (in analogy to **Staudinger et al. 2012**, see Application to plant research and Publications), and subsequently using the Mercator pipeline, as well as unpublished Python scripts, I've created a so-called „mapping“ file for tobacco. The latter links protein identifiers (Accession Numbers) to multiple functional categories. Using stand-alone BLAST in conjunction with unpublished Python scripts, I found homologues of tobacco to *A. thaliana*.

Published manuscript

Comprehensive Cell-specific Protein Analysis in Early and Late Pollen Development from Diploid Microsporocytes to Pollen Tube Growth*[S]

Till Ischebeck‡§, Luis Valledor‡, David Lyon‡, Stephanie Gingl‡, Matthias Nagler‡, Mónica Meijón¶, Volker Egelhofer‡, and Wolfram Weckwerth‡§

Pollen development in angiosperms is one of the most important processes controlling plant reproduction and thus productivity. At the same time, pollen development is highly sensitive to environmental fluctuations, including temperature, drought, and nutrition. Therefore, pollen biology is a major focus in applied studies and breeding approaches for improving plant productivity in a globally changing climate. The most accessible developmental stages of pollen are the mature pollen and the pollen tubes, and these are thus most frequently analyzed. To reveal a complete quantitative proteome map, we additionally addressed the very early stages, analyzing eight stages of tobacco pollen development: diploid microsporocytes, meiosis, tetrads, microspores, polarized microspores, bipolar pollen, desiccated pollen, and pollen tubes. A protocol for the isolation of the early stages was established. Proteins were extracted and analyzed by means of a new gel LC-MS fractionation protocol. In total, 3817 protein groups were identified. Quantitative analysis was performed based on peptide count. Exceedingly stage-specific differential protein regulation was observed during the conversion from the sporophytic to the gametophytic proteome. A map of highly specialized functionality for the different stages could be revealed from the metabolic activity and pronounced differentiation of proteasomal and ribosomal protein complex composition up to protective mechanisms such as high levels of heat shock proteins in the very early stages of development. ***Molecular & Cellular Proteomics 13: 10.1074/mcp.M113.028100, 295–310, 2014.***

Plants contain numerous specialized cell types, each of them expressing a specific set of proteins. In recent studies much effort has been put into isolating and analyzing proteins of these individual cell types (1) rather than whole organs. New emerging methods have led to the in-depth analysis of different plant cell types, including guard cells (2, 3), mesophyll cells (4), trichomes (5–9), root hair cells (10–12), and egg cells (13). Additionally, because of their easy availability, mature pollen and *in vitro*–grown pollen tubes are among the most frequently studied cell types. Pollen and pollen tube proteomes have been analyzed, for example, from *Arabidopsis* (14–16), lily (17, 18), tomato (19, 20), rice (21, 22), quercus and pine trees (23–27), and tobacco (28).

As pollen represents the severely reduced male gametophyte of higher plants, it expresses a very unique set of genes (29) required for the fast and energy-consuming polar outgrowth of the pollen tube during the fertilization process (30). Enzymes required for metabolism and energy generation are overrepresented, but there are also components of the exocytotic machinery, including signaling proteins (14) required for the deposition of pectin compounds at the tip of the growing pollen tube.

Although mature pollen and *in vitro*–grown pollen tubes have been the focus of research because of the ease of harvesting procedures and are widely used for cell biological studies (31–36), this is not the case for earlier stages of pollen development.

In angiosperms, mature pollen develops from microsporocytes in the anthers of the flower in a series of distinct stages (37). After the microsporocytes have completed meiosis, they form tetrads that release microspores with one central haploid nucleus. These microspores undergo polarization and asymmetric mitosis. The bigger vegetative cell internalizes the smaller cell, which later divides again and forms the two sperm cells. Finally, the pollen desiccates. When the pollen falls on the stigma, it rehydrates, and the vegetative cell forms a pollen tube that delivers the two sperm cells through the transmitting tract to the ovule (38).

Even though pollen development studies using electron microscopy date back to the 1960s (39) and many mutants are described that are disrupted in this process (14), informa-

From the ‡Department of Molecular Systems Biology, Faculty of Life Sciences, University of Vienna, Althanstrasse 14, A-1090, Vienna, Austria; ¶Gregor-Mendel-Institute for Molecular Plant Biology, Dr. Bohr-Gasse 3, 1030 Vienna, Austria

⌘ *Author's Choice*—Final version full access.

Received February 7, 2013, and in revised form, September 24, 2013

Published, MCP Papers in Press, September 27, 2013, DOI 10.1074/mcp.M113.028100

tion on the proteome of developing pollen remains relatively sparse and is mostly restricted to whole anthers (40, 41). Only very recently, a work on tomato pollen was conducted covering five developmental stages (42).

The transcriptome of *Arabidopsis* pollen has been analyzed from the microspore stage on (43), but the earlier stages of microsporocytes, meiosis, and tetrads were not studied, most likely because of a limitation of available material. However, this study was able to show dramatic changes in the transcriptome during the development from microspores to the mature pollen. Similar studies have been performed with *Brassica napus* (44) and rice pollen (45).

A comparative analysis of the proteome from these stages, as presented in our study, can have special relevance, because in pollen the proteome can greatly differ from the transcriptome not only quantitatively, but also qualitatively, as has been shown for *Arabidopsis* pollen (14). It seems that often the mRNA is degraded while the protein persists or mRNA is stored in desiccated pollen to be transcribed after rehydration (14).

Additionally, in our proteomic study, we extended the analysis to even earlier stages, including the stage of meiosis. We were able to compare, for the first time, the proteome of a total of eight stages: the diploid microsporocytes, cells undergoing meiosis, tetrads, microspores, polarized microspores (undergoing mitosis I), bipolar pollen, desiccated pollen, and finally pollen tubes. We found that the proteome underwent great changes during development, especially during the polarized microspore stage.

EXPERIMENTAL PROCEDURES

Plant Growth and Pollen Collection—Tobacco was grown under greenhouse conditions (12 h of light, 120 μmol m^{-2} s^{-1}, 23 °C during the day, 20 °C at night, 60% humidity). Flowers of different sizes were collected, and the anthers of individual flowers were sampled in 200 μl of 10% mannitol. Anthers were gently squeezed open and vortexed, and the supernatant including the released pollen was transferred to a new tube. Pollen was spun down at 100 $\times$ *g* for 1 min and washed twice with 10% mannitol. A subfraction of the pollen of each individual flower was analyzed under a microscope to determine the developmental stage. Samples not representing a stage with at least 90% of their pollen were discarded.

Pollen tubes were grown for 5 h in pollen tube medium (10% sucrose, 15 mM MES-KOH pH 5.9, 1 mM $CaCl_2$, 1 mM KCl, 0.8 mM $MgSO_4$, 1.6 mM H3BO3, 30 μM CuSO4) slightly modified according to Read *et al.* (46).

Young leaves and roots were ground in liquid nitrogen, and proteins were extracted accordingly.

Microscopy—Pollen samples were fixed in 10% mannitol and 4% formaldehyde overnight, collected via centrifugation, and resuspended in 1 μg/ml DAPI and 1% triton X-100 5 min prior to microscopy.

Images were recorded with an upright point laser scanning confocal microscope (LSM780, Zeiss, Oberkochen, Germany) using a 405-nm diode laser for excitation and a band-pass filter ranging from 450–550 nm. Acquired images were processed using Fiji software.

Quantitative Proteome Analysis (GeLC-LTQ-Orbitrap MS)—For each sample, pollen from between 5 and 30 flowers (depending on the stage) was pooled, freeze-dried, cooled in liquid nitrogen, and ground for 3 min in a shaking mill using three 2-mm steel balls per tube. The pollen fragments were resuspended in 200 μl of protein extraction buffer (62.5 mM Tris-HCl pH 6.5, 5% SDS (w/v), 10% glycerol (v/v), 10 mM DTT, 1.2% (v/v) plant protease inhibitor mixture (Sigma P9599)) and incubated for 5 min at room temperature. After this time, the samples were mixed again by pipetting, incubated for 3 min at 90 °C, and then centrifuged at 21,000 $\times$ *g* for 5 min at room temperature. Supernatants were carefully transferred to a new tube. After the addition of an equal volume of 1.4 M sucrose, proteins were extracted twice with Tris-EDTA buffer–equilibrated phenol. The combined phenolic phases were counter-extracted with 0.7 M sucrose and subsequently mixed with five volumes of 0.1 M ammonium-acetate in methanol to precipitate the proteins. After 16 h of incubation at −20 °C, samples were centrifuged for 5 min at 5000 $\times$ *g* at 5 °C. The pellet was washed twice with 0.1 M ammonium-acetate and once with acetone and then air-dried. Pellets were redissolved in 6 M urea, 5% SDS, and protein concentrations were estimated via bicinchoninic acid assay (47).

Proteins were analyzed via a new gel-LC-MS protocol (48). 40 μg of protein were loaded into a mini-protean cell and run for 1.5 cm. Gels were fixed and stained with 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 as previously described (49). Peptides were then desalted with the use of Bond-Elute C-18 stage tips (50) and concentrated in a SpeedVac. Prior to mass spectrometric measurement, protein digest pellets were dissolved in 4% (v/v) acetonitrile, 0.1% (v/v) formic acid. 10 μg of digested peptides were loaded per injection into a one-dimensional nano-flow LC-MS/MS system equipped with a pre-column (Eksigent, Redwood City, CA, USA). Peptides were eluted using a monolithic C18 column Chromolith RP-18r (Merck, Darmstadt, Germany) of 15-cm length and 0.1-mm internal diameter during an 80-min gradient from 5% to 50% (v/v) acetonitrile/0.1% (v/v) formic acid with a controlled flow rate of 500 nl/min.

MS analysis was performed on an Orbitrap LTQ XL mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). Specific tune settings for the MS were as follows: the spray voltage was set to 1.8 kV using a needle with a 30-μm inner diameter (PicoTip Emitter, New Objective, Woburn, MA), and the temperature of the heated transfer capillary was set at 180 °C. Fourier transform MS was operated as follows: full scan mode, centroid, resolution of 30,000, covering the range of 300–1800 *m/z*, and cyclomethicone used as a lock mass. Each full MS scan was followed by 10 dependent MS/MS scans performed in the ion trap, in which the 10 most abundant peptide molecular ions were dynamically selected with a dynamic exclusion window set to 90 s and an exclusion list set to 500. Dependent fragmentations were performed in collision-induced dissociation mode with a normalized collision energy of 35, an isolation width of 2.0, an activation Q of 0.250, and an activation time of 30 ms. Ions with an unassigned charge or a charge of +1 were excluded for fragmentation. The minimum signal threshold was set at 1000.

Raw data were searched with the SEQUEST algorithm present in Proteome Discoverer version 1.3 (Thermo, Germany) as described elsewhere (51). In brief, identification confidence was set at a 5% false discovery rate, and the variable modifications were set as acetylation of the N terminus, oxidation of methionine, and carbamidomethyl cysteine formation, with mass tolerances of 10 ppm for the parent ion and 0.8 Da for the fragment ion. Up to two missed cleavage sites were permitted. Three different databases were employed (tobacco 7.0, a CDNA library from the gene index project with 120,122 entries; a tobacco protein database from UniProt 09.2011 with 4826 entries; and a genomic sequence database from the Tobacco Genome Initiative 11.2008 with 349,877 entries, resulting in 2,099,262 entries after

six-frame translation). Databases were translated with an in-house tool, taking into consideration only the longest open reading frame of all reading frames. In the case of the genomic sequences, the longest open reading frames of all reading frames were considered.

When the database from the gene index project was used, additional variable modifications were allowed: phosphorylation of threonine, serine, and tyrosine; methylation and dimethylation of lysine and arginine; and acetylation and trimethylation of lysine.

Peptides were matched against these databases plus decoys, with a significant hit considered as one in which the peptide confidence was at least medium or high, and the xcorr score threshold was established at 2.5 for +2 ions and 3.5 for charge states of +3 or greater. The high thresholds were chosen to minimize false identifications based on the incomplete databases used.

The identified proteins were quantitated via a label-free approach based on peptide count followed by a normalized spectral abundance factor (NSAF)[1] normalization strategy (52),

$$(NSAF)_k = (PSM/L)_k / \Sigma_{i=1}^{N} (PSM/L)_i$$

in which the total number of spectra counts for the matching peptides from protein *k* (PSM) was divided by the protein length (*L*) and then divided by the sum of PSM/*L* for all *N* proteins.

Multivariate Statistical and Bioinformatic Data Analysis—Multivariate statistical analyses such as principal components analysis (PCA) and k-means clustering were performed with the statistical toolbox COVAIN (53). The software and parameter settings can be accessed online. Missing values were estimated from the dataset, and data were log transformed before the PCA. For cluster analysis, the mean NSAF value of each developmental stage was calculated and normalized for each protein, setting the total amount throughout the stages to 1.

All proteins in the three used databases were blasted for the closest *Arabidopsis* (TAIR10) homologue using an unpublished Python script in conjunction with stand-alone BLAST v2.2.26+ using the default matrix, and entries in the TAIR *Arabidopsis* MapMan mapping file (Ath_AGI_LOCUS_TAIR10_Aug2012) were replaced as previously described (54). This way, most tobacco protein accessions could be assigned to a functional bin and an *Arabidopsis* homologue.

Tobacco and *Arabidopsis* microarray results were binned according to the MapMan mapping files Ntob_AGILENT44K_mapping and Ath_AGI_LOCUS_TAIR10_Aug2012, respectively. Tobacco bin numbers were slightly adjusted to fit the tobacco protein bins.

Further blasting of the tobacco protein sequences *versus* the list of *Arabidopsis* proteins found in *Arabidopsis* pollen (14) and a list of pollen-affected *Arabidopsis* mutants (extended list from Ref. 14) was performed using the same Python script.

RESULTS AND DISCUSSION

Isolation and Proteomic Analysis of Early and Late Pollen Developmental Stages—Pollen from a total of eight developmental stages was harvested for proteomic analysis (Figs. 1 and 2).

These distinct stages were diploid microsporocytes (also referred to as stage A), meiosis (stage B), tetrad stage (stage C), microspores (stage D), polarized microspores (stage E), bipolar pollen (stage F), desiccated pollen (stage G), and pollen tubes (stage H).

Immature pollen was obtained by gently opening the anther buds of individual flowers and vortexing in 10% mannitol. In this way, pollen in the supernatant could be easily separated from larger cell debris via simple pipetting. Smaller cell debris and soluble proteins could be removed with the supernatant after low-speed centrifugation.

Although pollen from the microsporocyte and meiosis stages was obtained as large aggregates, which were associated with cell debris (Fig. 1), it was possible to isolate individual cells (or tetrads) from later stages (Figs. 1 and 2).

As the stage of pollen development cannot be easily determined by the size of the flower or anthers, especially in early stages, the developmental stage of the pollen of each individual flower was determined via microscopy, and a sufficient amount of pollen was pooled for protein extraction.

Desiccated pollen was harvested after anthesis, and pollen tubes were grown *in vitro* for 6 h.

For comparison, proteins from young tobacco leaves and roots were analyzed.

Three biological replicates of each stage (or tissue) were analyzed and separated into two fractions via SDS-PAGE prior to tryptic digestion and LC-MS/MS analysis.

The spectra of all identified peptides (supplemental Table S1) from the different stages can be reviewed online in the proteomics database PROMEX (http://promex.pph.univie.ac.at/promex/Experiment; Nic taba002 for pollen and Nic taba003 for roots and leaves). Additionally, the mass spectrometry proteomics data have been deposited in the ProteomeXchange Consortium via the PRIDE partner repository (55) with the dataset identifier PXD000469.

In total, 3817 protein groups were identified from all pollen stages (Table I, supplemental Table S2), with stages A–D and F–H showing the most overlap (Fig. 3*A*). When the results were compared with data on extracts from tobacco roots and leaves, a total of 4262 protein groups were identified: 1217 from leaves, 1285 from roots, and 3888 from pollen (Fig. 3*B*, Table I, supplemental Table S3; the increased number of pollen protein groups is due to different groupings of the proteins). The high number of identified pollen proteins was in part caused by the large tobacco genome (4.5 billion bp), leading to the finding of many homologue isoforms, but it is also attributable to the great changes that took place in the proteome during the development.

The protein groups represent a total of 12,728 putative protein accessions in pollen (supplemental Table S2) and 14,323 proteins in all the samples (supplemental Table S3). For easier reading, the protein groups are referred to as proteins hereinafter.

Protein abundances were quantified by peptide count and an NSAF normalization strategy (52). For further analysis, only proteins that were detected in all three biological replicates of at least one of the developmental stages (or tissues) were considered, leading to datasets of 1869 proteins when only pollen proteins were considered and 2135 proteins when leaves and roots were included. Proteins were classified by identifying the closest *Arabidopsis* homologue and assigning

[1] The abbreviations used are: NSAF, normalized spectral abundance factor; PCA, principal components analysis.

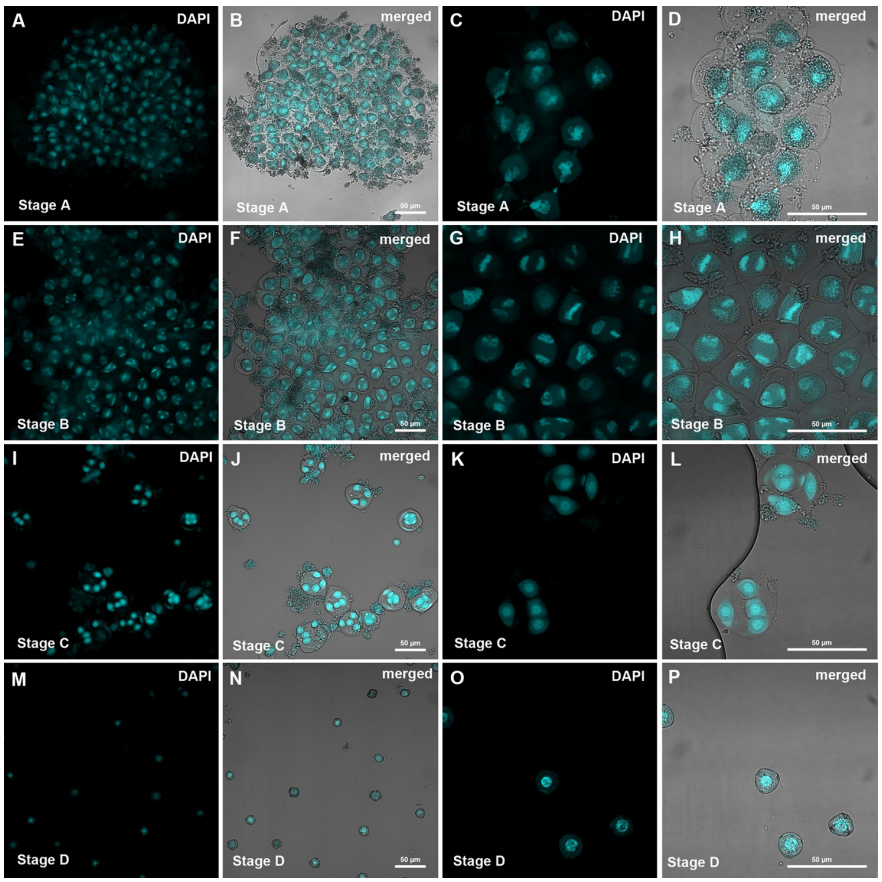


FIG. 1. **Confocal images of pollen purifications as were used for proteomic analysis.** *A–D*, microsporocytes; *E–H*, meiotic cells; *I–L*, tetrads; *M–P*, microspores. Pollen was stained with DAPI and images were obtained at 250× and 630× magnification.

a function according to functional Arabidopsis mapping for MapMan (supplemental Tables S2, S3, and S7).

Of the 2135 proteins used for quantification, 837 were not detected in any of the root or leaf samples (supplemental Table S3). It cannot be ruled out that these proteins are also present in minor amounts in these organs or in other non-analyzed tissues. However, the proteins showing high expression levels in one of the pollen stages (Table II) can be considered as especially strong candidates for being specifically expressed in developing pollen, or at least for serving specific purposes in these cell types. One example is the highly abundant ethanol dehydrogenase, which serves a specific function in the primary metabolism of pollen tubes (56) and is not needed in roots and leaves, at least under normal conditions, as well as cell wall degrading enzymes. Another protein that was also not found in roots and leaves is the Rab-GDP dissociation inhibitor, which is crucial for G-protein signaling and, thus, maintaining cell polarity during polar tip growth.

Multivariate Statistical Data Mining—A PCA of the pollen proteins alone revealed that, on the proteome level, tobacco pollen development could be separated into three major phases (Fig. 4*A*), with the first one including the stages from microsporocytes to microspores (A–D), the second one including only the polarized microspores (E), and the third one including the binuclear pollen stage to the pollen tubes (F–H).

This separation in the PCA and the stage specificity of the proteomes are clearly based on different cell functionalities. In the first four stages (A–D), the principal function of the pollen is its own transformation from diploid microsporocytes to microspores, whereas the obviously very different function of the rehydrated pollen is to produce and elongate a pollen tube. To facilitate a quick outgrowth, many proteins that support this function are apparently already synthesized prior to desiccation, which leads to the observed similarity of the last three stages (F–H). The polarized microspore stage (E) could be a transition stage. However, this stage also contains a unique set of proteins not present in any of the other stages.

The distinct composition of the proteome of this stage was also apparent in the individual principal components (supplemental Table S4). PC1 separated the samples according to their ongoing development, and PC2 separated stage E from all other stages (Fig. 4*A*).

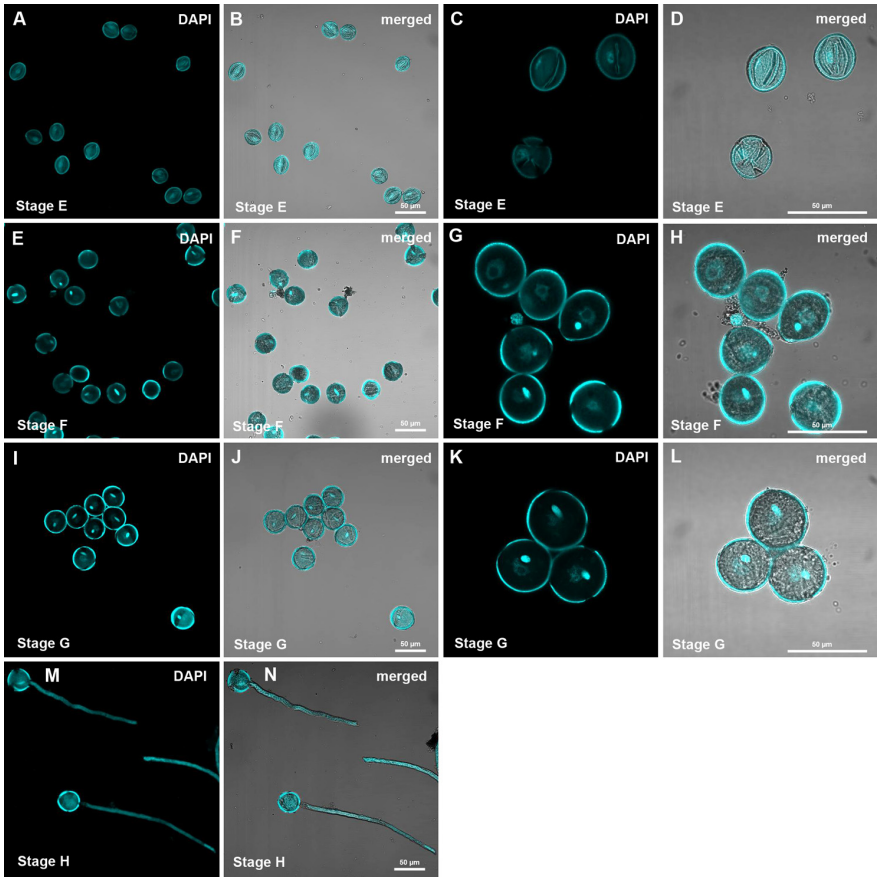


FIG. 2. **Confocal images of pollen purifications as used for proteomic analysis.** *A–D*, polarized microspores; *E–H*, binuclear pollen; *I–L*, desiccated and rehydrated pollen; *M–P*, *in vitro* grown pollen tubes. Pollen was stained with DAPI and images were obtained at 250× and 630× magnification.

TABLE I

Number of identified protein groups by stage/tissue. Values given in parentheses are the numbers of protein groups when only pollen proteins are considered, which leads to a smaller number of protein groups from the same number of proteins (due to different grouping)

Stage	Label	*n* proteins
Microsporocyte	A	1775 (1741)
Meiosis	B	1596 (1573)
Tetrad	C	1748 (1719)
Microspore	D	1288 (1264)
Polarized microspore	E	2004 (1956)
Binuclear pollen	F	1770 (1740)
Dessicated pollen	G	1604 (1580)
Pollen tubes	H	2526 (2485)
Pollen total		3888 (3817)
Leaves		1217
Roots		1285

The proteins with the highest loadings in PC2 showed comparatively high abundance in stage E (Fig. 4*B*), but they were also present in stages A, F, and H. Among these proteins were several subunits of the 26S proteasome.

In comparison, the proteins with the most negative loadings in PC2 showed an inverse expression pattern (Fig. 4*B*). The proteins with the highest PC2 loadings were ribosomal proteins, hinting at a severe rearrangement of the ribosomal complex during stage E: from a total of 155 detected ribosomal proteins, 71 were missing in stage E, and 44 of these were detected in all other stages. A specific set of 38 ribosomal proteins had higher loadings in stage E than the average over all samples.

Thus, we observed a pronounced reprogramming of the protein synthesis machinery that might prepare the ribosomal complex machinery for the high demands of protein synthesis during pollen tube growth.

As PC1 differentiated the samples according to development, the negative loadings represented proteins with high abundance in early stages that declined during development, whereas positive loadings represented proteins with high expression levels in the last stages (Fig. 4*B*).

Among the negative loadings, histones were especially well represented, probably because of their lower cell-volume-to-nucleus ratios in early stages relative to mature pollen.

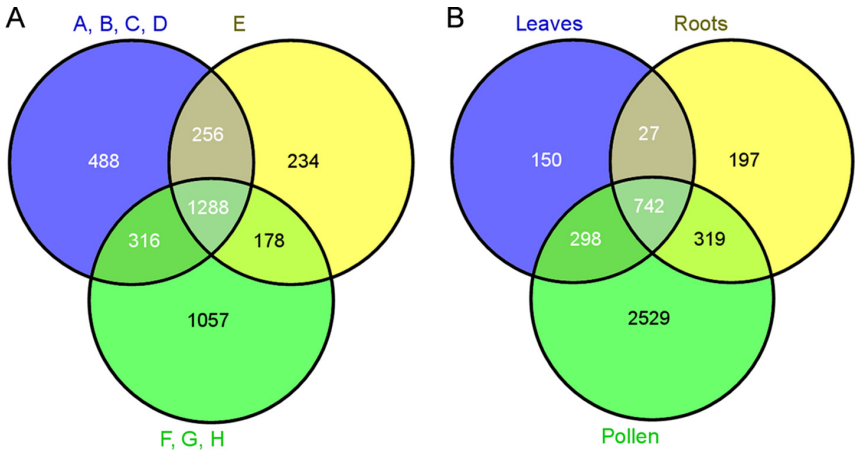


FIG. 3. **Venn diagrams.** *A*, number of proteins identified throughout development. A–D, microsporocytes, meiotic cells, tetrads, and microspores, respectively; E, polarized microspores; F–H, binuclear pollen, desiccated pollen, and pollen tubes, respectively. *B*, number of identified proteins in all pollen stages, leaves, and roots.

TABLE II
Proteins with the highest NSAF scores (maximum of eight stages) that were not detected in roots and leaves

Accession number	Proposed function	Closest *Arabidopsis* homologue	A	B	C	D	E	F	G	H
CN824898	Unknown	—					8.73			
AM795719	Unknown	—	3.29	0.44	5.27	7.58				
Q42953	Alcohol dehydrogenase	AT1G77120.1		0.06				3.46	4.98	2.94
TC125213	Globulin-like protein, storage	AT1G07750.1	0.06	0.32	0.06	0.25	4.73	3.28	4.27	2.38
191520275	Unknown	—					4.63			
AM837330	Fructokinase	AT4G10260.1		0.22			3.28	3.06	4.39	2.52
TC131738	Globulin-like protein, storage	AT1G07750.1	0.04	0.33	0.04	0.20	4.32	2.63	4.07	1.85
191442209	Polygalacturonase inhibitor protein precursor	AT5G06860.1	2.00	4.19	2.91	2.19	2.96	0.07		0.22
TC127895	Polygalacturonase inhibitor protein precursor	AT5G06860.1	2.10	3.98	3.32	2.70	2.80	0.20		0.15
191443577	Polygalacturonase inhibitor protein precursor	AT5G06860.1	2.11	3.85	3.02	2.74	2.75	0.21		0.16
TC156020	Glyceraldehyde 3-phosphate dehydrogenase	AT1G13440.2						3.74	1.86	2.38
O24625	Chalcone/stilbene synthase	AT1G02050.1		3.66	0.43	0.12	2.87	0.06		
D4I601	Alcohol dehydrogenase	AT1G77120.1		0.12				1.65	2.86	3.45
TC129072	Alcohol dehydrogenase	AT1G77120.1						1.48	3.37	2.72
TC132846	Late embryogenesis abundant protein	AT3G15670.1							3.27	
TC150130	Rab-GDP dissociation inhibitor	AT5G09550.1						0.66	3.11	0.71
190837846	Unknown	AT3G59510.1	1.40	3.09	1.63	1.22	2.29	0.60		
191572052	Unknown	AT1G12570.1		3.08			2.10			
CV021106	Ribosomal 60S subunit.L32	AT4G18100.1	1.35	1.41	1.05	2.95		0.17	0.28	0.26
AM806415	Membrane transporter	AT3G08580.2			0.16		0.37	2.70	1.33	1.06

Notes: Average NSAF scores over three biological replicates multiplied by 1000. A, microsporocytes; B, meiotic cells; C, tetrads and microspores; E, polarized microspores; F, binuclear pollen; G, desiccated pollen; H, pollen tubes.

The highest positive loadings included proteins required for pollen tube growth such as enzymes of the primary metabolism, ethanolic fermentation, and cell wall synthesis.

A PCA additionally including roots and leaves revealed a clear separation of the different tissues (Fig. 4*C*, supplemental Table S4). Whereas PC1 discriminated especially between the different pollen stages and the sporophytic tissues, PC2 discriminated between roots and leaves, with the earlier developmental pollen stages being more closely related to leaves. This closer connection of leaves and the early male gametophytes was also apparent from the correlation coefficients (Pearson's *R*, Fig. 4*C*).

In order to further group the pollen proteins according to their presence in the different stages, the NSAF scores were

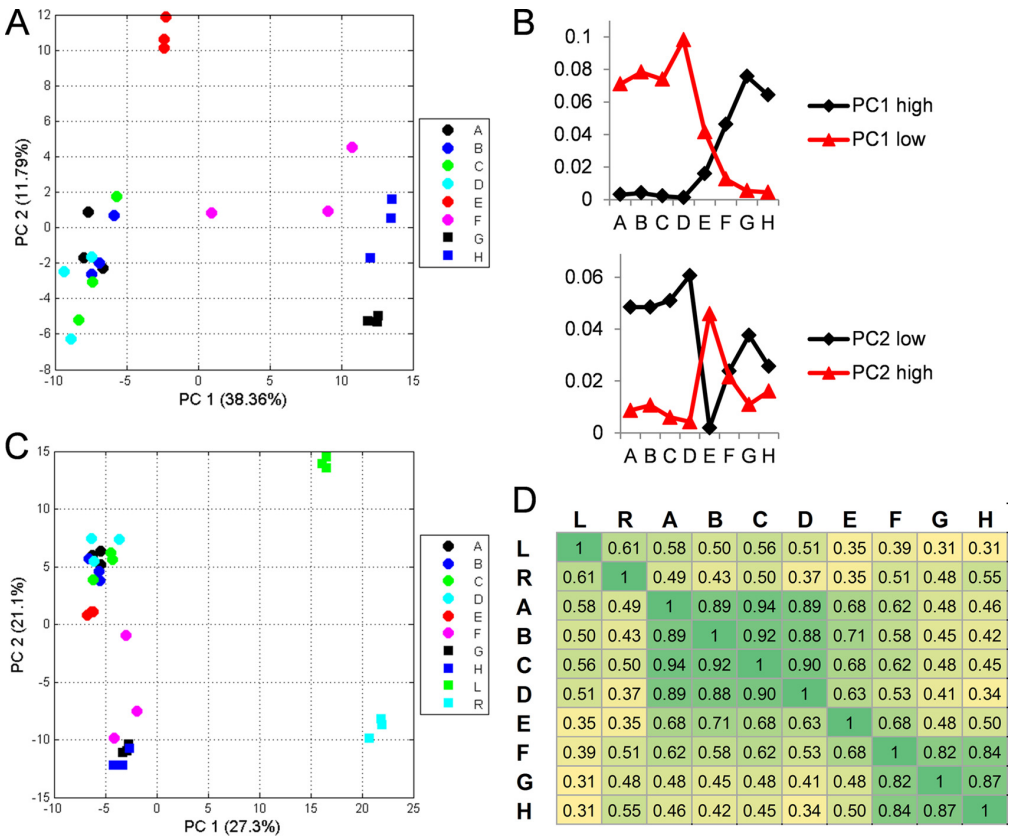


FIG. 4. **Multivariant statistic analysis.** *A*, PCA plot of the eight stages of pollen development. *B*, sum of the NSAF scores of the 30 highest and 30 most negative loadings of PC1 and PC2, respectively. *C*, PCA plot of the eight stages of pollen development including leaves and roots. PCA was based on log values of the NSAF scores. *D*, correlation coefficient according to calculated average NSAF values of three replicates. A, microsporocytes; B, meiotic cells; C, tetrads; D, microspores; E, polarized microspores; F, binuclear pollen; G, desiccated pollen; H, pollen tubes; L, leaves; R, roots.

TABLE III
Overview of stage-specific clusters identified via k means clustering

Stage	Cluster number	*N* proteins
A	10	22
B	27	5
C	33	15
D	2	7
E	5, 6, 21	59, 51, 29
F	30	10
G	13, 24	17, 33
H	11, 15	39, 47
B + E	23	33
F + H	22	37

normalized for each protein and the proteins were clustered using the k means algorithm (supplemental Table S6, supplemental Fig. S1). 12 of the 35 clusters showed proteins that were almost exclusively expressed in one of the stages (Table III).

Again, stage E stood out in terms of the number of specifically expressed proteins (clusters 5, 6, and 21; Fig. 5*A*). Among them were three proteins similar to Skp1 (BP531238, TC168823, and 191216821), a core component of the E3 ubiquitin ligase that targets protein for degradation by the 26S proteasome. The isoform expressed in stage E might interact with specific F-Box proteins, which could target a distinctive set of proteins for ubiquitination and breakdown, in order to adjust the proteome as required for the change in cellular function.

It is also possible that the skp1 proteins are directly involved in mitosis. The 26S proteasome is a key factor in the degradation of cell cycle proteins (57). Also, a skp1-like 1 (ASK1) of *Arabidopsis* is essential for meiosis in pollen (58), where it is essential for nuclear reorganization and homolog juxtapositioning (59). It was also found to be involved in mitosis (60).

Several proteases identified in stage E could also be in part responsible for the major rearrangement of the pollen proteome during in this transition stage.

Many potential cellulases, glucosidases, and mannosidases were expressed during stage E (supplemental Table S6). They

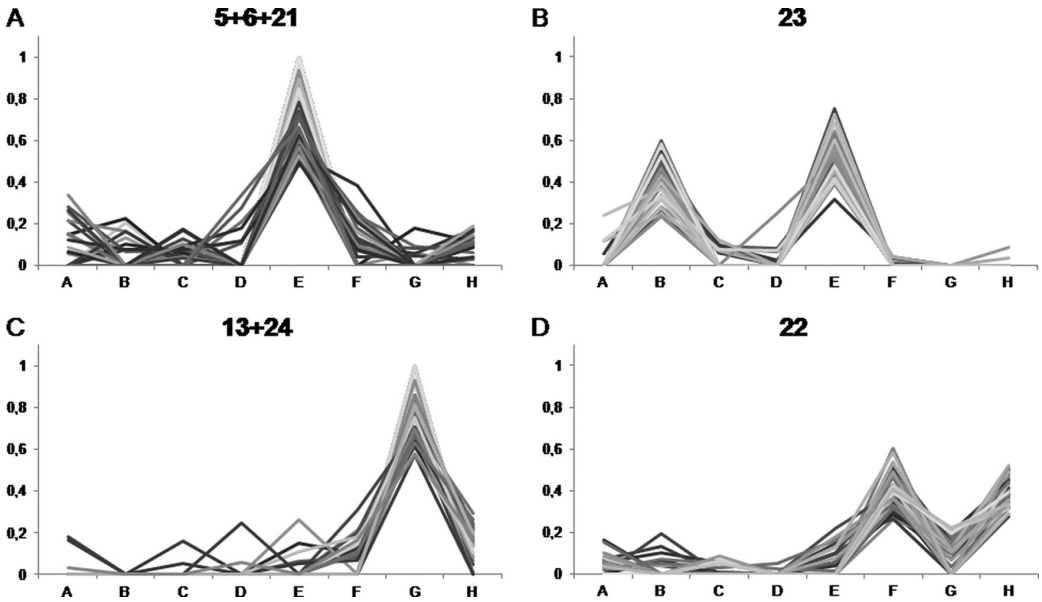


FIG. 5. **Cluster analysis.** Relative abundance of proteins in a selection of groups obtained via k means clustering. NSAF scores were averaged over three biological replicates and normalized for each protein to represent their proportion of the total abundance over all eight stages. A, microsporocytes; B, meiotic cells; C, tetrads; D, microspores; E, polarized microspores; F, binuclear pollen; G, desiccated pollen; H, pollen tubes.

have previously been proposed to support the loosening of the cell wall required for cell expansion taking place between stages D and F (61).

A subset of proteins grouped in cluster 23 (Fig. 5*B*) was expressed predominantly during meiosis (B) and mitosis (E), and these proteins might take part in the regulation of mitosis and meiosis, as they are specifically expressed in these stages (supplemental Table S6). One example is annexins (TC137724, BP533244, Q56D09), which might act in targeted secretion (62), required for cytokinesis. Another example is a set of potential subtilases (TC132351, TC133164, TC133288, 191501021) that could take part in signaling or specific protein cleavage and degradation.

Other proteins expressed during these two stages are predicted to play a role in secondary metabolism, which makes them unlikely to take direct part in the cell cycle. Three of them (191361943, CN949712, and O24625) show homologies to the anther expressed proteins less adhesive pollen 5 and 6, which show similarities to chalcone synthases and are essential for exine formation (63).

Before pollination, the pollen desiccates and has to drastically adjust its physiology to protect its membranes from breaking and its proteins from denaturation. In the cluster analysis, a set of proteins was grouped (clusters 13 and 24; Fig. 5*C*) that was almost exclusively expressed in the desiccated stage (G) and disappeared after rehydration and pollen tube growth (H). Among these proteins were late early abundant proteins (TC132846, TC146808, TC165472, 191501982) that also play a role in the desiccation of seeds (64) and are proposed to protect pollen during dehydration (65). Also, a homologue of an *Arabidopsis* tonoplast monosaccharide transporter (TC129132) and potential signaling proteins that could play a role in adaption to desiccation were grouped in this cluster.

Additionally, a set of proteins (cluster 22; Fig. 5*D*) was identified showing that some proteins might be degraded prior to desiccation and resynthesized after rehydration. Proteins in this cluster included enzymes of β-oxidation and of other primary metabolic pathways (supplemental Table S4). We can only speculate about the reason for this temporary degradation. The proteins might have a negative effect on the adaption to desiccation or be unstable under this condition.

Functional Remodeling of the Proteome During Pollen Development—During development, the pollen cells have to adjust their metabolism to suit their functions. In order to get a better overview of the functionality, proteins were matched against their closest *Arabidopsis* homologues and grouped according to their predicted functions (supplemental Table S7). The total NSAF scores were added (Fig. 6).

Some functional groups were predominantly present in specific stages, such as the already mentioned late early abundant proteins in desiccated pollen (stage G); the gluco-, galacto-, and mannosidases, factors of protein degradation in polarized microspores (stage E); and the enzymes of secondary metabolism during meiosis and mitosis (stages B and E).

Starch synthesis seems to occur in microsporocytes and binuclear pollen prior to desiccation (stages A and F), most likely to store energy for cell division and pollen tube growth

	L	R	A	B	C	D	E	F	G	H
Light reaction	61.0	0.0	8.6	4.1	3.5	3.2	8.5	0.6	0.5	0.8
Calvin cycle	103.7	17.9	22.1	21.5	23.7	22.8	22.9	18.7	15.1	10.8
Photorespiration	7.2	0.6	0.3	0.1	0.1	0.1	0.6	1.2	2.4	2.6
Glycolysis	49.3	86.6	68.3	60.7	73.8	60.3	62.1	114.5	112.9	117.1
Gluconeogenesis	0.0	1.2	1.7	1.2	0.9	0.6	2.9	6.3	6.6	6.7
Pentose Phosphate pathway	2.8	4.8	5.1	2.9	3.2	1.8	4.8	3.2	2.8	4.0
TCA cycle	22.5	37.9	53.5	50.4	51.9	43.4	50.7	50.0	61.4	55.2
Mitoch. electron transport	17.5	29.5	39.2	32.4	27.9	32.4	33.0	47.7	47.3	48.4
Sucrose synthesis	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.1	1.2	2.1
Ethanolic fermentation	0.0	0.1	0.0	0.2	0.0	0.0	0.0	7.2	13.1	11.0
Aldehyde dehydrogenases	0.0	1.9	7.7	15.2	7.3	8.7	8.2	2.1	2.1	1.8
Sucrose degradation	1.1	21.6	0.9	2.9	1.1	0.2	12.4	8.3	10.6	14.6
Starch synthesis	0.0	0.0	0.3	0.0	0.0	0.1	0.0	0.5	0.1	0.2
Fatty acid synthesis	1.9	1.1	5.0	6.6	4.2	4.4	7.4	6.4	7.3	9.2
Lipid degradation	0.6	3.0	0.1	0.6	0.0	0.0	1.2	2.4	2.4	5.6
N-metabolism	1.2	8.5	0.8	0.2	0.3	0.3	4.8	10.0	14.4	9.2
Amino acid synthesis	74.1	121.1	25.8	27.1	28.7	15.6	37.1	55.9	66.1	74.4
Amino acid degradation	4.5	9.3	3.5	5.9	4.4	2.6	4.7	11.0	15.9	16.8
Secondary metabolism	5.2	16.0	3.4	13.2	3.8	1.7	11.9	2.3	4.4	6.3
Hormone metabolism	0.3	9.6	1.2	1.5	1.3	1.6	1.9	2.0	3.4	2.3
C1 metabolism	3.6	10.0	3.3	4.4	3.4	2.1	3.3	4.1	8.0	10.8
Nucleotide metabolism	7.4	6.9	8.7	8.0	7.8	11.6	10.5	18.8	23.8	22.9
Polyamine metabolism	0.0	0.3	1.3	1.1	1.9	1.7	3.9	4.0	1.9	2.6
Cell wall	6.9	15.9	12.1	22.7	16.3	11.8	21.4	27.2	58.7	57.1
Gluco-, galacto- and mannosidases	0.4	5.8	1.2	1.7	2.8	0.3	8.7	1.3	0.2	2.2
RNA processing	3.8	1.3	3.7	3.9	4.5	3.4	3.0	1.5	0.8	1.8
Regulation of transcription	11.7	7.4	12.8	12.8	13.8	17.5	16.3	13.1	6.5	7.7
RNA binding proteins	6.1	0.3	9.7	7.7	9.1	8.7	7.6	1.8	2.6	1.2
Chromatine structure	122.3	71.7	87.2	77.0	82.6	116.6	34.1	23.8	19.6	9.1
Amino acid activation	3.4	1.9	3.2	5.2	5.1	4.9	4.3	3.3	3.1	2.9
Ribosomal proteins	137.4	89.1	168.2	151.9	142.1	184.5	95.0	142.8	124.9	109.2
Protein synthesis non- ribosomal	55.2	40.8	48.5	50.6	52.7	46.9	57.8	61.0	43.4	46.8
Protein targeting	6.5	6.1	8.0	5.9	6.1	3.6	7.3	7.0	8.5	12.0
Protein modification	2.0	0.5	2.4	3.5	4.5	2.7	4.9	2.8	1.2	3.0
Protein folding	22.8	8.2	31.9	24.0	28.4	24.5	19.9	11.0	6.0	9.4
Protein degradation	38.4	41.6	34.2	38.0	38.9	37.1	72.7	35.8	21.6	30.9
Signaling	19.8	26.8	28.0	29.9	30.6	26.6	36.8	37.4	23.8	38.7
Cell organisation	54.9	65.5	55.8	50.4	62.8	64.3	32.5	44.1	73.3	58.4
Cell division	1.0	1.1	1.1	1.0	1.2	0.9	3.5	1.3	3.4	2.2
Cell cycle	5.6	4.2	6.9	7.4	6.0	10.8	8.9	11.6	8.2	6.4
Vesicle transport	2.6	4.8	1.5	1.5	0.9	1.7	1.3	2.7	6.2	7.8
Transport	22.8	49.4	24.7	18.5	18.6	15.6	31.1	34.2	31.9	33.0
Storage proteins	0.0	0.2	5.0	7.9	5.5	4.9	25.8	26.1	25.6	15.3
LEA proteins	0.0	0.7	0.0	0.0	0.0	0.0	0.2	0.3	6.9	0.4
Heat stress	55.3	44.7	81.4	91.5	102.7	80.3	73.7	56.3	23.2	28.9
Cold stress	1.8	9.1	0.9	1.4	2.6	2.4	1.2	1.7	1.4	1.0
Redox	16.4	22.4	30.1	35.4	30.1	25.8	26.3	24.0	23.2	19.3
Other	39.1	92.8	80.6	89.8	82.9	89.1	112.9	49.4	52.0	60.2

FIG. 6. **Functional analysis.** Total NSAF scores of proteins grouped in functional groups. NSAF scores were averaged over three biological replicates and multiplied by 1000. Red indicates high abundance in relation to the other stages. Standard deviation can be assessed from supplemental Table S7. L, leaves; R, roots; A, microsporocytes; B, meiotic cells; C, tetrads; D, microspores; E, polarized microspores; F, binuclear pollen; G, desiccated pollen; H, pollen tubes.

(66), respectively. The synthesis of storage proteins starts during polarization of the microspore.

Many enzymes required for energy-consuming pollen tube growth are synthesized starting from the polarized microspore or binuclear pollen stage (stages E and F). These include enzymes required for ethanolic fermentation, which is performed by pollen tubes due to anoxia caused by rapid oxygen consumption during their growth (56). Furthermore, enzymes of sucrose, lipid, and amino acid degradation and proteins involved in cell wall metabolism and vesicle trafficking follow a similar expression pattern. These proteins are required to support pollen tube growth with sufficient energy (30) and the machinery to deposit large amounts of cell wall and membrane at the tip of the growing pollen tube (67).

Interestingly, the abundance of proteins involved in anabolic pathways like gluconeogenesis and sucrose synthesis is also increased in the later stages, which is somewhat surprising, as sucrose can be taken up by the pollen tube from the surrounding tissue (68) and was also present in large amounts in the medium used for cultivation in this study.

Proteins associated with cell division showed increased abundance during the polarized microspore stage, but their levels were also increased in desiccated pollen, probably in preparation for the mitosis of the generative cell, which takes place after pollen tube germination. One subgroup of proteins (C5MQG8, FG636560, Q1G0Z1, TC141620) in this functional group, the cell cycle controlling CDC48 (69), has numerous functions (70), including spindle disassembly at the end of mitosis (71). The heterozygous *Arabidopsis* mutant *cdc48a* (72) displayed incomplete pollen germination of the mutated pollen. Our data suggest an additional role of CDC48 in pollen mitosis, which was not affected in the described mutant, the reason being most likely the presence of two other isoforms of CDC48a in the *Arabidopsis* genome (73).

Hot and cold temperature stresses can be detrimental to all phases of pollen development and have major effects on sexual reproduction. Heat shock proteins have been described as strongly abundant in pollen during the later stages of development in comparison to other tissues (74). The analysis of earlier stages including microsporocytes, meiotic cells, and tetrads in our study revealed an even greater abundance of proteins associated with heat stress.

A total of 67 heat-stress-associated proteins were identified in our study, including isoforms of the chaperones HSP 70 and HSP 90 and luminal binding proteins (supplemental Table S7). Taken together, these proteins constituted up to 10% of the total protein abundance according to our NSAF calculation (Fig. 6), highlighting the importance of these groups of proteins for pollen development, especially in early stages.

The functional comparison of the pollen stages with roots and leaves displayed a number of functional groups including ethanolic fermentation, polyamine metabolism, and late early abundant and storage proteins, which were almost not found in the sporophytic tissues and which highlight the highly specialized functionality of the developing pollen (Fig. 6). The comparison also displayed the previously described high rate of anabolic metabolism, especially in the late pollen stages. Also, the synthesis of fatty acids seemed to be much higher in pollen tubes than in leaves and roots, as the abundance of the related proteins was much higher. This shows that pollen tubes do not rely solely on previously synthesized and oil-body stored fatty acids and also require *de novo* synthesis to cope with the rapidly expanding membranes.

The rate of protein synthesis, in contrast, did not seem to be strongly enhanced in developing pollen or, especially, pollen tubes when taking into account the abundance of proteins associated with protein synthesis (ribosomal and nonribosomal). This observation once more supports the idea that growing pollen tubes rely strongly on presynthesized proteins.

Analysis with Respect to Previous Transcriptomics, Proteomics, and Genetic Studies—In order to find out to what extent protein and transcript levels differed, the data in this study were compared with expression data from a previous study (75). The microarrays of the latter study were based on transcripts obtained from mature tobacco pollen grains and pollen tubes.

Because different accessions were used in this and the previous study and in order to simplify the comparison, the transcripts and proteins were grouped according to their MapMan bins, and the individual values were added (supplemental Table S8).

From the comparison, it is apparent that the transcripts of proteins involved in signaling were much higher than the protein levels, maybe because of high turnover rates (Table IV). In contrast, protein levels of enzymes of the primary metabolism were much higher relative to their transcript levels (Table IV). This could be because the translation rate of these transcripts is much higher or the turnover rate of the proteins is lower. Another possibility is that the proteins are synthesized in earlier stages of pollen development and persist while the mRNA is degraded. Unfortunately, no expression data on developing tobacco pollen are available to date.

Transcript data on *Arabidopsis* stages ranging from unicellular microspores to pollen tubes have been previously generated (43, 76), and we compared our dataset to these transcript levels. Again, the transcripts and proteins were grouped according to their MapMan bins (supplemental Table S9). Once more it was apparent that the abundance of proteins of the primary metabolism was greater than the corresponding transcript levels. Additionally, they followed a different pattern. The enzymes phosphoglycerate kinase and pyruvate decarboxylase, for example (Fig. 7), showed their greatest abundance in desiccated pollen, whereas the corresponding *Arabidopsis* transcripts peaked much earlier and were completely abolished in mature pollen and pollen tubes, respectively. It could be speculated that this difference is simply due to the different analyzed species. The detection of both proteins in substantial amounts in a proteomic survey of mature

TABLE IV
Comparison of tobacco pollen transcript and protein data

Bin	Function	Transcript		Protein NSAF							
		G	H	A	B	C	D	E	F	G	H
1.3.3 + 4.1.11	3-phosphoglycerate kinase	0.2	0.2	12.1	8.9	10.4	13.0	11.7	18.2	20.8	16.9
4.1.8	Glyceraldehyde 3-phosphate dehydrogenase	0.9	0.5	27.8	17.9	31.0	31.3	20.4	40.8	33.2	23.9
5.2	Pyruvate decarboxylase	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.3	2.9	2.1
5.3	Alcohol dehydrogenase	0.5	0.2	0.0	0.4	0.0	0.0	0.0	13.4	22.6	17.7
30.3	Calcium signaling	30.7	31.4	8.5	11.5	11.5	7.6	5.8	2.6	5.8	3.5
30.5	G-proteins	20.2	23.6	8.4	5.9	7.5	6.1	9.0	10.3	5.0	11.8
20.2.1	Stress abiotic heat	6.4	7.1	142.7	158.8	174.4	160.5	132.6	96.5	38.1	45.7

Notes: Transcript data were previously published by Hafidh et al. (75). Transcripts and proteins were binned according to MapMan. The complete dataset can be surveyed in supplemental Table S8. Normalized values were multiplied by 1000. A, microsporocytes; B, meiotic cells; C, tetrads; D, microspores; E, polarized microspores; F, binuclear pollen; G, desiccated pollen; H, pollen tubes.

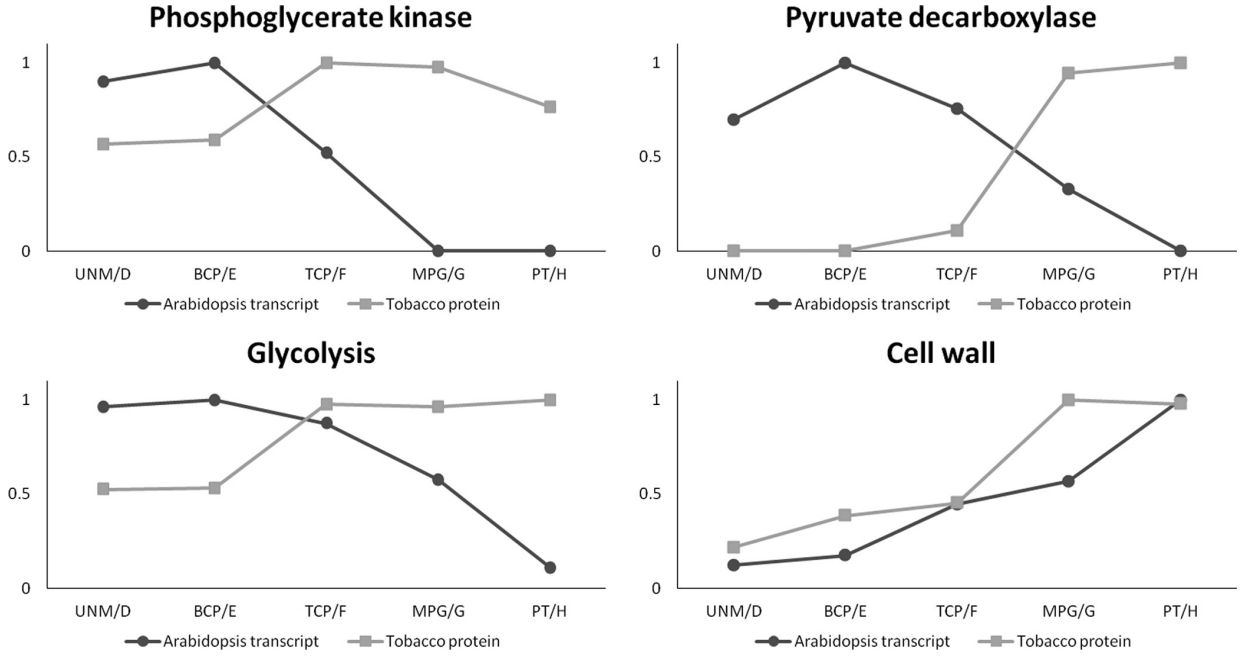


FIG. 7. **Protein and transcript comparison.** Tobacco protein abundances (NSAF scores over three replicates) and *Arabidopsis* transcript levels were grouped according to MapMan. All transcript data excluding pollen tubes were derived from Honys *et al.* (43). Pollen tube transcript data were derived from Wang *et al.* (76). Arabidopsis stages: UNM, unicellular microspores; BCP, bicellular pollen; TCP, tricellular pollen; MPG, mature pollen grains; PT, pollen tubes. Tobacco stages: D, microspores; E, polarized microspores; F, binuclear pollen; G, desiccated pollen; H, pollen tubes.

Arabidopsis pollen (14), however, makes a strong case that proteins are synthesized in earlier stages and persist though desiccation, rehydration, and pollen tube growth, by which time their transcripts are already degraded. This seems to be true for many enzymes of glycolysis, as the total protein and transcript abundance of this pathway follows a similar pattern. Another group that shows high protein levels in desiccated tobacco pollen and pollen tubes contains proteins associated with cell wall metabolism. Here, however, the *Arabidopsis* transcripts show a similar dynamic, maybe because the proteins in this group show a higher turnover rate and have to be resynthesized.

In order to be able to study the different dynamics of transcripts and proteins in better detail, it should be a future goal to generate either protein data from earlier *Arabidopsis* stages or transcript data throughout tobacco pollen development.

It must be concluded that the transcript levels in mature pollen can be very misleading when considering the importance of a specific gene for pollen tube growth or, even worse, pollen development, especially when only the transcript levels in mature pollen are considered, which is often the case (as, for example, in the commonly used open access version of GENEVESTIGATOR).

To find out how tobacco pollen might differ from *Arabidopsis* pollen, we compared the proteins in our study to the proteins found in the already mentioned proteomic survey of *Arabidopsis* (14), blasting all the sequences of the identified proteins from tobacco against the protein sequences from

Arabidopsis pollen. All matches with an E-value equal to or less than 10^{-10} were considered as homologues (supplemental Table S10).

Of the 3817 proteins in this study, only 1055 did not have a homologue in *Arabidopsis* pollen. Of the 1869 proteins considered for quantification, an even lower proportion (320 proteins) had no homologues in *Arabidopsis* pollen. Even though this indicates high similarity of the proteomes, there are some distinct differences.

The ortholog of *Arabidopsis* alcohol dehydrogenase (At1g77120), catalyzing the conversion from acetaldehyde to ethanol, was one of the enzymes with the greatest abundance in tobacco pollen tubes but was not found itself in *Arabidopsis* pollen. On the transcript level, this gene also showed only very weak expression in microspores and bicellular pollen and no expression in later stages. As the production of ethanol is one of the hallmarks of pollen tube metabolism in many species such as lily (77), tobacco, and petunia (56), this indicates a strong difference in primary metabolism between *Arabidopsis* and tobacco, probably based on the much shorter growing distance of *Arabidopsis* pollen tubes, which decreases the problem of anoxia. It is also possible that another enzyme takes this role in *Arabidopsis* pollen tubes, as several other proteins in *Arabidopsis* pollen showed a high similarity to the tobacco alcohol dehydrogenase. However, the protein with the highest similarity, ADH2, is already described as a glutathione reductase. It remains to be investigated whether *Arabidopsis* pollen tubes produce ethanol during growth.

Another possibility is that the acetaldehyde produced by pyruvate decarboxylase (which was found in *Arabidopsis* pollen) is directly converted to acetate and, later, acetyl-CoA, a pathway termed the pyruvate dehydrogenase bypass. An enzyme that is a strong candidate to perform the oxidation of acetaldehyde, aldehyde dehydrogenase (78), was detected in substantial amounts in *Arabidopsis*.

Another example of how the primary metabolism might differ is in the conversion of fructose-6-phosphate to fructose-1,6-bisphosphate. Whereas in *Arabidopsis* pollen only the phosphofructokinase was found, tobacco pollen and pollen tubes additionally contained several pyrophosphate-fructose-6-P phosphotransferase isoforms with a total abundance that was more than 8-fold greater than that of the phosphofructokinase (supplemental Table S9). This way, pyrophosphate can be used instead of ATP for the second activation step of glycolysis, increasing the ATP yield per hexose by one. Although this increase might not be so significant under aerobic conditions, it does make a big difference when ATP is generated via fermentation, which is the case in tobacco but might not be in *Arabidopsis*.

In *Arabidopsis*, many mutants are described that are affected in pollen development and pollen tube growth. After a survey of the literature, we updated a previously published list (14) of affected genes from 127 to 215 (supplemental Fig. S11). From these, we found 135 to have homologues (E-value equal to or less than 10^{-10}) in tobacco pollen. This supports the theory that most proteins with important functions in pollen development could be detected in tobacco pollen. However, from the 3817 proteins identified, only 320 homologues have been described so far in *Arabidopsis* mutant studies according to our literature survey, leaving tremendous room for future pollen research.

Post-translational Modifications—The identification of post-translational modifications was not the focus of this study, and the available material from early developmental stages was too limited for the enrichment of modified peptides. However, 655 potentially modified peptides were detected, including methylation of lysine and arginine; acetylation of lysine; and phosphorylation of serine, threonine, and tyrosine (supplemental Table S12). The protein with the greatest number of modifications was a homologue of elongation factor 1-α (Fig. 8). This protein is a member of the family of small G-proteins and serves a multitude of functions (79) including elongation of protein translation (80), regulation of the cytoskeleton (81, 82), and signaling (83–85). It has been previously shown to be methylated (86, 87) and acetylated (88). Multiple potential methylation sites were found; however, they must be considered with caution, as the mass of the peptide can be identical to a nonmodified peptide of a homologue protein, leading to ambiguous identifications (supplemental Table S12). This is an even bigger problem when the organism used in the study is not entirely sequenced, as potential ambiguous identifications might be missed because of the incomplete database. Also, methylated lysine and arginine residues lie at the N terminus of peptides cleaved by tryptic digest, making a confirmation of the modification based on the MS2 spectrum harder, as y-ions are too small to be detected. Therefore, only methylation sites can be considered as strong candidates (Fig. 8) if the modified amino acid lies in the middle of a miscleaved peptide and the MS2 spectrum shows the correct b- and y-ions of the modified amino acid.

The identification of additional modification sites and the study of their dynamics should be goals for the future. At least, the material available from the microspore stage on should be sufficient for metal oxide affinity chromatography enrichment of phosphopeptides, as has been performed recently for desiccated and activated pollen (28).

CONCLUSION

The comparative proteomic analysis of pollen development was, for the first time, extended to eight stages ranging from diploid microsporocytes to pollen tubes. In order to compare the data to results for sporophytic tissues, leaves and roots were also investigated, leading to the identification of a total of 4262 proteins.

Based on these data, pollen development can be divided into three phases (Fig. 9). The early phase that is still more closely related to leaves ranges from the microsporocytes to meiosis, extends to the formation of tetrads, and ends with

>TC141641

```
                                                  M           M                       A
                                                  D           D                       T
1   GIMGKEKVHI NIVVIGHVDS GKSTTTGHLI YKLGGIDKRV IERFEKEAAE MNKRSFKYAW VLDKLKAERE RGITIDIALW KFETTKYYCT VIDAPGHRDF

                                                                                      D
                                                                                      M
                                                                                M     T
101 IKNMITGTSQ ADCAVLIIDS TTGGFEAGIS KDGQTREHAL LAFTLGVKQM ICCCNKMDAT TPKYSKARYD EIVKEVSSYL KKVGYNPDKI PFVPISGFEG

                                                                M
201 DNMIERSTNL DWYKGPTLLE ALDQINEPKR PSDKPLRLPL QDVYKIGGIG TVPVGRVETG VLKPGMVVTF GPTGLTTEVK SVEMHHEALQ EALPGDMLGL

301 TLRMLLLRIS SVVLLPQTPR MTQPREHPTS PPRSSS
```

FIG. 8. **Post-translational modifications observed in a homologue of translation elongation factor 1-á (TC 141641).** Sequences highlighted in green were identified via LC-MS/MS analysis. Modification sites are marked above the sequence. M, methylation; D, dimethylation; T, trimethylation.

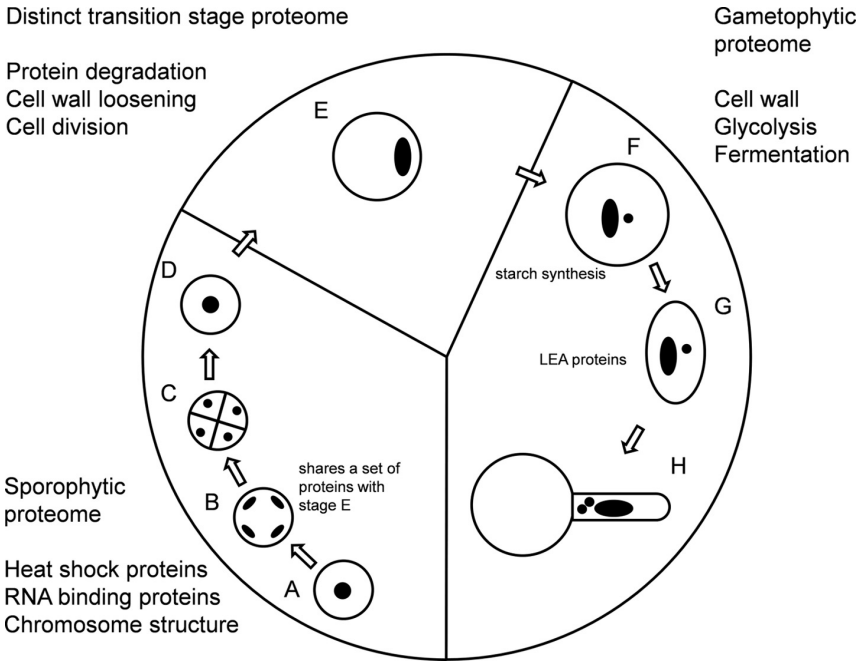


FIG. 9. **Key features of the tobacco pollen proteome through development.** A, microsporocytes; B, meiotic cells; C, tetrads; D, microspores; E, polarized microspores; F, binuclear pollen; G, desiccated pollen; H, pollen tubes.

the release of the microspores. The proteome of this phase is relatively static, with a high abundance of heat shock proteins to protect the cells in the process of meiosis and cell division. It appears that the "sporophytic proteome" synthesized in the microsporocytes is sustained throughout the development to early microspores.

The late phase ranges from binuclear pollen via desiccated pollen to pollen tubes and presents a "gametophytic proteome." Many proteins required for pollen tube growth, such as enzymes of the primary metabolism, and for cell wall synthesis are already produced prior to desiccation, so as to later allow a rapid outgrowth of the pollen tube without a bulk protein synthesis, as has also been previously observed (18, 24).

From the comparison of our protein data to *Arabidopsis* transcript data, it could also be concluded that many proteins, especially from the primary metabolism, do not seem to be further synthesized during pollen tube growth.

In between the early and the late phase, which are clearly very different in their cellular functionality, the pollen under-

goes an intermediate phase. During this polarized microspore stage, the "sporophytic proteome" is partially degraded, accompanied by ribosomal rearrangement and a strong increase in the abundance of proteins associated with protein degradation. However, this phase not only appears to be a turning point between the sporophytic and the gametophytic phase, but also seems to represent a phase on its own, because many proteins identified in this work were exclusively found during this stage.

Reasons for this could be the strong expansion in cell size, which is unique to this cell stage in pollen development, the performed polarization and asymmetric mitosis, and also the degradation of the sporophytic proteome, which would need distinct protein degradation machinery.

The great changes in the proteome observed during the three phases underlie the complexity of the protein networks required for male gametogenesis, which are just starting to get unraveled. As this work represents a first thorough proteomic map of pollen development, it could lay the base for a better understanding of these networks, especially of the early stages.

Acknowledgments—We thank EMBO for supporting Till Ischebeck with a long-term fellowship (EMBO-ALTF 299-2010). We thank Luis Recuenco-Munoz, Thomas Naegele, Stefanie Wienkoop, and Christiana Staudinger for helpful discussions. We are very grateful for the expert plant care by Thomas Joch and Andreas Schröfl. We also thank the PRIDE team for their great support.

* T.I. was supported by a long-term EMBO fellowship (EMBO-ALTF 299-2010)., L.V. was supported by the European Union through a Marie Curie IEF-grant (FP7-PEOPLE-2009-IEF-255109), and D.L. was funded by FWF Project Number P23441-B20.

S This article contains supplemental material.

§ To whom correspondence should be addressed: Univ.-Prof. Dr. Wolfram Weckwerth, Department of Molecular Systems Biology, Faculty of Life Sciences, University of Vienna, Austria, Althanstrasse 14, A-1090, Vienna, Austria. Tel.: 43-1-4277-76550; Fax: 43-1-4277-9577; E-mail: wolfram.weckwerth@univie.ac.at and Dr. Till Ischebeck, Department of Plant Biochemistry, Albrecht-von-Haller-Institute for Plant Sciences, Georg-August-University Göttingen, Justus-von-Liebig-Weg 11, 37077 Göttingen, Germany.

REFERENCES

1. Dai, S., and Chen, S. (2012) Single-cell-type proteomics: toward a holistic understanding of plant function. *Mol. Cell. Proteomics* **11,** 1622–1630
2. Zhao, Z., Zhang, W., Stanley, B. A., and Assmann, S. M. (2008) Functional proteomics of Arabidopsis thaliana guard cells uncovers new stomatal signaling pathways. *Plant Cell* **20,** 3210–3226
3. Zhu, M., Simons, B., Zhu, N., Oppenheimer, D. G., and Chen, S. (2010) Analysis of abscisic acid responsive proteins in Brassica napus guard cells by multiplexed isobaric tagging. *J. Proteomics* **73,** 790–805
4. Zhu, M., Dai, S., McClung, S., Yan, X., and Chen, S. (2009) Functional differentiation of Brassica napus guard cells and mesophyll cells revealed by comparative proteomics. *Mol. Cell. Proteomics* **8,** 752–766
5. Wienkoop, S., Zoeller, D., Ebert, B., Simon-Rosin, U., Fisahn, J., Glinski, M., and Weckwerth, W. (2004) Cell-specific protein profiling in Arabidopsis thaliana trichomes: identification of trichome-located proteins involved in sulfur metabolism and detoxification. *Phytochemistry* **65,** 1641–1649
6. Van Cutsem, E., Simonart, G., Degand, H., Faber, A. M., Morsomme, P., and Boutry, M. (2011) Gel-based and gel-free proteomic analysis of Nicotiana tabacum trichomes identifies proteins involved in secondary metabolism and in the (a)biotic stress response. *Proteomics* **11,** 440–454
7. Amme, S., Rutten, T., Melzer, M., Sonsmann, G., Vissers, J. P., Schlesier, B., and Mock, H. P. (2005) A proteome approach defines protective functions of tobacco leaf trichomes. *Proteomics* **5,** 2508–2518
8. Xie, Z., Kapteyn, J., and Gang, D. R. (2008) A systems biology investigation of the MEP/terpenoid and shikimate/phenylpropanoid pathways points to multiple levels of metabolic control in sweet basil glandular trichomes. *Plant J.* **54,** 349–361
9. Schilmiller, A. L., Miner, D. P., Larson, M., McDowell, E., Gang, D. R., Wilkerson, C., and Last, R. L. (2010) Studies of a biochemical factory: tomato trichome deep expressed sequence tag sequencing and proteomics. *Plant Physiol.* **153,** 1212–1223
10. Nestler, J., Schutz, W., and Hochholdinger, F. (2011) Conserved and unique features of the maize (Zea mays L.) root hair proteome. *J. Proteome Res.* **10,** 2525–2537
11. Brechenmacher, L., Lee, J., Sachdev, S., Song, Z., Nguyen, T. H., Joshi, T., Oehrle, N., Libault, M., Mooney, B., Xu, D., Cooper, B., and Stacey, G. (2009) Establishment of a protein reference map for soybean root hair cells. *Plant Physiol.* **149,** 670–682
12. Wan, J., Torres, M., Ganapathy, A., Thelen, J., DaGue, B. B., Mooney, B., Xu, D., and Stacey, G. (2005) Proteomic analysis of soybean root hairs after infection by Bradyrhizobium japonicum. *Mol. Plant Microbe Interact.* **18,** 458–467
13. Okamoto, T., Higuchi, K., Shinkawa, T., Isobe, T., Lorz, H., Koshiba, T., and Kranz, E. (2004) Identification of major proteins in maize egg cells. *Plant Cell Physiol.* **45,** 1406–1412
14. Grobei, M. A., Qeli, E., Brunner, E., Rehrauer, H., Zhang, R., Roschitzki, B., Basler, K., Ahrens, C. H., and Grossniklaus, U. (2009) Deterministic protein inference for shotgun proteomics data provides new insights into Arabidopsis pollen development and function. *Genome Res.* **19,** 1786–1800
15. Zou, J., Song, L., Zhang, W., Wang, Y., Ruan, S., and Wu, W. H. (2009) Comparative proteomic analysis of Arabidopsis mature pollen and germinated pollen. *J. Integr. Plant Biol.* **51,** 438–455
16. Holmes-Davis, R., Tanaka, C. K., Vensel, W. H., Hurkman, W. J., and McCormick, S. (2005) Proteome mapping of mature pollen of Arabidopsis thaliana. *Proteomics* **5,** 4864–4884
17. Han, B., Chen, S., Dai, S., Yang, N., and Wang, T. (2010) Isobaric tags for relative and absolute quantification-based comparative proteomics reveals the features of plasma membrane-associated proteomes of pollen grains and pollen tubes from Lilium davidii. *J. Integr. Plant Biol.* **52,** 1043–1058
18. Pertl, H., Schulze, W. X., and Obermeyer, G. (2009) The pollen organelle membrane proteome reveals highly spatial-temporal dynamics during germination and tube growth of lily pollen. *J. Proteome Res.* **8,** 5142–5152
19. Sheoran, I. S., Ross, A. R., Olson, D. J., and Sawhney, V. K. (2007) Proteomic analysis of tomato (Lycopersicon esculentum) pollen. *J. Exp. Bot.* **58,** 3525–3535
20. Lopez-Casado, G., Covey, P. A., Bedinger, P. A., Mueller, L. A., Thannhauser, T. W., Zhang, S., Fei, Z., Giovannoni, J. J., and Rose, J. K. (2012) Enabling proteomic studies with RNA-Seq: the proteome of tomato pollen as a test case. *Proteomics* **12,** 761–774
21. Dai, S., Li, L., Chen, T., Chong, K., Xue, Y., and Wang, T. (2006) Proteomic analyses of Oryza sativa mature pollen reveal novel proteins associated with pollen germination and tube growth. *Proteomics* **6,** 2504–2529
22. Dai, S., Chen, T., Chong, K., Xue, Y., Liu, S., and Wang, T. (2007) Proteomics identification of differentially expressed proteins associated with pollen germination and tube growth reveals characteristics of germinated Oryza sativa pollen. *Mol. Cell. Proteomics* **6,** 207–230
23. Wu, X., Chen, T., Zheng, M., Chen, Y., Teng, N., Samaj, J., Baluska, F., and Lin, J. (2008) Integrative proteomic and cytological analysis of the effects of extracellular Ca(2+) influx on Pinus bungeana pollen tube development. *J. Proteome Res.* **7,** 4299–4312
24. Fernando, D. D. (2005) Characterization of pollen tube development in Pinus strobus (Eastern white pine) through proteomic analysis of differentially expressed proteins. *Proteomics* **5,** 4917–4926
25. Chen, T., Wu, X., Chen, Y., Li, X., Huang, M., Zheng, M., Baluska, F., Samaj,

J., and Lin, J. (2009) Combined proteomic and cytological analysis of Ca2+-calmodulin regulation in Picea meyeri pollen tube growth. *Plant Physiol.* **149,** 1111–1126
26. Chen, Y., Chen, T., Shen, S., Zheng, M., Guo, Y., Lin, J., Baluska, F., and Samaj, J. (2006) Differential display proteomic analysis of Picea meyeri pollen germination and pollen-tube growth after inhibition of actin polymerization by latrunculin B. *Plant J.* **47,** 174–195
27. Valero Galvan, J., Valledor, L., Gonzalez Fernandez, R., Navarro Cerrillo, R. M., and Jorrin-Novo, J. V. (2012) Proteomic analysis of Holm oak (Quercus ilex subsp. ballota [Desf.] Samp.) pollen. *J. Proteomics* **75,** 2736–2744
28. Fila, J., Matros, A., Radau, S., Zahedi, R. P., Capkova, V., Mock, H. P., and Honys, D. (2012) Revealing phosphoproteins playing role in tobacco pollen activated in vitro. *Proteomics* **12,** 3229–3250
29. Becker, J. D., Boavida, L. C., Carneiro, J., Haury, M., and Feijo, J. A. (2003) Transcriptional profiling of Arabidopsis tissues reveals the unique characteristics of the pollen transcriptome. *Plant Physiol.* **133,** 713–725
30. Tadege, M., and Kuhlemeier, C. (1997) Aerobic fermentation during tobacco pollen development. *Plant Mol. Biol.* **35,** 343–354
31. Helling, D., Possart, A., Cottier, S., Klahre, U., and Kost, B. (2006) Pollen tube tip growth depends on plasma membrane polarization mediated by tobacco PLC3 activity and endocytic membrane recycling. *Plant Cell* **18,** 3519–3534
32. Ischebeck, T., Stenzel, I., and Heilmann, I. (2008) Type B phosphatidylinositol-4-phosphate 5-kinases mediate Arabidopsis and Nicotiana tabacum pollen tube growth by regulating apical pectin secretion. *Plant Cell* **20,** 3312–3330
33. Ischebeck, T., Stenzel, I., Hempel, F., Jin, X., Mosblech, A., and Heilmann, I. (2011) Phosphatidylinositol-4,5-bisphosphate influences Nt-Rac5-mediated cell expansion in pollen tubes of Nicotiana tabacum. *Plant J.* **65,** 453–468
34. Ischebeck, T., Vu, L. H., Jin, X., Stenzel, I., Lofke, C., and Heilmann, I. (2010) Functional cooperativity of enzymes of phosphoinositide conversion according to synergistic effects on pectin secretion in tobacco pollen tubes. *Mol. Plant* **3,** 870–881
35. Kost, B. (2008) Spatial control of Rho (Rac-Rop) signaling in tip-growing plant cells. *Trends Cell Biol.* **18,** 119–127
36. Dowd, P. E., Coursol, S., Skirpan, A. L., Kao, T. H., and Gilroy, S. (2006) Petunia phospholipase c1 is involved in pollen tube growth. *Plant Cell* **18,** 1438–1453
37. Berger, F., and Twell, D. (2011) Germline specification and function in plants. *Annu. Rev. Plant Biol.* **62,** 461–484
38. Boavida, L. C., Becker, J. D., and Feijo, J. A. (2005) The making of gametes in higher plants. *Int. J. Dev. Biol.* **49,** 595–614
39. Echlin, P., and Godwin, H. (1968) The ultrastructure and ontogeny of pollen in Helleborus foetidus L. II. Pollen grain development through the callose special wall stage. *J. Cell Sci.* **3,** 175–186
40. Kerim, T., Imin, N., Weinman, J. J., and Rolfe, B. G. (2003) Proteome analysis of male gametophyte development in rice anthers. *Proteomics* **3,** 738–751
41. Imin, N., Kerim, T., Weinman, J. J., and Rolfe, B. G. (2001) Characterisation of rice anther proteins expressed at the young microspore stage. *Proteomics* **1,** 1149–1161
42. Chaturvedi, P., Ischebeck, T., Egelhofer, V., Lichtscheidl, I., and Weckwerth, W. (2013) Cell-specific analysis of the tomato pollen proteome from pollen mother cell to mature pollen provides evidence for developmental priming. *J. Proteome Res.* **12,** 4892–4903
43. Honys, D., and Twell, D. (2004) Transcriptome analysis of haploid male gametophyte development in Arabidopsis. *Genome Biol.* **5,** R85
44. Whittle, C. A., Malik, M. R., Li, R., and Krochko, J. E. (2010) Comparative transcript analyses of the ovule, microspore, and mature pollen in Brassica napus. *Plant Mol. Biol.* **72,** 279–299
45. Wei, L. Q., Xu, W. Y., Deng, Z. Y., Su, Z., Xue, Y., and Wang, T. (2010) Genome-scale analysis and comparison of gene expression profiles in developing and germinated pollen in Oryza sativa. *BMC Genomics* **11,** 338
46. Read, S. M., Clarke, A. E., and Bacic, A. (1993) Stimulation of growth of cultured Nicotiana-Tabacum W-38 pollen tubes by poly(ethylene glycol) and Cu-(Ii) salts. *Protoplasma* **177,** 1–14
47. Smith, P. K., Krohn, R. I., Hermanson, G. T., Mallia, A. K., Gartner, F. H., Provenzano, M. D., Fujimoto, E. K., Goeke, N. M., Olson, B. J., and Klenk, D. C. (1985) Measurement of protein using bicinchoninic acid. *Anal. Biochem.* **150,** 76–85
48. Valledor, L., and Weckwerth, W. (2013) An improved detergent-compatible gel-fractionation LC-LTQ-Orbitrap workflow for plant and microbial proteomics. *Methods Mol. Biol*. **1072,** 347–358
49. Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V., and Mann, M. (2007) In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* **1,** 2856–2860
50. Ishihama, Y., Rappsilber, J., and Mann, M. (2006) Modular stop and go extraction tips with stacked disks for parallel and multidimensional peptide fractionation in proteomics. *J. Proteome Res.* **5,** 988–994
51. Valledor, L., Recuenco-Munoz, L., Egelhofer, V., Wienkoop, S., and Weckwerth, W. (2012) The different proteomes of Chlamydomonas reinhardtii. *J. Proteomics* **75,** 5883–5887
52. Paoletti, A. C., Parmely, T. J., Tomomori-Sato, C., Sato, S., Zhu, D., Conaway, R. C., Conaway, J. W., Florens, L., and Washburn, M. P. (2006) Quantitative proteomic analysis of distinct mammalian Mediator complexes using normalized spectral abundance factors. *Proc. Natl. Acad. Sci. U.S.A.* **103,** 18928–18933
53. Sun, X., and 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
54. Staudinger, C., Mehmeti, V., Turetschek, R., Lyon, D., Egelhofer, V., and Wienkoop, S. (2012) Possible role of nutritional priming for early salt and drought stress responses in Medicago truncatula. *Front. Plant Sci.* **3,** 285
55. Vizcaino, J. A., Cote, R. G., Csordas, A., Dianes, J. A., Fabregat, A., Foster, J. M., Griss, J., Alpi, E., Birim, M., Contell, J., O'Kelly, G., Schoenegger, A., Ovelleiro, D., Perez-Riverol, Y., Reisinger, F., Rios, D., Wang, R., and Hermjakob, H. (2013) The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* **41,** D1063–D1069
56. Bucher, M., Brander, K. A., Sbicego, S., Mandel, T., and Kuhlemeier, C. (1995) Aerobic fermentation in tobacco pollen. *Plant Mol. Biol.* **28,** 739–750
57. Peters, J. M. (2002) The anaphase-promoting complex: proteolysis in mitosis and beyond. *Mol. Cell* **9,** 931–943
58. Yang, M., Hu, Y., Lodhi, M., McCombie, W. R., and Ma, H. (1999) The Arabidopsis SKP1-LIKE1 gene is essential for male meiosis and may control homologue separation. *Proc. Natl. Acad. Sci. U.S.A.* **96,** 11416–11421
59. Zhao, D. Z., Yang, X. H., Quan, L., Timofejeva, L., Rigel, N. W., Ma, H., and Makaroff, C. A. (2006) ASK1, a SKP1 homolog, is required for nuclear reorganization, presynaptic homolog juxtaposition and the proper distribution of cohesin during meiosis in Arabidopsis. *Plant Mol. Biol.* **62,** 99–110
60. del Pozo, J. C., Boniotti, M. B., and Gutierrez, C. (2002) Arabidopsis E2Fc functions in cell division and is degraded by the ubiquitin-SCFAtSKP2 pathway in response to light. *Plant Cell* **14,** 3057–3071
61. Hruba, P., Honys, D., Twell, D., Capkova, V., and Tupy, J. (2005) Expression of beta-galactosidase and beta-xylosidase genes during microspore and pollen development. *Planta* **220,** 931–940
62. Konopka-Postupolska, D., Clark, G., and Hofmann, A. (2011) Structure, function and membrane interactions of plant annexins: an update. *Plant Sci.* **181,** 230–241
63. Dobritsa, A. A., Lei, Z., Nishikawa, S., Urbanczyk-Wochniak, E., Huhman, D. V., Preuss, D., and Sumner, L. W. (2010) LAP5 and LAP6 encode anther-specific proteins with similarity to chalcone synthase essential for pollen exine development in Arabidopsis. *Plant Physiol.* **153,** 937–955
64. Hanin, M., Brini, F., Ebel, C., Toda, Y., Takeda, S., and Masmoudi, K. (2011) Plant dehydrins and stress tolerance: versatile proteins for complex mechanisms. *Plant Signal. Behav.* **6,** 1503–1509
65. Wolkers, W. F., McCready, S., Brandt, W. F., Lindsey, G. G., and Hoekstra, F. A. (2001) Isolation and characterization of a D-7 LEA protein from pollen that stabilizes glasses in vitro. *Biochim. Biophys. Acta* **1544,** 196–206
66. Pacini, E., Guarnieri, M., and Nepi, M. (2006) Pollen carbohydrates and water content during development, presentation, and dispersal: a short review. *Protoplasma* **228,** 73–77
67. Franklin-Tong, V. E. (1999) Signaling and the modulation of pollen tube growth. *Plant Cell* **11,** 727–738

68. Jansen, M. A. K., Sessa, G., Malkin, S., and Fluhr, R. (1992) Pepc-mediated carbon fixation in transmitting tract cells reflects style pollen-tube interactions. *Plant J.* **2,** 507–515
69. Moir, D., Stewart, S. E., Osmond, B. C., and Botstein, D. (1982) Cold-sensitive cell-division-cycle mutants of yeast: isolation, properties, and pseudoreversion studies. *Genetics* **100,** 547–563
70. Yamanaka, K., Sasagawa, Y., and Ogura, T. (2012) Recent advances in p97/VCP/Cdc48 cellular functions. *Biochim. Biophys. Acta* **1823,** 130–137
71. Cao, K., Nakajima, R., Meyer, H. H., and Zheng, Y. X. (2003) The AAA-ATPase Cdc48/p97 regulates spindle disassembly at the end of mitosis. *Cell* **115,** 355–367
72. Park, S., Rancour, D. M., and Bednarek, S. Y. (2008) In planta analysis of the cell cycle-dependent localization of AtCDC48A and its critical roles in cell division, expansion, and differentiation. *Plant Physiol.* **148,** 246–258
73. Rancour, D. M., Dickey, C. E., Park, S., and Bednarek, S. Y. (2002) Characterization of AtCDC48. Evidence for multiple membrane fusion mechanisms at the plane of cell division in plants. *Plant Physiol.* **130,** 1241–1253
74. Mascarenhas, J. P., and Crone, D. F. (1996) Pollen and the heat shock response. *Sex. Plant Reprod.* **9,** 370–374
75. Hafidh, S., Breznenova, K., Ruzicka, P., Fecikova, J., Capkova, V., and Honys, D. (2012) Comprehensive analysis of tobacco pollen transcriptome unveils common pathways in polar cell expansion and underlying heterochronic shift during spermatogenesis. *BMC Plant Biol.* **12,** 24
76. Wang, Y., Zhang, W. Z., Song, L. F., Zou, J. J., Su, Z., and Wu, W. H. (2008) Transcriptome analyses show changes in gene expression to accompany pollen germination and tube growth in Arabidopsis. *Plant Physiol.* **148,** 1201–1211
77. Obermeyer, G., Fragner, L., Lang, V., and Weckwerth, W. (2013) Dynamic adaption of metabolic pathways during germination and growth of lily pollen tubes after inhibition of the electron transport chain. *Plant Physiol.* **162,** 1822–1833
78. Wei, Y., Lin, M., Oliver, D. J., and Schnable, P. S. (2009) The roles of aldehyde dehydrogenases (ALDHs) in the PDH bypass of Arabidopsis. *BMC Biochem.* **10,** 7
79. Ransom-Hodgkins, W. D. (2009) The application of expression analysis in elucidating the eukaryotic elongation factor one alpha gene family in Arabidopsis thaliana. *Mol. Genet. Genomics* **281,** 391–405
80. Andersen, G. R., Nissen, P., and Nyborg, J. (2003) Elongation factors in protein biosynthesis. *Trends Biochem. Sci.* **28,** 434–441
81. Demma, M., Warren, V., Hock, R., Dharmawardhane, S., and Condeelis, J. (1990) Isolation of an abundant 50,000-dalton actin filament bundling protein from Dictyostelium amoebae. *J. Biol. Chem.* **265,** 2286–2291
82. Ohta, K., Toriyama, M., Miyazaki, M., Murofushi, H., Hosoda, S., Endo, S., and Sakai, H. (1990) The mitotic apparatus-associated 51-kDa protein from sea urchin eggs is a GTP-binding protein and is immunologically related to yeast polypeptide elongation factor 1 alpha. *J. Biol. Chem.* **265,** 3240–3247
83. Yang, W., and Boss, W. F. (1994) Regulation of phosphatidylinositol 4-kinase by the protein activator PIK-A49. Activation requires phosphorylation of PIK-A49. *J. Biol. Chem.* **269,** 3852–3857
84. Chang, Y. W., and Traugh, J. A. (1998) Insulin stimulation of phosphorylation of elongation factor 1 (eEF-1) enhances elongation activity. *Eur. J. Biochem.* **251,** 201–207
85. Umikawa, M., Tanaka, K., Kamei, T., Shimizu, K., Imamura, H., Sasaki, T., and Takai, Y. (1998) Interaction of Rho1p target Bni1p with F-actin-binding elongation factor 1alpha: implication in Rho1p-regulated reorganization of the actin cytoskeleton in Saccharomyces cerevisiae. *Oncogene* **16,** 2011–2016
86. Zobel-Thropp, P., Yang, M. C., Machado, L., and Clarke, S. (2000) A novel post-translational modification of yeast elongation factor 1A—methylesterification at the C terminus. *J. Biol. Chem.* **275,** 37150–37158
87. Coppard, N. J., Clark, B. F. C., and Cramer, F. (1983) Methylation of elongation-factor 1-alpha in mouse 3t3b and 3t3b Sv40 cells. *FEBS Lett.* **164,** 330–334
88. Wu, X., Oh, M. H., Schwarz, E. M., Larue, C. T., Sivaguru, M., Imai, B. S., Yau, P. M., Ort, D. R., and Huber, S. C. (2011) Lysine acetylation is a widespread protein modification for diverse proteins in Arabidopsis. *Plant Physiol.* **155,** 1769–1778

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

The annotation ("identification") of metabolites remains the biggest bottleneck in untargeted LC/MS-based metabolomics studies. Plant secondary metabolites, such as Flavonoid derivatives, can be composed of composite metabolites, known sub-structures, e.g. Glycosylated Flavonoids, consisting of the Aglycon Flavonoid backbone as well as the sugar moiety (Matsuda, Yonekura-Sakakibara, et al. 2009; Tohge et al. 2005). Environmental stress, such as cold stress, can result in e.g. Reactive Oxygen Species (ROS), which can be scavenged in order prevent damage, thus producing chemical transformations (Apel and Hirt 2004). Furthermore, the study of Redox regulation and signaling is a promising field of plant biology (Hartl and Finkemeier 2012; Schmidtmann et al. 2014; Schwarzländer and Finkemeier 2013). A method to inspect relations of molecules, extract putative chemical transformations of compounds and to find putative pathways in high Mass Accuracy LC/MS metabolomics data was implemented as an algorithm, called mzGroupAnalyzer. This algorithm in available as part of the freely available Matlab tool-box called COVAIN (Sun and Weckwerth 2012). mzGroupAnalyzer generates multiple potential pathways directly from raw-LC/MS data and visualizes these as networks, additionally van Krevelen plots composed of untargeted metabolomics data can validate structural familiarity between compounds through a metabolic pattern.

Declaration of authorship

The results of this chapter are presented in the form of a manuscript published in the journal „PLOS One". I have provided a critical contribution to the following publication, though the largest part the work was performed by the coauthors. As stated in the author contributions: "David Lyon (D. L.) contributed reagents/materials/analysis tools and wrote the paper". Specifically, part of the data generated for the **Doerfler et al. 2013** publication (see Publications above) was used within this publication. Furthermore, I have refined Mass Spectrometry methodology (e.g. ramped CE, micro Scans, and MS/MS/MS scan events) which was implemented within this publication and I have assist-

ed in writing the manuscript.

Published manuscript

OPEN ACCESS Freely available online

PLOS ONE

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

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 μmol m^{-2} 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/H_2O 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/H_2O 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 putive 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.

3. Putative transformations that were not listed in the rules-file.
4. A mzStructure file for *Pathway viewer*.
5. 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

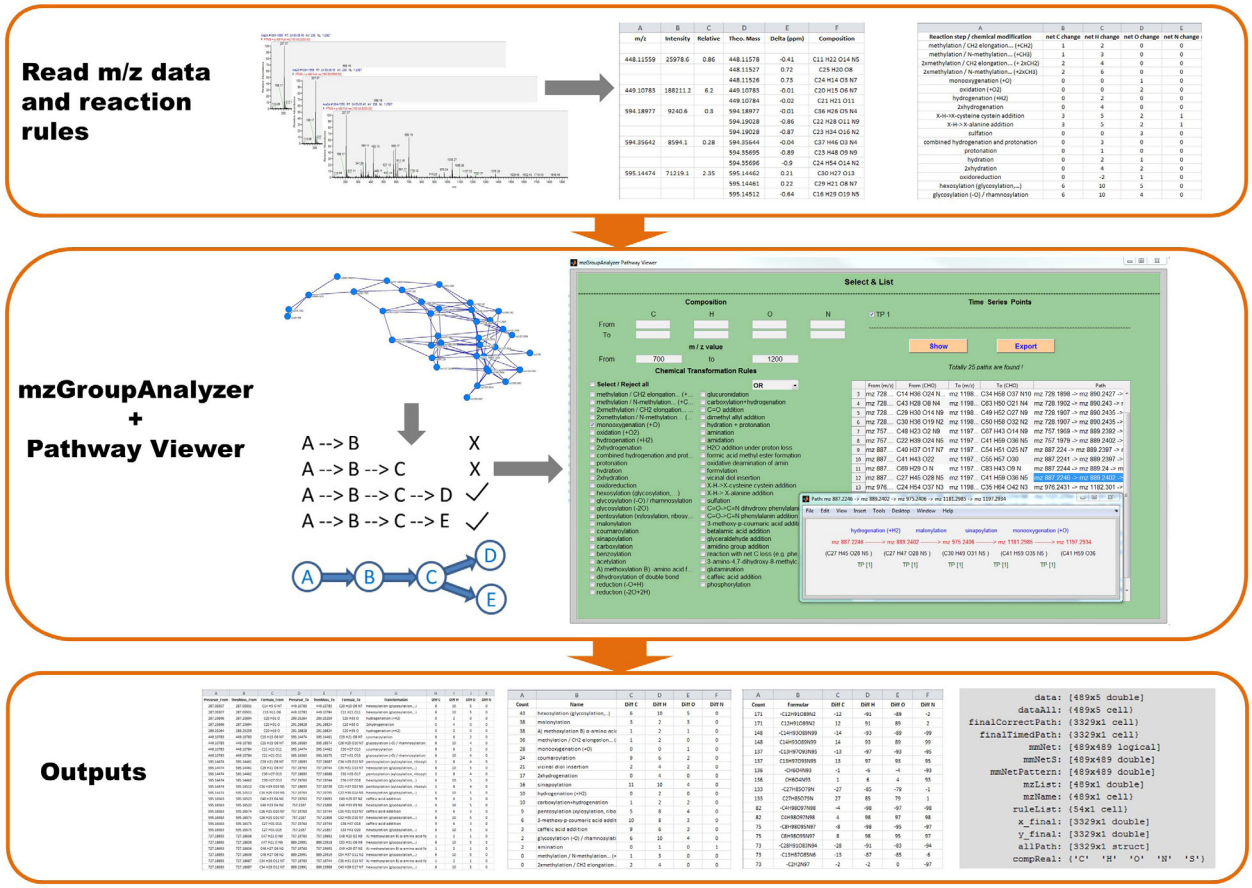


Figure 1. Scheme of the *mzGroupAnalyzer* and *Pathway Viewer* algorithm and GUI implementation. The program reads the *m/z* features which are extracted from Xcalibur, as well as the user predefined reaction rules. Then it finds transformations between all pairs of *m/z* features, and reports the frequency 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

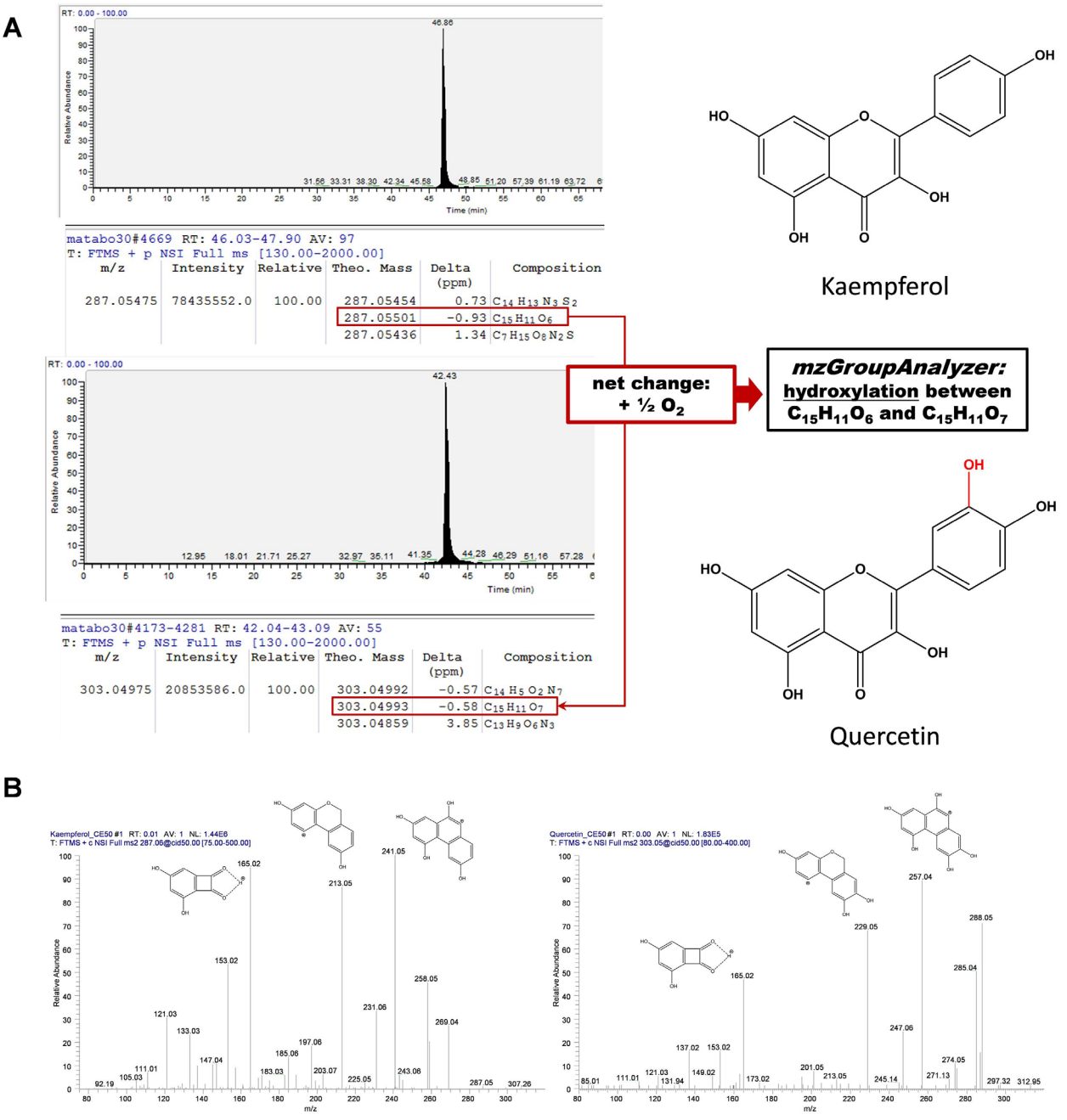


Figure 2. ***mzGroupAnalyzer*** **is able to detect possible metabolic steps out of various proposed sum formulas for a measured** ***m/z*** **feature. A** Kaempferol and Quercetin standards measured by LC-MS result into several sum formula suggestions for the measured mass-to-charge-ratio (*m/z*). Because 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** MS^2 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

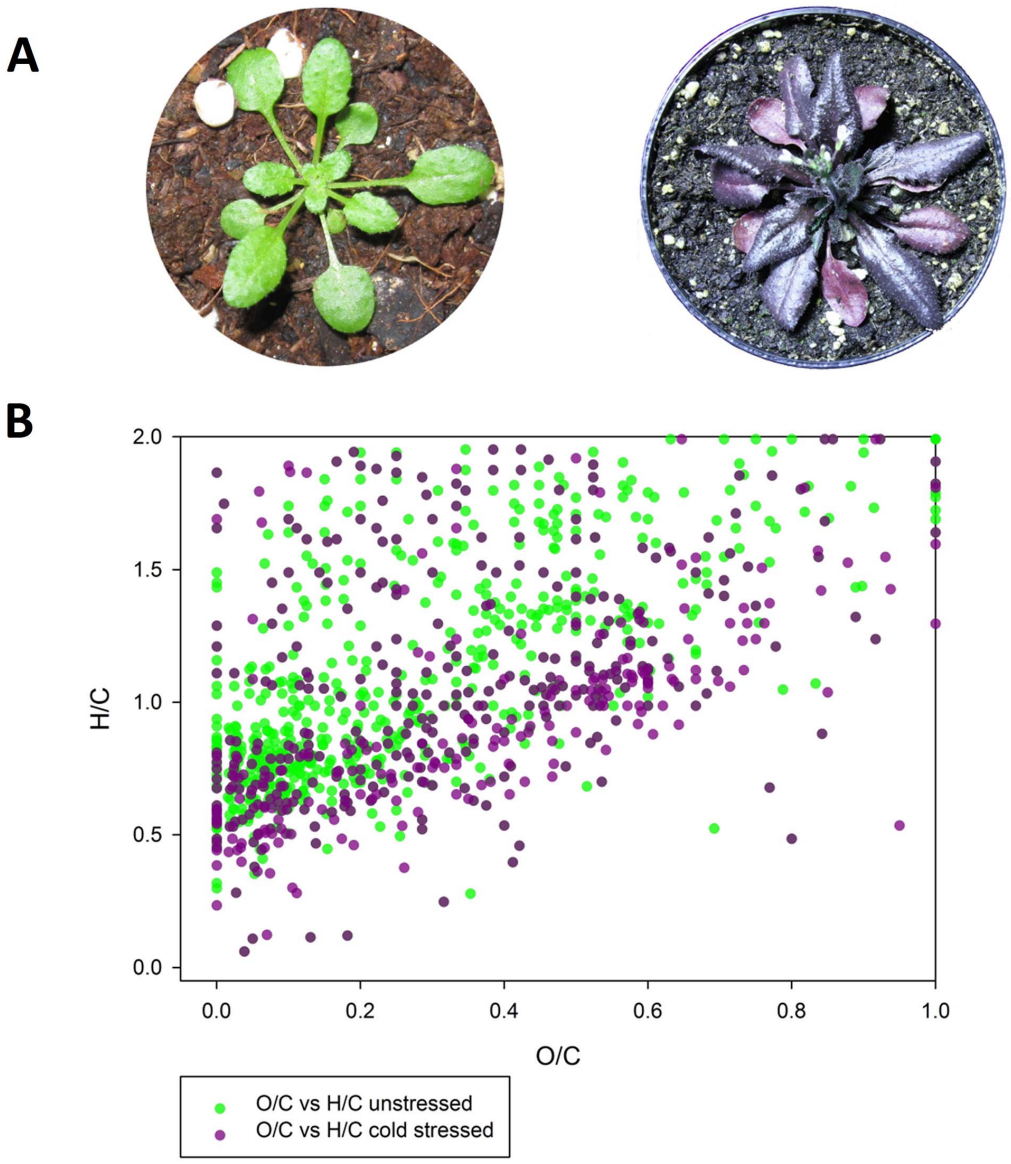


Figure 3. After oxidative stress the *Arabidopsis thaliana* plants turn from green into purple indicating a dramatic shift in metabolism, specifically elevated flavonoid biosynthesis involved in oxidative stress protection [6]**. A** Plants turns from green to purple under high light and cold temperature 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

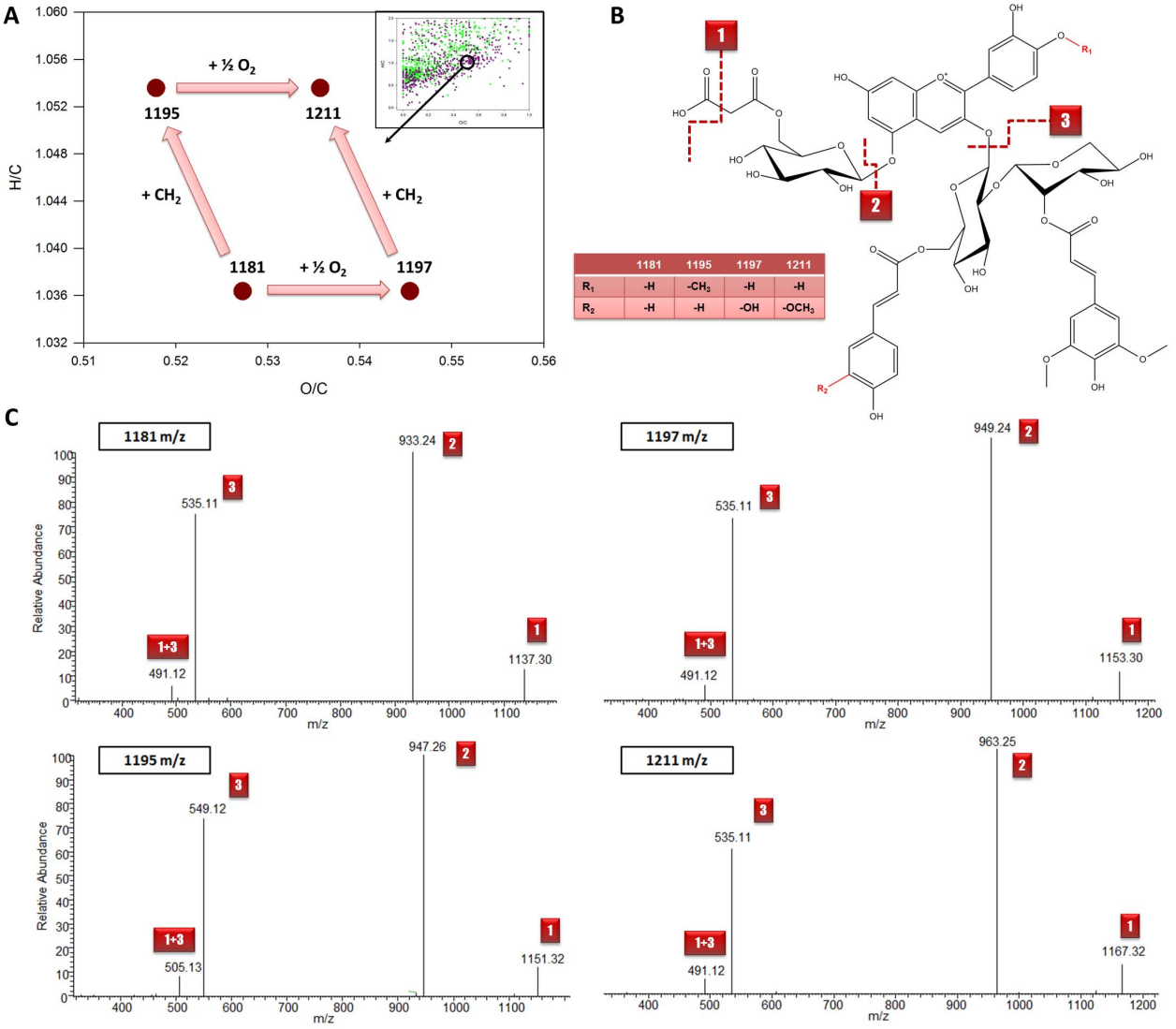


Figure 4. Exploration of the van Krevelen diagram created by sum formulas with chemical transformations detected by *mzGroupAnalyzer*. **A** *m/z* 1181, 1195, 1197 and 1211 are interconnected with net shifts of ½ O_2 and a CH_2 group and form a rhombic pattern. **B** Proposed fragmentation scheme of 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

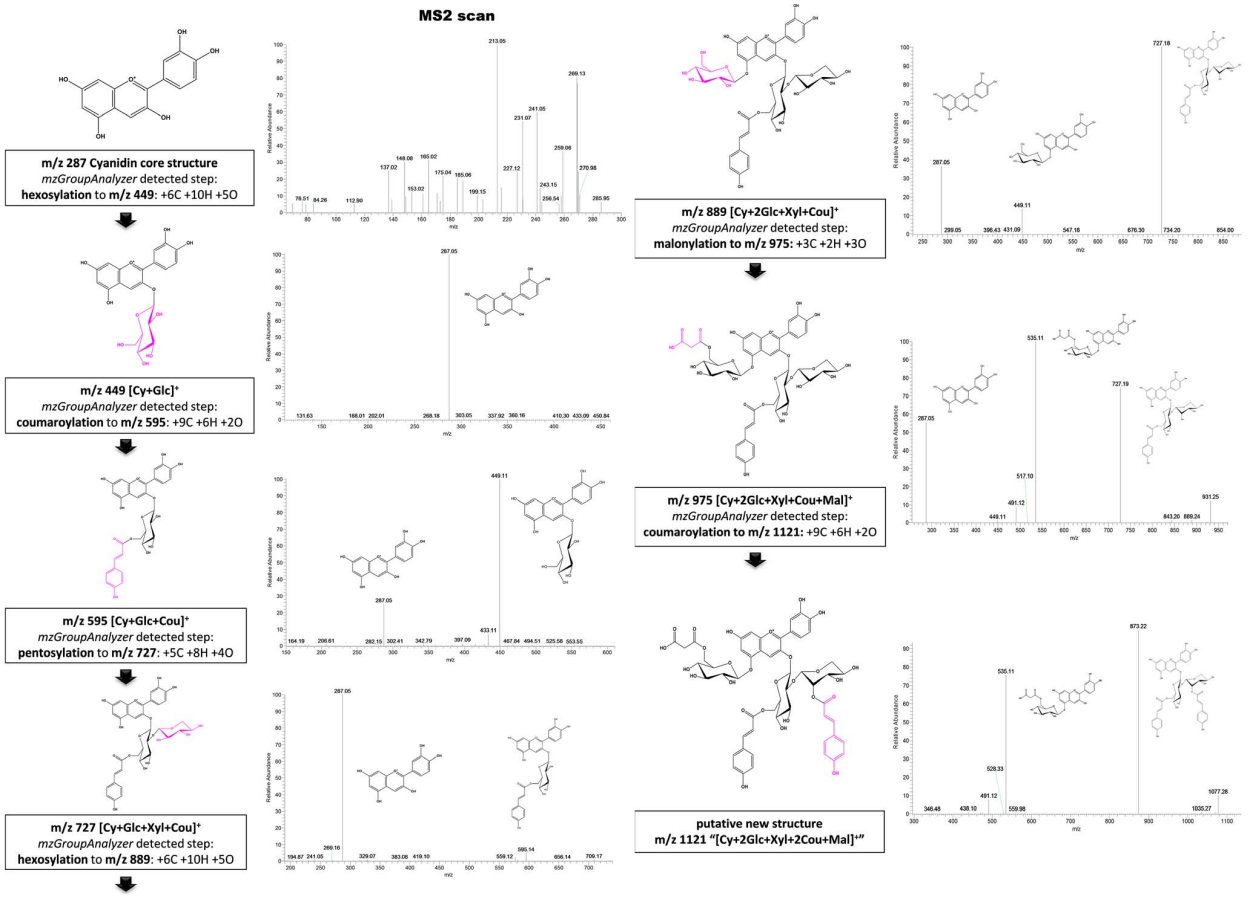


Figure 5. Identification of biochemical transformations of *in vivo* data using *mzGroupAnalyzer*. A metabolic pathway leading to a putative new compound *m/z* 1121 is revealed. Amongst several hundreds of other interlinked *m/z* values in the data, the figure shows metabolic transitions derived from sub-ppm accuracy measurements on the left side and their corresponding MS^2 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 MS^2 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 MS^2 product ion fragments.

name or *m/z* value	sum formula detected by *mzGroupAnalyzer*	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, oxido-reductions, 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 MS^2 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 MS^2 product ion spectra from the same data set. This pathway leads to

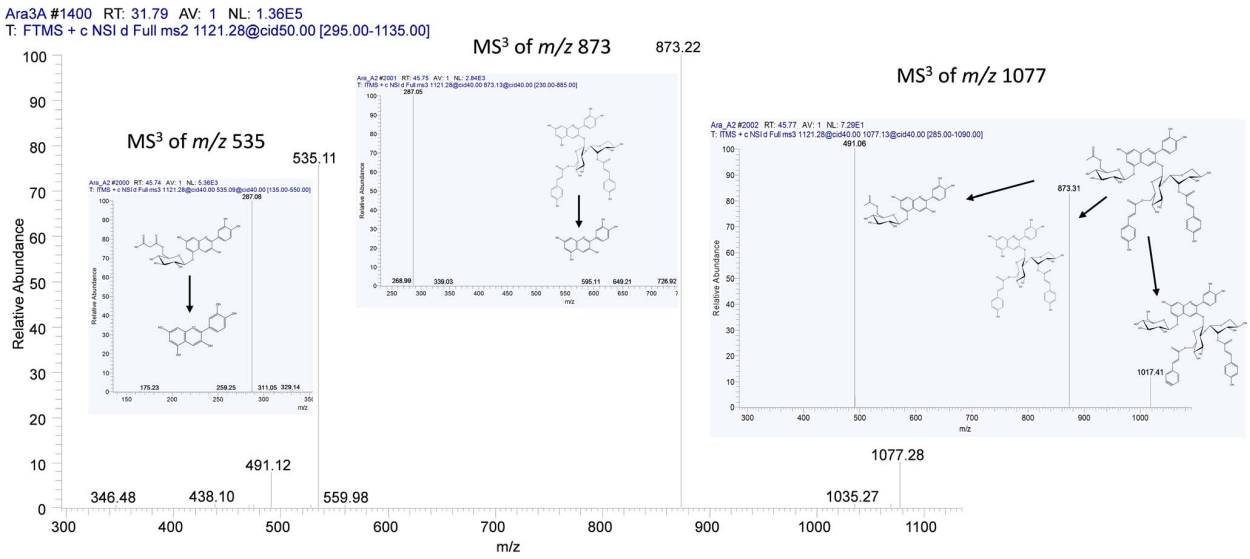


Figure 6. Structure validation of *m/z* 1121 by MS³ product ion scans. Both MS² fragments *m/z* 535 and 873 result in the core cyanidin structure by undergoing MS³ fragmentation. *m/z* 1077, the putatively decarboxylized form of 1121, yields *m/z* 491, as observed in the MS² spectrum already, by scission 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 MS² 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 MS² spectrum. To *m/z* 595, a 5-C sugar is then added to the already existing hexose group – judged from the MS² spectrum – forming compound *m/z* 727. In the MS² 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, MS² 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 MS² 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 MS² 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 MS² 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 MS² product ion scans supports the decarboxylation reaction of *m/z* 535. Figure 6 shows additional MS³ product ion scans of the isolated MS² product ions *m/z* 535, 873 and 1077. Both MS² peaks *m/z* 535 and 873 result in the core cyanidin structure *m/z* 287 by undergoing MS³ fragmentation. *m/z* 1077, the putatively decarboxylized form of *m/z* 1121, yields *m/z* 491, as observed in the MS² 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 MS² 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 MSⁿ scans has the power to validate known compounds and predict

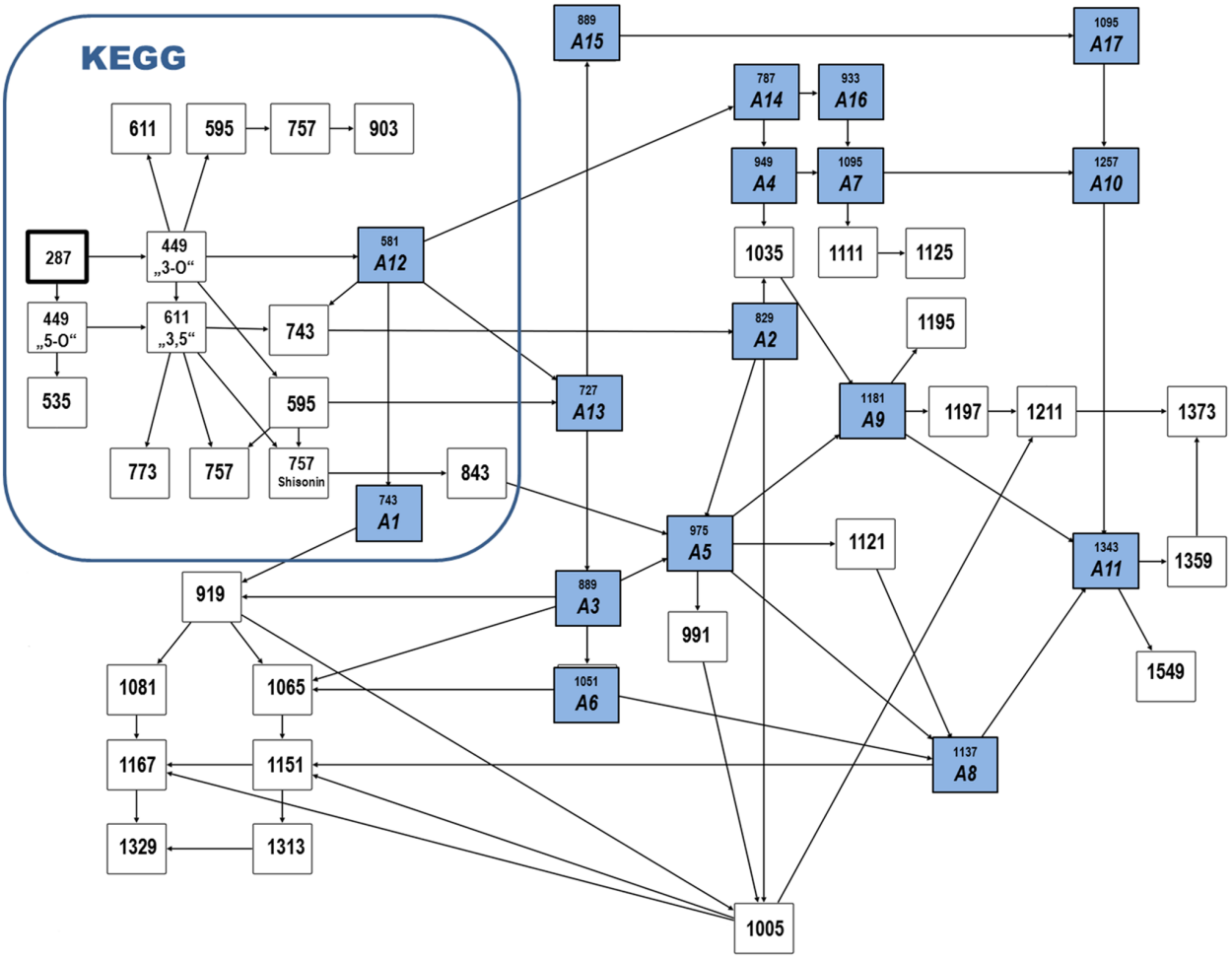


Figure 7. A proposed network of the detected anthocyanin family featuring putatively novel compounds as well as known structures including the KEGG pathway of anthocyanin biosynthesis [38,44]**.** Only compounds from the KEGG anthocyanin pathway are depicted, for which a suitable precursor mass was found 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.g007

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 data set, 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)

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.

References

1. Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN, et al. (2000) Metabolite profiling for plant functional genomics. Nature Biotechnology 18: 1157–1161.
2. Soga T (2007) Capillary electrophoresis-mass spectrometry for metabolomics. Methods Mol Biol 358: 129–137.
3. 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.
4. Scherling C, Roscher C, Giavalisco P, Schulze ED, Weckwerth W (2010) Metabolomics unravel contrasting effects of biodiversity on the performance of individual plant species. Plos One 5: e12569.
5. 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.
6. 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.
7. Dunn WB, Erban A, Weber RJM, Creek DJ, Brown M, et al. (2012) Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. Metabolomics.
8. 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.
9. Kind T, Fiehn O (2010) Advances in structure elucidation of small molecules using mass spectrometry. Bioanalytical reviews 2: 23–60.
10. Weckwerth W (2011) Green systems biology - From single genomes, proteomes and metabolomes to ecosystems research and biotechnology. Journal of Proteomics 75: 284–305.
11. Weckwerth W, Morgenthal K (2005) Metabolomics: from pattern recognition to biological interpretation. Drug Discov Today 10: 1551–1558.
12. 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.
13. Mittler R, Vanderauwera S, Gollery M, Van Breusegem F (2004) Reactive oxygen gene network of plants. Trends in plant science 9: 490–498.
14. Croteau R, Kutchan TM, Lewis NG (2000) Natural Products (Secondary Metabolites). Biochemistry and Molecular Biology of Plants: 1250–1318.
15. Lee KW, Lee HJ, Lee CY (2004) Vitamins, phytochemicals, diets, and their implementation in cancer chemoprevention. Crit Rev Food Sci Nutr 44: 437–452.
16. 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.
17. 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.
18. 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.
19. 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.
20. 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.
21. 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.
22. 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.
23. Asada K (2006) Production and scavenging of reactive oxygen species in chloroplasts and their functions. Plant physiology 141: 391–396.
24. Apel K, Hirt H (2004) Reactive oxygen species: metabolism, oxidative stress, and signal transduction. Annual review of plant biology 55: 373–399.
25. 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.
26. Seyoum A, Asres K, El-Fiky FK (2006) Structure-radical scavenging activity relationships of flavonoids. Phytochemistry 67: 2058–2070.
27. 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.
28. 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.
29. 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.
30. van Krevelen DW (1950) Graphical-statistical method for the study of structure and reaction processes of coal. Fuel: 269–284.
31. 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.
32. 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.
33. 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.
34. Giavalisco P, Hummel J, Lisec 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.
35. Rogers S, Scheltema RA, Girolami M, Breitling R (2009) Probabilistic assignment of formulas to mass peaks in metabolomics experiments. Bioinformatics 25: 512–518.
36. Kind T, Fiehn O (2007) Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC Bioinformatics 8: 105.
37. Karp PD, Riley M, Paley SM, Pellegrini-Toole A (2002) The MetaCyc Database. Nucleic acids research 30: 59–61.
38. Kanehisa M (2002) The KEGG database. Novartis Foundation symposium 247: 91–101; discussion 101–103, 119–128, 244–152.
39. 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.
40. Bloor SJ, Abrahams S (2002) The structure of the major anthocyanin in Arabidopsis thaliana. Phytochemistry 59: 343–346.
41. 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.
42. 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.
43. 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.
44. 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.
45. 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.
46. 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.

Using ProtMAX to create high-mass-accuracy precursor alignments from label-free quantitative mass spectrometry data generated in shotgun proteomics experiments

Various freely available, as well as commercial software for LC/MS data extraction exist. While commercial solutions are mostly easy to handle for the user, they often do not state the details of the applied methods, in order to protect their product. Freely available software, on the other hand, isn't always well-documented and often necessitates a more thorough understanding of the algorithms/filters to apply (specifically the order of filter application as well as which type of filter to apply, which can depend on the data at hand). Simple, robust, and fast solutions, capable of handling very large datasets, are needed.

Mass Accuracy Precursor Alignment (MAPA) is a freely available tool that offers the possibility to analyze and validate peptides without any previous sequence knowledge or dependency on a data base (Hoehenwarter et al. 2008; Egelhofer et al. 2013). Due to the high Mass Accuracy achieved by the "Orbitrap" Mass Spectrometer, the measured precursors can be identified and aligned with the ProtMAX algorithm, having the spectral counts or ion intensity counts of each precursor as a quantitative parameter. This means that the entire analysis can be performed on the peptide level, considerably increasing the amount of peptides detected. The database independency also permits the detection of yet unknown, modified, and mutated peptides that would not or only with considerable difficulty be recognizable through a database-dependent analysis; therefore, enabling the search for new putative biological markers. After de-novo-sequencing and through BLASTing against different databases, new evidence about cellular functions or analogies to similar proteins in other species can be derived.

Declaration of authorship

The results of this chapter are presented in the form of a manuscript published in the journal „Nature Protocols". I have provided a critical contribution to the following publication, though the largest part the work was performed by the coauthors. As stated in the publication: "David Lyon (D.L.) contributed to

the concept of the ProtMAX tool and writing of the manuscript“. Specifically, I contributed to the creation the figures, tested the software, and aided in writing the manuscript.

Published manuscript

Using ProtMAX to create high-mass-accuracy precursor alignments from label-free quantitative mass spectrometry data generated in shotgun proteomics experiments

Volker Egelhofer[1], Wolfgang Hoehenwarter[1,2], David Lyon[1], Wolfram Weckwerth[1] & Stefanie Wienkoop[1]

[1]Department of Molecular Systems Biology, University of Vienna, Vienna, Austria. [2]Present address: Proteome Analysis Research Group, Leibniz Institute for Plant Biochemistry, Halle, Germany. Correspondence should be addressed to S.W. (stefanie.wienkoop@univie.ac.at).

Published online 28 February 2013; doi:10.1038/nprot.2013.013

Recently, new software tools have been developed for improved protein quantification using mass spectrometry (MS) data. However, there are still limitations especially in high-sample-throughput quantification methods, and most of these relate to extensive computational calculations. The mass accuracy precursor alignment (MAPA) strategy has been shown to be a robust method for relative protein quantification. Its major advantages are high resolution, sensitivity and sample throughput. Its accuracy is data dependent and thus best suited for precursor mass-to-charge precision of ~1 p.p.m. This protocol describes how to use a software tool (ProtMAX) that allows for the automated alignment of precursors from up to several hundred MS runs within minutes without computational restrictions. It comprises features for 'ion intensity count' and 'target search' of a distinct set of peptides. This procedure also includes the recommended MS settings for complex quantitative MAPA analysis using ProtMAX (http://www.univie.ac.at/mosys/software.html).

INTRODUCTION

The genome-wide measurement of differences in protein abundance is the essence of proteomics. Thus far, the combination of liquid chromatography and mass spectrometry (LC-MS) has been indispensable for this endeavor. On-line or off-line multidimensional chromatography can achieve considerable separation of the several hundreds of thousands of peptides that are the result of enzymatic cleavage of the proteome. The latest generation of high-resolution, high-mass-accuracy mass spectrometers can resolve and acquire a substantial fraction of these peptide ions in a single 1D LC-MS experiment and target them for tandem mass spectrometry (MS/MS) analysis[1–3]. In the past decade, very accurate measurements of peptide and protein abundance have been achieved, despite the vastly differing electrospray ionization efficiencies of biomolecules[4–11]. As a result, state-of-the-art techniques have emerged, such as stable isotope dilution; metabolic labeling; and label-free quantitative measurements of proteomes of unicellular and multicellular organisms, as well as of their specific tissues[12–17].

To understand molecular mechanisms, it is often necessary to compare protein levels across cellular states. The inherent complexity of proteomics data requires statistical selection and testing to pinpoint significant features that may be of interest. In principle, shotgun proteomics approaches enable the detection of any protein in a given sample regardless of its concentration as long as the experiment is carried out a sufficiently large number of times[6,18]. High-quality proteomics profiling studies therefore often comprise tens to hundreds of shotgun proteomics analyses, thus producing hundreds of gigabytes to terabytes of data. This poses a considerable analytical challenge. Software must grasp and record the multivariate features of the data and the peptide ion signals, and it must map them to the data landscape for comparison. This will always come as a trade-off between the accuracy and comprehensiveness of feature detection and the computational cost.

Algorithms for analyzing MS data

Recently, new software tools have been developed for improved protein quantification using MS data. However, there are still limitations, especially in high-sample-throughput quantification methods, and most of these depend on extensive computational calculations in platforms such as MetAlign and MaxQuant[19,20]. We first published ProtMAX, a Windows Forms application in the Common Language Runtime (CLR) environment in 2008 (ref. 21). It was distinguished by feature detection with a single variable: the accurately measured peptide ion mass-to-charge ratio (m/z). This is the most informative feature parameter, and we used it to trace peptide ions in shotgun proteomics analyses for comparison. For the number of MS/MS spectra recorded for each m/z, we used the spectral count to quantify the peptide ions in the analysis. This concept, called mass accuracy precursor alignment (MAPA), proved to be very powerful for data-dependent MS combining feature detection and quantification at minimal cost, and thus made the comparison of nearly 200 shotgun proteomics analyses of 12 tissue states possible[22]. The MAPA approach was also successfully applied to the detection and quantification of *in vivo* phosphorylation sites[23]. The current version of ProtMAX can additionally be used for LC-MS/MS-based metabolomics[24,25].

Overview of ProtMAX

ProtMAX is a software tool that builds on the MAPA concept and includes several key features.

First, although the m/z measured with sub-p.p.m. error comes reasonably close to a unique definition of every tryptic peptide in the proteome of higher eukaryotes, it is overly simplistic. Indeed, it is not uncommon to observe two different baseline-separated peptide ion signals that share the same error-tolerated mass (m/z). ProtMAX provides a local retention time (Rt) window (Environment) in order to discriminate peptides that share the same error-tolerated mass (m/z)

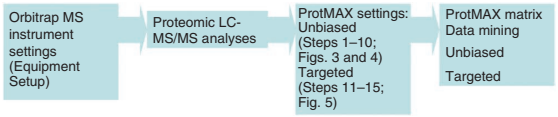


Figure 1 | Workflow diagram from MS analysis to a quantitative protein data matrix.

but that do not elute within an expected Rt window. Thus, peaks eluting outside of the expected local Rt window will be placed in a different 'bin'. Together with an absolute intensity–based noise filter, this is a simple yet effective means of incorporating the chromatographic Rt into the peptide definition without the necessity of complex peak detection algorithms and the accompanying marked increase in computational cost.

Spectral counting is a popular strategy for LC-MS/MS–based differential protein expression analyses[4–6]. It refers to the total number of MS/MS spectra assigned to a protein. We provide several options for expanding peptide quantification. In addition to counting dependent MS/MS spectra (spectral count), it is also possible to count the accurately measured precursor ions (m/z) of a given peak, again anchored by a dependent MS/MS spectrum. We call this the ion count. Here the peak is defined by the peak width setting (Environment Rt). The absolute signal intensity of the m/z can also be summed for intensity-based quantification, called ion intensity count.

These concepts in label-free quantification offer an increase in dynamic range and absolute signal. In a traditional spectral count approach, missing values can occur if the precursor has not been triggered for MS/MS in all replicates. As ion count and ion intensity count are independent of MS/MS spectral information, these features have a smaller number of missing values, thereby potentially improving quantification accuracy, particularly of low-abundance peptides. They also increase the resolution of a 'peak', resulting in more accurate quantitative ratio calculations. ProtMAX is an efficient, user-friendly and robust quantification tool. There is no need for complex preprocessing and knowledge of Java and R programming. It is optimized for huge proteomics data sets and is unique in its selection of m/z features on the basis of the MS/MS trigger. A workflow overview is given in **Figure 1**.

	A	B
1	779.88	46.6
2	516.26	30.7
3	639.78	23
4	829.43	51.9
5	821.4	36.7
6	510.24	17.8
7	715.86	42.7
8	484.73	6.4
9	458.25	33.7
10	761.89	31.4
11	613.3	30.5
12	697.85	48.1
13	692.84	58.1
14	680.35	39.6
15	574.79	20.4
16		

Figure 2 | Target list example. This file is a tab-delimited text file with two columns. Column A contains m/z values; column B contains Rt in minutes.

In principle, ProtMAX is not restricted to any MS instrument. However, the quality of the output depends on the quality of the input. For instance, for low-mass-accuracy data (>10 p.p.m.), the threshold for the mass-binning process is less stringent. As for protein identification, low mass accuracy will lead to less accurate results. Quantification comprises the comparison of different MS analyses. ProtMAX compares MS data derived from LC separations. High chromatographic variability between samples can impede comparison, or in worst cases, even render it impossible.

In the PROCEDURE, there are two approaches that are referred to as 'unbiased' (Steps 1–10) and 'targeted' (Steps 11–15). In addition to the unbiased approach, which creates a matrix from all of the precursors identified in the MS/MS data (Steps 1–10), ProtMAX also allows for known target-peptide extraction with specific m/z and Rt (Steps 11–15). Instead of using the MS/MS level for m/z precursor selection (unbiased approach), ProtMAX extracts all m/z precursors that have been uploaded using a target list (**Fig. 2**).

MATERIALS

EQUIPMENT

- Mass spectrometer: we recommend an LTQ Orbitrap family mass spectrometer (Thermo Scientific) and Xcalibur software (any other high-mass-accuracy MS instrument can be used); the instrument should be coupled to an HPLC system

Computer

- Operating system: XP, Vista or Windows 7
- Computer processor: Intel Core 2 Duo or better
- Computer memory: 2 GB or 4 GB or more recommended
- ProtMAX (http://www.univie.ac.at/mosys/software.html)
- RAW to mzXML file converter (MassMatrix MS Data File Conversion, v3.9, http://www.massmatrix.net/mm-cgi/downloads.py). According to MassMatrix, it is running on Windows PC (XP/7, 32/64 bit). The file converter does not need any particular options
- Excel (Microsoft), MATLAB (MathWorks), or other appropriate software depending on data analysis requirements

Input data files

- Data need to be uploaded as mzXML files (please find test files from **Fig. 3** for download at http://www.univie.ac.at/mosys/software.html)

EQUIPMENT SETUP

Instrument settings ProtMAX is not restricted to any specific mass spectrometer. However, the better the mass accuracy and LC performance, the better the results. Our recommended settings for quantitative analysis using an LTQ Orbitrap mass spectrometer are described in **Box 1** (**Fig. 4**).

Calibration In addition to standard instrument mass calibration according to the manufacturer's instructions, an additional real-time recalibration using internal Lock Mass calibrant molecules from an antiperspirant containing polydimethylcyclosiloxane may be used[26].

Figure 3 | ProtMAX output file. Samples 1 and 2 correspond to the test sample in the Equipment section. (**a**) Charge state (no information for target approach). (**b**) Sum across all samples. (**c**) Sum of ion intensity counts specific for sample 1 (can also be SC or Ion count, depending on the selected method). (**d**) Scan number of the most intense ion signal of the corresponding *m/z* value specific for sample 1. (**e**) Rt of the most intense ion signal of the corresponding *m/z* value specific for sample 1.

			Data for sample 1			Data for sample 2		
	a	b	c	d	e			
m/z	Charge state	Count [all]	al1_1 [Sum]	al1_1 [Scan]	al1_1 [Rt]	al2_1 [Sum]	al2_1 [Scan]	al2_1 [Rt]
300.63	2	0	0	0	0	0	2836	19.1
301.14	2	4.04E+07	1.94E+07	2486	17.3	2.10E+07	2535	17.4
301.14	2	2.79E+0 8	1.43E+08	466	4.7	1.36E+08	688	6.7
302.16	2	1187943	671294	1197	10.1	516649.2	974	8.6
302.67	2	1015228	243529	1652	12.6	771699.3	1712	12.8
303.17	2	1.78E+0 7	1.10E+07	1327	10.8	6882081	1369	10.9
303.19	2	0	0	2959	19.9	0	0	0
304.16	2	2204752	2078326	1278	10.5	126426.5	1582	12
304.17	2	4911157	2622588	723	7.1	2288569	777	7.4
305.18	2	2918113	1525773	150	1.5	1392340	438	4.3
305.68	2	4871819	2289013	625	6.3	2582806	441	4.3
306.13	2	2817526	1873275	925	8.5	944250.9	1000	8.8
307.15	2	329421.6	102788.5	1101	9.5	226633	1160	9.7
307.64	2	0	0	0	0	0	931	8.4
308.19	2	151484.2	61681.82	1628	12.5	89802.35	1666	12.5
308.19	2	1.76E+07	1.17E+07	692	6.9	5919995	910	8.2
308.64	2	183358.7	0	0	0	183358.7	1621	12.2
308.67	2	3423541	1637935	2312	16.3	1785607	1962	14.1
308.69	2	1590284	1244236	830	7.8	346048.4	638	6.3
308.69	2	1836817	631394.7	1527	11.9	1205423	1562	11.9
309.18	2	1.14E+07	4974495	405	4.1	6406557	420	4.2
309.18	2	3923815	1063879	981	8.9	2859936	1020	8.9
309.2	2	2900347	2454803	4011	26.1	445543.9	4020	25.9
309.67	2	459961.8	114685.2	2061	14.9	345276.5	2157	15.2
309.69	2	1950017	938587.5	918	8.5	1011430	949	8.5
310.17	2	869223.5	642399	2012	14.6	226824.5	2071	14.8
310.18	2	1703308	810197.1	1224	10.2	893110.6	1287	10.4
311.16	2	1909969	0	1030	9.1	1909969	1062	9.1
311.17	2	1127298	691543.3	1975	14.4	435754.6	2053	14.7
311.2	2	2690541	692244.7	2816	19.1	1998296	2878	19.3
311.68	2	6898012	2936522	1176	10	3961490	1219	10
312.15	2	186875.8	186875.8	3208	21.4	0	3325	21.9
312.17	2	1.45E+07	8890250	477	4.9	5649625	860	7.9
312.17	2	81385.92	0	2228	15.8	81385.92	2313	16.1
312.19	2	3706743	1800053	970	8.8	1906690	1027	8.9
312.21	2	8429528	3898936	1079	9.4	4530592	1109	9.4
312.62	2	0	0	2783	18.9	0	2855	19.2
312.69	2	2006294	761097.1	683	6.8	1245197	729	7.1
313.17	2	793937.9	501277.3	2428	16.9	292660.6	2633	17.9
313.71	2	254061.6	0	1632	12.5	254061.6	1664	12.5
314.15	2	3778490	624560.6	5993	38	3153929	5950	37.6
314.17	2	487654.6	44127.66	2202	15.7	443526.9	2482	17.1
314.65	2	845292.3	125796.6	2473	17.2	719495.6	2527	17.3
314.66	2	0	0	5525	35.1	0	5591	35.4
...								

Tissue compatibility To our knowledge, any tissue can be analyzed if protein extraction and sample handling are compatible with MS. In general, there is no optimal protocol for all tissues. However, there are recommended protocols[27]. A discussion about different MS-compatible plant protein extraction methods can be found here[28].

HPLC separation A shotgun approach using HPLC coupled to MS is required. For higher sensitivity, nano-flow HPLC systems using reversed-phase columns suitable for low flow rates (with an inner diameter of ~75 µm) are recommended. We usually use a monolithic column (150 mm × 0.1 mm) (Merck) that seems to be very robust for large-scale sample analyses[21–23].

Box 1 | Instrument settings for the protein shotgun LC-MS/MS analysis using an Orbitrap MS

The recommended LTQ Orbitrap MS settings are also shown in **Figure 4**.

(A) MS1 settings and internal Lock Mass calibration

▲ **CRITICAL** Before starting the LC-MS analysis, an Orbitrap mass calibration is recommended.

1. Set micro-scans to 3.
2. Check for spray stability.
3. Run a protein standard digest (e.g., BSA) to check chromatographic performance.
4. For each Fourier transform MS (FTMS) scan event (Scan Event 1), a resolution of 30,000 may be used. Scan range may be from 300 to 1,800 (**Fig. 4a**). Multiple replicate measurements are required for large-scale shotgun proteomics to achieve robust statistical significance. We usually use 6–15 independent MS analyses for each condition[21–23]. The replicate analysis must not be performed in sequence to exclude possible technical bias. For better reproducibility, experiments may be analyzed in one batch. At present, the most common normalization for label-free MS analysis may be the total ion current, but other methods are discussed[29]. In the future, ProtMAX will allow for automated internal standard normalization.
5. The Lock Mass can manually be set to any value (e.g., 445.12002 or 371.10123 *m/z* for polydimethylcyclosiloxane, see Equipment Setup and **Fig. 4a**), the difficulty being the known elemental composition of the respective molecule. To achieve a steady supply of calibrant molecules over the entire LC/MS run, we used Lock Mass protection according to Lee *et al.*[26].

(B) MS/MS Scan Events

1. For each MS/MS analysis, use 5–10 dependent ion trap MS scans (Scan Event 2–6 or 2–11) (**Fig. 4b**).

(C) Dependent ion trap MS Scan

The following settings are recommended if not default (**Fig. 4c**):

1. Dynamic Exclusion: set Repeat count 1, Repeat duration 20 s, Exclusion list size 500, Exclusion duration 60 s, Exclusion mass width 10 p.p.m. relative to reference mass.
2. Current Segment: set Enable preview mode for FTMS master scans.
3. Charge State: set Enable charge state screening, Enable monoisotopic precursor selection, Enable charge state rejection, reject charge state 1 and unassigned charge states.
4. Current Scan Event: set Minimum signal threshold to 1000 counts. Note: this threshold should be relatively high to allow for an MS/MS trigger closer to the peak apex.

(continued)

Box 1 | Instrument settings for the protein shotgun LC-MS/MS analysis using an Orbitrap MS (continued)

Instrument settings for the metabolite LC-MS/MS analysis

We suggest changing the following settings for metabolite analyses compared to those introduced in (A) (MS1 settings and internal Lock Mass calibration).

1. A low scan range limit for metabolomics experiments usually needs to be set, e.g., 140 to 380 *m/z*.
2. For very complex samples a resolution of 60.000 is recommended.
3. For metabolite analyses an accurate inclusion mass screening (AIMS)[29] may be used for confirmation of identity.
4. Only one data-dependent MS/MS scan using three micro scans is recommended.

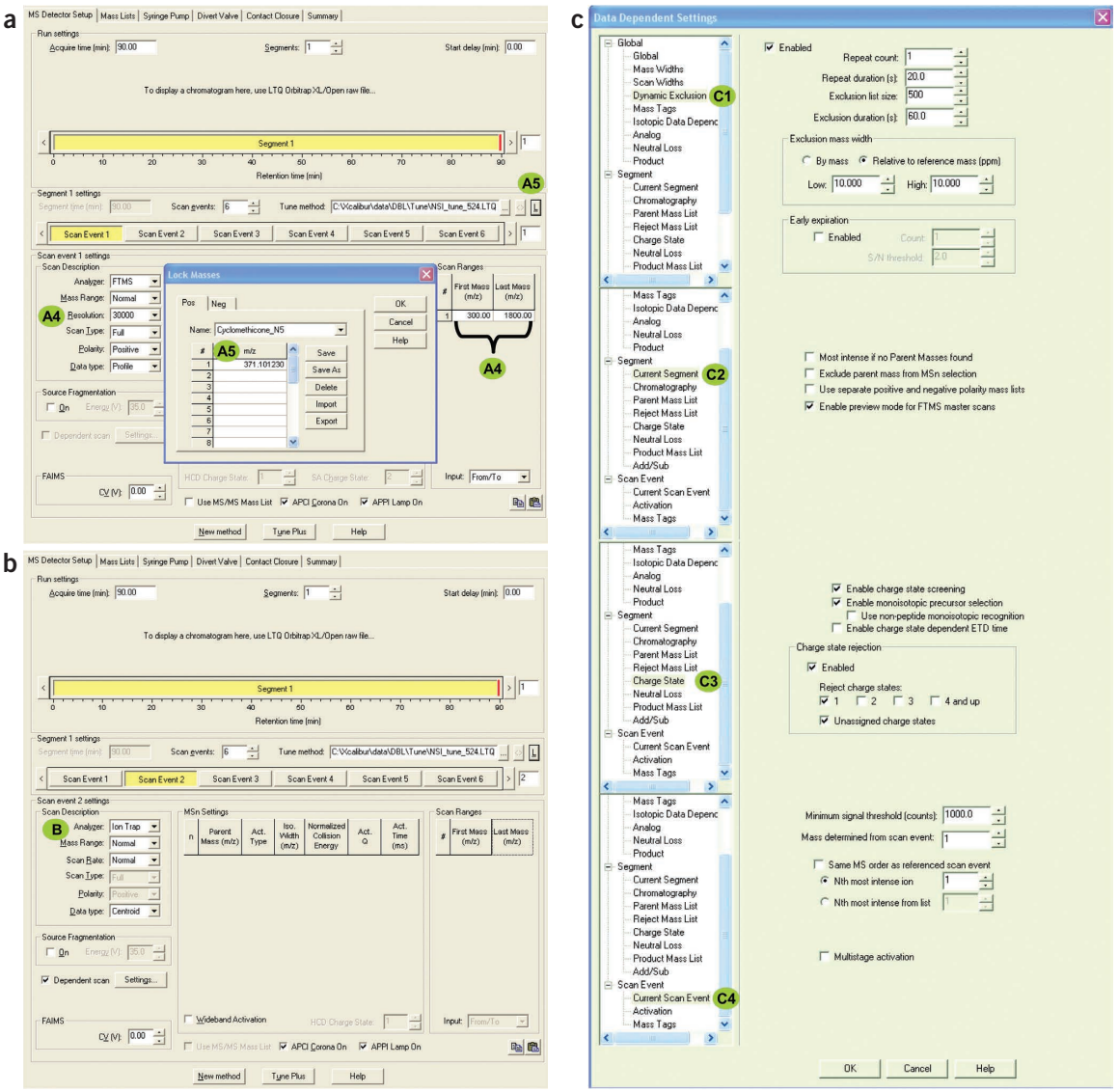


Figure 4 | Graphical user interface of the Xcalibur MS instrument setup. The settings for the protein analysis are indicated by green circles that are numbered according to the steps in **Box 1** (Instrument settings for the protein shotgun LC-MS/MS analysis using an Orbitrap) in which they are described. (**a**) MS1 settings and internal Lock Mass calibration. (**b**) MS/MS Scan Events. (**c**) Data-dependent settings for MS/MS scan events.

PROCEDURE

ProtMAX settings for unbiased protein matrix generation

▲ CRITICAL The ProtMAX settings are also shown in **Figure 5**.

▲ CRITICAL Steps 1–10 cover the settings for Preferences if the data accuracy is ≤10 p.p.m. In the case of less-accurate data, the 'Decimals' threshold (Step 2) needs to be adapted.

Figure 5 | Graphical user interface of ProtMAX. The settings for the unbiased peptide matrix generation are indicated by green circles that are numbered according to the PROCEDURE steps in which they are described.

1| For a traditional spectral count (summed number of all MS/MS spectra triggered for a specific *m/z* value)-based quantification, use Method 'Spectral Count' and Quantification 'Count'. For the ion count (summed number of all precursor ions belonging to a specific *m/z* value within a specified Rt window), select Method 'Ion Count' and Quantification 'Count', and for ion intensity count (summed intensity counts of all precursor ions belonging to one *m/z* value within a specified Rt window), use Method 'Ion Count' and Quantification 'Intensity'.

2| Set Decimals to the expected data accuracy. In our case (≤10 p.p.m.), set Decimals to '2' and 'Cut', which means that the *m/z* value is not rounded up or rounded down; it is merely cut off at the second digit.

3| Choose Unite Neighbors if you observe a mass shift (in this case at the second decimal, see Step 2). In other words, two precursor masses with a mass shift of ± 0.01 *m/z* eluting at the same time (occurring in the same Rt window, and thus most probably belonging to the same peptide) will be treated as one peptide. Neighboring precursor masses with a mass shift of ± 0.01 will be merged and treated as one peak (in the case of two decimals). Note: these mass shifts may also occur in rare cases because of differential mass rounding of the mzXML files compared with the RAW data.

4| It is possible to choose a Rt cutoff around the expected Rt, and thus to ensure that, for example, peaks eluting during equilibration phases will be excluded from further consideration.

5| It is possible to filter the charge state. ProtMAX will only select those monoisotopic precursor masses of acquired MS/MS scans with the selected charge state(s) if it was enabled in the MS analysis. In the case of protein quantification, we recommend leaving 'charge state 1+' unchecked.

6| Define the largest peak width. In the case of MS/MS-based precursor selection (untargeted), we recommend increasing the Environment setting by up to three times the observed peak width because the MS/MS (*m/z* anchor of the Rt window) might have been recorded at the beginning of the peak. The Environment setting allows discriminating between peptides of identical *m/z* ratio because it is not very likely that two peptides of the exact (± 10 p.p.m.) same mass and charge state elute at the same time. If Environment is used, it is a constant width, as it defines the peak width for the whole analysis. Note that it is better to use a higher peak-width setting than a lower peak-width setting.

7| The intensity expected describes the upper limit of the background noise. The default setting is 1% of the maximum peak intensity.

8| The default setting of the minimum number of counts required to accept a peak is 8.

9| Choose a path for the output file or use the default setting. The checked 'Launch Excel' application will automatically open an Excel result file after processing (**Fig. 2**). Uncheck this if Microsoft Excel is not installed. A full version of the output file for **Figure 4** can be downloaded at http://www.univie.ac.at/mosys/software.html.

10| RAW MS files must be converted to mzXML for upload into ProtMAX. Upload mzXML files by navigating to *File→Import*. Mark all desired files for import.

? TROUBLESHOOTING

ProtMAX settings for targeted protein or metabolite matrix generation

! CAUTION Because of the lack of information concerning the precursor charge state of the MS1 level within the mzXML file format after conversion of original MS RAW files, the target list quantification does not allow for charge-state matching. This also means that there is no charge-state restriction/filter for the targeted analysis available. Consequently, by using target mode, the mzXML format does not allow distinguishing between the monoisotopic and isotopomeric masses.

▲ CRITICAL As the target analysis is restricted to the MS1 level, it can be performed with both LC-MS/MS and LC-MS (full scan only). Compared with LC-MS/MS analysis, pure LC-MS may result in more data points (ion counts) for the MS1 level, as there is no loss in time for MS/MS scans. Thus, short gradients are possible. This approach may also be used to analyze MS data from metabolomics experiments.

▲ CRITICAL An Rt filter with each target is recommended, but this information is optional. When information for the Rt is given, ProtMAX only extracts quantitative information from this region of the chromatogram (± Environment settings). This will reduce the output file to the expected data and will thus facilitate data analysis.

11| Prepare a target precursor mass list containing all mass-to-charge ratios (*m/z*) of the peptide(s) of interest (**Fig. 2**).

12| Load the target list into ProtMAX: *File→Load target list*. The target list should be a tab-delimited text file without header, including the *m/z* values cut to the second decimal or more (rounding off the *m/z* data might degrade 1 p.p.m. accuracy) of the peptides of interest (targets) in row A.

13| In Preferences, select Method 'Target List' and Quantification 'Intensity'.

14| Rt can be set for each target if it is added to the target list: *m/z* values in row A (according to Step 11) and Rt values in row B (**Fig. 2**). If there is more than one peptide with identical *m/z* (e.g., ± 10 p.p.m.), the Rt specification narrows the output down to the target of interest. If the Rt of the target of interest is not given, several hits of identical *m/z* ratios but different Rt and scan numbers, and perhaps charge states, will be listed in the ProtMAX output file. Note that if the Environment setting is disabled, all *m/z* values will be binned together, regardless of the Rt. This setting can be used for data mining of direct infusion (flow injection) analysis.

15| See 'ProtMAX settings for unbiased protein matrix generation' Steps 1–10 for other settings.

Data output and mining

16| For further statistics (e.g., to examine the relationship between two or more variables), the data matrix generated by ProtMAX (**Fig. 3**) can be uploaded into the freely available MATLAB toolbox COVAIN[29] or analyzed with any other software of choice after removing the columns for Rt and scan number.

? TROUBLESHOOTING

The current version of ProtMAX has been extensively tested and is very stable. We welcome feedback and would like to hear of any bugs that you encounter. General errors you may experience include the following:

- *File Access Error.* This message appears when you are trying to import an mzXML file while the resulting file from the previous analysis is still open within Excel. Excel will lock it exclusively, and thus you must close the file before proceeding.
- *Error Target List.* This error occurs when the 'Target List' method is selected and the user tries to import the mzXML files before importing a target-list file.

● TIMING

The total duration of the protocol is dependent on the number of MS analyses and the file size. The time required for the initial MS instrument setup is about 10 min.

The nanoLC-MS/MS analysis time for a complex sample usually is ~90 min, including gradient and column equilibration. It is thus the limiting step. Conversion of RAW files (around 150 MB) to mzXML files takes about 10 s per file. The conversion leads to data reduction, but depending on the number of RAW files, it may require some additional disc space. The MAPA matrix generation using ProtMAX is typically in the range of minutes. For instance, the generation of the matrix for 183 LC-MS runs took less than 1 h on a quad-core personal computer[22].

ANTICIPATED RESULTS

ProtMAX will create a data matrix in a tab-delimited format. The files can be opened with Microsoft Excel. Each row of the data matrix will contain values of the chosen method corresponding to the *m/z* values of a given sample, as well

as the Rt and the scan number extracted from the LC-MS/MS data corresponding to the most intense MS signal from which the MS/MS was triggered. The ProtMAX result is database independent and does not provide protein or metabolite identifications.

After statistical filtering, the interesting *m/z* precursor masses of corresponding peptides can be matched against database-dependent search results. As precursor masses (*m/z*) of the ProtMAX output file depend on the measured and MS/MS-triggered masses of the MS analysis, they are identical to the list of masses used for identification (the measured, not calculated, *m/z*). MS/MS spectra of unidentified candidates can be extracted from the LC-MS/MS analysis for further identification.

ACKNOWLEDGMENTS We thank J. Hummel for interesting comments.

AUTHOR CONTRIBUTIONS V.E. developed the algorithm of the ProtMAX tool. W.H. contributed to the concept and optimization of the ProtMAX tool of protein analysis and writing of the manuscript. D.L. contributed to the concept of the ProtMAX tool and writing of the manuscript. W.W. conceived the concept of the MAPA strategy for proteomics. S.W. conceived Preferences and feature upgrading with the target approach of the ProtMAX tool and was responsible for project coordination and writing of the manuscript.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/nprot.2013.013. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Michalski, A., Cox, J. & Mann, M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J. Proteome Res.* **10**, 1785–1793 (2011).
2. Michalski, A. *et al.* Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol. Cell. Proteomics* **10**, M111.011015 (2011).
3. Thakur, S.S. *et al.* Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. *Mol. Cell Proteomics* **10**, M110.003699 (2011).
4. Griffin, N.M. *et al.* Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nat. Biotechnol.* **28**, 83–89 (2010).
5. Ishihama, Y. *et al.* Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell Proteomics* **4**, 1265–1272 (2005).
6. Liu, H., Sadygov, R.G. & Yates, J.R. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76**, 4193–4201 (2004).
7. Lu, P., Vogel, C., Wang, R., Yao, X. & Marcotte, E.M. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* **25**, 117–124 (2007).
8. Paoletti, A.C. *et al.* Quantitative proteomic analysis of distinct mammalian Mediator complexes using normalized spectral abundance factors. *Proc. Natl. Acad. Sci. USA* **103**, 18928–18933 (2006).
9. Schulze, W.X. & Usadel, B. Quantitation in mass-spectrometry-based proteomics. *Annu. Rev. Plant Biol.* **61**, 491–516 (2010).
10. Schwanhausser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
11. Silva, J.C., Gorenstein, M.V., Li, G.Z., Vissers, J.P.C. & Geromanos, S.J. Absolute quantification of proteins by LCMSE—a virtue of parallel MS acquisition. *Mol. Cell Proteomics* **5**, 144–156 (2006).
12. Baerenfaller, K. *et al.* Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. *Science* **320**, 938–941 (2008).
13. Brunner, E. *et al.* A high-quality catalog of the *Drosophila melanogaster* proteome. *Nat. Biotechnol.* **25**, 576–583 (2007).
14. de Godoy, L.M.F. *et al.* Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455**, 1251–1260 (2008).
15. Graumann, J. *et al.* Stable isotope labeling by amino acids in cell culture (SILAC) and proteome quantitation of mouse embryonic stem cells to a depth of 5,111 proteins. *Mol. Cell Proteomics* **7**, 672–683 (2008).
16. Malmstrom, J. *et al.* Proteome-wide cellular protein concentrations of the human pathogen *Leptospira interrogans*. *Nature* **460**, 762–U112 (2009).
17. Nagaraj, N. *et al.* Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.* **7**, 548 (2011).
18. Pavelka, N. *et al.* Statistical similarities between transcriptomics and quantitative shotgun proteomics data. *Mol. Cell Proteomics* **7**, 631–644 (2008).
19. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
20. De Vos, R.C.H. *et al.* Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. *Nat. Protoc.* **2**, 778–791 (2007).
21. Hoehenwarter, W. *et al.* A rapid approach for phenotype-screening and database independent detection of cSNP/protein polymorphism using mass accuracy precursor alignment. *Proteomics* **8**, 4214–4225 (2008).
22. Hoehenwarter, W. *et al.* MAPA distinguishes genotype-specific variability of highly similar regulatory protein isoforms in potato tuber. *J. Proteome Res.* **10**, 2979–2991 (2011).
23. Chen, Y., Hoehenwarther, W. & Weckwerth, W. Comparative analysis of phytohormone-responsive phosphoproteins in *Arabidopsis thaliana* using TiO2-phosphopeptide enrichment and mass accuracy precursor alignment. *Plant J.* **63**, 1–17 (2010).
24. Doerfler, H. *et al.* Granger causality in integrated GC-MS and LC-MS metabolomics data reveals the interface of primary and secondary metabolism. *Metabolomics.* http://dx.doi.org/10.1007/s11306-012-0470-0 (25 October 2012).
25. Mari, A. *et al.* Phytochemical composition of *Potentilla anserina* L. analyzed by an integrative GC-MS and LC-MS metabolomics platform. *Metabolomics.* http://dx.doi.org/10.1007/s11306-012-0473-x (17 November 2012).
26. Lee, K.A., Farnsworth, C., Yu, W. & Bonilla, L.E. 24-hour lock mass protection. *J. Proteome Res.* **10**, 880–885 (2011).
27. Isaacson, T. *et al.* Sample extraction techniques for enhanced proteomic analysis of plant tissues. *Nat. Protoc.* **2**, 769–774 (2006).
28. Sheoran, I.S. *et al.* Compatibility of plant protein extraction methods with mass spectrometry for proteome analysis. *Plant Sci.* **176**, 99–104 (2009).
29. 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* **8**, 81–93 (2012).

Phytochemical composition of *Potentilla anserina* L. analyzed by an integrative GC/MS and LC/MS metabolomics platform

Despite the knowledge of desirable effects of herbal remedies, the responsible bioactive compounds are often unknown. Modern analytical techniques, such as nanoLC/MS (specifically a Mass Spectrometer with high Mass Accuracy and Mass Resolving Power such as the LTQ-Orbitrap), enable the detection of a plethora of compounds. The phytochemical analysis and comparative characterization of various sources and extraction methods of the medicinal plant *Potentilla anserina* L. revealed the presence of compounds never reported for this species, as well as strongly varying concentrations of the bioactive ingredient Genistein. The Isoflavon Genistein, a so-called secondary compound, is known to occur in other plant genera such as Glycine and Trifolium. Generally, Flavonoids are ubiquitous compounds, and are amongst other properties well-known for their colorful chromophores and antioxidative properties, and are even used as dietary supplements.

Declaration of authorship

The results of this chapter are presented in the form of a manuscript published in the journal „Metabolomics“. I have provided a critical contribution to the following publication, though the largest part the work was performed by the coauthors.

I have contributed by creating the LC/MS/MS method (including Chromatography, Mass Spectrometry method, and autosampler settings), aiding in the LC/MS data acquisition, the subsequent data extraction (creating a data matrix from the raw data) and data transformation, analysis, and visualization (of the LC/MS data).

Published manuscript

Metabolomics (2013) 9:599–607
DOI 10.1007/s11306-012-0473-x

ORIGINAL ARTICLE

Phytochemical composition of *Potentilla anserina* L. analyzed by an integrative GC-MS and LC-MS metabolomics platform

Angela Mari · David Lyon · Lena Fragner · Paola Montoro · Sonia Piacente · Stefanie Wienkoop · Volker Egelhofer · Wolfram Weckwerth

Received: 20 July 2012 / Accepted: 16 October 2012 / Published online: 17 November 2012

Abstract *Potentilla anserina* L. (Rosaceae) is known for its beneficial effects of prevention of pre-menstrual syndrome (PMS). For this reason *P. anserina* is processed into many food supplements and pharmaceutical preparations. Here we analyzed hydroalcoholic reference extracts and compared them with various extracts of different pharmacies using an integrative metabolomics platform comprising GC-MS and LC-MS analysis and software toolboxes for data alignment (MetMAX Beta 1.0) and multivariate statistical analysis (COVAIN 1.0). Multivariate statistics of the integrated GC-MS and LC-MS data showed strong differences between the different plant extract formulations. Different groups of compounds such as chlorogenic acid, kaempferol 3-*O*-rutinoside, acacetin 7-*O*-rutinoside, and genistein were reported for the first time in this species. The typical fragmentation pathway of the isoflavone genistein confirmed the identification of this active compound that was present with different abundances in all the extracts analyzed. As a result we have revealed that different extraction procedures from different vendors produce different chemical compositions, e.g. different genistein concentrations. Consequently, the treatment may have different effects. The integrative metabolomics platform provides the highest resolution of the phytochemical composition and a mean to define subtle differences in plant extract formulations.

Keywords Medicinal plants · Metabolomics · GC-MS · LC-MS · Flavonoids · Genistein · Mass accuracy precursor alignment (MAPA)

1 Introduction

Potentilla anserina L. (silverweed) belongs to the family of Rosaceae and its extracts have been used for a long time in traditional medicine. The gynecological indication for *P. anserina* is based on pharmacological studies showing that the herb increases the tonus of the isolated uterus in various animal species (Schulz et al. 1998). Additionally, extracts of the aerial and/or underground parts have been applied in traditional medicine for the treatment of inflammations, wounds, certain forms of cancer, infections due to bacteria, fungi and viruses, diarrhoea, diabetes mellitus and other ailments (Bundesgesundheitsamt 1985, 1990). Tomczyk and Latté report that *P. anserina* (aerial parts or the whole plant) and other *Potentilla* species are generally used to prepare homeopathic medications (Tomczyk and Latté 2009) according to homeopathic pharmacopoeias like Homeopathic Pharmacopoeia of the United States (HPUS) and German Homeopathic Pharmacopoeia (HAB) (Hiller 1994). For this reason *P. anserina* is processed into many food supplements and pharmaceutical preparations such as teas, tinctures, capsules, tablets, and juice and is consumed by women in order to prevent the symptoms of pre-menstrual syndrome (PMS).

Electronic supplementary material The online version of this article (doi:10.1007/s11306-012-0473-x) contains supplementary material, which is available to authorized users.

A. Mari · P. Montoro · S. Piacente (✉)
Department of Pharmaceutical and Biomedical Sciences, University of Salerno, Salerno, Italy
e-mail: piacente@unisa.it

D. Lyon · L. Fragner · S. Wienkoop · V. Egelhofer · W. Weckwerth (✉)
Department of Molecular Systems Biology, University of Vienna, Vienna, Austria
e-mail: wolfram.weckwerth@univie.ac.at

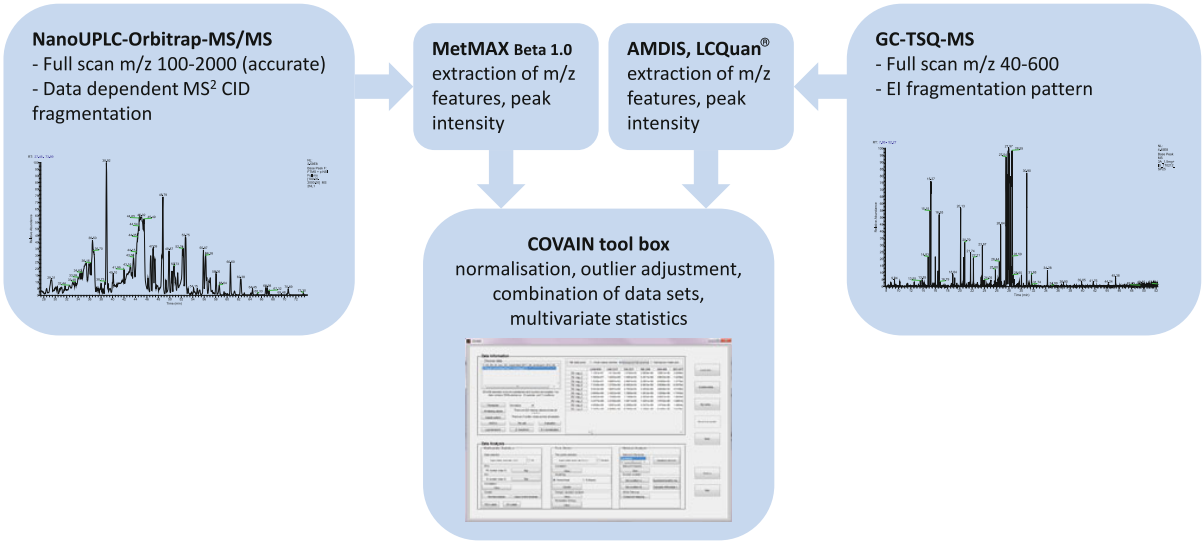


Fig. 1 Metabolomics platform for the characterization of medicinal plants integrating GC-MS, LC-MS, alignment tools and statistical analysis using COVAIN (Sun and Weckwerth 2012)

Despite all these positive effects, so far only limited analytical information of the chemical composition on *P. anserina* is available (Swiezewska and Chojnacki 1989; Kombal and Glasl 1995; Schimmer and Lindenbaum 1995; Tomczyk et al. 2010; Xu et al. 2010). In particular mass spectrometric data of the chemical composition of *P. anserina* are still lacking. These could however be helpful for the evaluation of physiological properties of individual plant secondary metabolites and for stability studies of pharmaceutical preparations. HPLC coupled to mass spectrometry (LC-MS) proved to be a very useful tool and is largely applied to the characterization of plant secondary metabolites. Gas chromatography coupled to mass spectrometry (GC-MS) provides complementary data to LC-MS analysis comprising small polar chemicals such as organic acids, sugars, amino acids, sugar alcohols and many more (Scherling et al. 2010; Weckwerth 2010). The aim of the present work is to characterize the phytochemical profile of hydroalcoholic extracts of *P. anserina* and its commercial products prepared by different pharmacies using a comprehensive metabolomics platform integrating GC-MS, LC-MS and multivariate statistics (Fig. 1).

2 Materials and methods

2.1 Chemicals

Chloroform, ethanol (absolute), methanol, *n*-butanol, petroleum ether (HPLC grade), were purchased from Sigma-Aldrich (Vienna, Austria) in Chromasolv® grade as well as pyridine over molecular sieve (GC-grade). Acetonitrile (HPLC grade) and formic acid were obtained from Merck (Darmstadt, Germany). Chlorogenic acid was purchased from Roth (Graz, Austria), genistein, quercetin-3-*O*-glucoside and D-sorbitol, β-alanine, arabinose, asparagine, citric acid, D(-)quinic acid, fructose, glucose, glyceric acid, glycerol, glycine, L-alanine, L-aspartic acid, L-leucine, L-serine, L-threonine, malic acid, mannitol, myo-inositol, phenylalanine, pinitol, proline, sucrose, trehalose, valine and, xylose were purchased from Sigma-Aldrich (Vienna, Austria), kaempferol-3-*O*-glucoside was obtained from Extrasynthese (Genay, France). Myricetin-3-*O*-rhamnoside and quercetin-3-*O*-glucuronide were isolated by size exclusion chromatography and semipreparative HPLC with UV detection from the hydroalcoholic extract of *P. anserina* and their structures were characterized by 1D and 2D nuclear magnetic resonance (NMR). For derivatization methoxyamine hydrochloride and *N*-methyl-*N*-trimethylsilyltrifluoroacetamide (MSTFA) were purchased from Sigma-Aldrich (Vienna, Austria), as well as retention index marker alkanes (all even C_{10}–C_{40}). HPLC grade water (18.2 mΩ) wËas prepared using a Millipore Milli-Q purification system (Millipore Corp., Bedford, MA, USA).

2.2 Plant material

Potenilla anserina air-dried plant parts were purchased from Minardi s.r.l. (Bagnacavallo, Ra, Italy). 500 g of whole plant parts were extracted with petroleum ether three times. After filtration the raw material was extracted three times with chloroform and finally with 70 % EtOH following the same procedure performed with petroleum ether. The collected alcohol-aqueous extract (PanserinaUniSa) was dried under vacuum.

A second extract (PanserinaUniVie) was prepared using a grinding mill system MM400 from Retsch (Haan, Germany). 250 mg of ground plant material was extracted with 25 mL of a solution of methanol/chloroform/water (2.5:1:0.5, v:v:v) (Weckwerth et al. 2004) and then vortexed for 10 min followed by 8 min incubation. The sample was then centrifuged for 4 min at 3,400×*g* and the supernatant was separated from the pellet. 5 mL of distilled water were added to the supernatant, followed by 10 s shaking on a vortex and 2 min centrifugation at 3,400×*g*. The alcoholic-aqueous phase was dried under vacuum.

Five mother tinctures were acquired from five drug-stores (# 1, 2, 3, 4 and 5) in Vienna. For each of them 1 mL was dried under vacuum.

All samples were analyzed by gas chromatography and liquid chromatography coupled to mass spectrometry. For data analysis (see below) all sample injections were normalized against corresponding extract dry weights.

2.3 Extract derivatization and GC-MS analysis

The protocol for GC-MS analysis was performed according to Weckwerth et al. (2004) with slight changes. Before derivatization 25 µL of ^{13}C-D-sorbitol (0.02 µg µL^{-1}) were added to all samples as internal standard. Samples were derivatized in two steps. First 20 µL methoxyamination mixture (40 mg mL^{-1} methoxyamine hydrochloride in dry pyridine) were added and incubated for 90 min at 30 °C in a thermo shaker. Then 80 µL of *N*-methyl-*N*-trimethylsilyltrifluoroacetamide (MSTFA) silylation mixture including retention index marker were added (30 µL of alkane mixture (even-numbered C10-C40-alkanes, each 50 mg L^{-1}) and incubated for 30 min at 37 °C.

Derivatized samples were centrifuged and 50 µL of supernatant was transferred to GC-vials with micro inserts and closed with crimp caps.

GC-MS analyses were performed on a ThermoFisher Trace gas chromatograph coupled to a Triple Quadrupole mass analyzer (Thermo Scientific TSQ Quantum GC^{TM}, Bremen, Germany). 1 µL of derivatized sample was injected at a constant temperature of 230 °C in splitless mode with a deactivated Siltek liner (Restek). Each sample was measured three times with the same conditions to get technical replicates.

GC separation was performed on a HP-5MS capillary column (30 m × 0.25 mm × 0.25 µm) (Agilent Technologies, Santa Clara, CA), at a constant flow 1 mL min^{-1} helium. Initial oven temperature was set to 70 °C and hold for 1 min, followed by a ramp to 76 °C at 1 °C min^{-1} and a second ramp at 6 °C min^{-1} to 350 °C hold for 1 min. Transfer line temperature was set to 340 °C and post run temperature to 325 °C for 10 min.

Table 1 Compounds occurring in *P. anserina* hydroalcoholic extracts measured by GC/EI(TSQ)MS

Compounds	Rt (min)	Fragments *m/z*
1. Glycolic acid (2TMS)	9.50	147
2. Alanine (2TMS)[a]	10.28	116
3. Unknown 1	11.36	133
4. Unknown 2	12.59	281
5. Valine (2TMS)[a]	13.53	144
6. Leucine (2TMS)[a]	15.03	158
7. Glycerol (3TMS)[a]	15.24	205
8. Proline (2TMS)[a]	15.59	142
9. Glycine (3TMS)[a]	15.81	174
10. Succinic acid (2TMS)	16.06	247
11. Glyceric acid (3TMS)[a]	16.60	189
12. Serine (3TMS)[a]	17.30	204
13. Threonine (3TMS)[a]	17.91	218
14. Malic acid (3TMS)[a]	20.18	233
15. Pyroglutamic acid (2TMS)	20.57	156
16. Threitol or erythritol (4TMS)	20.67	156
17. Aspartic acid (3TMS)[a]	20.73	232
18. 4-amino butyric acid (3TMS)	20.79	174
19. Unknown 3	21.73	292
20. Unknown 4	22.13	307
21. Phenylalanine (2TMS)[a]	22.67	192
22. Asparagine (3TMS)[a]	23.61	188
23. Arabinose (1MeOx) (4TMS)[a]	23.76	103
24. Xylose (1MeOx) (4TMS)[a]	24.05	217
25. Xylitol or ribitol (5TMS)	24.80	217
26. 2-desoxy-pentos-3ylose (2MeOx)(2TMS)	25.30	231
27. Unknown 5	25.79	257
28. Lyxonic acid (5TMS)	25.90	292
29. Shikimic acid (4TMS)	26.25	204
30. Carbohydrate	26.30	217
31. Citric acid (4TMS)[a]	26.43	273
32. Carbohydrate (5TMS)	26.50	204
33. Glucopyranoside (5TMS)	26.64	204
34. Pinitol (5TMS)[a]	26.76	217
35. Quinic acid (5TMS)[a]	27.24	345
36. Fructose (1MeOx) (5TMS)[a]	27.60	307
37. Hexose (5TMS)	27.78	191
38. Glucose (1MeOx) (5TMS)[a]	27.91	319
39. Mannitol (6TMS)[a]	28.43	319
40. Unknown (inositol isomer)	28.82	305
41. Glucopyranoside (5TMS)	29.27	217
42. Carbohydrate (glucopyranoside 5TMS)	29.28	204
43. Gluconic acid (6TMS)	29.59	333
44. Unknown 6	30.04	204
45. *Myo*-inositol[a] (6TMS)	30.89	305
46. Sucrose (8TMS)[a]	38.61	361

Table 1 continued

Compounds	Rt (min)	Fragments *m/z*
47. Trehalose (8TMS)[a]	39.83	361
48. Isomaltose (1MeOx) (8TMS)	40.89	361
49. Melibiose (1MeOx) (8TMS)	41.16	204
50. Unknown 7	41.77	369
51. Unknown 8	44.22	204
52. Unknown 9	44.35	647
53. Unknown 10	45.38	575

Given are (methoxime)-trimethylsilyl [(MeOx) (TMS)] derivatives of metabolites including their retention time (Rt) and EI-fragments taken for relative quantification

[a] Confirmed by comparison with corresponding reference standard

Mass analyzer was used in full scan mode scanning a range from *m/z* 40–800 at a scan time of 250 ms. Electron impact (EI) ionization was used at 70 eV and ion source temperature was set to 250 °C.

Metabolite derivatives were identified by matching retention time as well as mass spectra (see Table 1) with those of the corresponding reference standards and by comparison with an in house mass spectral library. Metabolites were considered identified with a spectral match factor higher than 850 and RI-deviation lower than 10. Deconvolution was performed with AMDIS (Stein 1999) and quantification with LC-Quan2.6.0 (Thermo Fisher Scientific Inc.). For statistical analyses a Matlab tool called COVAIN was used that provides a complete workflow including uploading data, data preprocessing, data integration and uni- and multivariate statistical analysis (Sun and Weckwerth 2012).

2.4 NanoLC-Orbitrap-MS/MS analyses

For all the samples described in the plant material section, 0.12 µg μL^{-1} water/acetonitrile (95:5, v:v) 0.1 % formic acid solutions were prepared and centrifuged at 13,000×*g* for 3 min. For each of them, 5 µL were used for LC-MS and MS/MS analysis in triplicates.

A 1D plus nanoUHPLC system (Eksigent, Dublin, Ireland) was equipped with an autosampler and the employed column was a Waters nanoAcquityHSS T_3, 1.8 µm, 100 µm × 100 mm. The mobile phases were water 0.1 % formic acid (A) and 90 % acetonitrile in water 0.1 % formic acid (B) at a flow rate of 500 µL min^{-1}. The LC conditions were 5 % B during 0–3 min, a linear increase from 5 to 20 % B during 3–25 min, from 20 to 40 % B during 25–40 min and from 40 to 50 % B during 40–55 min, finally from 50 to 95 % B during 55–63 min followed by 15 min of maintenance. A Thermo Electron LTQ-Orbitrap XL mass spectrometer equipped with a nano electrospray ion source (ThermoFisher Scientific, Bremen, Germany) and operated under Xcalibur 2.1 version software, was used in positive ionization mode for the MS analysis using data-dependent automatic switching between MS and MS/MS acquisition modes. The instrument was calibrated using the manufacturer's calibration standards. The scan was collected in the Orbitrap at a resolution of 30, 000 in a *m/z* range of 150–1,800. In order to achieve even higher mass accuracy a lock mass option was enabled in both MS and MS/MS mode and the cyclomethicone N5 ions generated in the electrospray process from ambient air ($m/z = 371.101230$) were used for internal recalibration in real time. This allowed mass accuracies of <1 ppm. The capillary voltage was 4.5 kV, the tube lens offset 160 V and the capillary temperature was set at 180 °C, no sheath gas and auxiliary gas were used.

Data deconvolution was performed with a modified ProtMAX version called MetMAX Beta 1.0 which provides mass accuracy precursor alignment of selected *m/z* signals in the LC-MS profile (Hoehenwarter et al. 2008). As for GC-MS, the COVAIN tool (Sun and Weckwerth 2012) was used for statistical analyses of the LC-MS data as well.

2.5 MetMAX Beta 1.0 processing and COVAIN analysis of LC-MS data

Raw data files were converted to mzXml format using the MassMatrix mass spectrometric data file conversion tool version 3.9 from the Case Western Reserve University (Cleveland, Ohio, USA; http://www.massmatrix.net/). MetMAX Beta 1.0 was used to process the mzXml files, generating a matrix of precursor ion intensities (Hoehenwarter et al. 2008). Each column vector contains the quantities of selected metabolites; each row vector describes the abundance of a respective metabolite ion over the entire set of analyses. Each column was normalized to its total spectral count. The .csv data table resulting from MetMAX Beta 1.0 were imported into COVAIN for statistical analysis (Sun and Weckwerth 2012). The values were then log-transformed. Principal component analysis (PCA) was performed for decomposition and visualization of data. The components of the column vectors, i.e. the precursor *m/z*, constitute the loadings of the independent components, and were identified by matching their retention times and mass spectra with those of the corresponding reference standards (see supplementary data).

3 Results and discussion

3.1 Qualitative and quantitative nanoLC-Orbitrap-MS/MS analyses of *P.anserina* crude extract

In order to obtain a metabolite profile of the crude extract of *P. anserina*, an analytical method based on

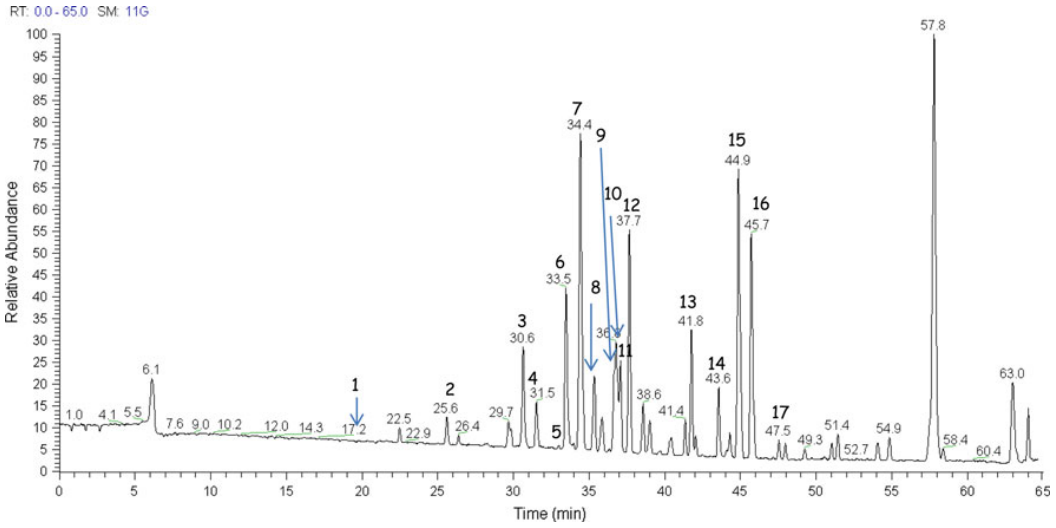


Fig. 2 NanoLC-Orbitrap-MS profile (full MS-mode) of the crude hydroalcoholic extract of *P. anserina* (positive ion mode) (see also Table 2)

Table 2 Retention time (Rt), precursor ions and product ions (for qualitative confirmation of the compound), of compounds occurring in *P. anserina* hydroalcoholic extracts by nanoLC-Orbitrap-MS/MS

Compound	Rt (min)	Precursor ion (*m/z*)	Product ions (*m/z*)	References
1. Chlorogenic acid[b]	19.6	377.0477	355; 163	–
2. *Chlorogenic acid isomer*	25.6	377.0477	355; 163	–
3. Myricetin 3-*O*-glucuronide[a]	30.6	495.0769	319	Kombal and Glasl (1995)
4. Quercetin -3-*O*-sambubioside[a]	31.5	597.1451	465; 303	Kombal and Glasl (1995)
5. Myricetin 3-*O*-rhamnoside[b]	33.5	465.1025	319	Kombal and Glasl (1995)
6. Quercetin 3-*O*-glucoside[b]	34.0	465.1027	303	Kim et al. (2004)
7. Quercetin 3-*O*-glucuronide[b]	34.4	479.0817	303	Merfort and Wendisch (1988)
8. Quercetin 3-*O*-xyloside[a]	35.3	435.0920	303	Zou et al. (2002)
9. Quercetin pentoside	36.6	435.0921	303	–
10. Kaempferol 3-*O*-glucoside[b]	36.8	449.1077	287	Kim et al. (2004)
11. Rutin (mass bank match)	37.1	611.1815	465; 303	Wang et al. (1999)
12. Isorhamnetin3-*O*-glucuronide[a]	37.7	493.0975	317	Kombal and Glasl (1995)
13. Acacetin-7-*O*-rutinoside (mass bank match)	41.7	593.1863	447; 285	–
14. Kaempferol 3-*O*-rutinoside (mass bank match)	43.6	595.1445	449; 287	–
15. *Unknown*	45.0	668.4370	489; 471; 453;435; 409	–
16. *Unknown*	45.7	668.4371	489; 471; 453;407; 316	–
17. Genistein[b]	47.0	271.0601	253; 243; 225;215; 197; 159; 153; 145	–

In *italised names* are putative identifications of compounds, without any comparison with the corresponding reference standard or with *Mass Bank*

[a] Compounds, without any comparison with the corresponding reference standard or with *Mass Bank* but already reported in *P. anserina*

[b] Confirmed by comparison with corresponding reference standard

nanoLC-Orbitrap-MS/MS was developed. The LC-MS profile highlighted the presence of a large group of compounds corresponding to the protonated molecular ions of different flavonoids and caffeoylquinic acids (Fig. 2).

Individual components were identified by comparison of their *m/z* values in the Total Ion Count (TIC) profile with those of the selected compounds described in literature (Table 2) or by matching their MS/MS spectra with those

reported in a public repository of mass spectral data called *Mass Bank* (Horai et al. 2010). According to our knowledge compounds 1, 2, 13, 14, 15, 16, 17 were never reported in this species. The positive HR-ESI-MS spectrum of compound 1 showed a $[M + Na]^+$ ion peak at *m/z* 377.0477 along with a less intense signal at *m/z* 355 corresponding to the protonated ion. The analysis of the MS/MS spectrum of the sodium adduct of compound 1, highlighted the presence of product $[(M\text{-}192) + H]^+$ at *m/z* 163 a.m.u. due to the loss of a quinic acid unit. By comparing the Rt, the mass and the MS/MS spectra of compound 1 with that of the commercial reference standard we unambiguously confirmed chlorogenic acid in *P. anserina* extracts (Table 2; Fig. 2).

The analysis of the HR-ESI-MS spectrum of compound 2 suggested it as a chlorogenic acid isomer showing the diagnostic $[M + Na]^+$ ion at *m/z* 377.0477 along with the $[M + H]^+$ ion at *m/z* 355 with a shift of 6 min in Rt. In particular, by the analysis of the tandem mass spectrum of the $[M + H]^+$ ion, compound 2 showed product ions accounting for the same composition of chlorogenic acid, originated by the neutral loss of 192 a.m.u. (Table 2; Fig. 2).

Full positive HR-ESI-MS profile of compound 13 was in agreement with a di-glycosylated acacetin structure, showing the diagnostic $[M + H]^+$ ion at *m/z* 593.1864. The analysis of the ESI-MS^2 spectrum of 13 allowed to determine the presence of a deoxy-hexose unit, besides product ion originated by the sequential neutral losses of 146 a.m.u leading to $[(M\text{-}146) + H]^+$ion at *m/z* 447 and of 162 a.m.u. leading to $[(M\text{-}146\text{-}162) + H]^+$ ion at *m/z* 285 corresponding to the acacetin-aglycone (Table 2; Fig. 2). For identification, the MS^2 spectrum of compound 13 was compared to that of acacetin 7-*O*-rutinoside present in *Mass Bank* library.

The HR-ESI-MS spectrum of compound 14 assigned it to a diglycosylated kaempferol according to the presence of the $[M + H]^+$ ion at *m/z* 595.1445. The tandem mass experiment on the $[M + H]^+$ ion allowed to observe a product ion at *m/z* 449, due to the neutral loss of one deoxy-hexose 146 a.m.u. and a product ion at *m/z* 287, due to the neutral loss of one hexose unit and corresponding to a kaempferol-aglycon (Table 2; Fig. 2). Identification of compound 14 as kaempferol 3-*O*-rutinoside was done by matching its tandem mass spectra with that of *Mass Bank* (data not shown).

According to the HPLC-ESIMS data, the positive ESIMS spectrum of compound 17 showed a minor $[M + H]^+$ ion peak at *m/z* 271.0601. Interestingly, the MS/MS spectrum of the $[M + H]^+$ ion showed a fragmentation pattern very similar to what was proposed by Lee et al. (2002) for the isoflavone genistein. By comparing compound 17 Rt and MS/MS spectra to that of the corresponding commercial standard we confirmed it as genistein (Table 2; Fig. 2). Although the presence of genistein and its glycosides is already reported in the family of Rosaceae (Jung et al. 2002; Lee et al. 2002; Ismail and Hayes 2005; Tohno et al. 2010) and in *Potentilla* genus (Şöhretoğlu and Sterner 2011), this is the first time that this isoflavone is reported in this particular species.

For compounds 15 and 16 no unambiguous identification was possible. These compounds could be isorhamnetin derivatives with two glucuronide units according to their fragmentation pattern in MS/MS. The structural elucidation is planned in future studies.

3.2 Multivariate statistical analysis of *Potentilla anserina* crude and commercial extracts from different pharmacies

In order to carry out a comparative study between our *P. anserina* reference extract and five hydroalcoholic extracts from different pharmacy vendors all were analyzed with the same GC- and LC-MS conditions.

In both cases, our results revealed that qualitative profiles of mother tinctures seem to be very similar to that of the crude extract shown in supplementary Fig. 1 and 2.

To better highlight the differences in metabolite profiling of the different extracts of *P. anserina*, unsupervised PCA was performed using COVAIN (Sun and Weckwerth 2012). Pre-processed GC-MS data sets (see "Materials and methods") from the different samples were analyzed. The PCA scores plot, shown in Fig. 3a, could be readily divided into two different groups indicating that the content and distribution of components were different between the *P. anserina* crude extracts (PanserinaUniSA and PanserinaUnivie) and the respective commercial products. The corresponding PCA loadings were utilized to identify the differential metabolic compositions accountable for the separation among groups (supplementary Fig. 3 and 4 and supplementary Table 1). In the loadings plot, the Rt and *m/z* values which point far away from zero represent characteristic markers with most confidence to each group. Unknown 1 (Rt 11.36 min, *m/z* 133), proline (2TMS) (Rt 15.59 min, *m/z* 142), 2-desoxy-pentos-3ylose-dimethoxyamine (2TMS) (Rt 25.30 min, *m/z* 231), carbohydrate (Rt 26.30 min, *m/z* 217), glucopyranoside (5TMS) (Rt 26.64 min, *m/z* 204) and unknown 7 (Rt 41.77 min, *m/z* 369) (Table 1) accounted primarily for the differences among our samples.

The nanoLC-Orbitrap-MS/MS data of all determined samples were processed and aligned with MetMAX Beta 1.0 software by selecting a target list containing all the 17 identified ions (Table 2). The resulting data matrix containing normalized intensities of the selected peaks was further exported into COVAIN for PCA (Fig. 3b). In this

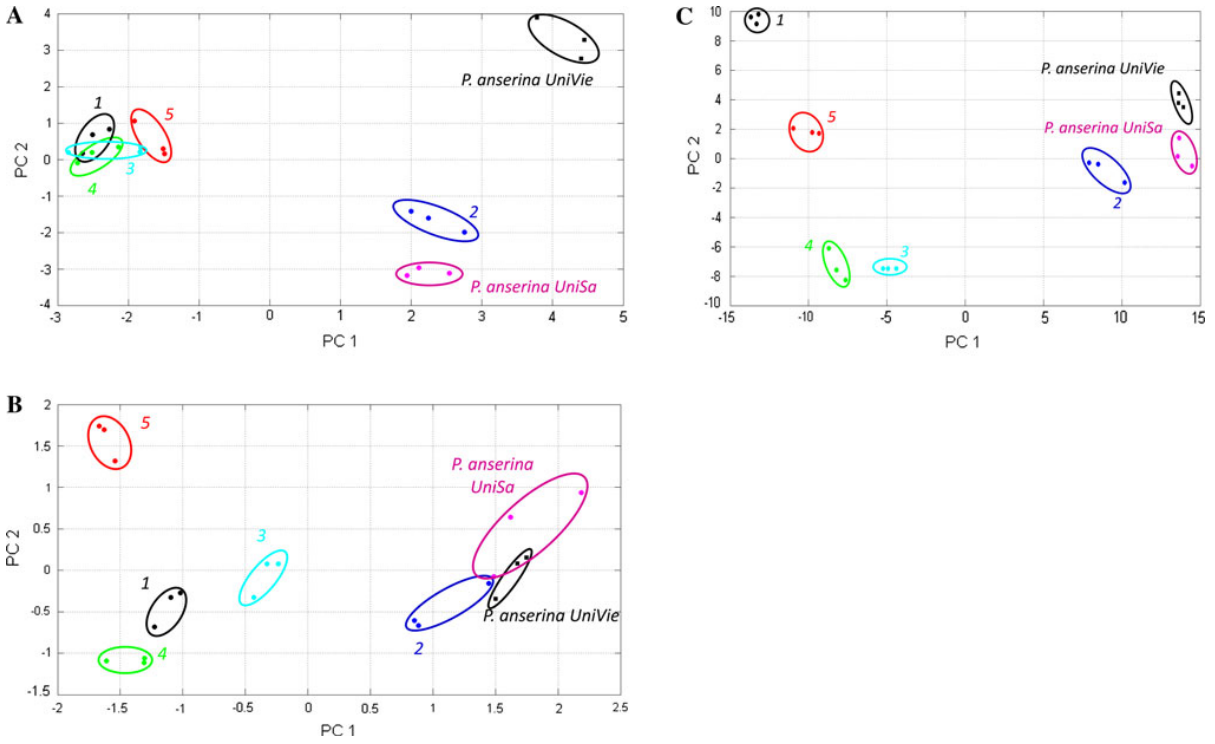


Fig. 3 **a** Sample patterns of the hydroalcoholic extracts of *P. anserina* analyzed by GC-MS. PC1 occupies 42 % and PC2 23 % of total variance. **b** Scores plot of the hydroalcoholic extracts of *P. anserina* analyzed by LC-MS (only identified compounds used as variables). PC1 occupies 39 % and PC2 19 % of total variance. **c** Sample patterns of the hydroalcoholic extracts of *P. anserina* analyzed by LC-MS (all the compounds with RSD <25 used as variables). PC1 occupies 52 % and PC2 27 % of total variance

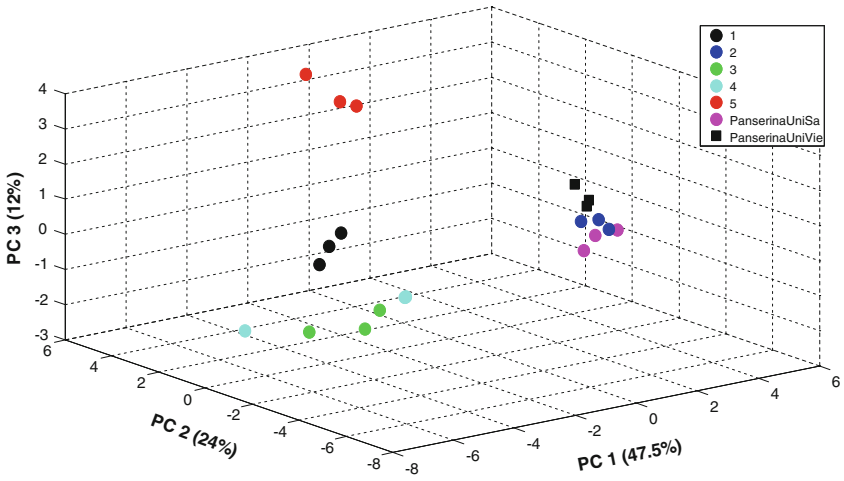


Fig. 4 Sample patterns of combined GC-MS and LC-MS data of the hydroalcoholic extracts of *P. anserina*

latter case chlorogenic acid (Rt 19.6 min, *m/z* 377.0477), myricetin-3-*O*-glucuronide (Rt 30.6 min, *m/z* 495.0769), acacetin-7-*O*-rutinoside (Rt 41.7 min, *m/z* 593.1863) and genistein (Rt 47.0 min, *m/z* 271.0601) (Table 2) are responsible for the differences among our samples. Both GC and LC-MS PCA plots showed the same tendencies between crude extracts and commercial samples (see Figs. 3, 4). In particular it is observed that "PanserinaUniSa" and "PanserinaUnivie" can be considered to be very similar and also very close to hydroalcoholic extract "#2". As shown in Fig. 3, extracts # 1, 3, 4 and 5 are far away from the other three samples.

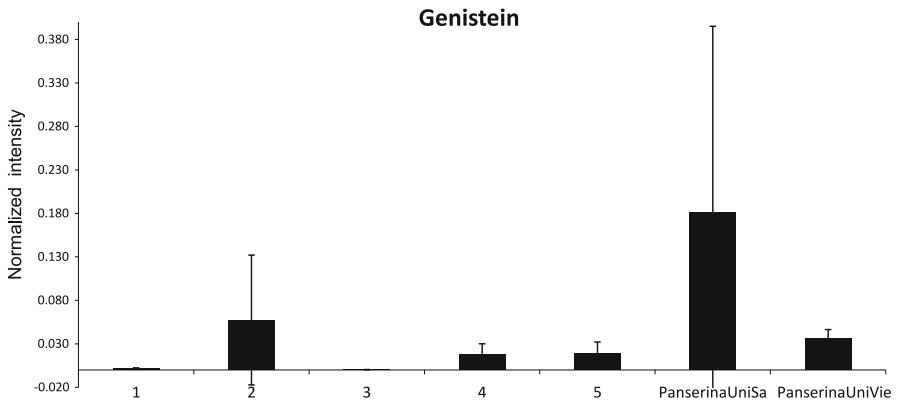


Fig. 5 Relative quantitative analysis of genistein in the commercial products (# 1, 2, 3, 4, 5) and in the hydroaloholic extracts PanserinaUnisa and UniVie. Values are mean of triplicates for each sample. *Error bars* indicate the standard deviation (SD±) values for each histogram

To obtain a more comprehensive view of the LC-MS data a second PCA was applied to a dataset pre-processed by MetMAX Beta 1.0. After calculating the intensity mean, standard deviation and the relative standard deviation (RSD) among three technical replicates for all the peaks, the RSD mean was estimated for each peak and only those with a value <25 were selected for multivariate statistical analysis. By application of this method we selected 1,866 variables for PCA (Fig. 3c). Both plots, Fig. 3b and c, are very similar indicating the robustness of the LC-MS MetMAX Beta 1.0 approach.

Eventually the integration of GC-MS and LC-MS data into one data matrix for PCA showed the clear separation of the hydroalcoholic extracts PanserinaUniSa, PanserinaUniVie and #2 from the other extracts (Fig. 4). The loadings (supplementary Table 1) of this PCA plot demonstrates the different importance of either GC-MS or LC-MS compounds for sample classification. Synergistic effects of data integration for sample pattern recognition were also recently revealed in studies for the integration of primary and secondary metabolism as well as due to the integration of metabolomic and proteomic data (Morgenthal et al. 2005; Wienkoop et al. 2008; Doerfler et al. 2012). The integration of GC-MS and LC-MS data enables the search of precursor-product correlations in biosynthetic pathways. This was recently shown in the study by Doerfler et al. (2012) using Granger causality analysis to reveal the biosynthetic interface of primary and secondary metabolism.

3.3 Comparative analysis of genistein in extracts from different pharmacy vendors

Figure 5 shows the relative evaluation of genistein in all samples. The values were obtained after normalizing the LC-MS intensities of this compound against the total counts of all variables within a sample (calculated with MetMAX Beta 1.0). The results show the higher amount of this isoflavone in the extracts of PanserinaUniSa and PanserinaUniVie as well as in the commercial product 2, thus confirming the similarities between these three samples already deduced from PCA analysis. Since genistein is considered as an active compound in estrogenic therapy (Ferrante et al. 2004; Hellstrom and Muntzing 2012), our results highlight that genistein intake may change depending on the origin of different commercial products, thereby having different effects on the treatment of PMS.

4 Conclusion

In this study we report for the first time a high resolution LC-MS method for the evaluation of the chemical composition of *P. anserina* polar extracts. By this accurate and sensitive analysis we revealed the presence of compounds never reported for *P. anserina*. Especially important is the identification of the isoflavone genistein which is considered as an active compound in the estrogenic therapy. This fact may explain the positive effect of *P. anserina* polar extracts in the treatment of premenstrual syndrome diseases.

Moreover our results showed the advantages of applying an integrated LC-MS, GC-MS metabolomics platform for the evaluation of the similarities between medicinal plant extracts and their commercial products. The unbiased assignment of *m/z* features to sample classification using Mass Accuracy Precursor Alignment (MAPA) and the corresponding MetMAX algorithm in combination with multivariate statistics [MAPA and COVAIN; (Hoehenwarter et al. 2008; Sun and Weckwerth 2012)] opens up opportunities to identify novel compounds in the medicinal plant extracts which were previously not detected. We have discussed two of these unknowns and will address these investigations in more detail in future studies.

Acknowledgements We thank the University of Vienna for generous support. DL is supported by the FWF grant P23441-B20. We thank the anonymous reviewers for very helpful suggestions

References

Bundesgesundheitsamt (1985). Monograph Anserinae herba. *Bundesanzeiger*, *223*.

Bundesgesundheitsamt (1990). Monograph Anserinae herba. *Bundesanzeiger*, *50*.

Doerfler, H., Lyon D., Nägele T., et al. (2012). Granger causality in integrated GC-MS and LC-MS metabolomics data reveals the interface of primary and secondary metabolism. *Metabolomics*. doi:10.1007/s11306-012-0470-0.

Ferrante, F., Fusco, E., Calabresi, P., & Cupini, L. M. (2004). Phytooestrogens in the prophylaxis of menstrual migraine. *Clinical Neuropharmacology, 27*, 137–140.

Hellstrom, A. C., & Muntzing, J. (2012). The pollen extract Femal: a nonestrogenic alternative to hormone therapy in women with menopausal symptoms. *Menopause-the Journal of the North American Menopause Society, 19*, 825–829.

Hiller, K. (1994). Potentilla. *Hager's Handbuch der Pharmazeutischen Praxis* (5th ed.), vol. 6, (pp. 254–269). Berlin: Springer.

Hoehenwarter, W., van Dongen, J. T., Wienkoop, S., 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.

Horai, H., Arita, M., Kanaya, S., et al. (2010). MassBank: A public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry, 45*, 703–714.

Ismail, B., & Hayes, K. (2005). Beta-glycosidase activity toward different glycosidic forms of isoflavones. *Journal of Agricultural and Food Chemistry, 53*, 4918–4924.

Jung, H. A., Kim, A. R., Chung, H. Y., & Choi, J. S. (2002). In vitro antioxidant activity of some selected Prunus species in Korea. *Archives of Pharmacal Research, 25*, 865–872.

Kim, H. Y., Moon, B. H., Lee, H. J., & Choi, D. H. (2004). Flavonol glycosides from the leaves of *Eucommia ulmoides* O. with glycation inhibitory activity. *Journal of Ethnopharmacology, 93*, 227–230.

Kombal, R., & Glasl, H. (1995). Flavan-3-Ols and flavonoids from Potentilla-Anserina. *Planta Medica, 61*, 484–485.

Lee, M. H., Son, Y. K., & Han, Y. N. (2002). Tissue factor inhibitory flavonoids from the fruits of *Chaenomeles sinensis*. *Archives of Pharmacal Research, 25*, 842–850.

Merfort, I., & Wendisch, D. (1988). Flavonoid Glucuronides from the flowers of *Arnica montana*1. *Planta Medica, 54*, 247–250.

Morgenthal, K., Wienkoop, S., Scholz, M., et al. (2005). Correlative GC-TOF-MS based metabolite profiling and LC-MS based protein profiling reveal time-related systemic regulation of metabolite-protein networks and improve pattern recognition for multiple biomarker selection. *Metabolomics, 1*, 109–121.

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.

Schimmer, O., & Lindenbaum, M. (1995). Tannins with antimutagenic properties in the herb of *Alchemilla* species and potentilla-anserina. *Planta Medica, 61*, 141–145.

Schulz, V., Hänsel R., & Tyler V. E. (1998). Rational phytotherapy. A physicians' guide to herbal medicine (3rd ed.) (p. 245). Berlin: Springer-Verlag.

Şöhretoğlu, D., & Sterner, O. (2011). Isoflavonoids, flavonoids and flavans from *Potentilla astracanica*. *Biochemical Systematics and Ecology, 39*, 666–668.

Stein, S. E. (1999). An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *Journal of the American Society for Mass Spectrometry, 10*, 770–781.

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.

Swiezewska, E., & Chojnacki, T. (1989). The occurrence of unique, long-chain polyprenols in the leaves of *Potentilla* species. *Acta Biochimica Polonica, 36*, 143–158.

Tohno, H., Horii, C., Fuse, T., et al. (2010). Evaluation of estrogen receptor beta binding of pruni cortex and its constituents. *Yakugaku Zasshi-Journal of the Pharmaceutical Society of Japan, 130*, 989–997.

Tomczyk, M., Bazylko, A., & Staszewska, A. (2010). Determination of polyphenolics in extracts of *Potentilla* species by high-performance thin-layer chromatography photodensitometry method. *Phytochemical Analysis, 21*, 174–179.

Tomczyk, M., & Latté, K. P. (2009). Potentilla: A review of its phytochemical and pharmacological profile. *Journal of Ethnopharmacology, 122*, 184–204.

Wang, M. F., Kikuzaki, H., Csiszar, K., et al. (1999). Novel trisaccharide fatty acid ester identified from the fruits of *Morinda citrifolia* (Noni). *Journal of Agricultural and Food Chemistry, 47*, 4880–4882.

Weckwerth, W. (2010). Metabolomics: An integral technique in systems biology. *Bioanalysis, 2*, 829–836.

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.

Wienkoop, S., Morgenthal, K., Wolschin, F., et al. (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*. *Molecular and Cellular Proteomics, 7*, 1725–1736.

Xu, J. F., Zheng, X. P., Liu, W. D., et al. (2010). Flavonol glycosides and monoterpenoids from *Potentilla anserina*. *Journal of Asian Natural Products Research, 12*, 529–534.

Zou, Y. N., Kim, A. R., Kim, J. E., et al. (2002). Peroxynitrite scavenging activity of sinapic acid (3,5-dimethoxy-4-hydroxycinnamic acid) isolated from *Brassica juncea*. *Journal of Agricultural and Food Chemistry, 50*, 5884–5890.

Outlook

^{15}N labeling, biological application

Subsequently, the dynamics of $^{15}N/^{14}N$ incorporation in time are the basis for calculating turnover rates. Studying *M. truncatula* in response to perturbation (see Research objectives) should permit one to follow fluxes of partial metabolic labeling and thus enable the inference of biological/physiological meaning. The time-dependent increased incorporation of ^{15}N of the drought-stressed group compared to the control group is apparent in Figure 3.A. The sponge-like re-hydration of the drought-stressed plants in the recovery phase causes an increased level of ^{15}N-enriched fertilizer, and thus leads to a pronounced ^{15}N-enrichment of amino acids and subsequently proteins, which in turn results in generally higher RIA values. Nevertheless, not all drought-stressed proteins show an increased RIA compared to the control group. Dependent on the plants' regulatory mechanisms, specific proteins display higher or lower turnover when comparing the two treatments. This is exemplarily demonstrated in Figure 3.B and C, where B shows a higher RIA in the drought-stressed group, whereas C shows a higher RIA in the control group. A peer-reviewed article with novel insights concerning the differential regulation of proteins during the recovery phase of drought stress, of N-fed *M. truncatula*, will be published (**Lyon and Castillejo et al. 2014** *in preparation*).

Graphical User Interface for the automated protein turnover program and computational speed improvements

In order to facilitate the use of the presented program (**Lyon et al. 2014**), a Graphical User Interface (GUI), using PyQt, is planned. A precompiled executable for the Windows as well as the Apple OSX Operating System will be freely available to download, making the installation of the Python programming language, as well as various dependent packages/modules unneces-

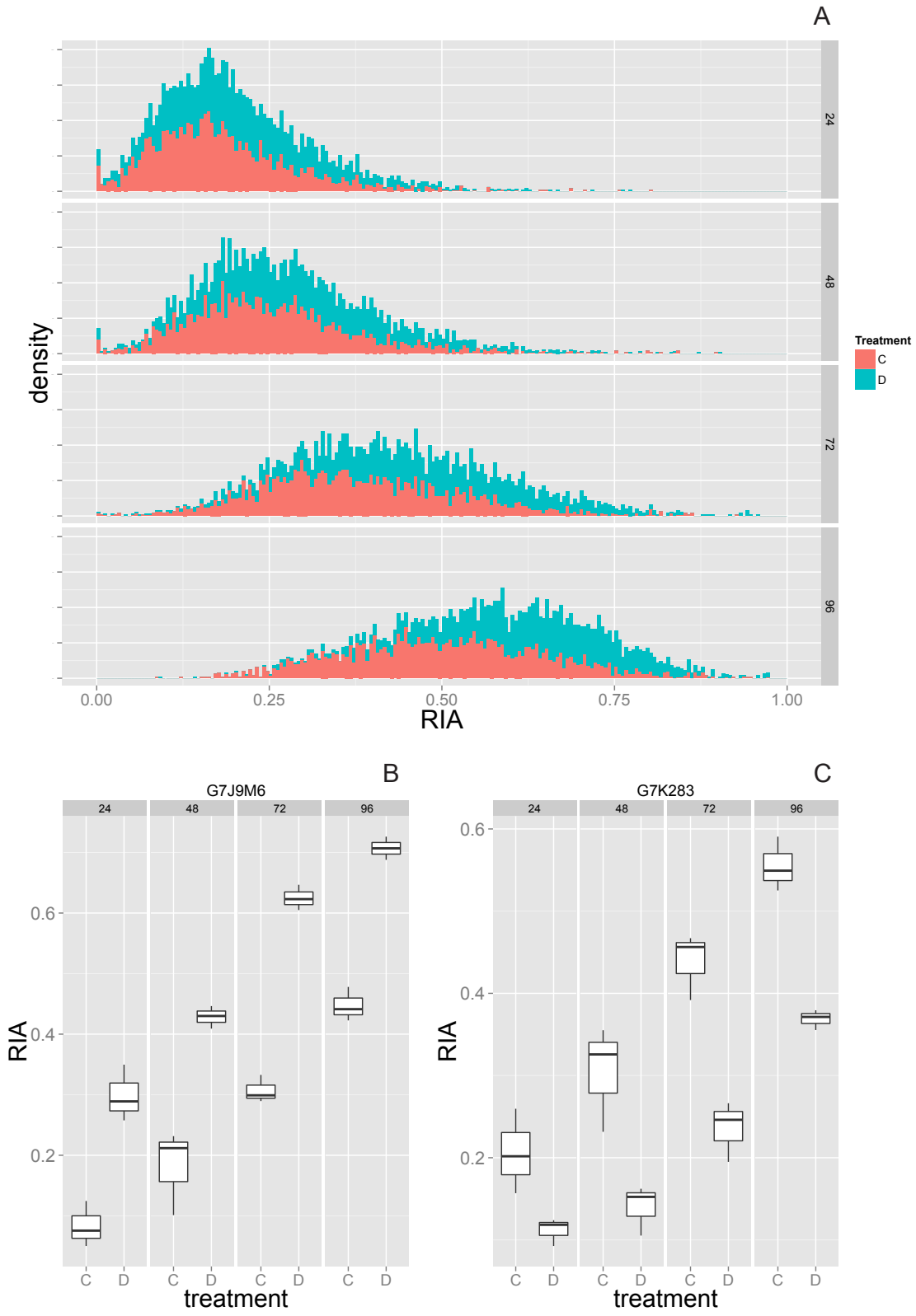


Figure 3 Legend:

Figure 3.A displays Histograms of the RIA versus the density of four TimePoints (from top to bottom 24, 48, 72, and 96 hours). The control group is colored in red and the drought-stress treated group in blue. Figure 3.B and C show Boxplots of the RIA comparing the control to the treatment group, sub-graphing the aforementioned TimePoints (from left to right). B shows the values for the Accession Number "G7J9M6", and C for the Accession Number "G7K283". Each individual Boxplot is composed of three biological replicates, which consist of averaged values from eight peptide signals from two technical replicates each. TimePoint 0h is omitted in all sub-graphs since the RIA is zero per definition (see Publications) and thus does not contain valuable information to display.

sary, since such a precompiled program can simply be executed (started) by the user without the need for installation. Also, an intuitive GUI will ease the handling of selecting source data and setting parameters for data processing. Additionally, various strategies will be employed in order to gain computational speed. Specifically, refactoring pure Python code, implementing Cython modules where deemed useful, and attempting to implement multithreading/multiprocessing. Furthermore, PTM support could be implemented. This work is planned to be subsumed in a publication in the near future.

Further publications submitted or in progress

I am the co-author of the following peer-reviewed publications, which are either accepted, submitted for review, or in preparation.

Recuenco-Munoz et al. 2014 (*accepted*), is a targeted study of RuBisCO of *Chlamydomonas reinhardtii* using proteomics and transcriptomics methodology. I have contributed to this publication by assisting in the development of an improved Mass Spectrometry method for the exact determination of stoichiometric complexes of proteins (see Publications: additional remarks of the Mass Western).

Gil et al. 2014 (*submitted*), investigates drought stress of the root nodule proteome of *M. truncatula* in comparison to *G. max*. I have contributed to the publication by preparing a merged non-redundant protein FASTA file, utilized the Mercator pipeline for functional annotation, and used BLAST for homology searches. This work is an extension of previously published work (**Staudinger et al. 2012**) (see Publications).

Lyon and Castillejo et al. 2014 (*in preparation*) (see Outlook: ^{15}N labeling biological application).

Meisrimler et al. 2014 (*in preparation*) is a study of long-term iron deficiency on *P. sativum* cultivars. I have contributed to the publication by preparing a merged non-redundant protein FASTA file, utilized the Mercator pipeline for functional annotation, and used BLAST for homology searches. This work is an extension of previously published work (**Staudinger et al. 2012**) (see Publications).

Concluding discussion

Consolidation of publications

The work within this thesis ranges from LC/MS method development and application to biological samples, to data analysis and application of available software and algorithms, to the creation of a novel algorithm. Understanding technological possibilities as well as their restraints is essential for method development and critical for the subsequent data analysis. These are intertwined processes and iteration is often necessary in order to acquire the desired results. Knowledge of and experience with the instruments in question is not only beneficial for troubleshooting, but more importantly, often explains variation of data due to technical issues, which can be solved/allayed on the instrument side of the experiment and/or at the resulting data. Particularly, batch to batch variance, nanoESI droplet formation, LC-pump problems, clogging of the ESI-tip, ion optics contamination, etc. can lead to gradual as well as to abrupt changes in the acquired data, which need to be differentiated from biological variation, in order to decide how to proceed with data analysis and to arrive at meaningful conclusions concerning biology. Small changes in MS settings can lead to dramatic differences in the resulting data, for better or worse. Understanding the logic of the instrument methods can only improve data evaluation and analysis. This expertise was crucial for the development of the ProtOver algorithm for automated data extraction of very particular and complex MS data.

Concluding the analysis, delineating the preamble and contribution to the scientific progress

There is not a single all-encompassing analytical platform to analyze all possible biochemical compounds. Generally, Mass Spectrometry is capable of detecting any compound that can be ionized and is known for its high sensitivity. Nevertheless, the multitude of biochemical compounds and their large dynamic range pose a great challenge to any analytical platform. Therefore, as previously mentioned, specialized Mass Spectrometry instruments exist in

conjunction with specialized methods, which need to be adapted and developed, also with regard to the rapid technical developments in the field of Mass Spectrometry (**Lyon, Weckwerth, and Wienkoop 2014**). Analogously to the data acquisition, there isn't a single process to follow, a method that can be applied, to integrate and analyze data of various sources that would find acceptance by the entire scientific community. Various strategies and statistics can be applied to gain biological insights, though they need to be reproducible and their application justifiable. The publications constituting this PhD thesis prove the successful application of aspects of this field of research and have led to derive meaningful biological insights. Integrating and linking data from primary and secondary metabolism is pivotal to deepen our understanding of molecular mechanisms. A successful implementation of such an approach has been shown by **Doerfler et al. 2013** (amongst others), constituting part of this thesis. Furthermore, slightly adapted methodologies were successfully used in other publications pertinent to this thesis and are a valuable resource for the scientific community (**Doerfler et al. 2013**; **Lyon, Weckwerth, and Wienkoop 2014**; **Doerfler et al. 2014**; **Mari et al. 2013**).

The functional characterization and visualization of data is not a singular event, in the sense that it can be performed and would thus be concluded, but an ongoing process. As information increases (e.g. improved genomic data and gene prediction thereof, influences translated proteomic data), iterations of this work need to be performed in order to refine functional annotations and visualize newly generated data (May et al. 2008). There are difficulties in keeping track of proteins from varying genome annotations of the same species, proving the importance of reliable and manually curated databases (such as SwissProt) and the need to continuously update them with novel data. There is a need for bioinformatics work which can be realized with scripting languages to perform simple but powerful tasks, accompanying almost every step of the scientific process. My varying contributions of this type of work led to three of the publications constituting this thesis **Staudinger et al. 2012**; **Ischebeck et al. 2014**; and **Lyon et al. 2014**. The MapMan mapping file, containing functional annotation, can be freely downloaded (http://www.univie.ac.at/mosys/databases.html) and thus serves as a valuable resource to the scientific community. Furthermore, the methodological processes illustrated in these publications

can be adapted for analogous studies.

Finally, the successful implementation of a novel algorithm for the extraction of LC/MS data, derived from partial metabolic labeling experiments, into an automated software, constitutes a major part of this thesis (**Lyon et al. 2014**). This work went beyond “simple scripting”. The software is freely available to the scientific community. The publication delineates the individual steps of the algorithm in written and mathematical form. The latter can be used as provided to extract MS data from other analogous experimental data of differing species or organs, but could also be adapted to individual goals/focus, if need be. Potentially, the data generated with the software in conjunction with other information could be used to calculate rate constants, which could be implemented in modelling biological systems. The utility of the ProtOver algorithm is exemplarily demonstrated by showing the differential regulation of protein turnover due to abiotic stress (see Outlook, Figure 3).

Bibliography

Adams, M D, J M Kelley, J D Gocayne, M Dubnick, M H Polymeropoulos, H Xiao, C R Merril, A Wu, B Olde, and R F Moreno. 1991. “Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project.” Science (New York, N.Y.) 252 (5013): 1651–56. http://www.ncbi.nlm.nih.gov/pubmed/2047873.

Aebersold, Ruedi, Alma L Burlingame, and Ralph a Bradshaw. 2013. “Western Blots versus Selected Reaction Monitoring Assays: Time to Turn the Tables?” Molecular & Cellular Proteomics : MCP 12 (9): 2381–82. doi:10.1074/mcp.E113.031658. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3769317&tool=pmcentrez&rendertype=abstract.

Aebersold, Ruedi, and Matthias Mann. 2003. “Mass Spectrometry-Based Proteomics.” Nature 422 (6928): 198–207. doi:10.1038/nature01511. http://www.ncbi.nlm.nih.gov/pubmed/12634793.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. “Basic Local Alignment Search Tool.” J. Mol. Biol., 403–10. doi:10.1016/S0022-2836(05)80360-2.

Apel, Klaus, and Heribert Hirt. 2004. “Reactive Oxygen Species: Metabolism, Oxidative Stress, and Signal Transduction.” Annual Review of Plant Biology 55 (January): 373–99. doi:10.1146/annurev.arplant.55.031903.141701. http://www.ncbi.nlm.nih.gov/pubmed/15377225.

Apweiler, Rolf, and The Uniprot Consortium. 2012. “Reorganizing the Protein Space at the Universal Protein Resource (UniProt).” Nucleic Acids Research 40 (Database issue): D71–5. doi:10.1093/nar/gkr981. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245120&tool=pmcentrez&rendertype=abstract.

Ashburner, M., C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, et al. 2000. “Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium.” Nature Genetics 25 (1): 25–29. doi:10.1038/75556. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3037419&tool=pmcentrez&rendertype=abstract.

Baerenfaller, Katja, Jonas Grossmann, Monica a Grobei, Roger Hull, Matthias Hirsch-Hoffmann, Shaul Yalovsky, Philip Zimmermann, Ueli Grossniklaus, Wilhelm Gruissem, and Sacha Baginsky. 2008. “Genome-Scale Proteomics Reveals Arabidopsis Thaliana Gene Models and Proteome Dynamics.” Science (New York, N.Y.) 320 (5878): 938–41. doi:10.1126/science.1157956. http://www.ncbi.nlm.nih.gov/pubmed/18436743.

Bald, Till, Johannes Barth, Anna Niehues, Michael Specht, and Michael Hippler. 2012. “pymzML - Python Module for High Throughput Bioinformatics on Mass Spectrometry Data and Christian Fufezan”, 1–2.

Bantscheff, Marcus, Markus Schirle, Gavain Sweetman, Jens Rick, and Bernhard Kuster. 2007. “Quantitative Mass Spectrometry in Proteomics: A Critical Review.” Analytical and Bioanalytical Chemistry 389 (4): 1017–31. doi:10.1007/s00216-007-1486-6. http://www.ncbi.nlm.nih.gov/pubmed/17668192.

Bourgeois, Michael, Françoise Jacquin, Florence Cassecuelle, Vincent Savois, Maya Belghazi, Grégoire Aubert, Laurence Quillien, Myriam Huart, Pascal Marget, and Judith Burstin. 2011. “A PQL (protein Quantity Loci) Analysis of Mature Pea Seed Proteins Identifies Loci Determining Seed Protein Composition.” Proteomics 11 (9): 1581–94. doi:10.1002/pmic.201000687. http://www.ncbi.nlm.nih.gov/pubmed/21433288.

Cargile, Benjamin J, Jonathan L Bundy, Amy M Grunden, and James L Stephenson. 2004. “Synthesis/degradation Ratio Mass Spectrometry for Measuring Relative Dynamic Protein Turnover.” Analytical Chemistry 76 (1): 86–97. doi:10.1021/ac034841a. http://www.ncbi.nlm.nih.gov/pubmed/14697036.

Chambers, Matthew C, Brendan Maclean, Robert Burke, Dario Amodei, Daniel L Ruderman, Steffen Neumann, Laurent Gatto, et al. 2012. “A Cross-Platform Toolkit for Mass Spectrometry and Proteomics.” Nature Biotechnology 30 (10): 918–20. doi:10.1038/nbt.2377. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3471674&tool=pmcentrez&rendertype=abstract.

Choi, Meena, Ching-Yun Chang, Timothy Clough, Daniel Broudy, Trevor Killeen, Brendan MacLean, and Olga Vitek. 2014. “MSstats: An R Package for Statistical Analysis of Quantitative Mass Spectrometry-Based Proteomic Experiments.” Bioinformatics (Oxford, England) 30 (17): 2524–26. doi:10.1093/bioinformatics/btu305. http://www.ncbi.nlm.nih.gov/pubmed/24794931.

Claydon, Amy J, and Robert Beynon. 2012. “Proteome Dynamics: Revisiting Turnover with a Global Perspective.” Molecular & Cellular Proteomics : MCP 11 (12): 1551–65. doi:10.1074/mcp.O112.022186. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3518130&tool=pmcentrez&rendertype=abstract.

Collins, Ben C, Ludovic C Gillet, George Rosenberger, Hannes L Röst, Anton Vichalkovski, Matthias Gstaiger, and Ruedi Aebersold. 2013. “Quantifying Protein Interaction Dynamics by SWATH Mass Spectrometry: Application to the 14-3-3 System.” Nature Methods 10 (12): 1246–53. doi:10.1038/nmeth.2703. http://www.ncbi.nlm.nih.gov/pubmed/24162925.

Cox, Jurgen, Marco Y Hein, Christian a Luber, Igor Paron, Nagarjuna Nagaraj, and Matthias Mann. 2014. “MaxLFQ Allows Accurate Proteome-Wide Label-Free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction.” Molecular & Cellular Proteomics : MCP, June. doi:10.1074/mcp.M113.031591. http://www.ncbi.nlm.nih.gov/pubmed/24942700.

Cox, Jürgen, and Matthias Mann. 2008. “MaxQuant Enables High Peptide Identification Rates, Individualized P.p.b.-Range Mass Accuracies and Proteome-Wide Protein Quantification.” Nature Biotechnology 26 (12): 1367–72. doi:10.1038/nbt.1511. http://www.ncbi.nlm.nih.gov/pubmed/19029910.

Cox, Jürgen, Nadin Neuhauser, Annette Michalski, Richard a Scheltema, Jesper V Olsen, and Matthias Mann. 2011. “Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment.” Journal of Proteome Research 10 (4): 1794–1805. doi:10.1021/pr101065j. http://www.ncbi.nlm.nih.gov/pubmed/21254760.

D’Auria, John C, Jonathan Gershenzon, and John C D Auria. 2005. “The Secondary Metabolism of Arabidopsis Thaliana: Growing like a Weed.” Current Opinion in Plant Biology 8 (3): 308–16. doi:10.1016/j.pbi.2005.03.012. http://www.ncbi.nlm.nih.gov/pubmed/15860428.

Deutsch, Eric W, Luis Mendoza, David Shteynberg, Terry Farrah, Henry Lam, Natalie Tasman, Zhi Sun, et al. 2010. “A Guided Tour of the Trans-Proteomic Pipeline.” Proteomics 10 (6): 1150–59. doi:10.1002/pmic.200900375. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3017125&tool=pmcentrez&rendertype=abstract.

Doerfler, Hannes, David Lyon, Thomas Nägele, Xiaoliang Sun, Lena Fragner, Franz Hadacek, Volker Egelhofer, and Wolfram Weckwerth. 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 (3): 564–74. doi:10.1007/s11306-012-0470-0.

Doerfler, Hannes, Xiaoliang Sun, Lei Wang, Doris Engelmeier, David Lyon, and Wolfram Weckwerth. 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. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4028198&tool=pmcentrez&rendertype=abstract.

Egelhofer, Volker, Wolfgang Hoehenwarter, David Lyon, Wolfram Weckwerth, and

Stefanie Wienkoop. 2013. “Using ProtMAX to Create High-Mass-Accuracy Precursor Alignments from Label-Free Quantitative Mass Spectrometry Data Generated in Shotgun Proteomics Experiments.” Nature Protocols 8 (3): 595–601. doi:10.1038/nprot.2013.013. http://www.nature.com/doifinder/10.1038/nprot.2013.013.

Eng, Jimmy K, Ashley L McCormack, John R Yates, K Jimmy, and R John. 1994. “An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database.” Journal of the American Society for Mass Spectrometry 5 (11): 976–89. doi:10.1016/1044-0305(94)80016-2. http://www.ncbi.nlm.nih.gov/pubmed/24226387.

Gehlenborg, Nils, Seán I O’Donoghue, Nitin S Baliga, Alexander Goesmann, Matthew a Hibbs, Hiroaki Kitano, Oliver Kohlbacher, et al. 2010. “Visualization of Omics Data for Systems Biology.” Nature Methods 7 (3 Suppl). Nature Publishing Group: S56–68. doi:10.1038/nmeth.1436. http://www.ncbi.nlm.nih.gov/pubmed/20195258.

Gillet, Ludovic C, Pedro Navarro, Stephen Tate, Hannes Röst, Nathalie Selevsek, Lukas Reiter, Ron Bonner, and Ruedi Aebersold. 2012. “Targeted Data Extraction of the MS/MS Spectra Generated by Data-Independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis.” Molecular & Cellular Proteomics : MCP 11 (6): O111.016717. doi:10.1074/mcp.O111.016717. http://www.ncbi.nlm.nih.gov/pubmed/22261725.

Glinski, Mirko, and Wolfram Weckwerth. 2006. “The Role of Mass Spectrometry in Plant Systems Biology.” Mass Spectrometry Reviews 25 (2): 173–214. doi:10.1002/mas.20063. http://www.ncbi.nlm.nih.gov/pubmed/16284938.

Goloborodko, Anton a, Lev I Levitsky, Mark V Ivanov, and Mikhail V Gorshkov. 2013. “Pyteomics--a Python Framework for Exploratory Data Analysis and Rapid Software Prototyping in Proteomics.” Journal of the American Society for Mass Spectrometry 24 (2): 301–4. doi:10.1007/s13361-012-0516-6. http://www.ncbi.nlm.nih.gov/pubmed/23292976.

Goodacre, Royston, Seetharaman Vaidyanathan, Warwick B Dunn, George G Harrigan, and Douglas B Kell. 2004. “Metabolomics by Numbers: Acquiring and Understanding Global Metabolite Data.” Trends in Biotechnology 22 (5): 245–52. doi:10.1016/j.tibtech.2004.03.007. http://www.ncbi.nlm.nih.gov/pubmed/15109811.

Görg, Angelika, Günther Boguth, Angelika Köpf, Gerold Reil, Harun Parlar, and Walter Weiss. 2002. “Sample Prefractionation with Sephadex Isoelectric Focusing prior to Narrow pH Range Two-Dimensional Gels.” Proteomics 2 (12): 1652–57. doi:10.1002/1615-9861(200212)2:12<1652::AID-PROT1652>3.0.CO;2-3. http://www.ncbi.nlm.nih.gov/pubmed/12469334.

Graham, Peter H, and Carroll P Vance. 2003. “Legumes: Importance and Constraints to Greater Use.” Plant Physiology 131 (3): 872–77. doi:10.1104/pp.017004. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1540286&tool=pmcentrez&rendertype=abstract.

Griss, Johannes, Andrew R Jones, Timo Sachsenberg, Mathias Walzer, Laurent Gatto, Jurgen Hartler, Gerhard G Thallinger, et al. 2014. “The mzTab Data Exchange Format: Communicating MS-Based Proteomics and Metabolomics Experimental Results to a Wider Audience.” Molecular & Cellular Proteomics : MCP, June, 1–28. doi:10.1074/mcp.O113.036681. http://www.ncbi.nlm.nih.gov/pubmed/24980485.

Gromski, Piotr S, Yun Xu, Helen L Kotze, Elon Correa, David I Ellis, Emily Grace Armitage, Michael L Turner, and Royston Goodacre. 2014. “Influence of Missing Values Substitutes on Multivariate Analysis of Metabolomics Data.” Metabolites 4 (2): 433–52. doi:10.3390/metabo4020433. http://www.mdpi.com/2218-1989/4/2/433/.

Hartl, Markus, and Iris Finkemeier. 2012. “Plant Mitochondrial Retrograde Signaling: Post-Translational Modifications Enter the Stage.” Frontiers in Plant Science 3 (November):

253. doi:10.3389/fpls.2012.00253. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3495340&tool=pmcentrez&rendertype=abstract.
Hebeler, Romano, Silke Oeljeklaus, Kai a Reidegeld, Martin Eisenacher, Christian Stephan, Barbara Sitek, Kai Stühler, et al. 2008. "Study of Early Leaf Senescence in Arabidopsis Thaliana by Quantitative Proteomics Using Reciprocal 14N/15N Labeling and Difference Gel Electrophoresis." Molecular & Cellular Proteomics : MCP 7 (1): 108–20. doi:10.1074/mcp.M700340-MCP200. http://www.ncbi.nlm.nih.gov/pubmed/17878269.
Hinkson, Izumi V, and Joshua E Elias. 2011. "The Dynamic State of Protein Turnover: It's about Time." Trends in Cell Biology 21 (5). Elsevier Ltd: 293–303. doi:10.1016/j.tcb.2011.02.002. http://www.ncbi.nlm.nih.gov/pubmed/21474317.
Hoehenwarter, Wolfgang, Joost T van Dongen, Stefanie Wienkoop, Matthias Steinfath, Jan Hummel, Alexander Erban, Ronan Sulpice, et al. 2008. "A Rapid Approach for Phenotype-Screening and Database Independent Detection of cSNP/protein Polymorphism Using Mass Accuracy Precursor Alignment." Proteomics 8 (20): 4214–25. doi:10.1002/pmic.200701047. http://www.ncbi.nlm.nih.gov/pubmed/18924179.
Hoehenwarter, Wolfgang, and Stefanie Wienkoop. 2010. "Spectral Counting Robust on High Mass Accuracy Mass Spectrometers." Rapid Communications in Mass Spectrometry, 3609–14. doi:10.1002/rcm.
Holzmann, Johann, Peter Pichler, Mathias Madalinski, Robert Kurzbauer, and Karl Mechtler. 2009. "Stoichiometry Determination of the MP1-p14 Complex Using a Novel and Cost-Efficient Method to Produce an Equimolar Mixture of Standard Peptides." Analytical Chemistry 81 (24): 10254–61. doi:10.1021/ac902286m. http://www.ncbi.nlm.nih.gov/pubmed/19924867.
http://en.wikipedia.org/
http://medicago.org/genome/IMGAG/
http://medicago.vbi.vt.edu/
http://proconsortium.org
http://www.ncbi.nlm.nih.gov/
http://www.uniprot.org
Hu, Qizhi, Robert J Noll, Hongyan Li, Alexander Makarov, Mark Hardman, and R Graham Cooks. 2005. "The Orbitrap: A New Mass Spectrometer." Journal of Mass Spectrometry : JMS 40 (4): 430–43. doi:10.1002/jms.856. http://www.ncbi.nlm.nih.gov/pubmed/15838939.
Hummel, Jan, Michaela Niemann, Stefanie Wienkoop, Waltraud Schulze, Dirk Steinhauser, Joachim Selbig, Dirk Walther, Wolfram Weckwerth, and Joachim Selbig Js. 2007. "ProMEX: A Mass Spectral Reference Database for Proteins and Protein Phosphorylation Sites." BMC Bioinformatics 8 (January): 216. doi:10.1186/1471-2105-8-216. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1920535&tool=pmcentrez&rendertype=abstract.
Ischebeck, Till, Luis Valledor, David Lyon, Stephanie Gingl, Matthias Nagler, Mónica Meijón, Volker Egelhofer, and Wolfram Weckwerth. 2014. "Comprehensive Cell-Specific Protein Analysis in Early and Late Pollen Development from Diploid Microsporocytes to Pollen Tube Growth." Molecular & Cellular Proteomics : MCP 13 (1): 295–310. doi:10.1074/mcp.M113.028100. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3879621&tool=pmcentrez&rendertype=abstract.
Kierszniowska, Sylwia, Dirk Walther, and Waltraud X Schulze. 2009. "Ratio-Dependent Significance Thresholds in Reciprocal ^{15}N-Labeling Experiments as a Robust Tool in Detection of Candidate Proteins Responding to Biological Treatment." Proteomics 9 (7): 1916–24. doi:10.1002/pmic.200800443. http://www.ncbi.nlm.nih.gov/pubmed/19260003.

Kind, Tobias, and Oliver Fiehn. 2006. “Metabolomic Database Annotations via Query of Elemental Compositions: Mass Accuracy Is Insufficient Even at Less than 1 Ppm.” BMC Bioinformatics 7 (January): 234. doi:10.1186/1471-2105-7-234. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1464138&tool=pmcentrez&rendertype=abstract.

Kind, Tobias, and Oliver Fiehn. 2007. “Seven Golden Rules for Heuristic Filtering of Molecular Formulas Obtained by Accurate Mass Spectrometry.” BMC Bioinformatics 8 (January): 105. doi:10.1186/1471-2105-8-105. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1851972&tool=pmcentrez&rendertype=abstract.

Klie, Sebastian, and Zoran Nikoloski. 2012. “The Choice between MapMan and Gene Ontology for Automated Gene Function Prediction in Plant Science.” Frontiers in Genetics 3 (June): 1–14. doi:10.3389/fgene.2012.00115. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3384976&tool=pmcentrez&rendertype=abstract.

Kline, Kelli G, and Michael R Sussman. 2010. “Protein Quantitation Using Isotope-Assisted Mass Spectrometry.” Annual Review of Biophysics 39 (January): 291–308. doi:10.1146/annurev.biophys.093008.131339. http://www.ncbi.nlm.nih.gov/pubmed/20462376.

Kohlbacher, Oliver, Knut Reinert, Clemens Gröpl, Eva Lange, Nico Pfeifer, Ole Schulz-Trieglaff, and Marc Sturm. 2007. “TOPP--the OpenMS Proteomics Pipeline.” Bioinformatics (Oxford, England) 23 (2): e191–7. doi:10.1093/bioinformatics/btl299. http://www.ncbi.nlm.nih.gov/pubmed/17237091.

Koulman, Albert, Gary Woffendin, Vinod K Narayana, Helen Welchman, Catharina Crone, and Dietrich A Volmer. 2009. “High-Resolution Extracted Ion Chromatography , a New Tool for Metabolomics and Lipidomics Using a Second- Generation Orbitrap Mass Spectrometer.” Rapid Communications in Mass Spectrometry, 1411–18. doi:10.1002/rcm.

Lange, Vinzenz, Paola Picotti, Bruno Domon, and Ruedi Aebersold. 2008. “Selected Reaction Monitoring for Quantitative Proteomics: A Tutorial.” Molecular Systems Biology 4 (222): 222. doi:10.1038/msb.2008.61. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2583086&tool=pmcentrez&rendertype=abstract.

Larrainzar, Estíbaliz, Stefanie Wienkoop, Christian Scherling, Stefan Kempa, Rubén Ladrera, Cesar Arrese-Igor, Wolfram Weckwerth, and Esther M González. 2009. “Carbon Metabolism and Bacteroid Functioning Are Involved in the Regulation of Nitrogen Fixation in Medicago Truncatula under Drought and Recovery.” Molecular Plant-Microbe Interactions : MPMI 22 (12): 1565–76. doi:10.1094/MPMI-22-12-1565. http://www.ncbi.nlm.nih.gov/pubmed/19888822.

Larrainzar, Estíbaliz, Stefanie Wienkoop, Wolfram Weckwerth, Rubén Ladrera, Cesar Arrese-Igor, and Esther M González. 2007. “Medicago Truncatula Root Nodule Proteome Analysis Reveals Differential Plant and Bacteroid Responses to Drought Stress.” Plant Physiology 144 (3): 1495–1507. doi:10.1104/pp.107.101618. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1914115&tool=pmcentrez&rendertype=abstract.

Lee, Kimberly a, Chris Farnsworth, Wen Yu, and Leo E Bonilla. 2011. “24-Hour Lock Mass Protection.” Journal of Proteome Research 10 (2): 880–85. doi:10.1021/pr100780b. http://www.ncbi.nlm.nih.gov/pubmed/21133379.

Lehmann, Ute, Stefanie Wienkoop, Hendrik Tschoep, and Wolfram Weckwerth. 2008. “If the Antibody Fails--a Mass Western Approach.” The Plant Journal : For Cell and Molecular Biology 55 (6): 1039–46. doi:10.1111/j.1365-313X.2008.03554.x. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2607522&tool=pmcentrez&rendertype=abstract.

Li, Lei, Clark J Nelson, Cory Solheim, James Whelan, and A Harvey Millar. 2012. “Determining Degradation and Synthesis Rates of Arabidopsis Proteins Using the

Kinetics of Progressive ^{15}N Labeling of Two-Dimensional Gel-Separated Protein Spots." Molecular & Cellular Proteomics : MCP 11 (6): M111.010025. doi:10.1074/mcp.M111.010025. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3433911&tool=pmcentrez&rendertype=abstract.

Li, Qingbo. 2010. "Advances in Protein Turnover Analysis at the Global Level and Biological Insights." Mass Spectrometry Reviews 29 (5): 717–36. doi:10.1002/mas.20261. http://www.ncbi.nlm.nih.gov/pubmed/19757418.

Lohse, Marc, Axel Nagel, Thomas Herter, Patrick May, Michael Schroda, Rita Zrenner, Takayuki Tohge, Alisdair R Fernie, Mark Stitt, and Björn Usadel. 2014. "Mercator: A Fast and Simple Web Server for Genome Scale Functional Annotation of Plant Sequence Data." Plant, Cell & Environment 37 (5): 1250–58. doi:10.1111/pce.12231. http://www.ncbi.nlm.nih.gov/pubmed/24237261.

Lommen, Arjen. 2009. "MetAlign: Interface-Driven, Versatile Metabolomics Tool for Hyphenated Full-Scan Mass Spectrometry Data Preprocessing." Analytical Chemistry 81 (8): 3079–86. doi:10.1021/ac900036d. http://www.ncbi.nlm.nih.gov/pubmed/19301908.

Lommen, Arjen, Arjen Gerssen, and J Efraim Oosterink. 2011. "Ultra-Fast Searching Assists in Evaluating Sub-Ppm Mass Accuracy Enhancement in U-HPLC / Orbitrap MS Data", 15–24. doi:10.1007/s11306-010-0230-y.

Lu, Wenyun, Michelle F Clasquin, Eugene Melamud, Daniel Amador-Noguez, Amy a Caudy, and Joshua D Rabinowitz. 2010. "Metabolomic Analysis via Reversed-Phase Ion-Pairing Liquid Chromatography Coupled to a Stand Alone Orbitrap Mass Spectrometer." Analytical Chemistry 82 (8): 3212–21. doi:10.1021/ac902837x. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2863137&tool=pmcentrez&rendertype=abstract.

Lyon, David, Maria Angeles Castillejo, Christiana Staudinger, Wolfram Weckwerth, Stefanie Wienkoop, and Volker Egelhofer. 2014. "Automated Protein Turnover Calculations from ^{15}N Partial Metabolic Labeling LC/MS Shotgun Proteomics Data." PloS One 9 (4): e94692. doi:10.1371/journal.pone.0094692. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3988089&tool=pmcentrez&rendertype=abstract.

Lyon, David, Wolfram Weckwerth, and Stefanie Wienkoop. 2014. „ Mass Western for absolute quantification of target proteins and considerations about the instrument of choice.", Plant Proteomics. Edited by Jesus V. Jorrin-Novo, Setsuko Komatsu, Wolfram Weckwerth, and Stefanie Wienkoop. Vol. 1072. Methods in Molecular Biology. Totowa, NJ: Humana Press. doi:10.1007/978-1-62703-631-3. http://link.springer.com/10.1007/978-1-62703-631-3.

Mann, Matthias, Nagarjuna Nagaraj, Jacek R Wisniewski, Tamar Geiger, Juergen Cox, Martin Kircher, Janet Kelso, Svante Pa, and Svante Pääbo. 2011. "Deep Proteome and Transcriptome Mapping of a Human Cancer Cell Line." Molecular Systems Biology 7 (548): 548. doi:10.1038/msb.2011.81. http://www.ncbi.nlm.nih.gov/pubmed/22068331.

Mari, Angela, David Lyon, Lena Fragner, Paola Montoro, Sonia Piacente, Stefanie Wienkoop, Volker Egelhofer, and Wolfram Weckwerth. 2013. "Phytochemical Composition of Potentilla Anserina L. Analyzed by an Integrative GC-MS and LC-MS Metabolomics Platform." Metabolomics : Official Journal of the Metabolomic Society 9 (3): 599–607. doi:10.1007/s11306-012-0473-x. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3651535&tool=pmcentrez&rendertype=abstract.

Martens, Lennart, Matthew Chambers, Marc Sturm, Darren Kessner, Fredrik Levander, Jim Shofstahl, Wilfred H Tang, et al. 2011. "mzML--a Community Standard for Mass Spectrometry Data." Molecular & Cellular Proteomics : MCP 10 (1): R110.000133. doi:10.1074/mcp.R110.000133. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3013463&tool=pmcentrez&rendertype=abstract.

Martin, Sarah F, Vijaya S Munagapati, Eliane Salvo-Chirnside, Lorraine E Kerr, and Thierry Le Bihan. 2012. “Proteome Turnover in the Green Alga Ostreococcus Tauri by Time Course ^{15}N Metabolic Labeling Mass Spectrometry.” Journal of Proteome Research 11 (1): 476–86. doi:10.1021/pr2009302. http://www.ncbi.nlm.nih.gov/pubmed/22077659.

Matsuda, Fumio, Yoko Shinbo, Akira Oikawa, Masami Yokota Hirai, Oliver Fiehn, Shigehiko Kanaya, and Kazuki Saito. 2009. “Assessment of Metabolome Annotation Quality: A Method for Evaluating the False Discovery Rate of Elemental Composition Searches.” PloS One 4 (10): e7490. doi:10.1371/journal.pone.0007490. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2761541&tool=pmcentrez&rendertype=abstract.

Matsuda, Fumio, Keiko Yonekura-Sakakibara, Rie Niida, Takashi Kuromori, Kazuo Shinozaki, and Kazuki Saito. 2009. “MS/MS Spectral Tag-Based Annotation of Non-Targeted Profile of Plant Secondary Metabolites.” The Plant Journal : For Cell and Molecular Biology 57 (3): 555–77. doi:10.1111/j.1365-313X.2008.03705.x. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2667644&tool=pmcentrez&rendertype=abstract.

May, Patrick, Stefanie Wienkoop, Stefan Kempa, Björn Usadel, Nils Christian, Jens Rupprecht, Julia Weiss, et al. 2008. “Metabolomics- and Proteomics-Assisted Genome Annotation and Analysis of the Draft Metabolic Network of Chlamydomonas Reinhardtii.” Genetics 179 (1): 157–66. doi:10.1534/genetics.108.088336. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2390595&tool=pmcentrez&rendertype=abstract.

McIlwain, Sean, Kaipo Tamura, Attila Kertesz-Farkas, Charles E Grant, Benjamin Diament, Barbara Frewen, J Jeffry Howbert, et al. 2014. “Crux: Rapid Open Source Protein Tandem Mass Spectrometry Analysis.” Journal of Proteome Research, September. doi:10.1021/pr500741y. http://www.ncbi.nlm.nih.gov/pubmed/25182276.

Metzker, Michael L. 2010. “Sequencing Technologies - the next Generation.” Nature Reviews. Genetics 11 (1). Nature Publishing Group: 31–46. doi:10.1038/nrg2626. http://www.ncbi.nlm.nih.gov/pubmed/19997069.

Michalski, Annette, Juergen Cox, and Matthias Mann. 2011. “More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority Is Inaccessible to Data-Dependent LC-MS/MS.” Journal of Proteome Research 10 (4): 1785–93. doi:10.1021/pr101060v. http://www.ncbi.nlm.nih.gov/pubmed/21309581.

Mitulovic, Goran, and Karl Mechtler. 2006. “HPLC Techniques for Proteomics Analysis--a Short Overview of Latest Developments.” Briefings in Functional Genomics & Proteomics 5 (4): 249–60. doi:10.1093/bfgp/ell034. http://www.ncbi.nlm.nih.gov/pubmed/17124183.

Moran, Deborah, Trevor Cross, Lewis M Brown, Ryan M Colligan, and David Dunbar. 2014. “Data-Independent Acquisition (MSE) with Ion Mobility Provides a Systematic Method for Analysis of a Bacteriophage Structural Proteome.” Journal of Virological Methods 195 (January). Elsevier B.V. 9–17. doi:10.1016/j.jviromet.2013.10.007. http://www.ncbi.nlm.nih.gov/pubmed/24129072.

Nagaraj, Nagarjuna, Nils Alexander Kulak, Juergen Cox, Nadin Neuhauser, Korbinian Mayr, Ole Hoerning, Ole Vorm, and Matthias Mann. 2012. “System-Wide Perturbation Analysis with Nearly Complete Coverage of the Yeast Proteome by Single-Shot Ultra HPLC Runs on a Bench Top Orbitrap.” Molecular & Cellular Proteomics : MCP 11 (3): M111.013722. doi:10.1074/mcp.M111.013722. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3316726&tool=pmcentrez&rendertype=abstract.

Natale, Darren a, Cecilia N Arighi, Judith a Blake, Carol J Bult, Karen R Christie, Julie Cowart, Peter D’Eustachio, et al. 2014. “Protein Ontology: A Controlled Structured Network of Protein Entities.” Nucleic Acids Research 42 (Database issue): D415–21. doi:10.1093/nar/gkt1173. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=

3964965&tool=pmcentrez&rendertype=abstract.
Nelson, Clark J, Lei Li, and a Harvey Millar. 2014. “Quantitative Analysis of Protein Turnover in Plants.” Proteomics 14 (4-5): 579–92. doi:10.1002/pmic.201300240. http://www.ncbi.nlm.nih.gov/pubmed/24323582.
Olsen, Jesper V, Lyris M F De Godoy, Guoqing Li, Boris Macek, Peter Mortensen, Reinhold Pesch, Alexander Makarov, 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 (12): 2010–21. doi:10.1074/mcp.T500030-MCP200. http://www.ncbi.nlm.nih.gov/pubmed/16249172.
Ong, Shao-En, Blagoy Blagoev, Irina Kratchmarova, Dan Bach Kristensen, Hanno Steen, Akhilesh Pandey, and Matthias Mann. 2002. “Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics.” Molecular & Cellular Proteomics : MCP 1 (5): 376–86. doi:10.1074/ mcp.M200025-MCP200. http://www.mcponline.org/cgi/doi/10.1074/mcp.M200025-MCP200.
Pappin, D J, P Hojrup, and a J Bleasby. 1993. “Rapid Identification of Proteins by Peptide-Mass Fingerprinting.” Current Biology : CB 3 (6): 327–32. http://www.ncbi.nlm.nih.gov/pubmed/15335725.
Parris, JK, TG Ranney, Halina T. Knap, and W. Vance Baird. 2010. “Ploidy Levels, Relative Genome Sizes, and Base Pair Composition in Magnolia.” J. AMER.SOC.HORT.SCI. 135 (6): 533–47. http://journal.ashspublications.org/content/135/6/533.short.
Perez, Fernando, and Brian E. Granger. 2007. “IPython: A System for Interactive Scientific Computing.” Computing in Science & Engineering 9 (3): 21–29. doi:10.1109/MCSE.2007.53. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4160251.
Pluskal, Tomás, Sandra Castillo, Alejandro Villar-Briones, and Matej Oresic. 2010. “MZmine 2: Modular Framework for Processing, Visualizing, and Analyzing Mass Spectrometry-Based Molecular Profile Data.” BMC Bioinformatics 11 (1). BioMed Central Ltd: 395. doi:10.1186/1471-2105-11-395. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2918584&tool=pmcentrez&rendertype=abstract.
Pratt, J. M., June Petty, Isabel Riba-Garcia, Duncan H. L. Robertson, Simon J. Gaskell, Stephen G. Oliver, and Robert J. Beynon. 2002. “Dynamics of Protein Turnover, a Missing Dimension in Proteomics.” Molecular & Cellular Proteomics 1 (8): 579–91. doi:10.1074/mcp.M200046-MCP200. http://www.mcponline.org/cgi/doi/10.1074/mcp.M200046-MCP200.
Pruitt, Kim, Garth Brown, Tatiana Tatusova, and Donna Maglott. 2002. “Chapter 18 : The Reference Sequence (RefSeq) Database Database Content : Background.” In The NCBI Handbook [Internet]. http://www.ncbi.nlm.nih.gov/books/NBK21091/.
Pruitt, Kim D, Garth R Brown, Susan M Hiatt, Françoise Thibaud-Nissen, Alexander Astashyn, Olga Ermolaeva, Catherine M Farrell, et al. 2014. “RefSeq: An Update on Mammalian Reference Sequences.” Nucleic Acids Research 42 (Database issue): D756–63. doi:10.1093/nar/gkt1114. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3965018&tool=pmcentrez&rendertype=abstract.
R Development Core Team, A. 2008. “R: A Language and Environment for Statistical Computing”. Vienna, Austria. http://www.r-project.org.
Reumann, Sigrun, Lavanya Babujee, Changle Ma, Stephanie Wienkoop, Tanja Siemsen, Gerardo E Antonicelli, Nicolas Rasche, et al. 2007. “Proteome Analysis of Arabidopsis Leaf Peroxisomes Reveals Novel Targeting Peptides, Metabolic Pathways, and Defense Mechanisms.” The Plant Cell 19 (10): 3170–93. doi:10.1105/tpc.107.050989. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2174697&tool=pmcentrez&rendertype=abstract.
Rice, Peter, Ian Longden, and Alan Bleasby. 2000. “EMBOSS: The European Molecular

Biology Open Software Suite." Trends in Genetics : TIG 16 (6): 276–77. http://www.ncbi.nlm.nih.gov/pubmed/10827456.
Rossum, Guido van, and F. L. Drake. 2001. "Python Reference Manual." http://www.python.org.
Röst, Hannes L, George Rosenberger, Pedro Navarro, Ludovic Gillet, Saša M Miladinović, Olga T Schubert, Witold Wolski, et al. 2014. "OpenSWATH Enables Automated, Targeted Analysis of Data-Independent Acquisition MS Data." Nature Biotechnology 32 (3): 219–23. doi:10.1038/nbt.2841. http://www.ncbi.nlm.nih.gov/pubmed/24727770.
Schaff, Jennifer E, Flaubert Mbeunkui, Kevin Blackburn, David McK Bird, Michael B Goshe, and Technical Advance. 2008. "SILIP: A Novel Stable Isotope Labeling Method for in Planta Quantitative Proteomic Analysis." The Plant Journal : For Cell and Molecular Biology 56 (5): 840–54. doi:10.1111/j.1365-313X.2008.03639.x. http://www.ncbi.nlm.nih.gov/pubmed/18665915.
Scherling, Christian, Christiane Roscher, Patrick Giavalisco, Ernst-detlef Schulze, and Wolfram Weckwerth. 2010. "Metabolomics Unravel Contrasting Effects of Biodiversity on the Performance of Individual Plant Species." PloS One 5 (9): e12569. doi:10.1371/journal.pone.0012569. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2935349&tool=pmcentrez&rendertype=abstract.
Schmidtmann, Elisabeth, Ann-Christine König, Anne Orwat, Dario Leister, Markus Hartl, and Iris Finkemeier. 2014. "Redox Regulation of Arabidopsis Mitochondrial Citrate Synthase." Molecular Plant 7 (1): 156–69. doi:10.1093/mp/sst144. http://www.ncbi.nlm.nih.gov/pubmed/24198232.
Schulze, Waltraud X, and Björn Usadel. 2010. "Quantitation in Mass-Spectrometry-Based Proteomics." Annual Review of Plant Biology 61 (January): 491–516. doi:10.1146/annurev-arplant-042809-112132. http://www.ncbi.nlm.nih.gov/pubmed/20192741.
Schwanhäusser, B, Dorothea Busse, and Na Li. 2011. "Global Quantification of Mammalian Gene Expression Control." Nature, no. 473: 337–42. doi:10.1038/nature10098. http://www.nature.com/nature/journal/v473/n7347/abs/nature10098.html.
Schwarzländer, Markus, and Iris Finkemeier. 2013. "Mitochondrial Energy and Redox Signaling in Plants." Antioxidants & Redox Signaling 18 (16): 2122–44. doi:10.1089/ars.2012.5104. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3698670&tool=pmcentrez&rendertype=abstract.
Smith, Colin a, Elizabeth J Want, Grace O'Maille, Ruben Abagyan, and Gary Siuzdak. 2006. "XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification." Analytical Chemistry 78 (3): 779–87. doi:10.1021/ac051437y. http://www.ncbi.nlm.nih.gov/pubmed/16448051.
Sperling, Edit, Anne E Bunner, Michael T Sykes, and James R Williamson. 2008. "Quantitative Analysis of Isotope Distributions in Proteomic Mass Spectrometry Using Least-Squares Fourier Transform Convolution." Analytical Chemistry 80 (13): 4906–17. doi:10.1021/ac800080v. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2502059&tool=pmcentrez&rendertype=abstract.
Staudinger, Christiana, Vlora Mehmeti, Reinhard Turetschek, David Lyon, Volker Egelhofer, and Stefanie Wienkoop. 2012. "Possible Role of Nutritional Priming for Early Salt and Drought Stress Responses in Medicago Truncatula." Frontiers in Plant Science 3 (December): 285. doi:10.3389/fpls.2012.00285. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3527748&tool=pmcentrez&rendertype=abstract.
Sun, Xiaoliang, and Wolfram 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 (S1): 81–93. doi:10.1007/s11306-012-0399-3. http://link.springer.com/10.1007/s11306-012-0399-3.

Suzek, Baris E, Hongzhan Huang, Peter McGarvey, Raja Mazumder, Cathy H Wu, and Baris E Suzek Ã. 2007. “UniRef: Comprehensive and Non-Redundant UniProt Reference Clusters.” Bioinformatics (Oxford, England) 23 (10): 1282–88. doi:10.1093/bioinformatics/btm098. http://www.ncbi.nlm.nih.gov/pubmed/17379688.

Sykes, Michael T, and James R Williamson. 2008. “Envelope: Interactive Software for Modeling and Fitting Complex Isotope Distributions.” BMC Bioinformatics 9 (January): 446. doi:10.1186/1471-2105-9-446. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2605472&tool=pmcentrez&rendertype=abstract.

Tatusova, Tatiana, Stacy Ciufo, Boris Fedorov, Kathleen O’Neill, and Igor Tolstoy. 2014. “RefSeq Microbial Genomes Database: New Representation and Annotation Strategy.” Nucleic Acids Research 42 (Database issue): D553–9. doi:10.1093/nar/gkt1274. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3965038&tool=pmcentrez&rendertype=abstract.

Tellström, Verena, Björn Usadel, Oliver Thimm, Mark Stitt, Helge Küster, and Karsten Niehaus. 2007. “The Lipopolysaccharide of Sinorhizobium Meliloti Suppresses Defense-Associated Gene Expression in Cell Cultures of the Host Plant Medicago Truncatula.” Plant Physiology 143 (2): 825–37. doi:10.1104/pp.106.090985. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1803732&tool=pmcentrez&rendertype=abstract.

Thimm, Oliver, Oliver Bläsing, Yves Gibon, Axel Nagel, Svenja Meyer, Peter Krüger, Joachim Selbig, Lukas a. Müller, Seung Y. Rhee, and Mark Stitt. 2004. “Mapman: A User-Driven Tool To Display Genomics Data Sets Onto Diagrams of Metabolic Pathways and Other Biological Processes.” The Plant Journal 37 (6): 914–39. doi:10.1111/j.1365-313X.2004.02016.x. http://doi.wiley.com/10.1111/j.1365-313X.2004.02016.x.

Tohge, Takayuki, Yasutaka Nishiyama, Masami Yokota Hirai, Mitsuru Yano, Jun-ichiro Nakajima, Motoko Awazuhara, Eri Inoue, 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 (2): 218–35. doi:10.1111/j.1365-313X.2005.02371.x. http://www.ncbi.nlm.nih.gov/pubmed/15807784.

Towbin, H, T Staehelin, and J Gordon. 1979. “Electrophoretic Transfer of Proteins from Polyacrylamide Gels to Nitrocellulose Sheets: Procedure and Some Applications. 1979.” Biotechnology (Reading, Mass.) 24 (9): 145–49. http://www.ncbi.nlm.nih.gov/pubmed/1422008.

Trevaskis, Ben, Gillian Colebatch, Guilhem Desbrosses, Maren Wandrey, Stefanie Wienkoop, Gerhard Saalbach, and Michael Udvardi. 2002. “Differentiation of Plant Cells during Symbiotic Nitrogen Fixation.” Comparative and Functional Genomics 3 (2): 151–57. doi:10.1002/cfg.155. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2447268&tool=pmcentrez&rendertype=abstract.

Usadel, Björn, Fabien Poree, Axel Nagel, Marc Lohse, Angelika Czedik-Eysenberg, and Mark Stitt. 2009. “A Guide to Using MapMan to Visualize and Compare Omics Data in Plants: A Case Study in the Crop Species, Maize.” Plant, Cell & Environment 32 (9): 1211–29. doi:10.1111/j.1365-3040.2009.01978.x. http://www.ncbi.nlm.nih.gov/pubmed/19389052.

Valledor, Luis, Takeshi Furuhashi, Anne-Mette Hanak, and Wolfram Weckwerth. 2013. “Systemic Cold Stress Adaptation of Chlamydomonas Reinhardtii.” Molecular & Cellular Proteomics : MCP 12 (8): 2032–47. doi:10.1074/mcp.M112.026765. http://www.ncbi.nlm.nih.gov/pubmed/23564937.

Valledor, Luis, Luis Recuenco-Munoz, Stefanie Wienkoop, Wolfram Weckwerth, and Volker Egelhofer. 2012. “The Different Proteomes of Chlamydomonas Reinhardtii.” Journal of Proteomics 75 (18). Elsevier B.V. 5883–87. doi:10.1016/j.jprot.2012.07.045. http://www.ncbi.nlm.nih.gov/pubmed/22967953.

Vincent, Catherine E, Gregory K Potts, Arne Ulbrich, Michael S Westphall, James A Atwood, Joshua J Coon, and D Brent Weatherly. 2013. “Segmentation of Precursor Mass Range Using ‘Tiling’ Approach Increases Peptide Identifications for MS1-Based Label-Free Quantification.” Analytical Chemistry 85 (5): 2825–32. doi:10.1021/ac303352n. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3607285&tool=pmcentrez&rendertype=abstract.

Vuckovic, Dajana, and Janusz Pawliszyn. 2011. “Systematic Evaluation of Solid-Phase Microextraction Coatings for Untargeted Metabolomic Profiling of Biological Fluids by Liquid Chromatography-Mass Spectrometry.” Analytical Chemistry 83 (6): 1944–54. doi:10.1021/ac102614v. http://www.ncbi.nlm.nih.gov/pubmed/21332182.

Walzer, Mathias, Lucia Espona Pernas, Sara Nasso, Wout Bittremieux, Sven Nahnsen, Pieter Kelchtermans, Peter Pichler, et al. 2014. “qcML: An Exchange Format for Quality Control Metrics from Mass Spectrometry Experiments.” Molecular & Cellular Proteomics : MCP, April, 1905–13. doi:10.1074/mcp.M113.035907. http://www.ncbi.nlm.nih.gov/pubmed/24760958.

Washburn, M P, D Wolters, and J R Yates. 2001. “Large-Scale Analysis of the Yeast Proteome by Multidimensional Protein Identification Technology.” Nature Biotechnology 19 (3): 242–47. doi:10.1038/85686. http://www.ncbi.nlm.nih.gov/pubmed/11231557.

Weckwerth, Wolfram. 2003. “Metabolomics in Systems Biology.” Annual Review of Plant Biology 54 (January): 669–89. doi:10.1146/annurev.arplant.54.031902.135014. http://www.ncbi.nlm.nih.gov/pubmed/14503007.

Metabolomics: Methods and Protocols. 2007. Edited by Wolfram Weckwerth. Biomedical Chromatography. Vol. 21. Totowa, NJ: Humana Press. doi:10.1002/bmc.853. http://doi.wiley.com/10.1002/bmc.853.

Weckwerth, Wolfram. 2011a. “Unpredictability of Metabolism--the Key Role of Metabolomics Science in Combination with next-Generation Genome Sequencing.” Analytical and Bioanalytical Chemistry 400 (7): 1967–78. doi:10.1007/s00216-011-4948-9. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3098350&tool=pmcentrez&rendertype=abstract.

Weckwerth, Wolfram. 2011b. “Green Systems Biology - From Single Genomes, Proteomes and Metabolomes to Ecosystems Research and Biotechnology.” Journal of Proteomics 75 (1). Elsevier B.V. 284–305. doi:10.1016/j.jprot.2011.07.010. http://www.ncbi.nlm.nih.gov/pubmed/21802534.

Weckwerth, Wolfram, Kathrin Wenzel, and Oliver Fiehn. 2004. “Process for the Integrated Extraction, Identification and Quantification of Metabolites, Proteins and RNA to Reveal Their Co-Regulation in Biochemical Networks.” Proteomics 4 (1): 78–83. doi:10.1002/pmic.200200500. http://www.ncbi.nlm.nih.gov/pubmed/14730673.

Wickham, Hadley. 2009. ggplot2: Elegant Graphics for Data Analysis. Springer New York. http://had.co.nz/ggplot2/book.

Wienkoop, Stefanie, Mirko Glinski, Nobuo Tanaka, Vladimir Tolstikov, Oliver Fiehn, and Wolfram Weckwerth. 2004. “Linking Protein Fractionation with Multidimensional Monolithic Reversed-Phase Peptide Chromatography/mass Spectrometry Enhances Protein Identification from Complex Mixtures Even in the Presence of Abundant Proteins.” Rapid Communications in Mass Spectrometry : RCM 18 (6): 643–50. doi:10.1002/rcm.1376. http://www.ncbi.nlm.nih.gov/pubmed/15052571.

Wienkoop, Stefanie, Estíbaliz Larrainzar, Mirko Glinski, Esther M González, Cesar Arrese-Igor, and Wolfram Weckwerth. 2008. “Absolute Quantification of Medicago Truncatula Sucrose Synthase Isoforms and N-Metabolism Enzymes in Symbiotic Root Nodules and the Detection of Novel Nodule Phosphoproteins by Mass Spectrometry.” Journal of Experimental Botany 59 (12): 3307–15. doi:10.1093/jxb/ern182. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2529246&tool=pmcentrez&rendertype=abstr

act.
Wienkoop, Stefanie, Christiana Staudinger, Wolfgang Hoehenwarter, Wolfram Weckwerth, and Volker Egelhofer. 2012. “ProMEX - a Mass Spectral Reference Database for Plant Proteomics.” Frontiers in Plant Science 3 (June): 125. doi:10.3389/fpls.2012.00125. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3368217&tool=pmcentrez&rendertype=abstract.
Wienkoop, Stefanie, and Wolfram Weckwerth. 2006. “Relative and Absolute Quantitative Shotgun Proteomics : Targeting Low-Abundance Proteins in Arabidobsis Thaliana”, 1–7. doi:10.1093/jxb/erj157.
Wienkoop, Stefanie, Julia Weiß, Patrick May, Stefan Kempa, Susann Irgang, Luis Recuenco-Munoz, Matthias Pietzke, et al. 2010. “Targeted Proteomics for Chlamydomonas Reinhardtii Combined with Rapid Subcellular Protein Fractionation, Metabolomics and Metabolic Flux Analyses.” Molecular bioSystems 6 (6): 1018–31. doi:10.1039/b920913a. http://www.ncbi.nlm.nih.gov/pubmed/20358043.
Winiewski, Jacek R, Marco Y Hein, Juergen Cox, and Matthias Mann. 2014. “A ‘Proteomic Ruler’ for Protein Copy Number and Concentration Estimation without Spike-in Standards.” Molecular & Cellular Proteomics : MCP, September. doi:10.1074/mcp.M113.037309. http://www.ncbi.nlm.nih.gov/pubmed/25225357.
Xu, Ying, Jean-François Heilier, Geoffrey Madalinski, Eric Genin, Eric Ezan, Jean-Claude Tabet, and Christophe Junot. 2010. “Evaluation of Accurate Mass and Relative Isotopic Abundance Measurements in the LTQ-Orbitrap Mass Spectrometer for Further Metabolomics Database Building.” Analytical Chemistry 82 (13): 5490–5501. doi:10.1021/ac100271j. http://www.ncbi.nlm.nih.gov/pubmed/20515063.

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.

Acknowledgements

First and foremost, I would like to thank my supervisor Dipl.-Biol. Dr. Stefanie Wienkoop for all her unconditional support (also previous to being my PhD supervisor), relaying her professional expertise, the opportunity to work in the field of bioinformatics, numerous fruitful discussions, her encouragement, for believing in me, and her curiosity about plant biology.

I would like to thank Univ.-Prof. Dr. Wolfram Weckwerth for giving me the opportunity to learn about MS-based Metabolomics and Proteomics and to start working in this field of research, for his encouragement to test novel methodologies, for his lightheartedness in stressful situations, and for many motivating and fruitful discussions.

I would like to thank Dr. Volker Egelhofer for his support and supervision of the ProtOver-algorithm development, for teaching me the importance and value of self-reliance, for emphasizing the value of my own work, for his affirming encouragement, and for many inspiring discussions.

I would like to thank Mag. Luis Recuenco-Munoz for his introduction to Proteomics lab-work, particularly his meticulous principle of operation and introduction to the QqQ.

I would like to thank Mag. Lena Fragner for her support in the laboratory (wet lab and MS) and lending me an ear in frustrating moments.

I would like to thank Christiana Staudinger for her support in the laboratory, being a great traveling companion, and many fruitful discussions.

I would like to thank Dr. Thomas Nägele for his help with mathematical and statistical questions as well as many fruitful scientific discussions.

I would like to thank Thomas Joch and Andreas Schröfl for their excellent plant cultivation and gardening expertise, which made our scientific work possible and also enabled unforgettable culinary excess (the garden).

I would also like to thank all the members of the MOSYS team for their support and for a good working atmosphere.

I would like to thank my mother, Dr. Nancy Amendt-Lyon, and father, Dr. Gert Lyon, for putting me into this world, all their love and compassion, their

spiritual and financial support and their encouragement.

I would like to thank Mag. Rosa Lyon, Mag. Sascha Schweditsch, and Michael Lyon for their love and support.

I would like to thank Univ.-Prof. Dr. Gerhard Amendt for his support (including travel stipends) and encouragement. I would like to thank Mag. Dr. Jutta Pauschenwein for her thoughtful advice and encouragement to learn how to program and her general support.

I would like to thank Mag. Sophie Wögenstein for going through thick and thin with me, and for all her love and support.

I would like to thank Stefan Karner for his support in programming ("bubu" I will never forget you), teaching me to take my eyes off the computer screen, his infectious love of the whiteboard, and his inspiring curiosity about almost everything.

I would like to thank the Python and scientific programming community for high-quality resources made available for free, and for their support in the community.

I would like to thank my family and friends for all their support and encouragement.

Abstract

Molecular Systems Biology aims to detect as many biochemical substances as possible utilizing and combining “omics” technologies. A specific part of this field of research focuses on the use of Mass Spectrometry for the detection of analytes in plants, such as primary and secondary metabolites as well as peptides and proteins. Specialized Mass Spectrometers, e.g. Triple Quadrupole or Linear Ion Trap/Orbitrap, can be utilized depending on the type of analyte and the focus of the research. In agreement with the instrument, corresponding data acquisition methods have to be applied. The generated data, depending on the acquisition method, can be large as well as complex, and its analysis necessitates the use of specialized computational methods.

An unbiased LC/MS approach was developed for the detection of plant secondary metabolites from crude extracts of *Arabidopsis thaliana*. In order to link biochemical pathways, plant primary and secondary metabolomics data was integrated. LC/MS data acquisition methodology, data extraction and data analysis are discussed within this thesis.

A Mass Spectrometry-based method for the absolute quantification of target proteins, termed Mass Western, is described, with particular regard to the choice of instrumentation. The latter was refined by the design of cross-concatenated standard peptides for exact stoichiometric quantification of protein complexes.

Medicago truncatula, a sequenced model organism of legumes, has the capacity to exchange sugars for ammonium with nitrogen-fixing bacteria, a process known as Symbiotic Nitrogen Fixation (SNF). This process is significant for sustainable agricultural systems worldwide, as legumes are among the most important plant families to humans. SNF is susceptible to environmental stresses, particularly drought, which inhibits SNF and subsequently reduces crop yields. *M. truncatula* served to study physiological, metabolic as well as proteomic responses to drought stress. The molecular mechanisms regulating the differential control of water relations during drought is of fundamental importance to plant physiology.

Mass Spectrometry-based shotgun-proteomics mostly relies on the transla-

tion of nucleotide to protein sequence databases for identification purposes. Therefore, a major bottleneck exists for un-sequenced or imperfectly sequenced organisms. Additionally, the functional characterization of many sequences is incomplete. Annotation information was transferred from well-characterized to un- or poorly-characterized species via automated pipelines (Mercator) and BLAST homology searches. These functional characterizations served to cluster detected compounds and to visualize omics data.

A ^{15}N partial metabolic labeling experiment was performed to study the molecular mechanisms regulating the differential control of water relations during drought. In order to extract complex spectral envelopes, arising due to the ^{15}N incorporation, a novel algorithm, enabling protein turnover calculations, was developed.

Zusammenfassung

Molekulare Systembiologie zielt darauf ab, unter Verwendung von „omics“ Technologien so viele biochemische Substanzen wie möglich zu detektieren. Ein spezifischer Teil dieser Forschung konzentriert sich auf die Verwendung von Massenspektrometrie zur Detektion von Analyten in Pflanzen, wie z.B. Primär- und Sekundärmetabolite, Peptide und Proteine. Spezialisierte Massenspektrometer (Tripel Quadrupol oder Lineare Ionen Falle/Orbitrap) werden in Abhängigkeit der zu untersuchenden Analyten und unter Berücksichtigung des Forschungsvorhabens verwendet. Instrumentelle Methoden zur Datenaufnahme müssen mit dem Gerät abgestimmt werden. Da die erzeugten Daten, abhängig von der Methode zur Datenaufnahme, sehr groß und komplex sein können, werden darauf angepasste computergestützte Methoden benötigt.

Es wurde eine nichtgerichtete LC/MS-Methode zur Detektion von pflanzlichen Sekundärmetaboliten aus Rohextrakten von *Arabidopsis thaliana* entwickelt. Um biochemische Stoffwechselwege zu vernetzen, wurden Metabolit-Daten aus dem primären und sekundären Stoffwechsel integriert. LC/MS-Datenaufnahme, Datenextraktion und Datenanalyse sind Gegenstand der Diskussion dieser Arbeit.

Der Mass-Western, eine massenspektrometriebasierte Methode zur Absolutquantifizierung von ausgewählten Proteinen, wird unter Berücksichtigung der Wahl der zu verwendenden Instrumente beschrieben. Diese Methode wurde mit dem Konzept von kreuzweise zusammenhängenden Standardpeptiden zur exakten stöchiometrischen Quantifizierung von Proteinkomplexen erweitert und angepasst.

Medicago truncatula, ein sequenzierter Modellorganismus von Leguminosen, besitzt die Fähigkeit mit stickstofffixierenden Bakterien Zucker gegen Ammonium auszutauschen. Dieser Prozess wird als symbiotische Stickstofffixierung (SSF) bezeichnet. Da Leguminosen zu den für Menschen wichtigsten Pflanzenfamilien zählen, ist SSF ein global wichtiger und wesentlicher Prozess für nachhaltige landwirtschaftliche Systeme. SSF reagiert empfindlich auf Umweltstress, insbesondere auf Trockenheit, da diese

SSF hemmt und folglich Ernteerträge reduziert. Die Reaktion von *M. truncatula* auf Stress wurde auf physiologischer und metabolischer Ebene und mittels Proteinexpression untersucht. Die molekularen Mechanismen, die den Wasserhaushalt unter Trockenstress regulieren, sind von zentraler Bedeutung für die Pflanzenphysiologie.

Massenspektrometriebasierte „shotgun-proteomics“ verwenden zur Identifizierung zumeist Proteindatenbanken, die durch Übersetzung von Nukleotidsequenzen erstellt werden. Dadurch besteht ein großer Engpass für nicht- bzw. unvollständig sequenzierte Organismen. Zusätzlich ist die funktionelle Charakterisierung vieler Sequenzen unvollständig. Annotationsinformationen wurden über automatisierte Methoden (Mercator) und BLAST Homologiesuchen von gut charakterisierten auf nicht bzw. schlecht charakterisierte Arten übertragen. Diese funktionellen Charakterisierungen wurden zur Gruppierung und zur Visualisierung von „omics“ Daten verwendet.

Ein partielles ^{15}N-Metabolitmarkierungsexperiment wurde zur Untersuchung des differentiell regulierten Wasserhaushalts mittels molekularer Mechanismen durchgeführt. Um komplexe zusammengesetzte Spektren zu extrahieren, die aufgrund von ^{15}N-Anreicherung entstehen, wurde ein neuer Algorithmus entwickelt, der die Berechnung von Proteinumsatzraten ermöglicht.

Curriculum vitae

Relevant education

2009 – present	Scientific research assistant and PhD student at the Department for Ecogenomics and Systems Biology (formerly Molecular Systems Biology, University of Vienna)
2002 – 2009	Master in Biology at the University of Vienna, Austria. Master Thesis at the Department of Comparative & Ecological Phytochemistry.
2004 – 2005	University of Lisbon, Department of Marine Biology Erasmus student exchange program
2001 – 2002	Community Service (alternative to serving in the Austrian army) at the youth centre
2000 – 2001	University of Vienna History as well as Social- and Cultural-Anthropology
1997 – 1998	AFS (American Field Service) foreign exchange student in Fresno, California
1992 – 2000	High School Diploma (Matura) at „Akademisches Gymnasium Wien“

Relevant practical work experience

2012 – present	FWF Project (Austrian Science Fund), “Legumes: Multilevel analysis towards drought stress tolerance”.
2010 – 2011	BIOCRATES Life Sciences AG - FFG Project: “Development of a diagnostic platform to analyze hypertension”.
2009 – 2010	BASF Plant Science Company GmbH - Project: “Identification and quantification of proteins encoded by transgenes in Potato and Canola using LC-MS”.

Training

- 7th European Summer School for Advanced Proteomics, 3. – 10.8.2013, Brixen, Italy.
- LTQ Orbitrap Biotech Operations, Thermo Scientific, 30.11.-2.12.2010, Biozentrum Althanstrasse, Vienna, Austria.
- LECO GCxGC TOFMS, LECO, 11.-14.10.2010, Mönchengladbach, Germany.
- „Chromatography and hyphenation with Mass Spectrometry“, ASAC (Austrian Society for Analytical Chemistry), 10.-13.9.2010, Schloß Seggau bei Leibnitz.
- “HPLC and sample preparation” by Sigma-Aldrich Chemie GmbH, 29.6.2010, Vienna, Austria.

Skills

Technical	LC-MS/MS method development and maintenance, data analysis, bioinformatics, algorithms	
Languages	German	native language
	English	native language
	French	fluent
	Portuguese	fluent
Programming	Python	fluent
	R	basics
	Git	basics
Soft skills	meticulous, ability to work well in teams, enthusiastic, adaptive, autonomous, resilient, curious, cooperative, innovative, attentive, diligent	

Teaching

- Proteomics in systems biology, practical course at the University of Vienna, summer semester 2013, course number 300094. I've given lectures on Mass Spectrometry techniques, instrumentation and interpretation and assisted in the supervision of students in practical lab work (wet lab and computer work).
- Ecosystem and Biogeochemistry Laboratory – Techniques in Systems Biology, practical course at the University of Vienna, summer semester 2014, course number 300434. I've given lectures and designed practical exercises in the field of Chromatography and assisted in the supervision of students in practical lab work (wet lab and computer work).
- Bioinformatics in Mass Spectrometry, practical course at the University of Vienna, winter semester 2012 and 2013, course number 300118, I've given a lecture on Mass Spectrometry techniques.

Publications

- **Lyon, D.**, Castillejo, M.A., Staudinger, C., Weckwerth, W., Wienkoop, S. Egelhofer, V. Automated protein turnover calculations from ^{15}N partial metabolic labeling LC/MS shotgun proteomics data. **PLOS One**, **2014.**
- Doerfler, H., Xiaoliang, S., Wang, L., Engelmeier, D., **Lyon, D.**, Weckwerth, W. mzGroupAnalyzer – Predicting pathways and novel chemical structures from untargeted high-throughput metabolomics data. **PLOS One**, **2014**.
- **Lyon, D.**, Weckwerth, W., Wienkoop, S. Mass Western for absolute quantification of target proteins and considerations about the instrument of choice. **Plant Proteomics: Methods in Molecular Biology,** vol. 1072, pp 199-208, **2014**.

- Ischebeck, T., Valledor, L, **Lyon, D.**, Gingl, S., Nagler, M., Meijon, M., Egelhofer, V., Weckwerth, W. Comprehensive cell-specific protein analysis in early and late pollen development from diploid microsporocytes to pollen tube growth. **Molecular & Cellular Proteomics, 2013**.
- Egelhofer, V., Hoehenwarter, W., **Lyon, D.**, Weckwerth, W., and Wienkoop, S. Using ProtMAX to create high mass accuracy precursor alignments from label-free quantitative mass spectrometry data generated in large shotgun proteomics experiments. **Nature Protocols, 2013**.
- Doerfler, H., **Lyon, D.**, Nägele, T., Sun, X., Fragner, L., Hadacek, F., Egelhofer, V., Weckwerth, W. Granger causality in integrated GC–MS and LC–MS metabolomics data reveals the interface of primary and secondary metabolism. **Metabolomics, 2012**. (Doerfler, **Lyon**, Nägele, Sun **contributed equally to this work**.)
- Mari, A**., Lyon, D.**, Fragner, L., Montoro, P., Piacente, S., Wienkoop, S., Egelhofer, V., Weckwerth, W. Phytochemical composition of *Potentilla anserina* L. analyzed by an integrative GC/MS and LC/MS metabolomics platform. **Metabolomics, 2012**.
- Staudinger, C., Mehmeti, V., Turetschek, R., **Lyon, D.**, Egelhofer, V. and Wienkoop, S. Frontiers in Plant Science: Rhizobial Symbiosis, Role of nutritional priming for early salt and drought stress responses in *Medicago truncatula*. **Frontiers, Plant Proteomics, 2012**.

Conferences

- "American Society for Mass Spectrometry" (ASMS), 9.-13.6.2013., Minneapolis, USA. Title of presented poster: "Automated Spectral extraction from partial metabolic labeling in Planta experiments enabling Protein turnover calculations".
- "Proteomic Forum" 17.-21.3.2013, Freie Universität Berlin, Germany. Title of presented poster: "Protein turnover calculations from partial metabolic labeling in planta experiments".
- "Late summer practical proteomics seminar", 29.-30.9.2011, IMP/IMBA, Vienna, Austria.
- "4th Central and Eastern European Proteomics Conference meets International Metabolomics Austria", 29.8.-3.9.2010, TU, Vienna, Austria.

Grants

- "PhD Completion Grant" from the University of Vienna, Austria, May, 2014.
- "ASMS Student Travel Stipend" from the "American Society for Mass Spectrometry" (ASMS) for the 61st ASMS Conference.
- "DGPF-Travel-Fellowship" from the "Deutsche Gesellschaft für Proteom Forschung" ("German Society for Proteome Science") for the 7th European Summer School in Brixen, Italy, 2013.
- "AuPA-Reisestipendium" (Travel stipend) from the "Austrian Proteomics Association" (AuPA) for the 7th European Summer School in Brixen, Italy, 2013.