Metabolomics and biochemical omics data – integrative approaches

Dissertation
zur Erlangung des
Doktorgrades der Naturwissenschaften (Dr. rer. nat.)
der
Naturwissenschaftlichen Fakultät I
– Biowissenschaften –
der Martin-Luther-Universität Halle-Wittenberg,

vorgelegt von
Frau Susann Mönchgesang
geb. am 07.06.1988 in Erfurt

Gutachter:
1. Prof. Dr. Dierk Scheel (Leibniz Institute of Plant Biochemistry)
2. Prof. Dr. Sascha Baginsky (Martin-Luther-University Halle-Wittenberg)
3. Prof. Dr. Nicole van Dam (Friedrich-Schiller University Jena)

Tag der Verteidigung: 04.04.2017

Zusammenfassung

Durch technische Fortschritte ist die Erfassung von Hochdurchsatzdaten umfassend und kostengünstig geworden. Die Herausforderung jedoch liegt in der Auswertung der generierten Datenmengen. Integrative Ansätze zur Analyse von Omics-Datensätzen erlauben tiefere Einblicke als die Auswertung der einzelnen Omics-Ebenen. Häufig werden z.B. genomische und metabolische Daten in quantitativen genetischen Analysen kombiniert, um die pflanzliche Stressantwort in *Arabidopsis thaliana* zu untersuchen. Wie die Informationsweitergabe vom Genom über Transkriptom und Proteom zum Metabolom erfolgt und welche Mechanismen diesen Prozess regulieren ist ebenso von großem Interesse.

In dieser Arbeit wurde die Analyse von Hochdurchsatzdaten der Pflanzenbiochemie auf das entsprechende experimentelle Design angepasst. Zunächst wurde die Supervised Penalized Canonical Correlation Analysis (spCCA) als überwachtes statistisches Verfahren für Experimente mit mehreren untersuchten Faktoren genutzt. In einer vergleichenden Studie des Transkriptoms und Proteoms der Phosphatmangelantwort stellten sich Peroxidasen als Schalter der oxidativen Stressantwort auf beiden Omics-Ebenen heraus. Für eine andere Studie wurde ein funktionaler Ansatz gewählt, um Einzelnukleotidpolymorphismen mit dem Sekundärmetabolimus zu assoziieren. Für 19 Akzessionen von *A. thaliana* konnte die Abwesenheit bestimmter Exsudatmetabolite auf vorzeitige Stopcodons in den Genen biosynthetischer Enzyme zurückgeführt und für drei Substanzen experimentell validiert werden. Des Weiteren wurden metabolische Ähnlichkeiten einzelner Akzessionen im Clustering der codierenden Sequenzen wiedergespiegelt. Um die Reproduzierbarkeit von biochemischen Omics-Daten zu untersuchen, wurden die beobachteten Varianzen für einen Proteomics- und einen Metabolomics-Datensatz in ihre Komponenten zerlegt. Die einzelne Pflanze hatte einen erheblichen Einfluss auf die Varianz; dieser Trend wird vom Protein zum Metaboliten hin verstärkt. Bei einer weiteren Studie wurde die pflanzliche Antwort auf biotischen Einfluss am Beispiel des Wurzelendophyten *Piriformospora indica* in Exsudaten, Wurzeln und Blättern von *A. thaliana* untersucht. Das Transkriptom deutete bereits auf Veränderungen des Sekundärmetabolismus und hormonresponsive Prozesse hin. Der wachstumsfördernde Effekt ging einher mit wenigen Veränderungen im überirdischen Teil und einer deutlicheren Hochregulation des unterirdischen Pflanzenmetabolismus.

Diese Arbeit zeigt an ausgewählten Beispielen maßgeschneiderte Lösungen für integrative Fragestellungen. Zukünftige Ansätze könnten die Systembiologie und Netzwerkmodellierung als Weiterentwicklung für die ganzheitliche Datenanalyse nutzen.

Abstract

Thanks to technical advances the acquisition of high throughput data has become comprehensive and affordable, but the challenge remains in the analysis of the generated bulk of data. Integrative approaches for omics data analysis allow for deeper insights than the evaluation on a single omics level. For example, genomic and metabolic data are often combined in quantitative genetic analyses to investigate the plant stress response in *Arabidopsis thaliana*. The flow of information from genome through transcriptome and proteome down to the metabolome and which mechanisms regulate this process are of great interest.

In this thesis, the analysis of high throughput data was optimized to fit the respective experimental design. Supervised penalized canonical correlation analysis (spCCA) was applied as a supervised statistical method for experiments with multiple factors that were investigated. In a comparative study between transcriptome and proteome of the phosphate deficiency response, peroxidases were regulated moderately on both omics levels. A more functional analysis was chosen to associate secondary metabolites with single nucleotide polymorphisms. For 19 accessions of *A. thaliana*, the absence of certain exudate metabolites due to premature stop codons in genes encoding biosynthetic enzymes could be validated for three substances. Moreover, metabolic similarity of some accessions was reflected in the clustering of coding sequences. To investigate the reproducibility of biochemical omics data, the total observed variances in a proteomics and a metabolomics experiment were dissected. A single plant substantially influences the variance and this trend increases from protein to metabolite. Last, the plant's response to a biotic interaction was examined exemplarily for the root endophyte *Piriformospora indica* with the metabolic analysis of exudates, roots and leaves of *A. thaliana*. The transcriptome had already pointed towards secondary metabolism and hormone-responsive processes. The growth-promoting effect was accompanied by the upregulation of the belowground, but not aboveground plant metabolism.

This thesis demonstrates customized solutions for integrative research questions for selected examples. Future approaches could utilize systems biology and network modeling as an advancement in holistic data analysis.

Contents

1 Introduction

1.1 Motivation

This thesis seeks to integratively analyze big data in the context of plant biology using the model plant *Arabidopsis thaliana*. Hereby, four topics were focused on, as illustrated in Figure 1: First, a newly developed statistical method named supervised penalized canonical correlation analysis (spCCA) was applied to investigate the abiotic stress factor phosphate deficiency. Second, further investigations in the rhizosphere pointed out the association between metabolites in root exudates and the genetic background in a collection of 19 naturally occurring accessions that differ by single nucleotide polymorphisms (SNPs) and small insertions and deletions (INDELs). The root metabolic profiles of these 19 accessions were third analyzed with regard to natural variation as well as plant-to-plant variability and its substance-class dependency. A proteomics study also investigated the components software and plant-to-plant variability as contributors to the overall variance. Fourth, the metabolic response of *A. thaliana* to a microbe was investigated upon colonization with the root endophytic fungus *Piriformospora indica* and integrated with previously existing transcriptomics data.

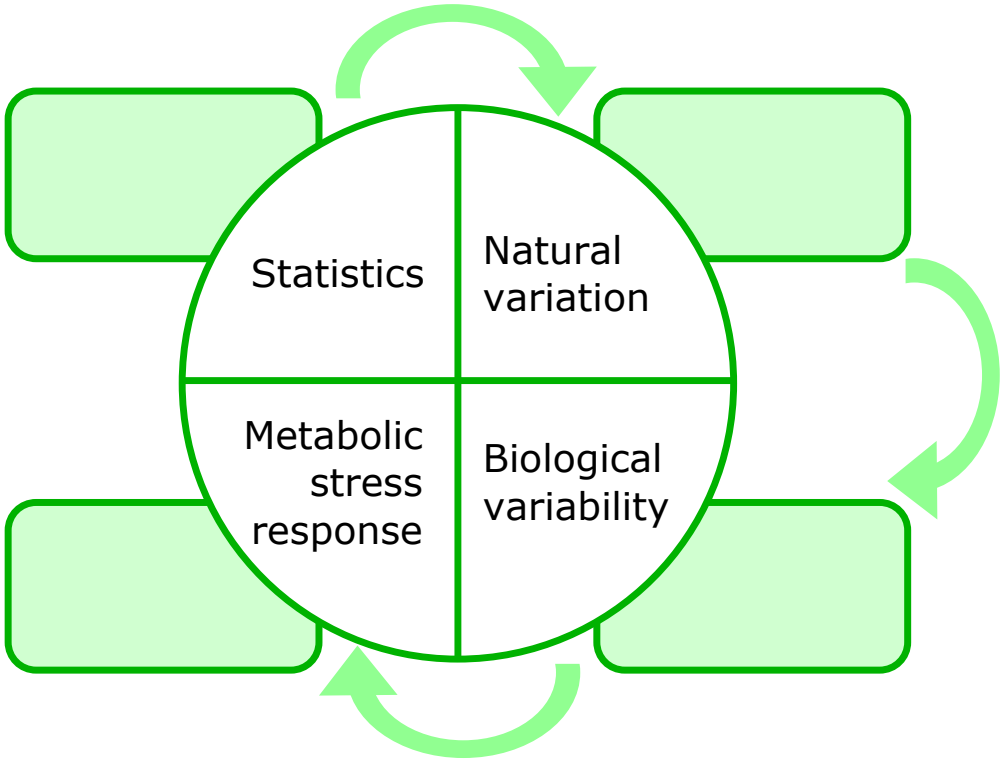


Figure 1: Graphical abstract. The topics in this thesis started off with the application of a general statistical method (upper left circular segment), which was not suitable to analyze natural variation in metabolic patterns. During the development of a customized workflow to investigate natural variation (upper right circular segment) substantial biological variability was noticed, which was subsequently analyzed in-depth (lower right circular segment). The biological relevance of metabolomics studies was examplified for the metabolic response to a microbe (lower left circular segment). The rectangles attached to each circular segment will be filled with the manuscripts covering the respective topic in the following thesis.

1.2 High throughput omics technologies and the challenges in data analysis

The information flow from gene through transcript to protein is known as the central dogma of biology. The term omics refers to the entirety of genes (genomics), transcripts (transcriptomics), proteins (proteomics) and metabolites (metabolomics). Omics technologies aim to analyze all biological molecules of the same kind in a single study.

Arabidopsis thaliana, thale cress, is a well-studied model organism in the family *Brassicaceae* and is related to cabbage and mustard [1]. Its small genome with five chromosomes and 25,498 genes was one of the first genomes that was fully sequenced in 2000 [2]. Moreover, *A. thaliana* has a short life cycle of approximately 6 weeks allowing crossing experiments in a reasonable time frame [3]. Databases, like The Arabdiopsis Information Resource (TAIR) [4–6], are well curated and thanks to relatively high homology, other Brassica species can be more easily inferred [7].

With the decreasing costs of Sanger sequencing and emerging shot-gun technologies, genomics has made huge progress. The availability of Next Generation Sequencing facilitated the high throughput analysis of DNA and RNA sequences, expanding the transcriptomics view from gene-encoding transcripts reflected by microarrays to regulatory, non-coding RNAs. Nucleic acid-based omics deal with four building blocks differing in the nucleobases adenosine (A), thymine (T) or uracil (U), cytosine (C) and guanine (G). The combinatorial spectrum of short oligonucleotides, as obtained by shot-gun technologies, increases the computational effort compared to traditional sequencing. Reference genomes and a variety of mapping tools are available [8]. Microarrays have been well established and cover all reported gene-coding transcripts. Due to their ease of use, they are often used to profile the transcriptome of known genes. By now, workflows are well established in the next generation technologies and hence, nucleic acid-based omics are furthest advanced.

The proteome consists of more building blocks, namely 20+ proteinogenic amino acids, and is also subject to posttranslational modifications. Proteomics readouts are closer to physiology than genomics or transcriptomics [9] and protein abundances cover a large dynamic range that have to be captured by the instrument without prior amplification. Baerenfaller *et al.* [10] provided a comprehensive proteome map of *A. thaliana* with 13,029 identified proteins. Variation at the proteome level can be utilized to derive biomarkers for accelerated breeding of crop cultivars [7].

Regulatory effects at each level will eventually be integrated at the level of metabolites, which constitute the biochemical phenotype as shown in Figure 2. Results from studies in upper omics levels should be validated by metabolite analysis. Metabolomics is the most complex of these four omics technologies, as the building blocks are vast and can be combined to create an enormous diversity of metabolites but it is also the omics level that is closest to cellular physiology.

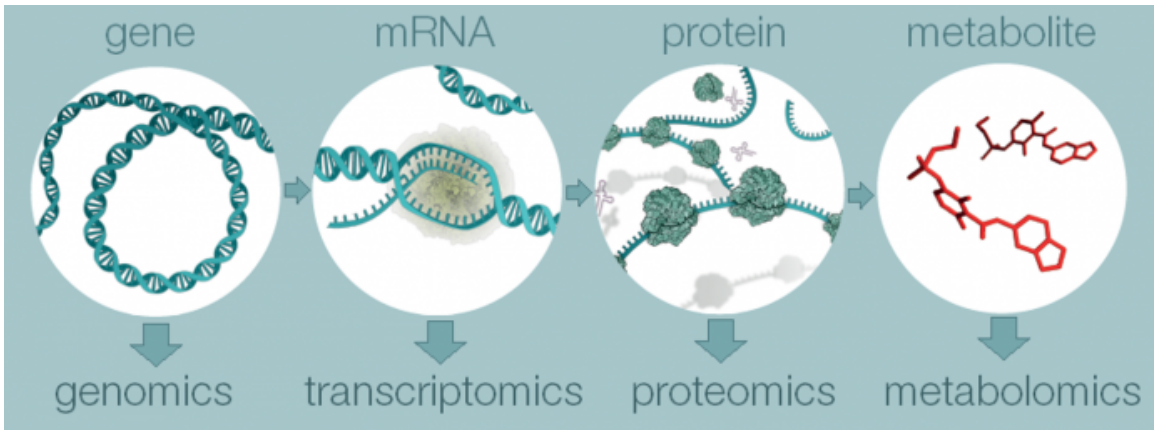


Figure 2: Omics hierarchy. Genomics analyses deal with mutations at DNA level. These alterations alongside epigentic modifications influence the transcription into mRNA, which is analyzed by transcriptomics. Proteomics captures altered protein abundances and can also be optimized to investigate post-translational modifications. All influences are eventually captured at the level of metabolites; modified from [11].

The omics levels address different research questions: The genome contains information, i.e. the potential for all biological processes, but not all of it is transmitted. The trancriptome reflects the strategy for a particular developmental stage, tissue or stress response. Proteins are the molecules that carry out the biological process and determine the physiological state. The products of these enzyme-catalyzed processes are metabolites that act as functional entities [12, 13].

Large omics datasets allow an explorative data analysis approach in addition to the classical hypothesis-driven analysis. To find the distinguishing molecules, multivariate methods are commonly used for visualization of a single dataset. These methods have to cope with many variables (features) and little observations (samples) likely resulting in ill-conditioned matrices for statistics [14]. Principal component analysis (PCA) is a non-supervised method to identify sources of variation between samples, i.e. genotype, treatment or batch effects. PCA aims at dimensionality reduction and projects the samples (scores) or metabolites (loadings) along the axis of greatest variance or, mathematically speaking, performs a singular value decomposition. Other non-supervised representations are multidimensional scaling (MDS) and hierarchical clustering (HCA). Partial least squares (PLS) is a supervised multivariate technique to detect covariances between the predicted and observed variables by applying a regression model. Thereby, both variable sets are projected to latent structures [14].

The challenge with omics data is not primarily data acquisition, but their analysis. The combination of multiple omics technologies can provide a more comprehensive understanding of biological processes. An integrative analysis goes beyond the reductionalistic approach and provides insights that might not have been inferred from a single omics study. To capture long-range and complex interactions, the system as a whole must be understood as more than the sum of its parts. Commonly used integrative methods are canonical correlation analysis (CCA) or PLS2, an extension of PLS, to find the variables in two or more datasets that are associated [15].

To study genotype-phenotype-associations, quantitative genetic methods can be often used in population genetics. Quantitative trait loci (QTL) mapping analyzes the genetic linkage in a population of related individuals, i.e. inbred offspring of contrasting parent lines. Genome-wide association studies (GWAS) identify associations between a phenotypic trait and SNPs in a population of non-related individuals [16]. Novel methods include multivariate population genetics approaches and search for a combination of phenotypic traits and genetic alleles [17, 18]. All approaches require a large number of individuals to draw statistically valid conclusions. If the experimental set-up limits the number of samples, an approach based on more prior knowledge can be suitable to analyze the genotype-phenotype association, as done in section 2.2 for nonsense mutations and qualitative metabolite differences in 19 accessions of *A. thaliana*.

1.3 Plant metabolism, metabolomics and analytical techniques

The metabolism comprises all enzyme- and non-enzyme catalyzed chemical transformations in a living cell [19]. Metabolites are low-molecular-weight compounds within the mass range 50-1,500 Da [20]. Primary metabolism is vital for growth and survival. Plant secondary metabolism allows the sessile organisms to respond to external stimuli and modulate their environment [21].

Metabolomics is the comprehensive analysis of all metabolites in a biological sample. More than 200,000 types of metabolites have been described for the plant kingdom [22, 23]. Major secondary compound classes occuring in *A. thaliana* are glucosinolates, flavonoids, phenylpropanoids, benzenoids, fatty acid derivatives and terpenoids, of which selected examples are shown in Figure 3 [24, 25].

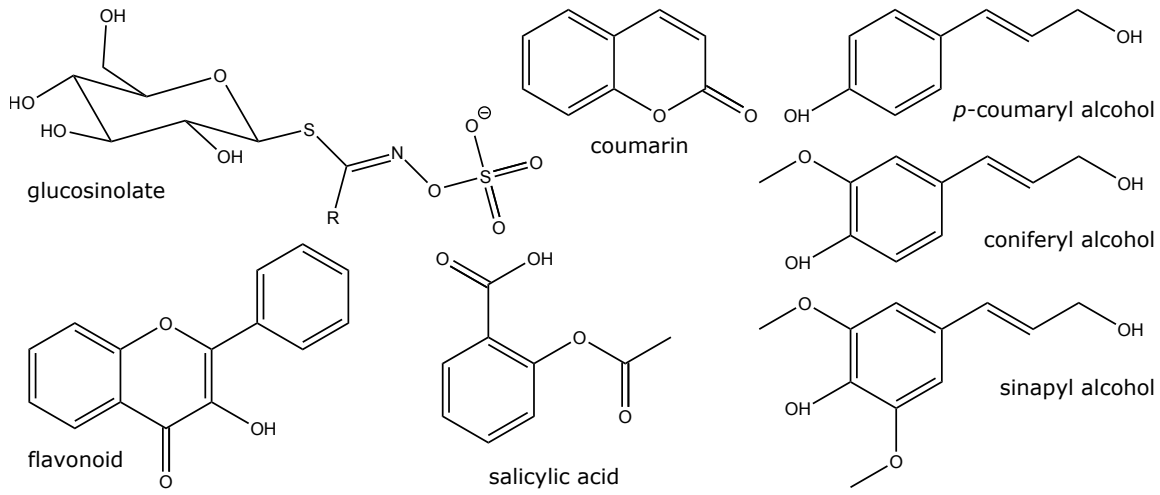


Figure 3: Secondary metabolite classes. Glucosinolates (R = amino acid derivative)and their aglycones, lignols derived from coumaryl, coniferyl or sinapyl alcohol monomers, flavonoids and coumarins as well as few phytohormones like salicylic acid are commonly detected in *A. thaliana* extracts with the applied LC/MS method.

The great diversity in secondary metabolism is achieved by conjugation of core structures.

Thereby, the bioactivity of these molecules is also regulated. Defense metabolites like mustard oils can be glycosylated and thus, these non-toxic forms can be stored by the plants. Similarly, hormone activity is controlled by modifications like hydroxylation and glycosylation. Salicylic acid is regulated by further hydroxylation at the 2' or 3' position and subsequent glycosylation to dihydroxybenzoic acid glycosides [26].

There are two approaches to a metabolic analysis: Targeted metabolomics monitors a set of known compounds to investigate its role in a biological system. Untargeted metabolomics can comprehensively profile all metabolites in a sample and hence, can reveal novel biomarkers for e.g. natural variation and defense responses. The general untargeted workflow is illustrated in Figure 4.

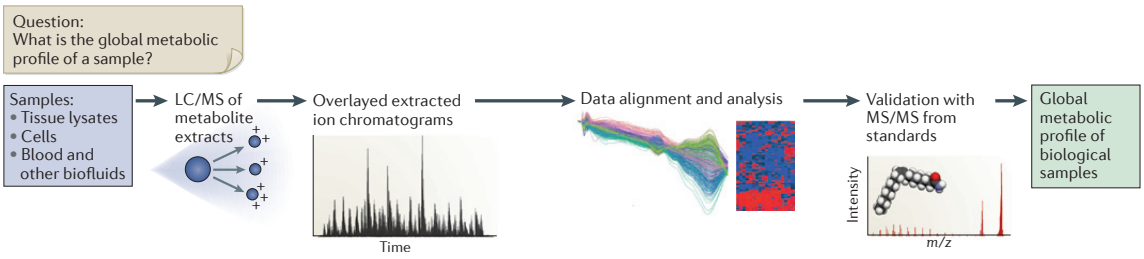


Figure 4: Untargeted metabolomics workflow. The untargeted metabolomics workflow aims at a global metabolic profile of a sample. Extracts are subjected to liquid chromatography coupled to mass spectrometry (LC/MS) and the derived feature matrix is statistically analyzed for the sample class discriminating features. Putative candidates can be structurally annotated with tandem MS (MS/MS) and are validated against the fragment spectrum of an authentic standard; modified from [12].

Chromatography-coupled mass spectrometry (MS) with high sensitivity and resolution is commonly used in proteomics and metabolomics. The separation of molecules in the chromatographic domain relies on the interaction with stationary and mobile phase. Depending on the mobile phase of the chromatography, liquid (LC) and gas chromatographic (GC) techniques are distinguished. In LC, hydrophobic or hydrophilic columns are in use and molecules with similar polarity as the stationary phase are retained. Two or more solvents that are gradually mixed constitute the mobile phase. In GC, an inert gas is used as a mobile and mostly quartz as a stationary phase. Depending on polarity and evaporation pressure, molecules adsorb to the column. A GC runs at very high temperatures (GC "oven") and the gradient is a temperature increase over the run time. GC/MS is suited for volatile and derivatized non-volatile compounds [27].

Upon elution, compounds have to be ionized before injection into the mass spectrometer. Whereas electron ionization (EI) is widely used to ionize molecules for GC/MS, the softer electrospray ionization (ESI) is often applied in LC/MS. The mobile phase is drawn into a capillary with a high voltage (determining the polarity ESI(+) or ESI(-)). Together with the nebulizer gas, charged droplets turn into a fine mist with evaporating solvent along the nebulizer unit. The ions eventually vaporize before they reach the MS inlet. ESI is suitable for thermolabile and semipolar to polar compounds [27].

In LC/MS metabolomics, often reversed-phase chromatography with a hydrophobic stationary phase, e.g. C18, is used [27]. Polar metabolites are eluted first, less polar molecules in the extract adsorb to the column and are eluted with decreasing polarity of the mobile phase. Acidified solvents provide protons and result in clearer peak shapes. Upon elution, molecules are subjected to ESI and injected into the mass analyzer, e.g. a time-of-flight (TOF) mass spectrometer. Within the electric field of the quadrupole, ions are accelerated according to their mass-to-charge (m/z) ratios, low-molecular-weight ions reach the detector earlier, ions with higher m/z later.
Databases like KEGG, ChemSpider and PubChem allow the matching of exact masses with the contained structures. GC/EI-MS spectra are reproducable and can be annotated by the comparison to NIST and the Golm Metabolome Database [28]. LC/ESI-MS spectra are subject to the applied analytical conditions and therefore, less suitable for mass spectral library matching.
In proteomics, ions are further fragmented resulting in MS/MS spectra allowing to decipher the amino acid composition of the oligopeptides. This is also done in metabolomics to elucidate the structure of selected compounds. Metabolite identification according to the Metabolomics Standards Initiative level 1 requires the analysis of an authentic standard [29].
Nuclear Magnetic Resonance (NMR) spectroscopy can also be used for confirmation of structural proposals and is highly selective. The main limitation is its low sensitivity compared to MS-based techniques.

1.4 Roots and chemical communication in the rhizosphere

Although plants are sessile organisms that interact with their surroundings, belowground interactions are still not elucidated comprehensively. Roots are important for nutrient acquisition and interactions with other organisms in the soil. They additionally provide physical strength and are vital to cope with abiotic and biotic stress, such as nutrient deficiency, salinity, drought and other soil organisms.
The narrow zone surrounding plant roots is known as the rhizosphere [30] and is the interface between the plant root, soil, microorganisms, invertebrates and roots of other plants. The rhizosphere is shaped by processes like root growth, uptake of water and nutrients as well as rhizodeposition leading to distinct bio-physico-chemical properties compared to bulk soil. Roots interact with their surroundings via chemical communication by rhizodeposition. These rhizodeposits are low molecular weight compounds and derive from sloughed-off root cap cell lysates, root border cells or are released from intact root cells as exudates. This occurs either via simple or facilitated diffusion. ATP-binding cassette (ABC) as well as multidrug and toxic compound extrusion (MATE) transporters also allow the active secretion of molecules [31, 32]. A complex mixture of compounds is deposited into the rhizosphere. Whereas plant mucilages,

mucigel and large carbohydrates belong to the water-insoluble fraction, secondary metabolites, organic acids and amino acids constitute interesting compound classes of the more hydrophylic fraction of root exudates. As briefly mentioned in section 1.3, many secondary metabolites are important for plant defense. Coumarins have additionally been shown to be involved in iron acquisition [33, 34]. Furthermore, roots and root exudates harbor a variety of lignolic compounds that are regulated upon phosphate depletion ultimately resulting in altered lignification.
Organic acids are involved in various processes. Besides others, citrate, malate and oxalate facilitate the uptake of insoluble minerals like inorganic phosphorous from soil through chelation and/or ligand exchange [35–37]. Malate and oxalate have also been reported to detoxify metals [38, 39].
A large range of amino acids and dipeptides has been previously described in exudates of *A. thaliana* [40] and together with the knowledge about peptide transporters like PTR1 and PTR5 with dipeptide affinity [41], the role and origin of dipeptides in the rhizosphere could be elucidated.
Advances in analytical techniques deliver sufficient sensitivity for the measurement of low abundance compounds in exudates. Several methods have been described to analyze rhizodeposition: exudates can either be collected from hydroponically or from plants grown in sand or soil. Hydroponic systems usually result in large sample volumes and hence, a dilution effect. For exudate collection from soil-grown plants, either adsorbing materials may sample directly or extensive washing to remove soil particles is required for subsequent exudation into trap solutions. Thereby, rhizotrons or rhizoboxes are often used to separate rhizosphere from bulk soil. Oburger *et al.* [42] compared various exudate collection methods and benchmarked a set-up of rhizoboxes combined with micro-suction cups distinguished by a high spatial resolution. Mathieu *et al.* [43] introduced a "rhizoponics" approach incorporating the advantages of rhizotrons and hydroponic systems.
A central issue in exudate collection is the sterility of the procedure because microbial organisms alter the chemical composition of the rhizosphere, e.g. via the degradation of primary metabolites [44, 45]. Strehmel *et al.* [40] developed a sterile hydroponic system to monitor exudation of *A. thaliana*. However, the low concentration and localized deposition of compounds remains a challenge in rhizosphere analytics.

1.5 Factors in experimental design as a basis of data analysis

When designing an experiment, researchers usually take several factors with multiple levels into consideration, as illustrated in Figure 5. A common experimental set-up in plant physiology is the comparison of a wild type plant with a mutant. An average study examines several factors with few levels and is hence suitable for matrix operations-based statistical analyses, despite the issue of a larger number of observations (features) than observables (samples) in omics

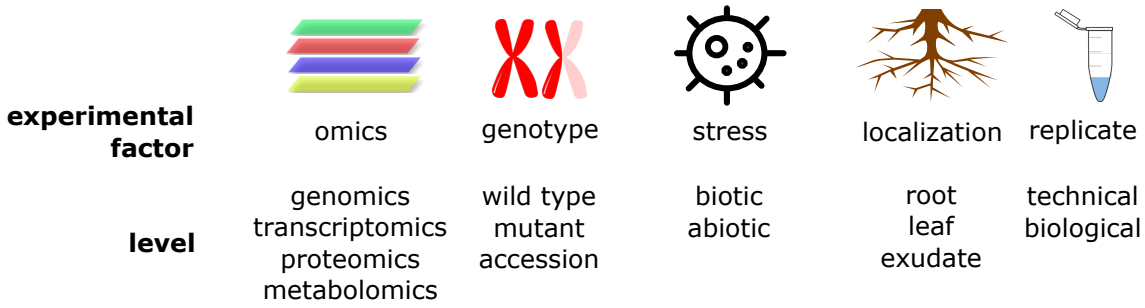


Figure 5: Examples for experimental design factors with possible levels. A study typically investigates the influence of an experimental factor, such as genotype and treatment, by contrasting at least two of its levels, such as a mutant vs. wild type or a treated vs. non-treated scenario.

data.

In addition to mutant scenarios, naturally occurring strains of the same species collected from different locations, known as ecotypes or nowadays as accession, can be analyzed. Natural variation refers to small, but numerous genomic alterations like SNPs between these accessions. These accessions can be utilized to identify the genetic origin of phenotypic traits. As elaborated in section 1.2, several levels of the omics hierarchy may be integrated to elucidate the genotype-phenotype-interplay.

Another typical experimetal scenario is the comparison of a stressed state vs. its control condition. However, the exploration of the general physiological state detects genes controlling this state, which is the base to comprehend perturbations of this system. Three studies of this thesis (sections 2.2, 2.3.1 and 2.3.2) invesigate the metabolic and proteomic composition in an unperturbed state. The exposure to stress allows to dissect the effect of a single gene, transcript, protein or metabolite. Different growth conditions are applied to go beyond the general physiological state and explore the roles of certain players in the stress response. Nutrient deficiencies as well as salinity, drought, light and temperature constitute commonly investigated abiotic stresses [46–48]. Interactions with microorganisms such as bacteria and fungi as well as herbivores can constitute biotic forms of stress challenging plants [49]. Two publications in this thesis analyze the plant's response to the abiotic stress factor phosphate starvation in section 2.1.2 and to the biotic factor *P. indica*, a root endophyte, in section 2.4. The spCCA method manuscript in section 2.1.1 also demonstrates the power of the analysis on a dataset of *Arabidopsis*' response to the oomycete *Phytophthora infestans*.

To investigate localization characteristics, multiple cell types or tissues of the same plant can be analyzed to reveal a more systemic picture like demonstrated for roots, leaves and exudates in the *P. indica* study (section 2.4).

One statistically important factor in experimental design is type and number of replicates. There are uncertainities about the nature of biological and technical replicates and how to integrate the latter into the analysis. The studies in section 2.3 analyze different replicate types and reveal interesting subsets of proteins and metabolites with high plant-to-plant variability.

1.6 Aims and questions addressed

Metabolomics is one of the omics technologies with the largest required effort in data analysis as the building blocks are most diverse. The combined analysis of biochemical phenotypes at the metabolomics and proteomics levels together with higher omics levels can facilitate interpretation and allow for overarching conclusions.
The general aim of this thesis was to find appropriate integrative approaches for different experimental design of omics, especially MS-based omics, technologies. I hereby focused on the model plant *A. thaliana*, which has been subjected to many omics analyses but still leaves room for investigations on roots, exudates, plant-to-plant variability and particular stress responses. The specific aims of this thesis were the following:

1. **Application of general statistical methods for combining multiple experimental factors:**
   SpCCA is a supervised statistical approach to integrate several omics datasets and its experimental design factors (section 2.1.1). It was applied onto transcriptomics and proteomics to decipher the phosphate response in the root tip. Hereby, two omics levels were combined with three genotypes and two growth conditions (section 2.1.2).

2. **Exploration of secondary metabolism by integrating genomics:**
   We demonstrate a direct genotype-phenotype association by linking stop codons in genes encoding biosynthetic enzymes with metabolite absences in root exudates in section 2.2. Hereby, the factor "accession" had a substantial number of levels. Since the sample preparation did not allow high throughput screens and thus sufficient power for a GWAS experiment, an alternative approach was followed to combine two omics levels for 19 accessions. Since the metabolite abundances were quite variable between the replicates of one accession, a qualitative measure was chosen.

3. **Investigation of biological variability at omics levels downstream of genomics:**
   In a proteomics and a metabolomics study (section 2.3), the total observed variance in abundances was decomposed into the fractions attributable to plant-to-plant variability, experimental batch, sample preparation and data processing. Hereby, an emphasis was laid on multi-level experimental designs and the variances at each level. Different kinds of replicates are integrated into a statistical analysis and the question was asked whether plant-to-plant variability is biologically meaningful.

4. **Deciphering the metabolic response to microbes:**
   To study the interaction between the endophytic fungus *P. indica* with *A. thaliana*, metabolic profiles of roots, exudates and leaves were interpreted in the context of previously reported root transcriptomics (section 2.4). Hereby, plants were grown in two

conditions, exposed to biotic stress and the control condition. The aboveground and belowground part of the plant as well as root exudates were analyzed for changes in both primary and secondary metabolism and integrated with another omics level revealing a comprehensive metabolic picture.

An outline of the topics covered in this thesis is illustrated in Figure 6.

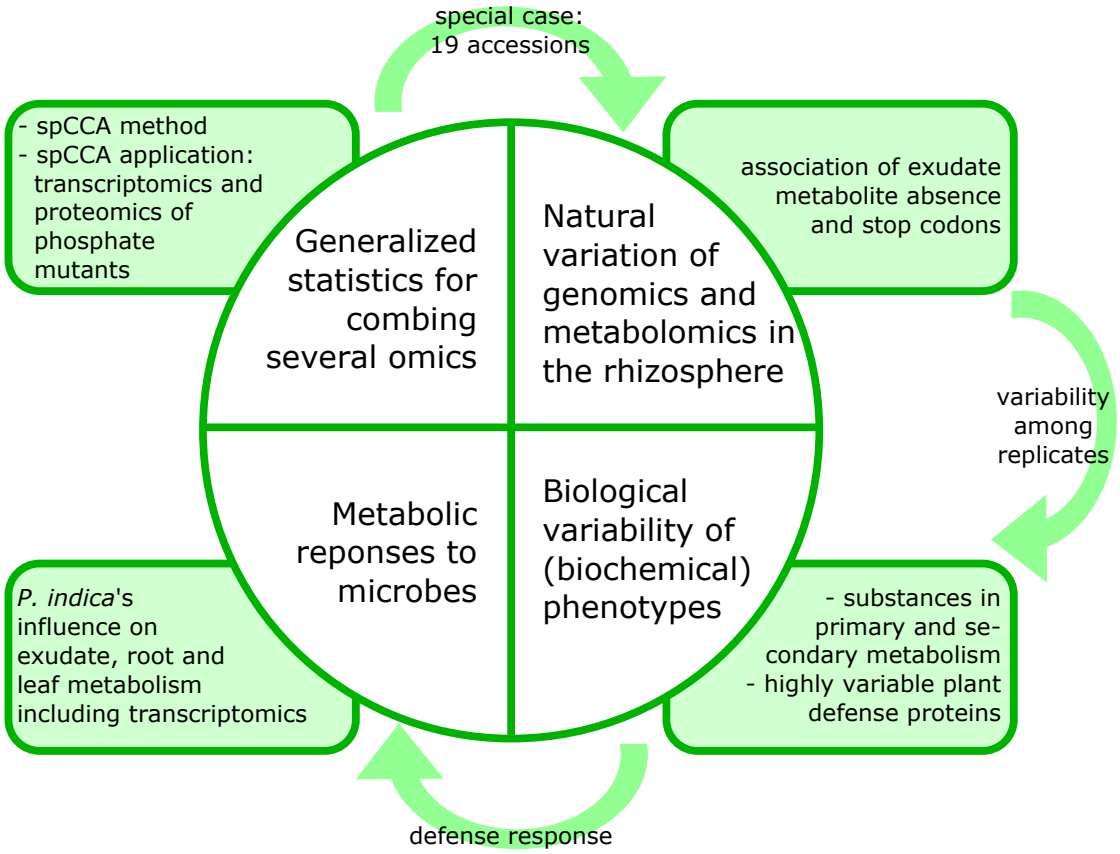


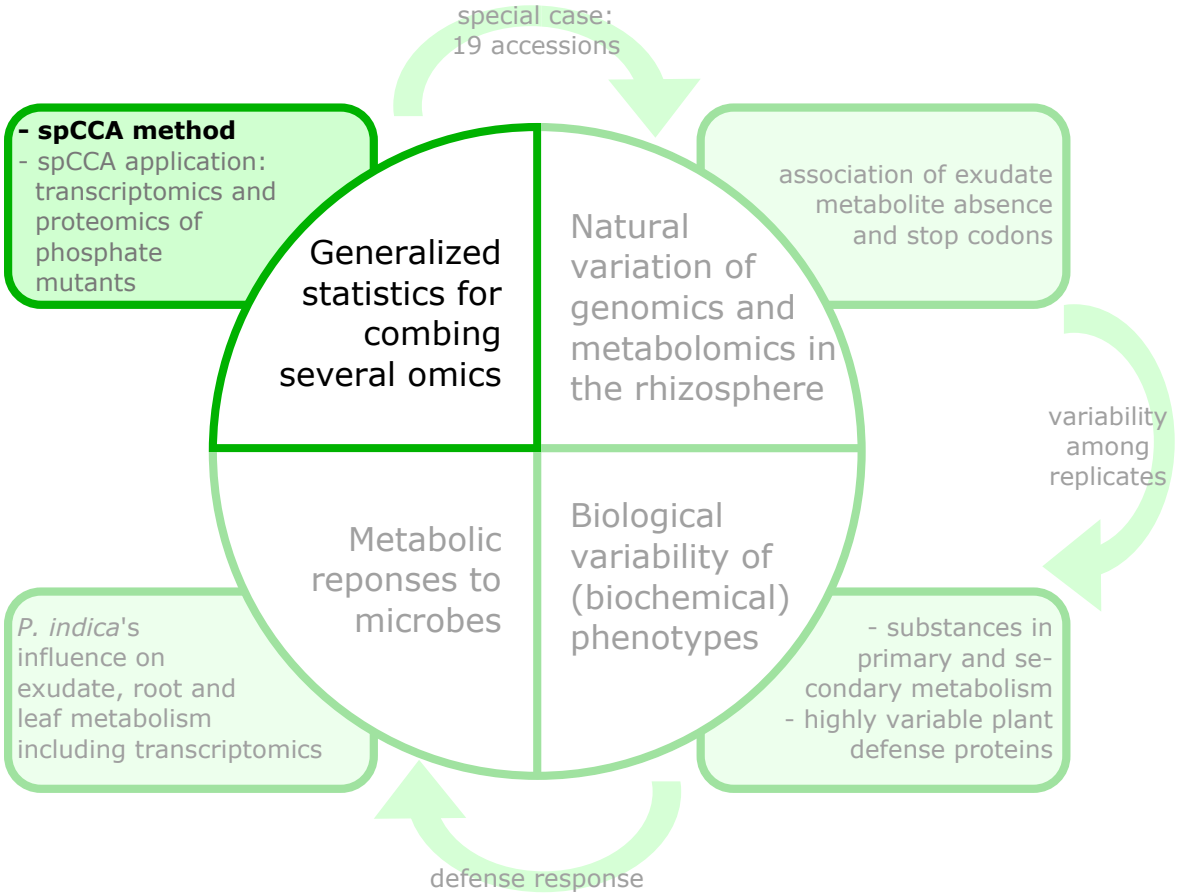
Figure 6: Graphical outline. The four topics were investigated in either one or two examplary studies. The arrows illustrate the observations from one topic that directed research towards the topic in the next circular segment.

2 Publications

2.1 General statistical methods for combining multiple omics

2.1.1 Supervised Penalized Canonical Correlation Analysis

Thum, A.; Mönchgesang, S.; Westphal, L.; Lübken, T.; Rosahl, S.; Neumann, S.; Posch, S., Supervised Penalized Canonical Correlation Analysis. *arXiv* **2014**, *1405.1534*.



Supervised Penalized Canonical Correlation Analysis

Andrea Thum, Susann Mönchgesang, Lore Westphal, Tilo Lübken, Sabine Rosahl, Steffen Neumann and Stefan Posch

May 8, 2014

Abstract

Motivation: The canonical correlation analysis (CCA) is commonly used to analyze data sets with paired data, e.g. measurements of gene expression and metabolomic intensities of the same experiments. This allows to find interesting relationships between the data sets, e.g. they can be assigned to biological processes. However, it can be difficult to interpret the processes and often the relationships observed are not related to the experimental design but to some unknown parameters.

Results: Here we present an extension of the penalized CCA, the *supervised penalized* approach (spCCA), where the experimental design is used as a third data set and the correlation of the biological data sets with the design data set is maximized to find interpretable and meaningful canonical variables.

The spCCA was successfully tested on a data set of *Arabidopsis thaliana* with gene expression and metabolite intensity measurements and resulted in eight significant canonical variables and their interpretation. We provide an R-package under the GPL license.

Availability: R package spCCA at http://msbi.ipb-halle.de/msbi/spCCA/

Contact: andrea.thum@informatik.uni-halle.de

1 Introduction

Systems biology aims to understand living organisms, often by combining multi-factorial experiments and multiple assay techniques to obtain, e.g., gene expression, protein or metabolite levels. To unravel the interactions between genes, proteins, or metabolites, statistical methods are used to discover dependencies among the data.

Many genes are already known to be involved in the control of metabolism and activation of pathways. Correlations between genes and specific metabolites have been used to assign signaling functions to the metabolites (Hannah *et al.* (2010)). Moreover, genes encoding enzymes for secondary metabolite synthesis have been identified by specifically looking for expression profiles of possible candidate genes (Muroi *et al.* (2009)).

In order to detect linear correlations between two data sets, the canonical correlation analysis (CCA, Hotelling (1936)) can be used. The CCA returns a pair of linear combinations of the features (e.g. gene or metabolite levels) in each of the two data sets, which correlate maximally: the first pair of canonical variates, which is the first canonical variable. Orthogonal to this pair, the second pair of combinations with second largest correlation coefficient can be found, and so forth.

The canonical variables for the observed biological data contain information about the underlying processes in the organism, where a large weight for one feature in the linear combination corresponds to large influence

of this feature to the process. A process can be deduced from the pattern of the course of the canonical variable across the samples.

For large-scale data sets, there is often an imbalance between the large number of features and the much smaller number of biological samples measured. A robust and sparse solution is neccessary to deal with the underdetermination and to extract the most relevant features. This can be achieved by penalized CCA (pCCA, Waaijenborg and Zwinderman (2009)), where an elastic net solution is implemented which combines two penalty terms. The ridge regression term is used to eliminate the singularity. A lasso regression term is implemented which forces small weights within the linear combinations to zero, thus essentially removing them.

A biological system is influenced by many factors, which can be caused by the experimental design, but also by parameters beyond the control of the experimentalist. Often it is difficult to recognize which processes are associated with the canonical variable as there may be processes that interfere with each other resulting in a complicated pattern of the canonical variable. Furthermore, processes independent from the experimental design are difficult to interpret. Therefore, it is desirable to associate the canonical variables with the experimental design.

We extend the pCCA to more than two data sets and use the experimental design as an additional data set to obtain a *supervised penalized* CCA (spCCA). In this case, the sum of the correlation coefficients between the canonical variables of each biological data set with the experimental design data set is maximized. Thus, a high correlation can be achieved between the linear combination of the features of each biological data set and the linear combination of the vectors of the experimental design. These combinations can be interpreted easily in terms of the experimental design.

The paper is structured as follows: in Section 2.1 we first introduce standard and penalized canonical correlation analysis for two data sets, as well as the generalized CCA for more than two data sets. In 2.2 we show how we combined generalized and penalized CCA and adopted this approach to obtain the new supervised penalized CCA. Section 2.3 gives details of the biological experiments including two assays obtained from *Arabidopsis thaliana*, where both gene expression and metabolite levels were determined. In the results section 3 we apply the new spCCA to these data sets. We identify several well-explainable processes, and compare the results to standard pCCA to demonstrate the potential of the supervised approach.

2 Materials and methods

We consider n experiments with two sets of features. The features in the sets are generally of different type, e.g., metabolite intensities and gene expression. The p_1 and p_2 features are collected in a $n \times p_1$-matrix X_1 and a $n \times p_2$-matrix X_2, where rows represent the experiments. The variables of the matrices are normalized columnwise to have zero means and unit variance.

2.1 Canonical Correlation Analysis

In order to maximize the correlation between the linear combinations $X_1 w_1$ and $X_2 w_2$ of the columns of the matrices X_1 and X_2 the CCA determines weight vectors w_1 and w_2. The resulting linear combinations are the pair of canonical variates for the first canonical variable.

2.1.1 Standard CCA

To compute the standard CCA, the correlation coefficient of the linear combinations is to be maximized with respect to w_1 and w_2 :

$$\mathrm{corr}(X_1 w_1, X_2 w_2) = \frac{w_1^T X_1^T X_2 w_2}{\sqrt{w_1^T X_1^T X_1 w_1 w_2^T X_2^T X_2 w_2}} \to \max,$$

where $X_1^T X_1$ and $X_2^T X_2$ are the variance matrices and $X_1^T X_2$ is the co-variance matrix of data sets X_1 and X_2.

This is equivalent to:

$$w_1^T {X_1}^T X_2 w_2 \to \max$$

with constraint:

$$w_1^T {X_1}^T X_1 w_1 = w_2^T {X_2}^T X_2 w_2 = 1$$

This leads to a generalized eigenvalue problem, where the maximal eigenvalue corresponds to the correlation coefficient between the first canonical variates for X_1 and X_2, and the corresponding eigenvector yields the weights of the canonical variate for data set X_1. The weights for X_2 can be inferred (Hardoon *et al.* (2003)). The weights of the canonical variates indicate the contribution of each feature, e.g. the gene or metabolite, to the correlation.

So far we considered the combination with the highest correlation coefficient achievable. Further combinations of features with lower correlation can be inferred from the remaining eigenvectors. There are $\min(\mathrm{rank}(X_1), \mathrm{rank}(X_2))$ canonical variables, which are orthogonal to each other with decreasing correlation coefficients.

2.1.2 Penalized CCA

In data sets from biological experiments the number of experiments is often much smaller than the number of features. For data matrices with $p_1 > n$ or $p_2 > n$, the variance matrix is singular and thus not invertible, and the CCA-problem is ill-posed. Parkhomenko *et al.* (2009) as well as Waaijenborg and Zwinderman (2009) propose the pCCA solution to address this problem by incorporating the elastic net approach.

The elastic net (Zou and Hastie (2005)) is a combination of two regression penalties: ridge regression and lasso. Ridge regression is implemented by regularizing a matrix M, to make them full rank and thus invertible (Hastie *et al.* (2009)). For this purpose, γI is added to the matrix, where γ is a positive scalar parameter and I is the identity matrix.

Since not all features are expected to be involved in the underlying process, the lasso penalty aims at eliminating unimportant features. Weights less than a parameter $\frac{1}{2}\lambda$ are set to zero (Tibshirani (1996)). To avoid a double shrinkage by this two-stage procedure the coefficients are multiplied by $(1 + \gamma)$.

The elastic net can deal with data sets with more features than experiments, produces sparse results and shows a 'grouping effect', i.e. it assigns similar weights to highly correlated features within each data set.

The elastic net has no analytical solution. To integrate this approach into the standard CCA framework, Waaijenborg and Zwinderman (2009) translated this problem into a coupled regression framework. This framework can be solved iteratively by the power method, which determines the eigenvectors w_1 and w_2 for the canonical variates $X_1 w_1$ and $X_2 w_2$ for the dominant eigenvalue. To compute the next pair, which is

orthogonal to the first, the data sets $X_k^{'}$ are constructed orthogonally to the first variable by the subtraction of the canonical variates from the data sets. This can be done by regressing each column of X_k on the canonical variate $X_k w_k$ and keeping the orthogonal residual (Waaijenborg and Zwinderman (2009)).

For each data set, two parameters are required: the ridge regression parameter γ_k to control the strength of regularization and lasso parameter λ_k for the sparsity. It is not obvious how to set these parameters. Parkhomenko *et al.* (2009) suggest a strong ridge regression regularization and set the ridge regression parameter to infinity. The lasso parameters are determined via resampling, maximizing the test sample correlation.

2.1.3 Generalized CCA

As we aim to analyze more than two data sets by a penalized CCA we consider first the standard generalized canonical anlysis.

The standard generalized canonical correlation analysis (gCCA) computes the CCA for $m > 2$ data sets with n experiments and p_k features, $k = 1, \ldots, m$. Here, different optimization criterions are possible. We used the sum of the correlation coefficients of the canonical variates between all pairs X_k, X_l of the data sets, which is to be maximized (SUMCOR formulation in Kettenring (1971)):

$$\frac{1}{m(m-1)} \sum_{\substack{k,l=1,\\ k \neq l}}^{m} w_k^T X_k^T X_l w_l \rightarrow \max$$

with constraint:

$$w_k^T X_k^T X_k w_k = 1, \; k = 1, \ldots m$$

For the standard CCA, a canonical variable is a pair of two maximally correlated variates. Now, a generalized canonical variable for m data sets consists of m variates, where the sum of the correlation coefficients of all pairs of these variates is maximal.

The SUMCOR optimization problem has no closed form solution. The eigenvalue problem can be translated into a regression framework. As described in Vía *et al.* (2006) this results in m coupled regression problems, which can be solved iteratively.

For iteration step t this yields:

$$w_k^{(t)} = X_k^+ \cdot \sum_{i=1}^{m} X_i w_i^{(t-1)}, \; k = 1, \ldots, m$$

where $X_k^+ = (X_k^T X_k)^{-1} X_k^T$ is the pseudoinverse of X_k. The $w_k^{(t)}$ have to be normalized to length 1 before the next iteration step.

2.2 Supervised pCCA

In addition to $m-1$ biological data sets $X_1, \ldots, X_{m-1}$, we include the design data set X_m to the CCA. This design data set is semantically similar to the design matrix of the experiments and describes the experimental setup, for example the growth condition, mutations or treatment. The information about the experimental design can be encoded in binary design vectors of size n, the number of experiments. These vectors describe the group membership of each experiment to different experimental conditions. Depending on the experiments this yields p_m vectors each in analogy to the measurements of one feature in all n

experiments. For example, the feature vector for $n = 10$ and a treatment vs. control setup with five replicates each is given as: $(0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 1)^T$. The combination of these vectors as columns yields the $n \times p_m$ design matrix X_m.

We call the generalized pCCA applied to these m data sets *supervised penalized CCA (spCCA)* as knowledge about the experimental setup is directly intergrated and exploited. The weights of the canonical variate $X_m w_m$ yield a combination of the experimental conditions and facilitate the interpretation of the underlying processes.

If the sum of the pairwise correlation coefficients is maximized as proposed by the SUMCOR approach, high correlation coefficients between pairs of biological data sets might compensate low correlation between biological data sets and the design data set. Thus, as we are especially interested in a high correlation with the design data set, we only consider the correlation coefficients for each biological data set with the design data set. This does not take the correlation between the biological data sets into account and maximizes the following problem:

$$\frac{1}{m-1} \sum_{k=1}^{m-1} w_k^T X_k^T X_m w_m \rightarrow \max,$$
$$w_k^T X_k^T X_k w_k = 1,\ k = 1, \dots m$$

where X_m denotes the design data set.

This results in a reduced iterative solution:

$$w_k^{(t)} = w_k^{(t-1)} + X_k^+ X_m w_m^{(t-1)},\ k = 1, \dots, m-1$$
$$w_m^{(t)} = w_m^{(t-1)} + X_m^+ \cdot \sum_{i=1}^{m-1} X_i w_i^{(t)}$$

We used a strong regularization for the elastic net, which sets the co-variances in the pseudoinverses $X_k^+$of each data matrix to zero (Parkhomenko *et al.* (2009)).

The lasso penalty is included by setting weights below a threshold $\frac{1}{2}\lambda_k$ to zero in each iteration step.

It is difficult to adjust the lasso parameters λ_k to adequately control the sparsity of the penalized CCA.

We used a resampling technique to determine the parameters. Using a grid search on the λ_k the following algorithm for the spCCA is repeated several times for different training data sets and for each combination of λ_k. For our Arabidopsis data set, ten training data sets were sampled. For each training data set, one eighth of the experiments was randomly drawn.

ALGORITHM FOR spCCA

1. INPUT: $m-1$ biological data sets $X_1, \ldots, X_{m-1}$, normalized to zero mean and unit variance; one design data set X_m, normalized to zero-mean and unit variance, extracted from the experimental design; m sparsity parameters λ_k
2. Compute strong regularized pseudoinverses for the biological data sets $X_1^+, \ldots, X_{m-1}^+$
3. Compute pseudoinverse of the design data set X_m. X_m does not have to be regularized, if there are no redundant vectors.
4. Set initial normalized values for $w_k^{(0)},\ k = 1, \ldots, m$
5. In iteration step t:
   (a) for $1 \leq k < m$ (biological data sets):

$$v_k = w_k^{(t-1)} + X_k^+ X_m w_m^{(t-1)}$$

   for $k = m$ (design data set):

$$v_m = w_m^{(t-1)} + X_m^+ \cdot \sum_{i=1}^{m-1} X_i w_i^{(t)}$$

   (b) Normalize $v_k = \frac{v_k}{|v_k|}$, $k = 1, \ldots, m$
   (c) Set all components of each v_k to zero, which are smaller or equal to the sparsity parameter $\frac{1}{2}\lambda_k$ ($\widehat{=}$ lasso penalty)
   (d) Normalize again to obtain updated weigth vector $w_k^{(t)} = \frac{v_k}{|v_k|}$, $k = 1, \ldots, m$
   (e) if convergence criterium reached: break

Due to the lasso penalty in step (c), which enforces the sparseness, the algorithm converges to an eigenvector. Depending on the initial weights, it does not necessarily converge to the vector associated with the dominant eigenvalue. Thus we repeat the iteration in step 5. for different initializations with random values (step 4.) ten times and keep the solution with the largest eigenvalue for the training data set. The eigenvectors with the median eigenvalue of all training data sets is used as the weight vectors for the canonical variable.

Again, this determines only the first canonical variable. To compute further variables the variates are subtracted from the biological data sets X_k to produce orthogonal data sets. The design data set remains unchanged.

2.2.1 Significance of correlation

To decide whether a canonical variable for three data sets is significant, a permutation test was used. We found that for our data set the correlation coefficient needs to exceed 0.605 for a level of significance $\alpha = 0.05$ and needs to be larger than 0.635 for $\alpha = 0.01$.

2.3 Data

We demonstrate the power of the supervised pCCA approach using data from an experiment on the response of the model plant *Arabidopsis thaliana* to the pathogen *Phytophthora infestans*. The data set consists of microarray gene expression data and LC/MS based metabolite profiles.

The oomycete *P. infestans* is the causal agent of late blight, the most devastating potato disease. In contrast to potato, *A. thaliana* is able to successfully prevent colonization of the pathogen due to a multi-layered nonhost resistance.

Several mutants have been isolated which are impaired in penetration resistance. A mutation in the gene *PEN2*, which encodes an enzyme involved in indole glucosinolate metabolism (Bednarek *et al.* (2009)), results in the loss of penetration resistance against *P. infestans* (Lipka *et al.* (2005)). Despite its ability to penetrate epidermal cells of *pen2* mutant plants, *P. infestans* is still not able to colonize these plants. Additional mutants were isolated by Kopischke *et al.* (2013) which show enhanced defense responses upon infection with *P. infestans*: *pen2erp1* and *pen2erp2*, and backcrossed mutants *erp140* and *erp2D*.

We used six different plant lines, the wildtype-like *gl1*, and the five different mutants (*pen2, pen2erp1, pen2erp2, erp2D, erp140*). The plants were either infected with *P. infestans* spores or treated with water as control, and harvested 6h and 12h after treatment. The experiment was repeated three times with different *P. infestans* cultures, resulting in biological triplicates, for an overall of $6 \times 2 \times 2 \times 3 = 72$ samples.

These samples from the same plant material were used on 72 Affymetrix microarrays and for LC/MS-metabolite profiling (see supplemental material for details). For each of the 72 samples, three LC/MS-runs were performed. We obtained 202 LC/MS measurements (72 samples with one to three technical replicates each) of the abundance of polar metabolites. We used the centWave algorithm (Tautenhahn *et al.* (2008)) to extract features from the LC/MS raw data, and used xcms (Smith *et al.* (2006)) to group them into a rectangular data matrix. The technical replicates were averaged and the metabolomic data resulted in a 72×5896 data matrix. The metabolomic data was reduced to 3007 putative pseudomolecular ions with help of the R-package CAMERA (Kuhl *et al.* (2012)). The microarray data was processed and normalized with the R-package simpleaffy (Wilson and Miller (2005)) and resulted in a 72×22810 data matrix.

We reduced the data sets by excluding features with low variance (threshold chosen $\sigma < 1$ for genes and $\sigma < 0.4$ for metabolites), resulting in a 72×1277 gene expression matrix, and a 72×252 LC/MS signal intensity matrix.

3 Results

The reduced data sets were analysed by the supervised pCCA. The supervised solution included the experimental information in Fig. 1 as the design data set.

To describe six mutants non-redundantly only five vectors are required, likewise only two vectors to code three replicates and the oomycete cultures. In consequence, the design data set contains 11 design vectors of length 72 for the 72 experiments.

To determine the sparsity parameters for our data set we performed a ten-fold repeated hold-out sampling with a grid of size $16 \times 21 \times 31 = 10416$. This requires between 20 to 120 minutes using an AMD Athlon 64 Processor 4000+ with 2400 MHz and 2 GB RAM.

The first canonical variable (Fig. 2) is a combination of genotype and pathogen response, which is elevated in all three *pen2*-mutants upon infection. One *pen2*-mutant shows exceedingly high expression values and metabolite intensities for both treatment and control. The largest weighted metabolites include camalexin as well as flavonoids, which play a role in defense. This is in agreement with biological knowledge: the plant senses the attack by *P. infestans* and immediately releases camalexin as first defense reaction.

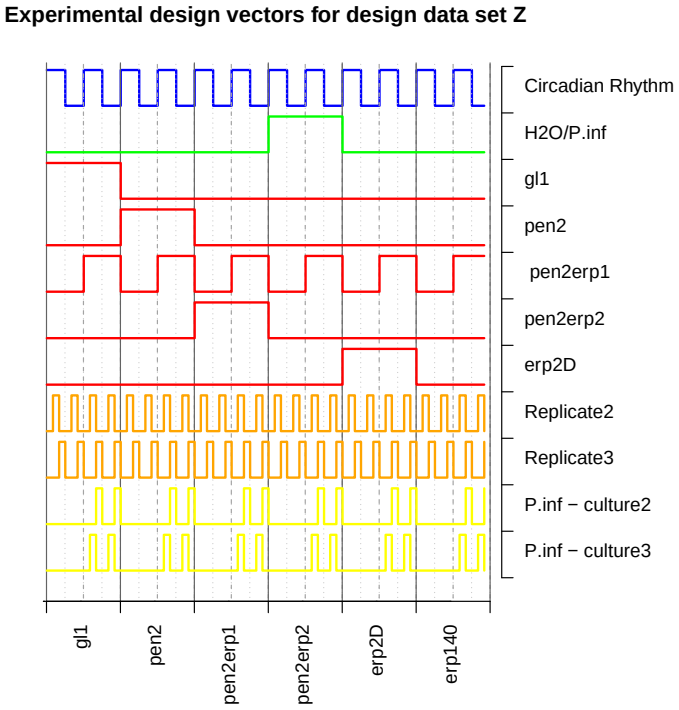


Figure 1: The experimental design vectors for the 72 experiments. The biological replicates are consecutive, the solid vertical lines separate the six mutants, the dotted lines seperate the time points (6h and 12h) and the dashed lines the treatment (H_2O and *P. infestans*).

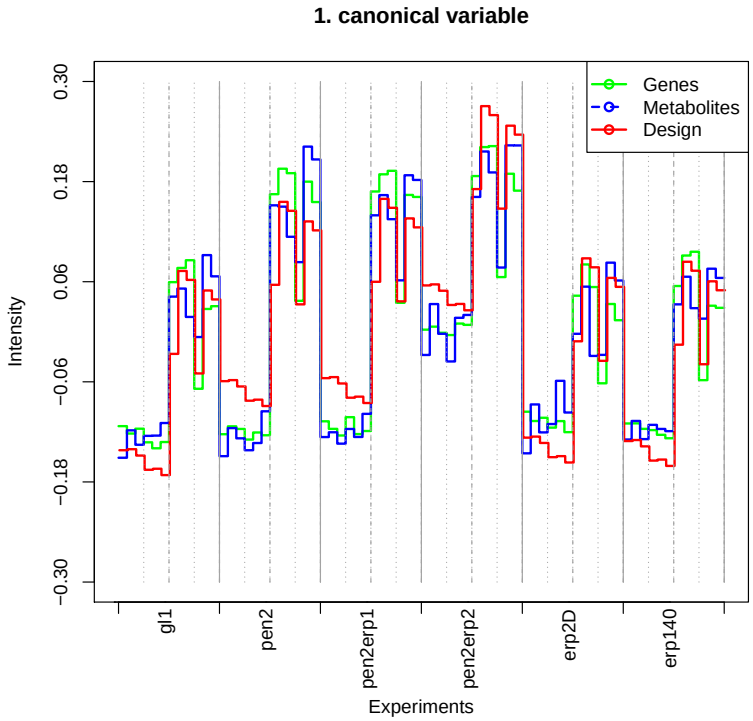


Figure 2: Supervised pCCA results: First canonical variable for the genes ($X_1 w_1$), metabolites ($X_2 w_2$) and the experimental design data set ($X_m w_m$) for all 72 experiments.

The genotype *pen2erp2* is the explanation of the second canonical variable (for this and all further variables: see supplemental material). The metabolites salicylic acid glucoside and dihydroxybenzoic acid receive the largest weights for this variable. Salicylic acid appears to be present at constitutively high levels in *pen2erp2*.

The third variable resembles the circadian rhythm, since the sample collection time was 6h and 12h after inoculation (at noon and late afternoon, respectively). Not surprisingly, a large number of genes show the circadian rhythm, but only a few metabolites in our data set. This might be due to the fact that the metabolites associated with circadian rhythm (especially primary metabolism and sugars), cannot be detected by LC/MS. A combination of two mutations is found in the fourth variable: the *erp2*-mutation and the *pen2*-mutation. As the *pen2erp2*-component was already subtracted in the second variable, mainly mutant *erp2D* is increased and *pen2* and *pen2erp1* are decreased in the canonical variable. The gene *PEN2* can be found among the largest weighted features.

An interesting effect is quality control of the experiment. Despite the best efforts to sustain reproducible experimental conditions, the fifth, the sixth and the ninth canonical variable could possibly be explained by unknown environmental factors or slight variations in the sample processing procedures which influenced the replicates, as well as the influence of different oomycete cultures to the plants. The spCCA allows to decompose these effects.

There are two further canonical variables (seventh and eighth) with a quite low but still significant correlation coefficient. They suggest a similarity between the mutants *pen2erp1*, *pen2* and *erp2D*, as well as differences

between the mutants *pen2* and *pen2erp1*.

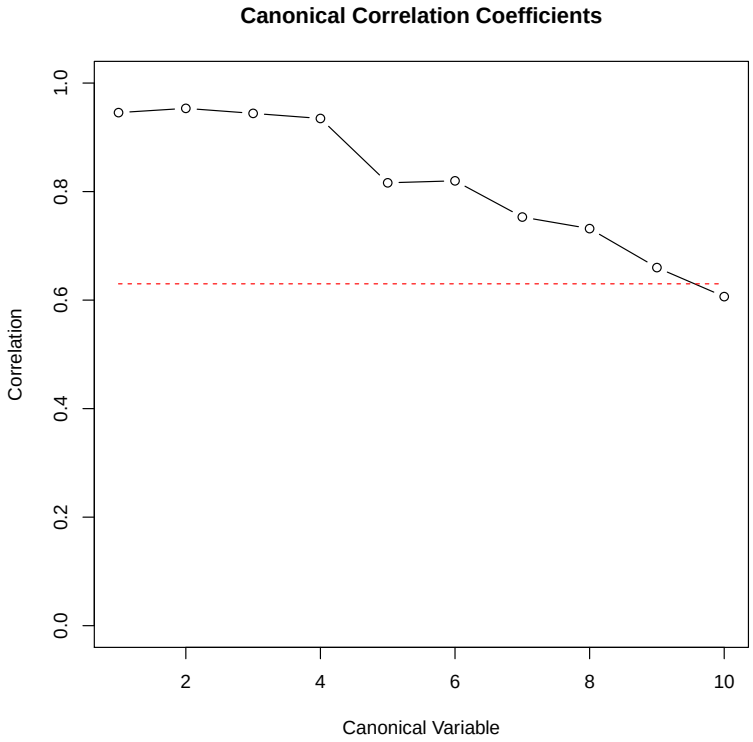


Figure 3: The correlation coefficients for the first ten canonical variables. The dashed line indicates a significance of $\alpha = 0.01$.

Our supervised approach was able to recognize nine significant variables for significance level $\alpha = 0.01$ and to give an explanation to them. The main explainable processes in the plants seem to be the defense against the oomycete, as well as the circadian rhythm. The effects of two of the mutations, as well as influences of the replicates show a slightly lower correlation. Further variables are not significantly correlated (Fig. 3).

For the standard CCA, which does not use sparse weight vectors, the correlation coefficients of the canonical variables are monotonously decreasing. This may not be the case in penalized CCA. This effect can be seen in Fig. 3, where the correlation coefficients for the first ten variables are shown. The second and the sixth correlation coefficient are larger than their respective predecessors.

If standard non-supervised penalized CCA is used, ten significant combinations of genes and metabolites were found, but only two were easy to interpret.

The first canonical variable (see supplemental material) is similar to the first variable of the spCCA and shows the reaction to the infection, which is increased for the *pen2*-mutants. The second variable corresponds to mutant *pen2erp2*, and the main metabolite is salicylic acid. A number of further canonical variables are found which are difficult to interpret. The canonical variables for the *pen2*-mutant as well as the *erp1*-mutation (fourth supervised pCCA variable) were not found.

4 Discussion

Discovery and interpretation of complex relationships between gene activity and metabolites is still a challenge in systems biology. A penalized canonical correlation analysis is a useful tool for this purpose. Still, the main question to a canonical variable is: which process does it resemble, what does it biologically mean? Although some processes are easy to identify, most of the canonical variables are difficult to interpret. One solution is to check the corresponding genes and metabolites of the canonical variable – but this is elaborate and very complicated since many genes and gene functions are still unknown, or the metabolites might not yet be identified.

Furthermore, many correlations are based on an unknown genotype effect or unobserved growth conditions. These effects are usually not interesting for the experimentalist and it is hard or even impossible to interpret the pattern. A standard pCCA can only search for high correlation, and thus well-explainable variables with lower correlation coefficient will be missed.

The supervised pCCA provides additional information as it explicitly incorporates the design of the experiment into the analysis, and thus the underlying biological questions. This assists in interpreting the biological processes and guides the spCCA to find interpretable processes of interest. We showed that this method was very useful to unravel relevant relationships between two data sets of gene expression and metabolite levels of *Arabidopsis thaliana* subjected to pathogen infection. To extract further processes a standard penalized CCA can be applied subsequently to unveil additional correlation in the residual data sets.

In the analyses described in this work the supervised pCCA was applied to two biological data sets and a third design data set. The methods can be easily extended to more than three data sets, but the determination of suitable sparsity parameters λ_k becomes very costly.

We created an R-package, which includes functions, examples and visualizations for two biological and one design data set. It is available at http://msbi.ipb-halle.de/msbi/spCCA under GPL license.

5 Acknowledgements

The gene expression experiments were funded by DFG-SPP1212 Plant Micro.

References

Ahn, J. H. (2009). RNA extraction from Arabidopsis for Northern blots and reverse transcriptase-PCR. *Cold Spring Harb Protoc*, **2009**(9), pdb.prot5295.

Bednarek, P., Pislewska-Bednarek, M., Svatos, A., Schneider, B., Doubsky, J., Mansurova, M., Humphry, M., Consonni, C., Panstruga, R., Sanchez-Vallet, A., Molina, A., and Schulze-Lefert, P. (2009). A glucosinolate metabolism pathway in living plant cells mediates broad-spectrum antifungal defense. *Science*, **323**, 101–106.

Hannah, M., Caldana, C., Steinhauser, D., Balbo, I., Fernie, A., and Willmitzer, L. (2010). Combined transcript and metabolite profiling of arabidopsis grown under widely variant growth conditions facilitates the identification of novel metabolite-mediated regulation of gene expression. *Plant Physiol.*, **152**, 2120–2129.

Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. (2003). Canonical correlation analysis; an overview with application to learning methods. Technical Report CSD-TR-03-02, Department of Computer Science, University of London.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, New York.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, **28**, 321–377.

Kettenring, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika*, **58**(3), 433–451.

Kopischke, M., Westphal, L., Schneeberger, K., Clark, R., Ossowski, S., Wewer, V., Fuchs, R., Landtag, J., Hause, G., Dörmann, P., Lipka, V., Weigel, D., Schulze-Lefert, P., Scheel, D., and Rosahl, S. (2013). Impaired sterol ester synthesis alters the response of *Arabidopsis thaliana* to *Phytophthora infestans*. *Plant J*, **73**(3), 456–468.

Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R., and Neumann, S. (2012). CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal Chem*, **84**(1), 283–289.

Lipka, V., Dittgen, J., Bednarek, P., Bhat, R., Wiermer, M., Stein, M., Landtag, J., Brandt, W., Rosahl, S., Scheel, D., Llorente, F., Molina, A., Parker, J., Somerville, S., and Schulze-Lefert, P. (2005). Pre- and postinvasion defenses both contribute to nonhost resistance in arabidopsis. *Science*, **310**, 1180–1183.

Muroi, A., Ishihara, A., Tanaka, C., Ishizuka, A., Takabayashi, J., Miyoshi, H., and Nishioka, T. (2009). Accumulation of hydroxycinnamic acid amides induced by pathogen infection and identification of agmatine coumaroyltransferase in arabidopsis thaliana. *Planta*, **230**, 517–527.

Parkhomenko, E., Tritchler, D., and Beyene, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, **8**(1).

Smith, C., Want, E. J., O'Maille, G., Abagyan, R., and Siuzdak, G. (2006). Xcms: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem*, **78**(3), 779–787.

Tautenhahn, R., Boettcher, C., and Neumann, S. (2008). Highly sensitive feature detection for high resolution lc/ms. *BMC Bioinformatics*, **9:504**.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.

Vía, J., Santamaría, I., and Pérez, J. (2006). A learning algorithm for adaptive canonical correlation analysis of several data sets. *Neural Networks*, **20**, 139–152.

Waaijenborg, S. and Zwinderman, A. H. (2009). Correlating multiple snps and multiple disease phenotypes: penalized non-linear canonical correlation analysis. *Bioinformatics*, **25**(21), 2764–2771.

Wilson, C. L. and Miller, C. J. (2005). Simpleaffy: a bioconductor package for affymetrix quality control and data analysis. *Bioinformatics*, **21**(18), 3683–3685.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, **67**, 301–320.

A Materials and Methods

A.1 Experiments

Six different Arabidopsis lines (*gl1* as wildtype, single mutants *pen2*, *erp140* and *erp2D*, double mutants *pen2erp1* and *pen2erp2*) were used for the experiment. Plants were treated with water or *P. infestans* spores as described at http://bar.utoronto.ca/efp/cgi-bin/efpWeb.cgi. Leaves were harvested at 6 hours and 12 hours after inoculation. The leaves of 6 different plants per genotype were pooled for the isolation of total RNA and for metabolite profiling. The whole experiment was repeated another two times resulting in 72 samples (6 plant lines $\times$ 2 treatments $\times$ 2 time points $\times$ 3 repeats.

A.1.1 Gene expression data

Total RNA was isolated from Arabidopsis leaves according to Ahn (2009) and purified using the RNeasy Plant Mini Kit (Qiagen). Hybridization of the samples to Affymetrix ATH1 GeneChips was performed by AROS APPLIED BIOTECHNOLOGY (Aarhus, Denmark).

A.1.2 LC/MS metabolite profiling and data processing

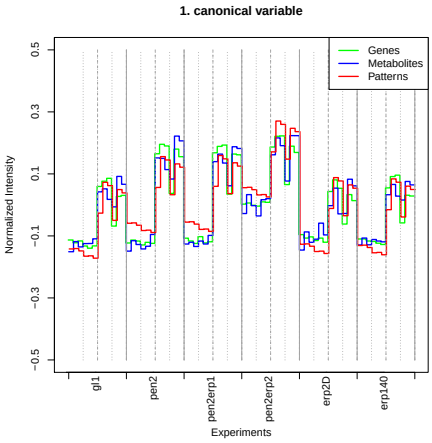
Chromatographic separations were performed on an Acquity UPLC system (Waters) equipped with a HSS T3 column (100×1.0 mm, particle size 1.8 μm; Waters) with a flow rate of 200 μL/min at 40 °C column temperature using the following gradient program: 0 – 60 s, isocratic 95% A (water/formic acid, 99.9/0.1 (v/v)), 5% B (acetonitrile/formic acid, 99.9/0.1 (v/v)); 60 – 360 s, linear from 5 to 30% B; 360 – 600 s, linear from 30 to 95% B; 360 – 720 s isocratic 95% B. The injection volume was 2.0 μL (full loop injection). Eluted compounds were detected at a spectra rate of 3 Hz from *m/z* 100 – 1000 using a MicrOTOF-Q-II (Bruker, Daltonics) equipped with an Apollo II electrospray ion source in positive ion mode with the following instrument settings: nebulizer gas, nitrogen, 1.2 bar; dry gas, nitrogen, 8 L/min, 190 °C; capillary, -4500 V; end plate offset, -500 V; funnel 1 RF, 200 Vpp; funnel 2 RF, 200 Vpp. Mass calibration of individual raw data files was performed on lithium formate cluster ions obtained by automatic infusion of 20 μL 10 mM lithium hydroxide in isopropanol/water/formic acid, 49.9/49.9/0.2 (v/v/v) at a gradient time of 720 s using a diverter valve.

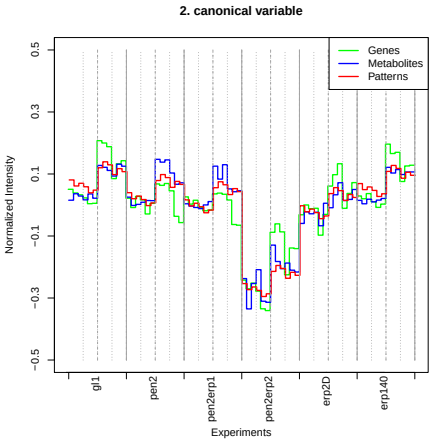
XCMS settings for processing LC/MS data were prefilter=3,500; snthr=3; ppm=25, peakwidth=5,12. For alignment group.density function with parameters minfrac=0.75 and bw=5 was used.

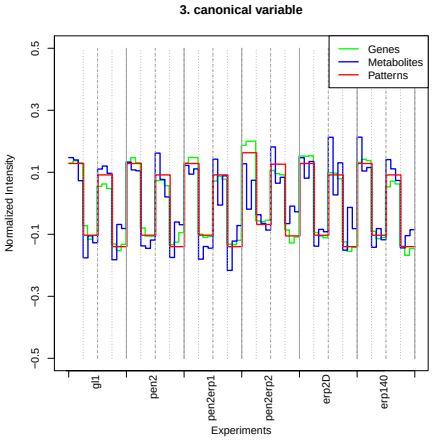
B Canonical variables for supervised pCCA for data assay of *Arabidopsis columbiana* with infection with *P. infestans*

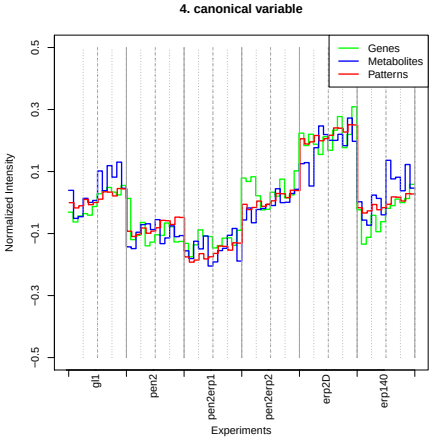
Below the eight significant variables are described and are all depicted in the figures below.

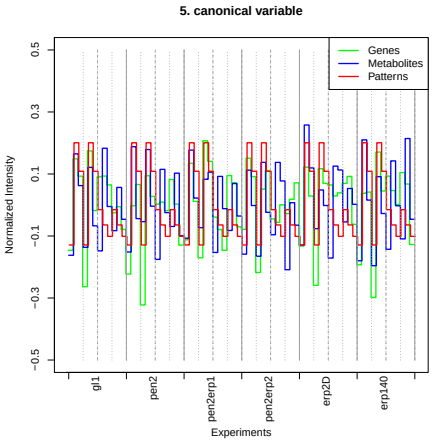
- First variable: Reaction to *P. infestans*.
- Second variable: Mutant *pen2erp2*. Constantly high abundance of salicylic acid in *pen2erp2*-mutants. Because the infection with *P. infestans* was subtracted in the previous canonical variable by regression, a negative image of this variable was created, resulting in this pattern.
- Third variable: Circadian rhythm.
- Forth variable: Combination of *pen2*-mutation and *erp2*-mutation. One associated gene is the *PEN2*-gene. *pen2erp2* was already subtracted in the second variable.
- Fifth variable: Replicates and influence of different oomycete cultures on plants.
- Sixth variable: Replicates.
- Seventh variable: Similarities between mutants *pen2erp1*, *pen2* and *erp2D*.
- Eighth variable: Mutant *pen2erp1*.
- Ninth variable: Influence of different oomycete cultures on plants.

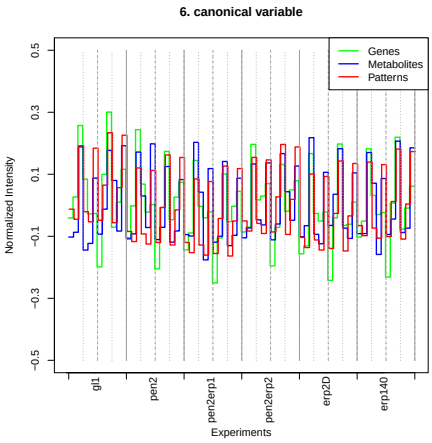


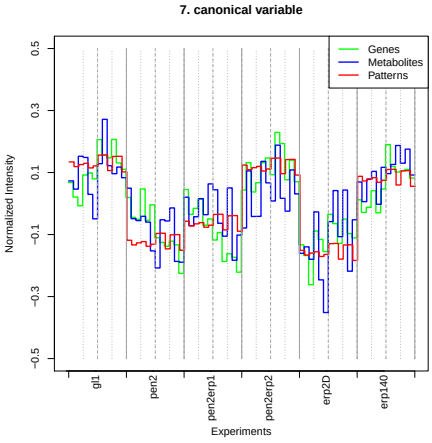

2. canonical variable
Genes
Metabolites
Patterns
Normalized Intensity
0.5
0.3
0.1
−0.1
−0.3
−0.5
gl1
pen2
pen2erp1
pen2erp2
erp2D
erp140
Experiments

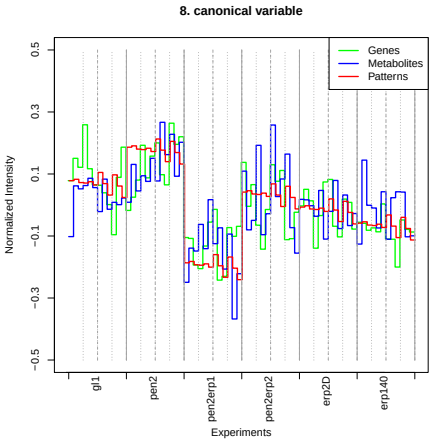

3. canonical variable
Genes
Metabolites
Patterns
Normalized Intensity
0.5
0.3
0.1
−0.1
−0.3
−0.5
gl1
pen2
pen2erp1
pen2erp2
erp2D
erp140
Experiments

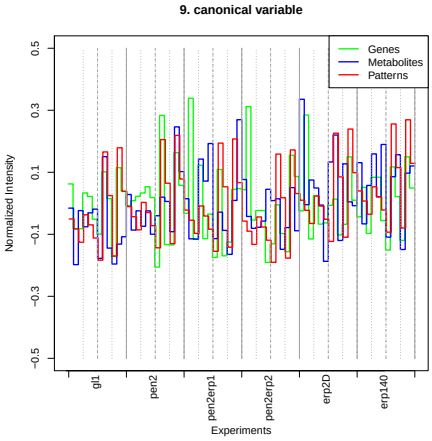

4. canonical variable
Genes
Metabolites
Patterns
Normalized Intensity
0.5
0.3
0.1
−0.1
−0.3
−0.5
gl1
pen2
pen2erp1
pen2erp2
erp2D
erp140
Experiments


5. canonical variable
Genes
Metabolites
Patterns
Normalized Intensity
0.5
0.3
0.1
−0.1
−0.3
−0.5
gl1
pen2
pen2erp1
pen2erp2
erp2D
erp140
Experiments


6. canonical variable
Genes
Metabolites
Patterns
Normalized Intensity
0.5
0.3
0.1
−0.1
−0.3
−0.5
gl1
pen2
pen2erp1
pen2erp2
erp2D
erp140
Experiments


7. canonical variable
Genes
Metabolites
Patterns
Normalized Intensity
0.5
0.3
0.1
−0.1
−0.3
−0.5
gl1
pen2
pen2erp1
pen2erp2
erp2D
erp140
Experiments

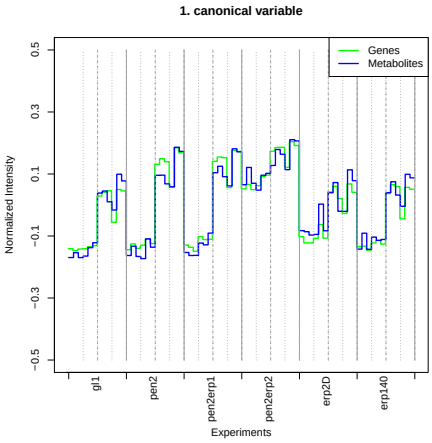


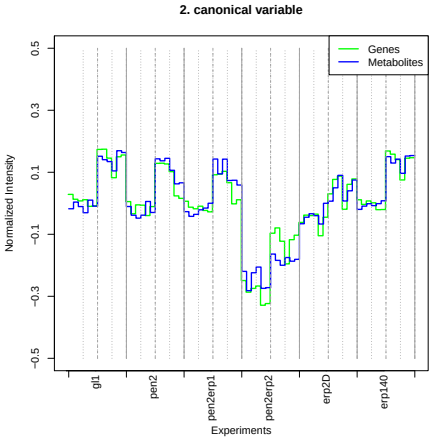


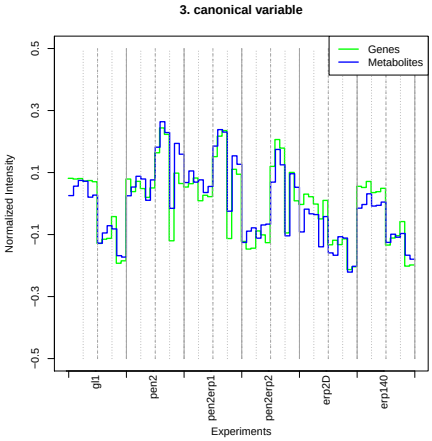
C Canonical Variables for Standard Penalized CCA

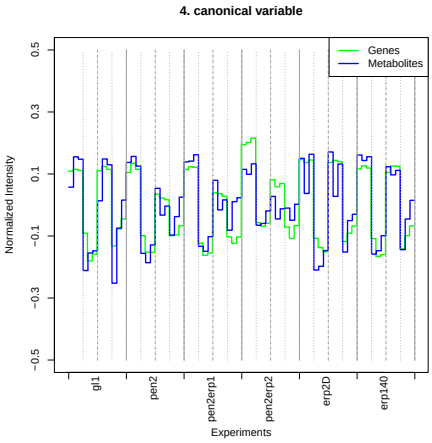
Below, the seven significant variables are described and are all depicted in the figures below.

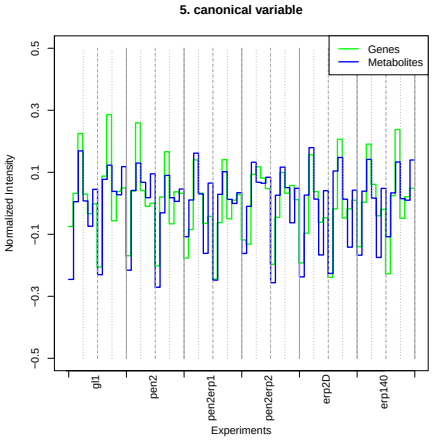
- First variable: Reaction to *P.infestans*, enhanced reaction for *pen2*-mutants. Camalexin is the main metabolite.
- Second variable: Mutant *pen2erp2*. Constantly high abundance of salicylic acid in *pen2erp2*-mutants. Because the infection with *P. infestans* was subtracted in the previous canonical variable by regression, a negative image of this variable was created, resulting in this pattern.
- Fourth variable: Circadian rhythm.
- Further variables: Unknown processes

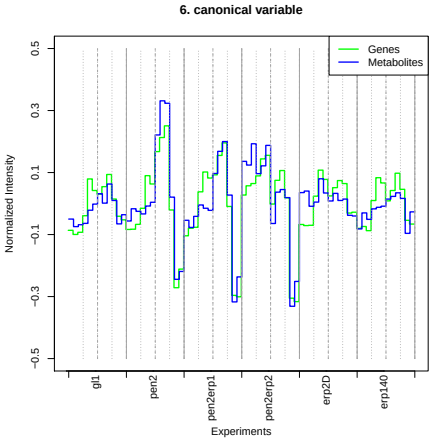

1. canonical variable
Genes
Metabolites
Normalized Intensity
0.5
0.3
0.1
−0.1
−0.3
−0.5
gl1
pen2
pen2erp1
pen2erp2
erp2D
erp140
Experiments

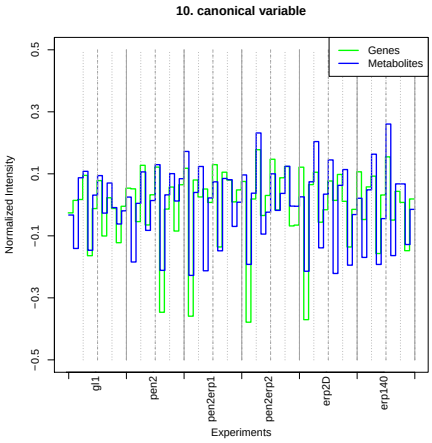

2. canonical variable
Genes
Metabolites
Normalized Intensity
0.5
0.3
0.1
−0.1
−0.3
−0.5
gl1
pen2
pen2erp1
pen2erp2
erp2D
erp140
Experiments


3. canonical variable
Genes
Metabolites
Normalized Intensity
0.5
0.3
0.1
−0.1
−0.3
−0.5
gl1
pen2
pen2erp1
pen2erp2
erp2D
erp140
Experiments


4. canonical variable
Genes
Metabolites
Normalized Intensity
0.5
0.3
0.1
−0.1
−0.3
−0.5
gl1
pen2
pen2erp1
pen2erp2
erp2D
erp140
Experiments


5. canonical variable
Genes
Metabolites
Normalized Intensity
0.5
0.3
0.1
−0.1
−0.3
−0.5
gl1
pen2
pen2erp1
pen2erp2
erp2D
erp140
Experiments

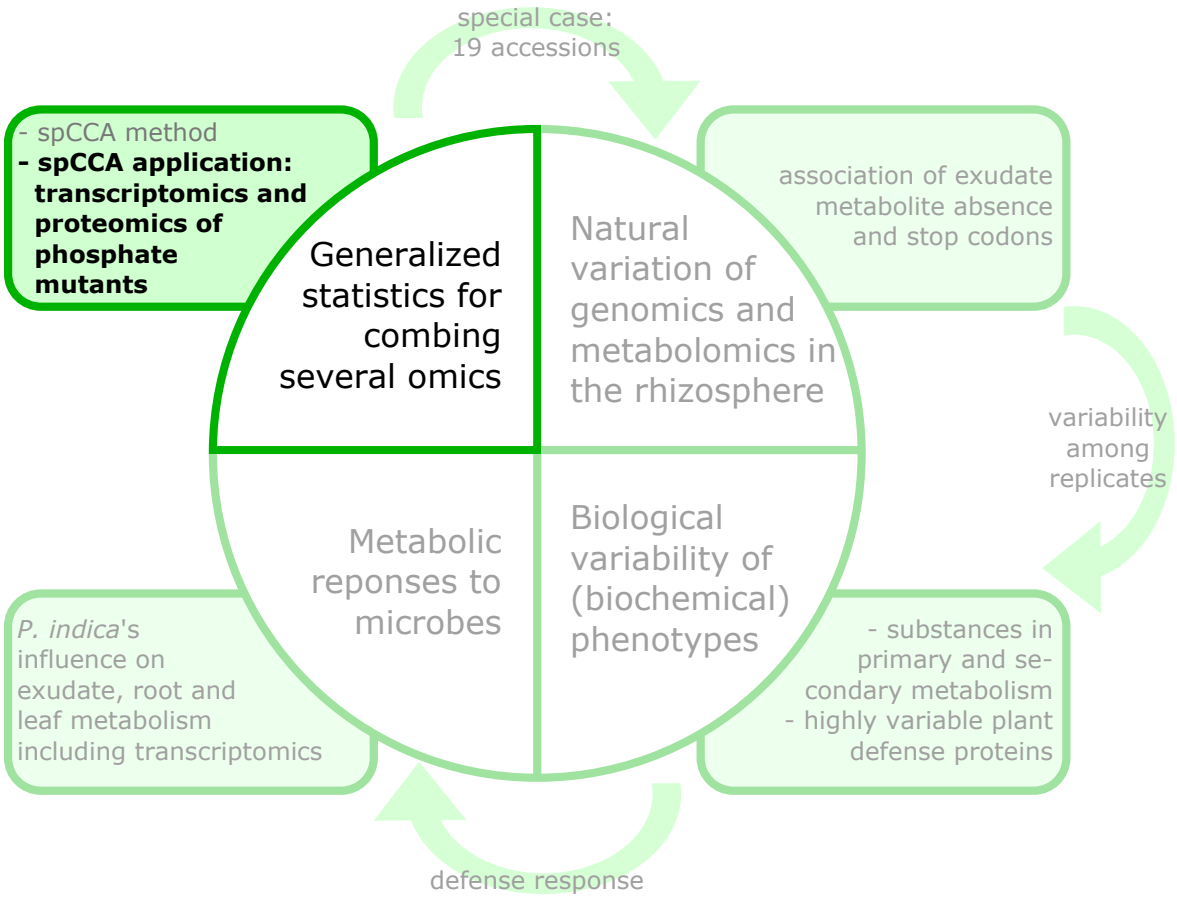

6. canonical variable
Genes
Metabolites
Normalized Intensity
0.5
0.3
0.1
−0.1
−0.3
−0.5
gl1
pen2
pen2erp1
pen2erp2
erp2D
erp140
Experiments


10. canonical variable
Genes
Metabolites
Normalized Intensity
0.5
0.3
0.1
−0.1
−0.3
−0.5
gl1
pen2
pen2erp1
pen2erp2
erp2D
erp140
Experiments

2.1.2 Comparative expression profiling reveals a role of the root apoplast in local phosphate response

Hoehenwarter, W.; Mönchgesang, S.; Neumann, S.; Majovsky, P.; Abel, S.; Müller, J. Comparative expression profiling reveals a role of the root apoplast in local phosphate response. *BMC Plant Biol* **2016**, *16*, 106.

equal contributions



Hoehenwarter *et al. BMC Plant Biology* (2016) 16:106
DOI 10.1186/s12870-016-0790-8

BMC Plant Biology

RESEARCH ARTICLE Open Access

Comparative expression profiling reveals a role of the root apoplast in local phosphate response

Wolfgang Hoehenwarter[1†], Susann Mönchgesang[2†], Steffen Neumann[2], Petra Majovsky[1], Steffen Abel[3,4,5] and Jens Müller[3*†]

Abstract

Background: Plant adaptation to limited phosphate availability comprises a wide range of responses to conserve and remobilize internal phosphate sources and to enhance phosphate acquisition. Vigorous restructuring of root system architecture provides a developmental strategy for topsoil exploration and phosphate scavenging. Changes in external phosphate availability are locally sensed at root tips and adjust root growth by modulating cell expansion and cell division. The functionally interacting *Arabidopsis* genes, *LOW PHOSPHATE RESPONSE 1* and *2* (*LPR1/LPR2*) and *PHOSPHATE DEFICIENCY RESPONSE 2* (*PDR2*), are key components of root phosphate sensing. We recently demonstrated that the *LOW PHOSPHATE RESPONSE 1 - PHOSPHATE DEFICIENCY RESPONSE 2 (LPR1-PDR2)* module mediates apoplastic deposition of ferric iron (Fe^{3+}) in the growing root tip during phosphate limitation. Iron deposition coincides with sites of reactive oxygen species generation and triggers cell wall thickening and callose accumulation, which interfere with cell-to-cell communication and inhibit root growth.

Results: We took advantage of the opposite phosphate-conditional root phenotype of the *phosphate deficiency response* 2 mutant (hypersensitive) and *low phosphate response 1* and *2* double mutant (insensitive) to investigate the phosphate dependent regulation of gene and protein expression in roots using genome-wide transcriptome and proteome analysis. We observed an overrepresentation of genes and proteins that are involved in the regulation of iron homeostasis, cell wall remodeling and reactive oxygen species formation, and we highlight a number of candidate genes with a potential function in root adaptation to limited phosphate availability. Our experiments reveal that *FERRIC REDUCTASE DEFECTIVE 3* mediated, apoplastic iron redistribution, but not intracellular iron uptake and iron storage, triggers phosphate-dependent root growth modulation. We further highlight expressional changes of several cell wall-modifying enzymes and provide evidence for adjustment of the pectin network at sites of iron accumulation in the root.

Conclusion: Our study reveals new aspects of the elaborate interplay between phosphate starvation responses and changes in iron homeostasis. The results emphasize the importance of apoplastic iron redistribution to mediate phosphate-dependent root growth adjustment and suggest an important role for citrate in phosphate-dependent apoplastic iron transport. We further demonstrate that root growth modulation correlates with an altered expression of cell wall modifying enzymes and changes in the pectin network of the phosphate-deprived root tip, supporting the hypothesis that pectins are involved in iron binding and/or phosphate mobilization.

Keywords: *Arabidopsis thaliana*, Phosphate deficiency, Root growth, Proteomics, Transcriptomics, Iron transport, Cell wall, Pectin

* Correspondence: Jens.Mueller@ipb-halle.de
†Equal contributors
[3]Department of Molecular Signal Processing, Leibniz Institute of Plant Biochemistry, D-06120 Halle (Saale), Germany
Full list of author information is available at the end of the article

Background

Inorganic phosphate (Pi) is an essential macronutrient for plant growth and development. Despite its high abundance in the rhizosphere, bioavailability of Pi is typically limited because its majority is bound in organic compounds or complexed with metal ions such as Ca (alkaline soils), Fe or Al (acidic soils) [1]. Thus, plants evolved strategies to enhance Pi acquisition and to conserve or remobilize Pi from internal sources to adapt to Pi limiting conditions. Previous efforts elucidated some of these adaptive responses, including the identification of high-affinity Pi transport systems, the characterization of diverse metabolic bypass reactions, the reutilization of Pi from phospholipids, and many more [2]. Most of the Pi starvation response (*PSR*) genes involved in these systemic adjustments are regulated by the myb transcription factor PHR1 (PHOSPHATE STARVATION RESPONSE1) [3–6].

Dynamic redesign of the root system architecture (RSA) provides another strategy to maintain cellular Pi supply. In *Arabidopsis*, low external Pi availability is locally sensed by the growing root tip, which causes reduction of cell elongation and meristematic activity at the site of Pi depletion. The resultant inhibition of root growth is accompanied by accelerated formation of root hairs and development of lateral roots to increase the absorptive surface for topsoil exploration [7, 8]. The development of a densely branched and/or shallow root systems increases Pi starvation tolerance in several plant species, including agronomically important crops such as barley, lupin, soybean or common bean [9]. Several *Arabidopsis* mutants with altered Pi dependent root growth responses have been described [10–18]. However, for most of the underlying genes only little information is available how they affect Pi sensing and root growth modulation. *LPR1* (*LOW PHOSPHATE ROOT1*), its closely related paralog *LPR2*, and *PDR2* (*PHOSPHATE DEFICIENCY RESPONSE2*) have been identified as central players in local root Pi sensing [11, 13, 19]. *PDR2*, which codes for the single P5-type ATPase of unknown substrate-specificity (AtP5A), and *LPR1*, encoding a multicopper oxidase, are expressed in overlapping domains of the root apical meristem (RAM). *LPR1* and *PDR2* interact genetically and are required for meristem maintenance and cell elongation in Pi-deprived roots. Importantly, the *lpr1lpr2* mutation impedes local root growth inhibition under Pi limitation and suppresses the hypersensitive short-root phenotype of *pdr2* plants, indicating that they act in the same pathway [11, 13].

Previous work revealed that external Fe availability modifies local Pi sensing [11, 13, 20]. A number of studies observed that Pi-starved *Arabidopsis* and rice plants accumulate elevated levels of Fe in the root and the shoot [20–23], which has been suggested as a proactive strategy to mobilize Pi from insoluble Fe complexes [8]. Fe participates in the formation of reactive oxygen species (ROS) and it has been proposed that Fe toxicity causes local root growth inhibition [20]. We recently provided evidence for apoplastic LPR1 ferroxidase activity and uncovered a major role of the *LPR1-PDR2* module for root tip-specific deposition of Fe^{3+} in cell walls (CW) of the RAM and elongation zone (EZ) during Pi limitation [19]. We further showed that Fe accumulation in the RAM is massively enhanced in Pi-starved *pdr2* roots, but suppressed in the insensitive *lpr1lpr2* line. Fe deposition coincides with sites of ROS generation and triggers CW thickening and callose accumulation, which interferes with cell-to-cell communication, RAM maintenance, and cell elongation.

In recent years, a set of transcriptome profiling studies provided significant insights into the transcriptional changes upon Pi deficiency in *Arabidopsis* [6, 21, 24–28]. In addition, a complementary transcriptome and proteome study highlighted the convergence of mRNA and protein expression profiles on lipid remodeling and glucose metabolism upon Pi-deprivation [25]. In this study, we performed comparative transcriptome and proteome expression profiling on roots of Pi-replete and Pi-starved wild-type (Col-0), *pdr2*, and *lpr1lpr2* plants in combination with a set of physiological and cell biological experiments. Our analysis emphasizes the importance of root Fe uptake and redistribution under Pi limitation. We highlight the potential role of so far unknown players in the regulation of Pi-dependent Fe-redistribution and demonstrate that apoplastic but not intracellular Fe accumulation triggers Pi-dependent root growth modulation. Consistently, we observed regulation of several CW modifying enzymes, which correlates with an increased deposition of pectin at sites of Fe accumulation. The potential role of pectin in Pi-dependent root Fe storage and Pi mobilization is discussed.

Results

Differential gene expression correlates with genotype-specific Pi sensitivity

For transcriptome analysis, wild-type, *pdr2* and *lpr1lpr2* seedlings were germinated on + Pi agar (4 days) and transferred to + Pi or −Pi medium for 20 h, a period during which Pi limitation alters global gene expression [28] as well as root meristem activity [19]. RNA was extracted from roots of three biological replicates and prepared for hybridization with ATH1 Affymetrix chips. Data were analyzed using ARRAYSTAR (Version 4.1.0) and further processed (Additional file 1: Table S1). Hierarchical clustering (Fig. 1a) confirmed high homogeneity within each replicate set because the biological replicates clustered together for each genotype and Pi condition (as indicated by the short branches at the

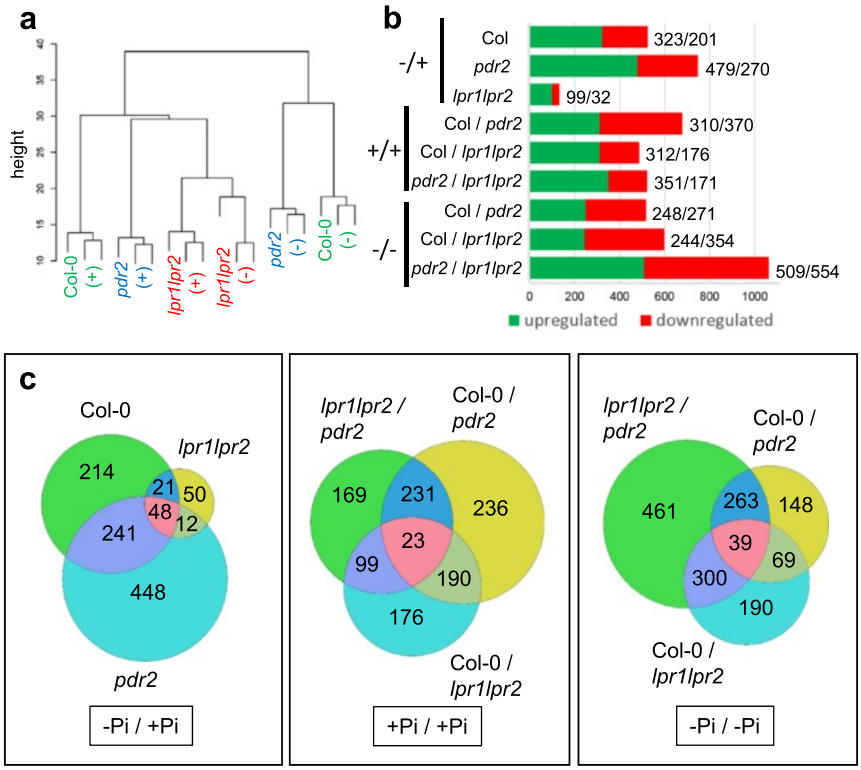


Fig. 1 Statistical analysis of comparative gene expression analysis. **a** Hierarchical clustering of all ATH1 datasets, including wild-type (Col-0), *pdr2* and *lpr1lpr2* samples, the two growth regimes (+Pi and –Pi), and three biological replicates. Distances in the dendrogram illustrate the degree of relationship between samples. Note the short distance between biological replicate sets (*lowest branches*) compared to the relatively long distance between + Pi and –Pi conditions. **b** Number of up- and downregulated genes ($p \leq 0.05$, Student's *t*-test; $0.66 \geq FC \geq 1.5$) in wild-type, *pdr2* and *lpr1lpr2* roots upon transfer from + Pi to –Pi (–/+), and in the three genotypes under + Pi (+/+) or –Pi (–/–) conditions. **c** Venn diagrams illustrating the number of differentially regulated genes in pairwise comparisons for all three genotypes under both growth regimes

bottom of the dendrogram). It also revealed a clear separation between + Pi and –Pi samples for the wild-type and the hypersensitive *pdr2* mutant (long branches between the + Pi and –Pi samples), but less pronounced differences for the insensitive *lpr1lpr2* line (shorter branches between the + Pi and –Pi samples). Pairwise comparisons using a fold-change cutoff value of ≥ 1.5 for increased and of ≤0.66 for decreased transcript levels ($p \leq 0.05$; Student's *t*-test) revealed 2292 differentially expressed genes across all genotypes and the two growth conditions. Low Pi exposure altered the expression of 749, 524, and 131 genes in *pdr2*, wild-type, and *lpr1lpr2* roots, respectively (Fig. 1b). Thus, the genotype-specific sensitivity of root growth inhibition in response to Pi depletion positively correlates with the number of differentially regulated genes.

Identification of genotype-independent Pi-responsive genes

We generated Venn diagrams to illustrate the distribution of differentially expressed genes between the three genotypes (Fig. 1c). Wild-type shared a subset of 289 and 69 Pi-responsive genes with *pdr2* and *lpr1lpr2*, respectively, and all three lines had in common a core set of 48 genes (Fig. 1c). Hierarchical clustering of this core set revealed similar expression changes in all genotypes in response to –Pi with high positive correlation (Additional file 2: Figure S1 A, B). The core set comprises two partially overlapping groups that consist of at least 19 *PSR* and 23 metal-responsive genes (Table 1, Additional file 3: Table S2). Members of the first group (e.g., *SPX1*, *PAP17/ACP5*, *SRG3*, *CAX3*) are known targets of the Pi-regulated myb transcription factor PHR1 [5, 6, 29–31], suggesting that the systemic response to Pi deficiency is maintained in *pdr2* and *lpr1lpr2* mutants.

In the second group, Fe-related genes are overrepresented (17 members) and comprise the majority of repressed genes (Table 1). The most strongly suppressed gene in all three genotypes (>10-fold repression) codes for IRT1, the major feedback-regulated Fe-uptake system in *Arabidopsis* [32, 33]. Many *IRT1* co-regulated genes (http://atted.jp) are induced under Fe deficiency [34–36].

Table 1 Pi-dependent transcriptional changes of commonly regulated genes

Locus	Name	fc (-Pi/+Pi)			responsiveness
		Col	*pdr2*	*lpr1 lpr2*	
At1g08430	ALMT1	6.0	7.1	10.9	Al[1,2] responsive
At3g59930	defensin-like protein	4.5	6.8	2.7	Zn[3] responsive
At1g73220	AtOCT1	4.3	6.1	3.6	Pi[4] responsive
At5g20150	SPX1	4.3	3.9	5.0	Pi[5,6] responsive
At5g20790	hypothetical protein	3.7	4.7	4.2	Pi[4] and As[7] responsive
At5g06860	PGIP1	3.4	2.0	2.2	Pi[8] and pathogen[9,10] responsive
At1g05340	hypothetical protein	3.4	-1.9	2.6	Al[1] / oxidative stress[11]
At3g17790	PAP17/ACP5	3.3	4.4	3.2	Pi[4,12] responsive
At1g10970	ZIP4	2.8	3.6	1.7	Zn[13] and Fe[13] responsive
At3g02040	SRG3	2.7	2.2	1.8	Pi[4,5] and As[7] responsive
At5g38930	germin-like protein	2.6	1.8	2.9	
At1g05000	PFA-DSP1	2.6	2.5	2.9	Pi[14] responsive
At2g46600	unknown	2.5	2.2	1.8	N[34] responsive
At2g04460	transposable_element_gene	2.5	2.3	3.8	As[7] responsive
At1g80240	ATGDI1	2.5	1.9	1.8	
At5g22890	STOP2	2.3	1.8	2.2	Al[33] responsive
At5g38710	proline dehydrogenase 2	2.2	2.3	1.7	
At4g03960	PFA-DSP4	2.2	2.4	2.2	Pi[14] and pathogen[15] responsive
At4g30110	HMA2	2.2	3.0	1.7	metal[17] responsive
At2g34180	CIPK13	2.2	2.2	3.0	
At2g41380	methyltransferase-like	2.2	1.8	2.4	Cd[35] responsive
At3g49160	pyruvate kinase-like	2.2	3.0	1.6	
At3g51860	CAX3	2.1	1.9	1.7	Pi[18] responsive
At3g29810	COBL2	2.1	1.6	1.7	
At5g01600	FER1	2.0	2.3	1.7	Pi[5] and Fe[16] responsive
At3g47420	PS3	2.0	2.7	3.1	Pi[4] and Fe[21] responsive
At4g25100	FSD1	1.9	2.5	1.6	oxidative stress[22] responsive
At1g60960	IRT3	1.7	2.2	1.6	Zn[3,23] and Fe[3] responsive
At5g47740	protein coding	1.7	1.6	2.1	Pi[4] responsive
At1g05300	ZIP5	1.5	2.2	1.9	Zn[3] and Fe[3] responsive
At3g56980	BHLH039	-1.5	-4.6	-2.1	Fe[24] responsive
At5g38820	protein coding	-1.6	-2.5	-1.9	Fe[21] responsive
At3g07720	galactose oxidase	-1.7	-2.3	-1.8	Zn[25] and Fe[21] responsive
At3g61930	unknown	-1.8	-2.6	-2.6	Fe[21] responsive
At3g22231	PCC1	-1.9	-1.7	-1.6	pathogen / circadian[26] responsive
At5g03570	ATIREG2	-2.0	-2.9	-1.8	Pi[5], Zn[3], Fe[3,21] and Ni[27] responsive
At3g02610	protein coding	-2.0	-2.0	-1.6	
At2g40750	WRKY54	-2.2	-2.2	-1.5	
At5g45070	AtPP2-A8	-2.4	-2.1	-1.9	
At3g58060	MTPc3	-2.6	-3.3	-1.6	Pi[5], Zn[3] and Fe[3,21] responsive
At3g58810	MTPA2	-3.0	-5.0	-2.7	Pi[5], Zn[3] and Fe[3,21] responsive
At3g46900	COPT2	-3.0	-4.2	-2.4	Pi[5,29], Cu[28] and Fe[21,29] responsive
At5g62420	oxidoreductase	-3.4	-2.4	-1.6	
At5g02780	GSTL1	-5.6	-7.0	-2.9	Fe[21] responsive
At3g12900	2OG-Fe(II) oxygenase family	-5.7	-12.5	-4.9	Pi[5], Zn[3] and Fe[3,21] responsive
At1g73120	unknown protein	-9.9	-5.1	-2.9	
At4g31940	CYP82C4	-14.5	-29.4	-8.2	Pi[5], Zn[3] and Fe[3,21,30] responsive
At4g19690	IRT1	-26.3	-23.8	-8.9	Pi[5] and metal[31,32] responsive

Shown is the fold change expression (FC) of all 48 Pi-responsive genes that are regulated in each of the tested genotypes (wild-type, *pdr2* and *lpr1lpr2*). Grey and white boxes denote genes that are significantly suppressed or induced, respectively ($p \leq 0.05$, student's *t*-test; $0.66 \geq FC \geq 1.5$). All genes were interrogated for published responsiveness to Pi-starvation and/or metal-ions. References are indicated in superscript numbers and listed in Additional file 3: Table S2

Interestingly, 13 of the top 25 co-regulated genes are repressed in Pi-starved roots irrespective of the genotype (Table 2). Intriguingly, Pi-replete *pdr2* roots show higher expression of at least 12 Fe-related genes (Table 2), including a group of transcription factors (BHLH039, BHLH101, MYB10, MYB72) known to promote Fe-uptake under Fe deficiency [37–39]. The remaining Fe-related genes of this group are similarly induced in all three genotypes and encode the Fe storage protein FERRITIN1 (FER1) and various Fe-responsive metal transporters thought to be involved in transition metal detoxification and homeostasis (Table 1, Additional file 3: Table S2).

Pi depletion alters expression of cell wall-related genes

We identified 241 Pi-responsive genes that are shared between the wild-type and the hypersensitive *pdr2* mutant, but not with the insensitive *lpr1lpr2* line (Fig. 1c). Surprisingly, only 10 genes of unknown functions in Pi starvation response were significantly deregulated in *pdr2* compared with the wild-type (>2-fold), whereas the remaining genes showed a high positive correlation ($r = 0.88$) between both genotypes (Additional file 2: Figure S1C, Additional file 4: Table S3). GO term analysis revealed high overrepresentation of gene products associated with the extracellular region (GO:0005576). An extended analysis for enriched GO terms within a group of 1680 genes (Additional file 5: Table S4), which are either regulated by –Pi in one or more genotypes or are differentially expressed in at least one of the lines in + Pi ($p < 0.05$; BH corrected), confirmed overrepresentation of genes (322) annotated to encode extracellular proteins (Additional file 2: Figure S1D, Additional file 6: Table S5). In this group, we identified a subset of 66 genes with putative functions in CW remodeling (Table 3). A similar number of genes were differentially expressed in *pdr2* (27) and wild-type (33) but only one-third (11) in *lpr1lpr2* roots. As noted for Fe-related genes, many CW-modifying genes (31) were deregulated in Pi-replete *pdr2* roots. Within the subset of 66 genes, 29 encoded proteins could be assigned a potential function in pectin modification, predominantly pectin methylesterification. In addition, we noted several expansins and xyloglucan endotransglycosylases (XTH) as well as a set of carbohydrate hydrolyzing enzymes. Intriguingly, all these proteins are predicted to regulate CW extensibility [40, 41].

GO term analysis also revealed overrepresentation of genes encoding tetrapyrrole- and heme-binding proteins (GO:0046906 and GO:0020037) with oxidoreductase activity (GO:0016491) (Additional file 2: Figure S1D). This group codes for 29 peroxidases and most of those (28) belong to the 73 member-family of class III peroxidases (CIII Prx) (Additional file 7: Table S6), which are extracellular enzymes with partly antagonistic functions in ROS formation and CW dynamics [42]. While Pi-responsive expression of 8 CIII Prx-encoding genes was similar between wild-type and *pdr2* roots, 7 genes were regulated independently in each line under low Pi, and only three CIII Prx genes responded significantly to Pi

Table 2 Pi-dependent regulation of the top 25 genes co-regulated with *IRT1* (ATTEDII)

Locus	(-Pi/+Pi)			(+Pi/+Pi)		Gene
	Col	*pdr2*	*lpr1 lpr2*	*pdr2*	*lpr1 lpr2*	
At4g19690	-26.3	-23.8	-8.9	1.2	-1.1	IRT1 (IRON-REGULATED TRANSPORTER 1)
At3g58810	-3.0	-5.0	-2.7	-1.1	1.0	MTPA2 (METAL TOLERANCE PROTEIN A2)
At4g31940	-14.5	-29.4	-8.2	2.7	-1.3	CYP82C4 (cytochrome P450-like)
At1g74770	-1.2	-2.0	-1.5	1.7	1.3	zinc ion binding protein
At5g56080	-1.1	1.3	1.2	1.1	-1.6	NAS2 (NICOTIANAMINE SYNTHASE 2)
At3g12900	-5.7	-12.5	-4.9	1.8	1.0	2OG (2OG-Fe(II) oxygenase family protein)
At1g73120	-9.9	-5.1	-2.9	1.1	1.0	hypothetical protein
At3g56980	-1.5	-4.6	-2.1	3.5	1.3	BHLH039; transcription factor
At3g46900	-3.0	-4.2	-2.4	1.6	-1.3	COPT2 (copper ion transmembrane transporter)
At5g38820	-1.6	-2.5	-1.9	1.7	1.1	putative amino acid transporter
At1g62280	-1.1	-1.8	-1.0	4.4	1.1	SLAH1 (SLAC1 HOMOLOGUE 1); transporter
At3g07720	-1.7	-2.3	-1.8	1.2	-1.1	galactose oxidase/kelch repeat-containing protein
At5g03570	-2.0	-2.9	-1.8	2.0	1.8	ATIREG2 (IRON-REGULATED PROTEIN 2)
At5g02780	-5.6	-7.0	-2.9	1.0	-1.2	GSTL1 (glutathione transferase lambda 1)
At5g04150	1.0	-2.5	-1.1	2.9	1.8	BHLH101; transcription factor
At1g56160	-1.2	-4.4	-1.4	3.2	1.1	MYB72 (MYB DOMAIN PROTEIN 72)
At3g12820	-1.2	-2.3	-1.6	2.2	1.1	MYB10 (MYB DOMAIN PROTEIN 10)
At5g45070	-2.4	-2.1	-1.9	1.2	-1.2	AtPP2-A8 (Phloem protein 2-A8)
At4g09110	1.0	-1.1	-1.2	1.0	1.2	putative RING-H2 finger protein ATL35
At4g00910	-1.2	-1.4	-1.1	2.3	1.0	aluminum activated malate transporter family
At4g19680	-1.1	-1.6	1.1	1.3	1.0	IRT2 (IRON-REGULATED TRANSPORTER 2)
At3g01260	1.1	-1.1	-1.1	1.2	1.2	aldose 1-epimerase-like prtotein
At5g04950	-2.0	-1.3	-1.4	-1.1	-1.9	NAS1 (NICOTIANAMINE SYNTHASE 1)
At3g58060	-2.6	-3.3	-1.6	1.4	-1.4	putative metal tolerance protein C3
At1g32450	-1.1	1.2	-1.1	1.3	1.0	NRT1.5 (NITRATE TRANSPORTER 1.5)

Shown is the fold change expression in wild type, *pdr2* and *lpr1lpr2* after transfer to –Pi or the fold change expression of Pi-replete *pdr2* and *lpr1lpr2* plants compared to the wild-type. Red and green boxes denote genes that are significantly suppressed or induced ($p \leq 0.05$, student's *t*-test; $0.66 \geq FC \geq 1.5$)

Table 3 Pi-dependent regulation of cell wall modifying enzymes

Locus	(-Pi/+Pi)			(+Pi/+Pi)		Name
	Col-0	*pdr2*	*lpr1 lpr2*	*pdr2*	*lpr1 lpr2*	
Pectin modification						
At1g53830	0.56	0.83	0.76	0.67	1.10	PME02
At1g05310	0.94	0.89	0.88	0.63	1.15	PME08 (probable)
At2g26440	2.06	1.90	0.91	1.01	1.19	PME12 (probable)
At2g43050	0.91	0.57	0.87	0.93	1.12	PME16 (probable)
At2g47550	1.99	1.98	1.29	0.64	0.86	PME20 (probable)
At3g10720	1.71	1.25	1.10	0.67	0.97	PME25
At3g43270	2.32	2.09	0.92	1.11	0.98	PME32 (probable)
At3g47400	0.78	1.05	0.91	0.63	0.93	PME33 (probable)
At4g02330	3.55	0.79	0.59	2.08	1.15	PME41 (probable)
At5g04970	0.63	0.80	1.03	0.58	0.82	PME47 (probable)
At5g19730	1.02	1.61	1.05	0.64	0.76	PME53 (probable)
At5g51500	0.62	1.07	0.94	0.55	0.94	PME60 (probable)
At5g55590	1.35	1.68	1.06	1.03	0.94	PME62 / QRT1
At5g06860	3.37	2.01	2.24	1.59	1.51	PGIP1(polygalacturonase inhibitor 1)
At5g14650	0.70	0.87	0.75	0.48	1.11	pectin lyase-like protein
At1g05650	0.35	0.60	0.87	0.52	0.75	pectin lyase-like protein
At4g22080	1.37	1.14	1.99	0.68	0.87	putative pectate lyase 17
At1g11920	1.02	1.47	1.96	0.94	0.76	putative pectate lyase 2
At3g17130	1.15	1.47	1.52	0.55	0.66	pectin methylesterase inhibitor
At5g62340	0.94	0.93	0.82	1.26	1.72	pectin methylesterase inhibitor
At3g47380	2.95	1.71	1.21	0.66	0.89	pectin methylesterase inhibitor
At1g23205	0.88	0.97	0.85	0.62	1.01	pectin methylesterase inhibitor
At3g09410	1.55	0.85	0.97	2.00	1.33	pectinacetylesterase family
At5g23870	0.65	0.93	0.98	1.05	1.13	pectinacetylesterase family
At2g23630	0.55	0.88	0.77	0.59	0.92	sks16 / pectinesterase
At4g01890	0.72	0.64	1.07	0.89	0.88	putative polygalacturonase
At2g43870	0.86	1.45	1.20	0.76	0.65	putative polygalacturonase
At2g43880	0.59	0.89	1.07	0.67	0.78	putative polygalacturonase
At2g43890	0.45	0.71	0.91	0.83	0.70	putative polygalacturonase
Cell wall relaxation						
At3g45970	1.16	1.41	0.73	0.62	0.94	ATEXLA1
At4g38400	0.52	0.87	0.92	0.59	0.79	ATEXLA2
At4g17030	1.21	1.63	0.94	1.23	1.44	ATEXLB1
At1g26770	0.91	1.27	1.11	0.72	0.60	ATEXPA10
At3g15370	2.15	2.05	1.08	0.80	0.92	ATEXPA12
At5g56320	0.58	0.80	1.04	0.66	0.86	ATEXPA14
At4g38210	1.51	1.64	1.06	0.94	0.85	ATEXPA20
At2g28950	1.21	1.28	0.95	1.51	1.26	ATEXPA6
At2g20750	0.44	0.50	0.88	0.63	0.85	ATEXPB1
Hemi-/Cellulose modification						
At5g57530	1.43	0.98	2.18	1.01	0.85	XTH12
At5g57540	1.55	1.05	2.70	0.96	0.68	XTH13
At3g23730	0.49	0.52	0.82	0.92	1.00	XTH16
At1g65310	0.47	0.58	1.04	0.95	0.98	XTH17
At2g18800	0.68	0.67	0.70	0.60	0.91	XTH21
At5g57560	3.81	3.61	1.71	1.56	0.97	XTH22
At4g28850	5.54	0.95	6.41	1.98	1.27	XTH26
At2g36870	1.61	1.74	1.34	0.93	0.69	XTH32
At4g37800	1.18	1.79	1.19	1.11	1.09	XTH7
At1g11545	0.58	0.70	1.02	0.74	0.76	XTH8
At4g03210	0.92	0.90	0.95	1.69	1.34	XTH9
At4g25810	1.67	1.20	1.87	0.70	0.61	XTH23
Carbohydrate hydrolization						
At3g10740	0.83	1.04	0.75	0.71	1.53	ALPHA-L-ARABINOFURANOSIDASE 1
At5g08380	1.32	1.57	0.84	1.29	1.39	ALPHA-GALACTOSIDASE 1
At2g32990	1.36	1.36	0.94	0.71	0.64	GLYCOSYL HYDROLASE 9B8
At5g61250	0.62	0.55	0.67	0.97	1.23	GLUCURONIDASE 1
At5g11920	1.40	1.45	0.82	1.26	1.56	6-&1-FRUCTAN EXOHYDROLASE
At3g52840	1.02	0.72	0.80	1.26	2.16	BETA-GALACTOSIDASE 2
At5g56870	0.61	0.60	0.73	1.31	1.66	BETA-GALACTOSIDASE 4
At1g02850	1.98	2.71	0.99	1.05	1.22	BETA GLUCOSIDASE 11
At3g03640	1.04	0.72	1.02	1.60	0.99	BETA GLUCOSIDASE 25
At1g47600	0.81	1.53	0.74	0.46	1.09	BETA GLUCOSIDASE 34
At1g26560	1.10	1.52	1.01	0.97	0.86	BETA GLUCOSIDASE 40
At3g18080	1.10	1.19	0.84	1.87	1.69	B-S GLUCOSIDASE 44
At1g61810	1.35	1.12	1.28	1.20	1.57	BETA-GLUCOSIDASE 45
At5g49360	0.65	0.77	0.91	0.95	1.22	BETA-XYLOSIDASE 1
At1g02640	1.00	0.69	0.81	1.02	2.17	BETA-XYLOSIDASE 2
At3g47040	0.97	0.75	0.78	1.05	1.55	glycosyl hydrolase family

Shown is the fold change expression of selected CW modifying enzymes in wild-type, *pdr2* and *lpr1lpr2* after transfer to –Pi or the fold change expression of Pi-replete *pdr2* and *lpr1lpr2* plants compared to the wild-type. Candidates were selected from a set of regulated genes annotated to be localized in the extracellular region (see also Additional File 6: Table S5). Red and green boxes denote significantly suppressed or induced ($p \leq 0.05$, student's *t*-test; $0.66 \geq FC \geq 1.5$) genes. PME, pectin methyl esterase; EXP, expansin; EXL, expansin-like; XTH, xyloglucan endotransglucosylase/hydrolyse

limitation in *lpr1lpr2* plants (Additional file 7: Table S6). Again, 19 CIII Prx genes were deregulated in *pdr2* under + Pi. Thus, peroxidases may be an important link between ROS formation and CW remodeling upon Pi starvation.

Proteomics supports regulation of Pi-responsive genes in *pdr2* and *lpr1lpr2* mutants

Genotype-specific changes in the root proteome upon Pi deficiency were monitored in an unlabeled approach using a fast scanning high resolution accurate mass (HRAM) LC-MS system. Three biological and three technical replicates were measured for each genotype under + Pi and –Pi conditions (54 samples) yielding 3,328,368 MS/MS spectra (individual peptide measurements). 726,944 spectra could be annotated to a peptide sequence (peptide spectral match, PSM) with a global false discovery rate (FDR) threshold of 0.01 %. These PSMs were used to identify 5110 protein groups (unique proteins), each with at least one unique peptide and a global FDR threshold of 1 % (Additional file 8: Table S7). Protein abundance was inferred based on peptide abundance determined by peptide ion signal peak integration using the PROGENESIS software. Pairwise comparison of all genotypes under both growth regimes revealed 2439 differentially regulated proteins ($p \leq 0.05$). Based on this list, we identified 1304 proteins that were either Pi-responsive in at least one genotype or which were already deregulated in one of the mutant lines grown on Pi-replete conditions ($0.769 \geq FC \geq 1.3$) (Additional file 9: Table S8).

Multidimensional scaling (MDS) analysis of ANOVA filtered ($p < 0.05$) samples revealed low variance between biological replicates but significant differences between genotypes and Pi conditions (Fig. 2a). The levels of 108 proteins were increased or decreased in the wild-type upon Pi depletion (Fig. 2b). As expected, the highest number of proteins (451) were regulated in hypersensitive *pdr2* mutant, probably reflecting changes in root morphology. We also identified a high number of Pi-responsive proteins (265) in the insensitive *lpr1lpr2* line. Of these, 214 proteins were unique to *lpr1lpr2* (Fig. 2c),

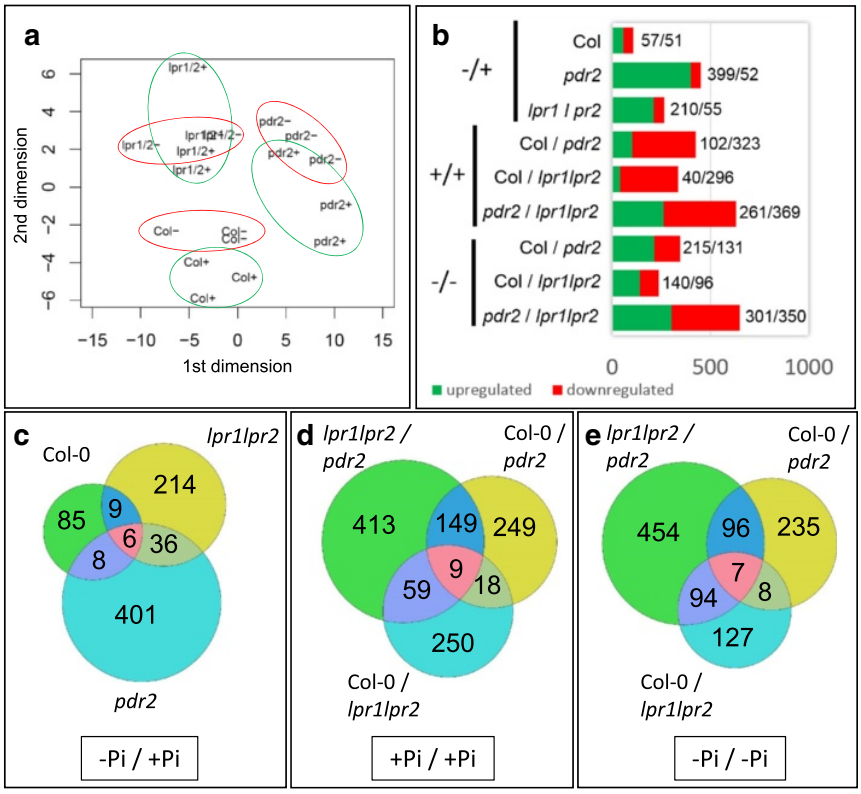


Fig. 2 Statistics of comparative protein expression analysis. **a** Multidimensional scaling (MDS) analysis of all biological replicate samples. **b** Illustration of the number of up- and downregulated proteins ($p \leq 0.05$, Student's *t*-test; $0.769 \geq FC \geq 1.3$) in wild-type, *pdr2* and *lpr1lpr2* roots upon transfer from + Pi to –Pi (–/+), and the number of differentially regulated proteins in pairwise comparisons of the three genotypes under + Pi (+/+) or –Pi (–/–) conditions. **c**, **d**, and **e** Venn diagrams illustrating the number of regulated proteins in pairwise comparisons for all three genotypes under both growth regimes (see also Additional file 8: Table S7, Additional file 9: Table S8, Additional file 10: Table S9)

indicating that the adjustment of protein expression might contribute to the decreased Pi responsiveness. Both mutant lines showed differential regulation of more than 300 proteins under Pi-replete conditions. This relatively high value is reminiscent of what we observed in the transcript dataset, supporting the assumption that PDR2 and LPR proteins may also regulate Pi independent processes.

Venn diagrams identified a group of 6 proteins that were similarly regulated in all lines upon Pi depletion (Fig. 2c, d, e). Notably, 4 of these proteins were positively correlated with our transcript data, showing induction on both mRNA and protein level (Table 4). Two members of this group were FER1 and the pectin modifying enzyme POLYGALACTURONASE INHIBITING PROTEIN1 (PGIP1) [43, 44], which further indicates that changes in Fe distribution and CW modification are associated with the response to low Pi.

Correlation of proteome and transcriptome analysis

Next, we performed GO term analysis to identify groups of proteins involved in genotype-specific Pi responsiveness. Most proteins have assigned metabolic functions in wild-type and *lpr1lpr2*, probably reflecting processes related to Pi recycling and mobilization. Strikingly, in + Pi condition and upon transfer to −Pi, the *pdr2* line showed a significant regulation of proteins assigned as *response to metal ion* (GO:0010038) and *oxidoreductase activity* (GO:0016491). A closer examination revealed repression of 15 peroxidases in *pdr2* in + Pi and induction of 9 peroxidases in −Pi condition. Within the group of repressed proteins we identified 14 CIII Prxs of which 3 enzymes were regulated at the transcript level. Only one and six Pi-responsive CIII Prx were detected in wild-type and *lpr1lpr2* root extracts, respectively (Additional file 10: Table S9).

To compare the proteome and transcriptome data sets, we plotted all significantly regulated proteins ($p \leq 0.05$, Student's *t*-test) against their cognate transcript. For those differentially expressed proteins, the percentage of detected transcripts was 91.6 % for wild-type (152/166), 94.3 % for *pdr2* (541/574) and 92.1 % for *lpr1lpr2* (351/381) roots. We observed only a low, but highly significant, positive correlation of transcript and protein abundance for all three genotypes ($R \geq 0.2$, $p \leq 0.001$) (Fig. 3a, b). We generated a list of significantly altered transcripts, which we compared to the list of significantly altered proteins ($p < 0.05$). We identified 26 cognate genes for wild-type, 22 for *lp1lpr2* and 211 for *pdr2*. The correlation coefficient markedly increased when we plotted these genes against their cognate proteins (Fig. 3b, c, d, e; Additional file 11: Table S10).

We identified the 4 genes, including *FER1* and *PGIP1*, that were co-regulated on mRNA and protein level across all genotypes in response to Pi depletion (Additional file 11: Table S10). In wild-type, we noticed induction of PPa4 (PYROPHOSPHORYLASE 4), a candidate for Pi mobilization, and PCK1 (PHOSPHOENOLPYRUVATE CARBOXYKINASE 1), which is involved in metabolic adjustment to Pi deprivation [45]. We further identified two hemicellulose modifying enzymes, XTH8 and XTH31 (XYLOGLUCAN ENDOTRANSGLUCOSYLASE/HYDROLASE), which were slightly decreased in low Pi. Interestingly, both enzymes were previously shown to be regulated by SIZ1 [46], a SUMO E3-ligase involved in Pi dependent root growth remodeling [47, 48].

GO term analysis of the 211 mRNA/protein pairs altered in *pdr2* revealed an overrepresentation of metabolic processes. The second most significant term (*response to metal ion*) is consistent with altered metal homeostasis in *pdr2* plants [19]. For example, we

Table 4 Pi-dependent Protein/mRNA regulation

	Locus	(-Pi/+Pi)			(+Pi/+Pi)		Name
		Col-0	*pdr2*	*lpr1 lpr2*	*pdr2*	*lpr1 lpr2*	
PO	AT5G01600	1.80	2.43	1.61	0.68	1.18	FER1 (FERRETIN 1)
TC		2.01	2.32	1.70	1.07	1.12	
PO	AT2G41380	2.30	1.81	1.91	0.99	0.81	S-adenosyl-L-methionine-dependent methyltransferases superfamily
TC		2.17	1.83	2.43	1.49	1.05	
PO	AT2G23540	1.39	1.36	1.34	1.14	0.70	GDSL-like Lipase
TC		1.46	1.66	1.05	0.89	0.92	
PO	AT2G31670	1.46	1.51	1.33	0.83	1.16	Stress responsive alpha-beta barrel domain protein
TC		0.99	0.93	0.98	1.01	0.97	
PO	AT5G06860	2.73	2.29	1.77	1.26	1.48	PGIP1
TC		3.37	2.01	2.24	1.59	1.51	
PO	AT2G01520	0.73	1.45	0.72	0.83	0.76	MLP-like protein 328
TC		1.20	1.07	0.88	1.27	1.57	

Shown is the fold change expression of the 6 proteins (PO) that are Pi-responsive in all lines (wild-type, *pdr2* and *lpr1lpr2*) or the fold change expression of Pi-replete *pdr2* and *lpr1lpr2* plants compared to the wild-type. Protein expression is compared to transcript changes (TC). Red and green boxes denote genes that are significantly suppressed or induced ($p \leq 0.05$, student's *t*-test; $0.76 \geq FC \geq 1.3$)

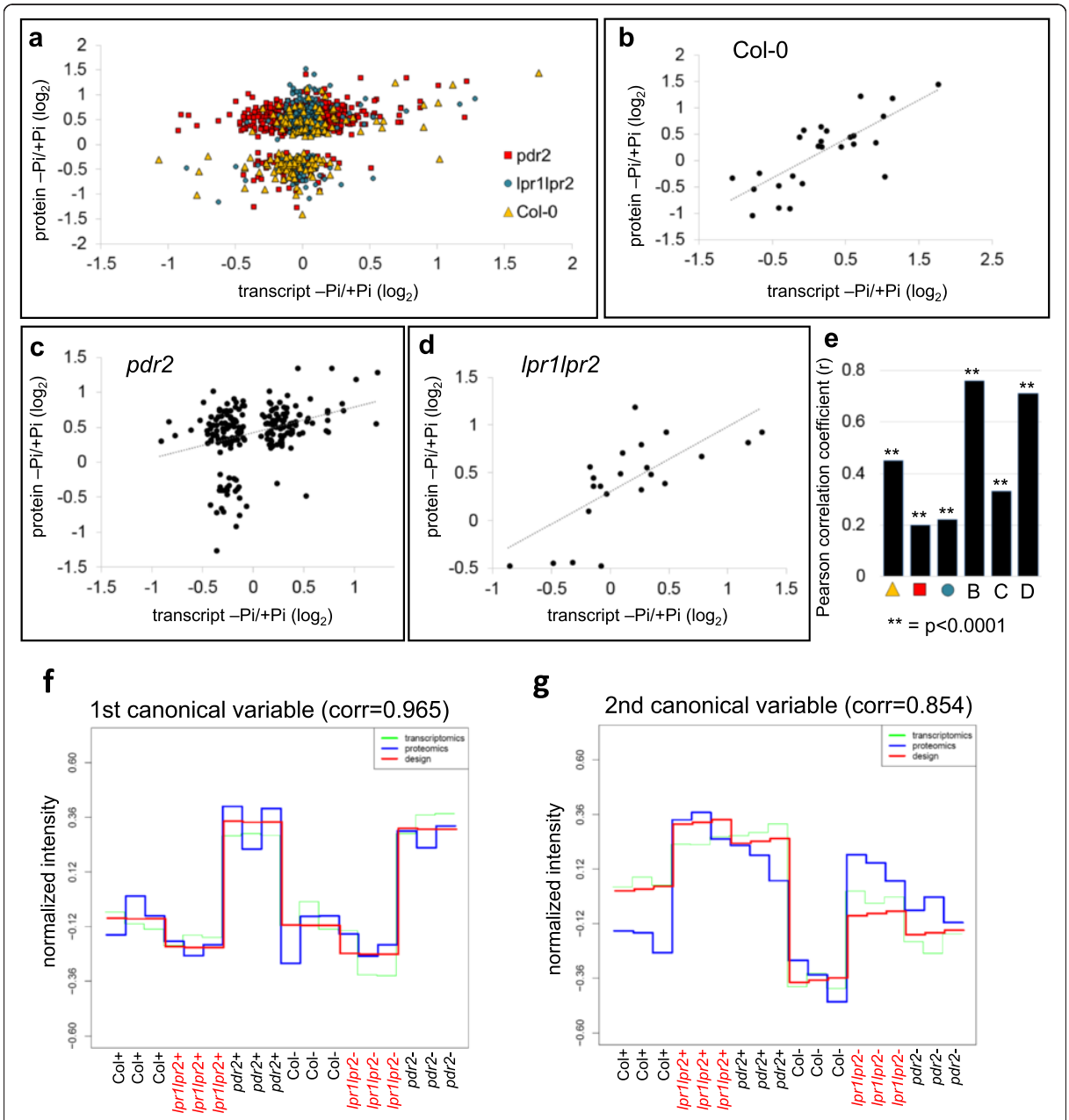


Fig. 3 Comparative analysis of transcriptome and proteome data and spCCA analysis. **a**, **b**, **c**, and **d** Correlation between transcript and protein fold-changes upon Pi-deficiency. **a** For each genotype, all significantly regulated proteins ($p \leq 0.05$) were plotted against its cognate transcript, if present on the ATH1 chip. **b**, **c**, and **d** Significantly ($p \leq 0.05$) regulated protein/mRNA pairs were plotted against each other. Scatter plots show 26 protein/mRNA pairs regulated in wild-type upon Pi-deficiency, 211 pairs identified for *pdr2*, and 21 pairs identified in *lpr1lrp2*. **e** Correlation (r) values for each pairwise comparison. Asterisks indicate significance for each correlation analysis ($p < 0.0001$). **f**, **g** Canonical variables of the spCCA analysis representing a subset of transcripts/proteins which showed maximum correlation with the illustrated patterns that were generated by the spCCA algorithm. (see also Additional file 1: Figure S1, Additional file 11: Table S10. Additional file 13: Table S11)

noticed induction of FER3 and proteins potentially involved in detoxification of metal ion-induced ROS formation, including several GLUTATHIONE-*S*-TRANSFERASEs (GSTs) (Additional file 11: Table S10). We also identified F6'H1 (feruloyl-CoA 6'-hydroxylase 1), which is involved in coumarin biosynthesis and Fe-mobilization in alkaline soils [49–51]. Our datasets revealed anti-correlation of F6'H1 expression in *pdr2*,

showing elevated protein but decreased transcript levels in −Pi and an inverse relation in + Pi (Additional file 1: Table S1, Additional file 9: Table S8), which indicates stringent regulation of F6'H1 expression in *pdr2*. In addition, protein level of CCoAOMT1 (caffeoyl coenzyme A O-methyltransferase 1), which converts caffeoyl-CoA to feruloyl-CoA, the substrate of F6'H1 [52], was also elevated in *pdr2* (Additional file 9: Table S8). Thus, coumarin-mediated mobilization of Fe may be involved in Pi dependent Fe accumulation.

Integrative spCCA analysis supports Pi-dependent metal redistribution

We integrated the two −omics approaches to uncover relationships that are supported by both individual datasets. We performed a supervised penalized canonical correlation analysis (spCCA), which searches for correlations between a set of transcripts and proteins [53]. The experimental design was integrated into the analysis to allow for biological interpretation of the derived canonical variables. The experimental factors (i.e., genotype, Pi condition, replicate sample) were provided as a binary matrix of design vectors that uniquely characterize each sample (Additional file 12: Figure S2). The supervised correlation approach seeks a linear combination of a feature subset from each -omics dataset that correlates maximally with a subset of experimental design factors. To maximize stringency, only varying transcripts and proteins were considered for spCCA. For transcriptomics, we choose a list of 1143 ANOVA filtered genes ($p \leq 0.05$, var ≥ 0.12) and for proteomics a list of 47 proteins ($p \leq 0.05$, var ≥ 0.4). Our analysis revealed distinct canonical variables (CVs), each representing a specific pattern correlating with a subset of proteins and/or transcripts. The first two CVs revealed structured patterns (Fig. 3f, g), while a third CV was disordered and therefore not further examined (Additional file 12: Figure S2B). The first CV mainly represented genes/transcripts (g/t) that were differentially expressed in *pdr2* compared to wild-type and *lpr1lpr2* independently of Pi status (Fig. 3f). We examined the top 100 g/t in this variable and found several Fe-related candidates (Additional file 13: Table S11), such as Fe chelate reductase 3 (FRO3) [54] and MYB10, which is required for growth in Fe deficiency [37]. MYB10 and MYB72 mediate Fe-dependent induction of NICOTIANAMINE SYNTHASE 4 (NAS4) [37], which is also present in this group. NAS proteins synthesize nicotianamine, a Fe-chelator essential for Fe-remobilization in the root [55]. We further identified a member of the ALUMINUM ACTIVATED MALATE TRANSPORTER (ALMT) family. It is of note that *ALMT1* is most highly induced in all three genotypes during Pi depletion (Table 1).

The second CV mainly represented g/t that were similarly expressed in Pi-replete *pdr2* and *lpr1lpr2* roots but slightly differed from the wild-type. In contrast to the first CV, the majority of these g/t were Pi responsive in all genotypes. As expected, we found several known Pi acquisition g/t, including SPX1, CAX3, the phosphate transporter PT2, and the Pi starvation-inducible inorganic pyrophosphatase 1 (Additional file 13: Table S11). Many other g/t are implicated in metal homeostasis, e.g., the Fe/Zn transporters IRT1 and IRT3, the Ni transporter IREG2, the Zn/Cd transporter HMA2 or the NA transporter YSL2, further supporting our observation that metal homeostasis is strictly controlled in all genotypes upon Pi starvation.

Root growth inhibition in low Pi is independent of general Fe uptake and cellular storage

We previously reported that LPR1-dependent Fe accumulation and distribution in root tips controls RAM activity in response to low Pi [19]. Our comparative transcriptomics and proteomics analysis of entire roots revealed Pi-responsive expression of Fe-related genes, notably *FER1* and *IRT1* (Table 1, Table 4), which correlated with Fe overload in Pi-starved roots of the three genotypes under study [19] (Additional file 14: Figure S3). To further investigate the role of Fe during the local response of roots to Pi availability, we analyzed the impact of *FER1* and *IRT1* loss-of-function mutants on Fe-distribution and root growth inhibition upon Pi deprivation.

Ferritins are located in plastids and can be visualized by Perls/DAB Fe staining as dot-like structures in root cells of wild-type plants, which are not detectable in *fer1-3-4* roots lacking *FER* expression [56]. Using semi-thin sections from Perls/DAB-stained wild-type roots, we observed similar dot-like structures in Pi-replete root tips, which strongly increased in number and staining intensity upon transfer to −Pi medium. These punctuate structures are associated with the symplast and are clearly distinctive from apoplastic Fe staining (Additional file 15: Figure S4A). We next performed root growth assays using the *fer1-3-4* triple and *fer1-2-3-4* quadruple mutant. Primary root growth rates of the *fer* mutants were indistinguishable from the wild-type on both + Pi or −Pi medium (Fig. 4a). Thus, ferritins do not affect the local root growth response to −Pi.

Similarly, we performed Perls/DAB Fe-staining to examine Fe distribution in wild type and *irt1* roots. Compared with Pi-replete wild-type seedlings, the *irt1* mutant showed more intense Fe staining on the root surface of the mature root zone (Additional file 15: Figure S4B), which is in accordance with impaired Fe uptake from the rhizosphere. However, both lines displayed similar Fe staining in the RAM and EZ, which is consistent with predominant *IRT1* expression in the

Hoehenwarter *et al. BMC Plant Biology* (2016) 16:106

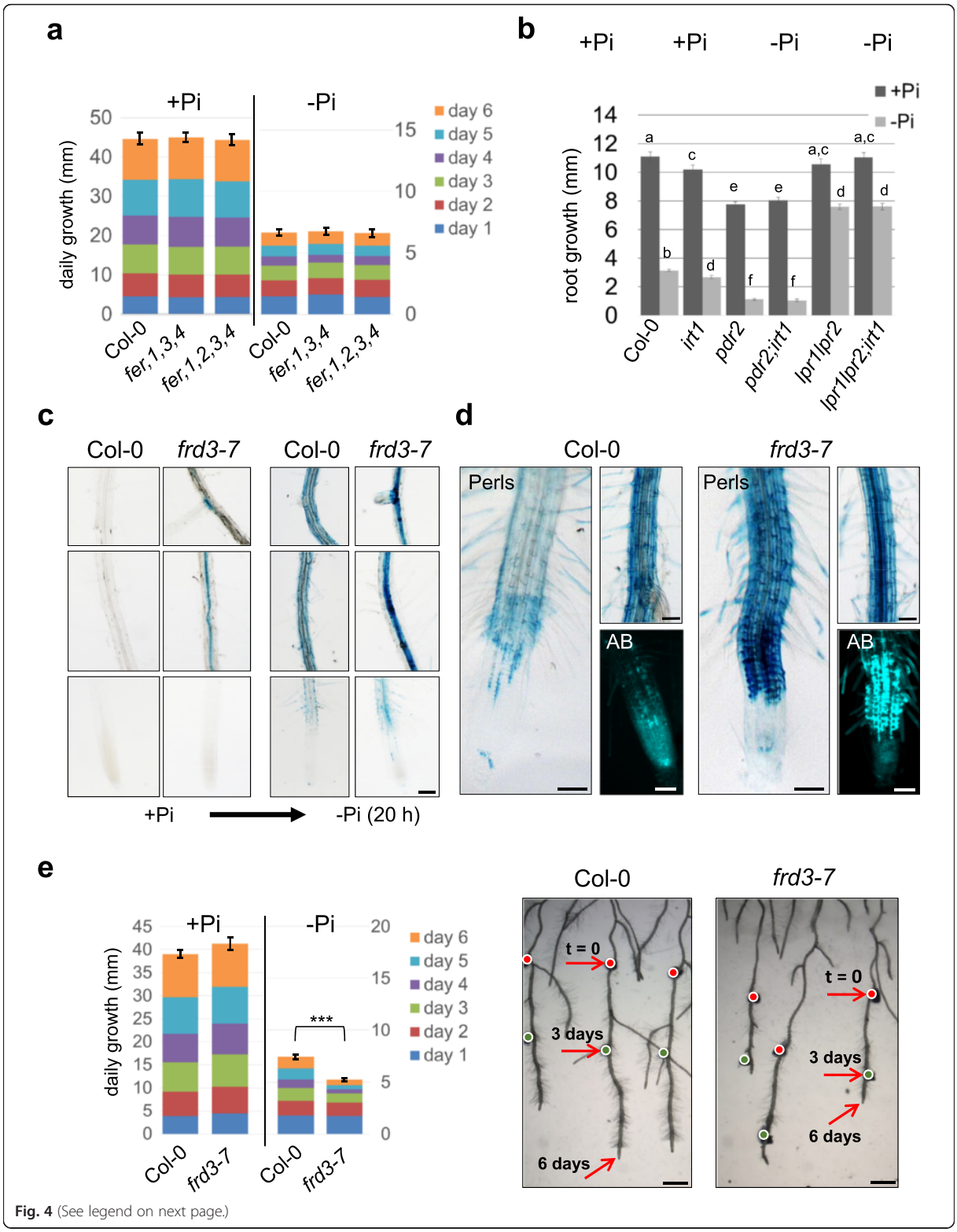


Fig. 4 (See legend on next page.)

(See figure on previous page.)
Fig. 4 Root growth in *fer* and *irt1* mutant plants and phenotypes of *frd3* roots. **a** 4-days-old seedlings were transferred from + Pi to + Pi or −Pi medium for up to 6 days. Daily increase in root growth was measured and illustrated in segmented boxes within the bar graph (±SE, $n \geq 15$). Standard error was calculated from the average total root growth. **b** Total increase in root length after transfer from + Pi to either + Pi or −Pi medium *t*-test; $p < 0.05$ (±SE, $n \geq 20$). **c**, **d**, and **e** Fe staining and root growth assays of wild-type and *frd3-7* seedlings. 4-days-old plants were transferred from + Pi to + Pi or −Pi medium for up to 6 days. **c** Perls staining in different root segments 20 h after transfer to + Pi or −Pi medium. Upper and middle panels show mature root segments. The lower panels show the RAM and early differentiation zone. Scale bar 200 μm. **d** Fe (Perls) and aniline blue (AB) callose staining of root tips and differentiated root segments 6 days after transfer to −Pi medium. Scale bar 100 μm. **e** Root growth of wild-type and *frd3-7* seedlings within 6 days after transfer to Pi-depleted medium. The bar graph shows the daily increase in root growth, illustrated in segmented boxes. Standard error was calculated from the average total root growth. *** *t*-test; $p = 1.85^{-8}$ (±SE, $n \geq 20$). Overview images show the root growth after 3 days and 6 days on −Pi medium . Arrows indicate the position of the root tip, directly after transfer to −Pi (t = 0), as well as 3 days and 6 days after transfer. Scale bar 1000 μm. (See also Additional file 14: Figure S3)

differentiation zone [32] and confirms our previous study [19]. Under Pi depletion, Fe staining increased strongly and comparably in all segments of wild-type and *irt1* roots, indicating that Fe accumulation and distribution in root tips is independent of IRT1. We generated homozygous *pdr2irt1* double and *lpr1lpr2irt1* triple mutants and monitored primary root growth on + Pi and −Pi agar. As expected, the *irt1* mutation did not affect the Pi-dependent root growth response of *pdr2* and *lpr1lpr2* plants (Fig. 4b), indicating IRT1-independent Fe accumulation in the root tip in response to low Pi.

Apoplastic Fe redistribution modifies Pi-dependent root growth adaptation

Long distance apoplastic Fe transport and distribution in symplastically disconnected tissues are mediated by the citrate exporter FERRIC REDICTASE DEFECTIVE 3 (FRD3) [57, 58]. Intriguingly, a previous study reported that *frd3* plants display a hypersensitive short-root phenotype when grown on −Pi medium [20]. To examine a potential role of FRD3 for mediating Pi-dependent Fe distribution via Fe-citrate complexes, we performed Perls Fe-staining (without DAB intensification to avoid oversaturation) on wild-type and *frd3* roots. As previously reported [58–60], Pi-replete *frd3* roots overaccumulated Fe in the vascular tissue (Fig. 4c). Within 20 h after transfer to −Pi, wild-type plants accumulated Fe in the outer cell layers, whereas *frd3* roots showed enhanced Fe staining in the vasculature, particularly in differentiated root segments. Importantly, only minor differences were noted in the root tip, where Fe accumulation was slightly increased in *frd3* (Fig. 4c); However, extended growth on −Pi (up to 6 days) progressively increased this difference, finally causing massive overaccumulation of Fe within the EZ and early differentiation zone of *frd3* roots (Fig. 4d).

We previously showed that Pi-dependent Fe accumulation correlates with callose formation at the sites of Fe deposition (<2 days) [19]. After transfer to −Pi (2 days), callose deposition at sites of Fe accumulation and resultant root growth inhibition were similar for wild-type and *frd3* plants (Additional file 16: Figure S5A). However, extended exposure (6 days) caused callose overproduction in *frd3* roots which correlated with an enhanced growth inhibition (Fig. 4d, e).

Based on our observations, we assumed that mobilization of apoplastic Fe-citrate complexes might be involved in the Pi dependent modulation of root growth. To test this, we transferred wild-type plants from + Pi conditions to + Pi or −Pi medium, supplemented with citrate, which was previously shown to restore Fe mobilization on *frd3* mutants [57] and monitored their growth behavior. Indeed, addition of 100–250 μM citrate promotes root growth within the first two days after transfer to −Pi. However, this effect was transient and external supply of citrate eventually suppressed root growth on low Pi (Additional file 16: Figure S5B, C).

Pi deprivation modifies pectins at Fe accumulation sites

Our comparative expression profiling pointed to a role for pectin-modifying enzymes. Therefore, we studied Pi-dependent changes in the pectineous CW by using Ruthenium Red (RR), an inorganic dye that stains unesterified pectins [61, 62].

Roots of wild-type, *pdr2* and *lpr1lpr2* showed a similar RR staining pattern on + Pi medium. One day after transfer to −Pi, we observed a strong increase in RR staining intensity in wild-type root tips, particularly within the differentiating EZ (Fig. 5a). Compared with wild-type, *pdr2* seedlings showed a more intense staining in this region while the RR staining in the *lpr1lpr2* mutant was unaltered. Interestingly, the site of enhanced pectin staining correlated well with the site of low Pi induced Fe deposition in wild-type and *pdr2* roots (Fig. 5a, c).

We also visualized the distribution of methyl-esterified pectin by using the hydroxylamine ferric chloride (HFC) reagent, which specifically reacts with methyl esters of pectin and results in a yellow to red coloration [62–64]. Only weak staining was evident in roots on + Pi (Fig. 5b) and transfer to −Pi did not significantly change the staining pattern in the differentiating EZ. However, higher magnification images revealed increased staining in the RAM of wild-type, with the highest intensity in the quiescent center (QC) and the cortical cell layer at

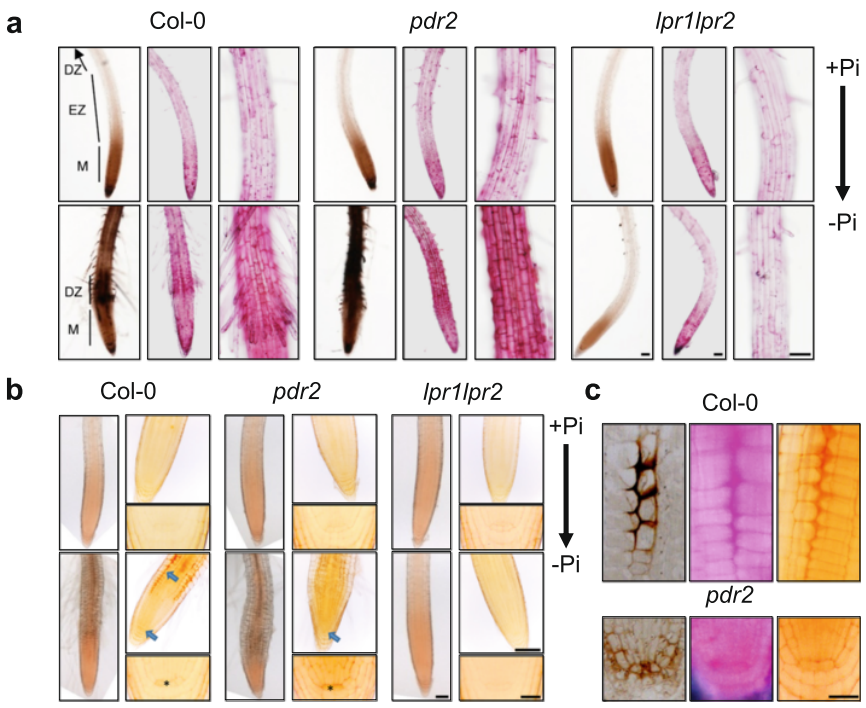


Fig. 5 Fe and pectin staining of roots after seedling transfer from + Pi agar to + Pi or –Pi medium for up to 2 days. **a** Perls/DAB Fe staining (*left panels*) and Ruthenium Red (RR) staining of non-methylesterified pectins (*center and right panels*). Shown are overview images of Perls/DAB and RR stained roots as well as detail images of the elongation/early differentiation zone. Scale bar 100 μm. **b** Hydroxylamine ferric chloride (HFC) staining of methylesterified pectins. Shown are overview images of the root (scale bar 100 μm) as well as detail images of the root meristem (scale bar 100 μm) and the stem cell niche (scale bar 20 μm). Differences in coloration between overview and detail result from the use of different imaging devices. *Arrows* point to the SCN and cortical cell layer in the transition zone, asterisks label the quiescent center cells in the SCN. **c** Details of the Perls/DAB, RR and HFC staining in Pi-depleted cortical cells of Col plants and stem cells of *pdr2* plants. Scale bar 20 μm

the transition zone, which demarcates the border between the RAM and EZ. In contrast, *pdr2* seedlings showed enhanced staining in the RAM, particularly within the QC region, but no distinct labeling of the cortical cell layer. No differences in pectin staining were detected in *lpr1lpr2* roots after transfer to –Pi medium. High magnification images of RR- and HFC-stained roots revealed simultaneous accumulation of acidic and methyl-esterified pectin in the meristem of the two sensitive lines. In particular, after transfer to –Pi, strong HFC and RR staining was evident in the cortex cell layer of wild-type and in the QC region of *pdr2* roots (Fig. 5c), which co-localized with major sites of Fe deposition.

Discussion

Plant adaptation to Pi limitation depends on coordinated transcriptional and translational regulation of gene expression [6, 21, 24–28]. While comparative transcriptome analysis proved to be a viable approach to distinguish between local and systemic regulation in Pi-starved plants [5, 6], only little information is available on the regulation of genes and proteins associated with the Pi-dependent adaption of root system architecture. Previous work revealed that *PDR2* and *LPR* genes act together in the local response to Pi availability by regulating cell type-specific deposition of Fe and callose in the root tip [11, 13, 19]. Here, we took advantage of the contrasting Pi-dependent root phenotype of *pdr2* and *lpr1lpr2* plants to investigate the associated changes in steady-state transcript and protein levels in a comparative approach. Genotype independent regulation of several *PSR* genes demonstrated the validity of our experiments and revealed that *pdr2* and *lpr1lpr2* mutants are likely not affected in the systemic response to Pi limitation (Table 1). Further analysis of our dataset revealed a number of candidate genes that are possibly involved in the Pi-dependent regulation of Fe storage and Fe redistribution as well as in the modulation of CW dynamics and/or ROS formation within the root.

Pi depletion modulates root Fe distribution

Our study revealed genotype-independent repression of numerous Fe-responsive and *IRT1*-coregulated genes upon transfer to Pi limitation, which likely reflects feedback regulation as a consequence of elevated Fe accumulation in Pi-starved differentiated roots. On the other

hand, de-repression of Fe-related genes in Pi-replete *pdr2* plants may sensitize Fe overaccumulation in limiting Pi [19] (Table 2, Additional file 12: Figure S2).

FER1 and related ferritins are plastid-localized Fe storage proteins protecting cells from Fe-mediated oxidative stress [65]. Using Perls/DAB Fe staining, Reyt et al. [56] recently reported dot-like structures in root cells of wild-type plants that likely display ferritin-bound Fe because they are absent in *fer1-3-4* roots [65]. A previous study showed that *FER1* is induced by PHR1 in low Pi independent of external Fe [66], indicating that FER1 may play a role in Pi-dependent Fe distribution.

Our comparative analysis revealed induction of FER1 expression on mRNA and protein level in all three lines (Table 4). Detection of Fe accumulation in dot-like structures supports the notion of intracellular Fe storage under Pi limitation, possibly as ferritin Fe (Figure S4). Importantly, Pi-dependent root growth was not affected by loss of ferritins (*fer1-[2]-3-4* mutants) or loss of *IRT1* in *pdr2* (*pdr2irt1)* and *lpr* (*lpr1lpr2irt1)* mutants (Fig. 4), indicating that Pi dependent root growth modulation is independent of intracellular Fe accumulation. Our data are consistent with a recent study reporting indistinguishable primary root growth of *fer1-3-4* and wild-type plants on high Fe [56].

Fe mobilization from the rhizosphere is facilitated by chelators such as carboxylates (e.g., citrate and malate) and coumarins, and apoplastic long distance Fe trafficking is mediated by Fe-citrate complexes [49–51, 67]. FRD3 exports citrate and the *frd3* mutant is defective in apoplastic Fe translocation, causing Fe hyperaccumulation in root stele tissues [58–60]. Importantly, *frd3* mutants show a hypersensitive short root phenotype in low Pi [20] and we demonstrated Fe overaccumulation in Pi-deprived *frd3* roots (Fig. 4), which indicates that citrate secretion is required for proper Fe-distribution under Pi limitation. Interestingly, citrate application transiently promoted *frd3* root growth in low Pi (Additional file 16: Figure S5), indicating that the Pi-dependent short root phenotype of *frd3* is likely a consequence of altered Fe redistribution in the growing root.

Transcript analysis and spCCA (Additional file 1: Table S1, Additional file 13: Table S11) revealed regulation of *ALMT* genes, including a strong Pi-dependent induction of *ALMT1* (Table 1), which was previously shown to exude malate into the rhizosphere as a strategy to cope with aluminum toxicity [68]. Earlier studies revealed PHR1-dependent accumulation of malate and citrate in Pi-depleted plants [24, 69]. Interestingly, exudation of both carboxylates into the rhizosphere was shown to facilitate mobilization of Pi and Fe in several plant species that do not form mycorrhiza [67].

We also noticed deregulation of coumarin biosynthesis-related genes, F6'H1 and CCoAOMT1, in *pdr2* roots (Additional file 1: Table S1, Additional file 9: Table S8). Several studies showed that coumarins (scopoletin and esculetin) are exuded into the rhizosphere to mobilize Fe in alkaline soils [49–51]. A recent report showed that esculetin accumulates in roots of Pi-starved wild-type plants but was suppressed in the *phr1* mutant, which lacks the induction of *PSR* genes upon Pi deficiency [69]. Moreover, using a non-targeted approach to identify metabolites from Pi-starved Arabidopsis root exudates, we recently confirmed Pi-dependent regulation of coumarin secretion [70]. Thus, our analysis implicates additional Fe-chelators in the regulation of Pi-dependent Fe accumulation and/or distribution in roots.

Pi depletion modulates root pectins

Inhibition of root cell elongation, formation of root hairs and induction of lateral roots are the most robust local responses to Pi deficiency [7, 8], which all require extensive reorganization of the CW. Our analysis revealed Pi-dependent regulation of CW-modifying enzymes, particularly in the sensitive wild-type and *pdr2* plants and to a lesser extent in *lpr1lpr2* roots (Table 3). Consistent with a previous transcriptome study [28], we identified several putative pectin esterases and esterase inhibitors. Pectins are secreted into the apoplast in a highly methylesterified state. In the CW, pectin methylesterases (PME) may remove methyl groups, generating free carboxylate functions on the surface of pectin polymers. Crosslinking of these carboxylate-groups by Ca^{2+} reduces CW extensibility and regulates cell expansion [41]. Our experiments revealed low Pi-induced accumulation of non-methylated pectin, specifically within the EZ of wild-type and *pdr2* roots (Fig. 5), which might contribute to rapid inhibition of cell elongation in these lines. In addition, there is growing evidence that plants exchange Ca^{2+} ions for other divalent and trivalent metal ions to prevent metal uptake and ROS formation [71]. Gessa et al. [72] showed in vitro Fe^{3+} binding to carboxylate groups on polygalacturonic acids (PGA), and two studies in *Arabidopsis* and rice demonstrated the ability of PGA to mobilize Pi from $FePO_4$ complexes and clay [73, 74]. Interestingly, a decrease in pectins in the *Arabidopsis qua1-2* mutant causes a hypersensitive short root phenotype upon Pi depletion [74]. Here, we show that accumulation of pectin in the root meristem coincides with the sites of Fe accumulation (Fig. 5a, c). Local pectin deposition might be a strategy to mobilize Pi from Fe-phosphate complexes. The data support our previous observations of CW thickening and callose deposition at sites of Fe accumulation in the root tip [19].

A recent study of the *Arabidopsis* flower transcriptome revealed deregulation of PGIP1 and other CW-modifying enzymes in the *ferritin1-3-4* triple mutant [75]. PGIP1 is a member of the leucine-rich repeat

(LRR) protein superfamily and inhibits fungal and bacterial polygalacturonases, which cleave non-methylated pectin residues in infected tissues [43]. It further regulates germination by inhibiting the breakdown of seed coat pectins [44]. Intriguingly, our analysis revealed co-regulation of PGIP1 and FER1 on transcript and protein level in all lines upon Pi-depletion (Table 4), further indicating a potential link between the Pi-dependent regulation of Fe distribution and the modification of pectin in the CW.

Peroxidases may modulate ROS formation and cell wall dynamics

We identified 41 CIII Prxs (56 % of the 73-member family) that were regulated on the mRNA and/or protein level, either in response to Pi depletion (23 members) or as a consequence of the *pdr2* and *lpr* mutations (Additional file 7: Table S6, Additional file 10: Table S9). Interestingly, the majority of CIII Prx mRNAs/proteins (30) were deregulated in *pdr2* in Pi replete conditions. CIII Prxs are involved in superoxide formation by transferring electrons from NADH to O_2 as well as in the Fe catalyzed generation of hydroxyl radicals [76, 77]. ROS formation is likely responsible for the cleavage of CW polysaccharides to promote cell expansion. On the other hand, oxidation of monolignols by CIII Prxs is the predominant mechanism of monolignol polymerization (lignification) which rigidifies the CW and degrades H_2O_2 [78]. The potential role of CIII Prxs for modulating ROS levels and CW dynamics and their strong deregulation in *pdr2* mutants points to a function in local root growth adaptation. A comprehensive analysis of available transcriptome and proteome data revealed that most CIII Prxs are mainly expressed in the root [42]. Two of those, Prx33 and Prx34, bind to Ca^{2+} polygalacturonates and mediate root growth in *Arabidopsis* [79]. A more recent study demonstrated that *prx33* and *prx34* knockdown lines exhibited reduced ROS and callose formation upon treatment with microbe-associated molecular patterns (MAMPs), implicating a direct role of these gene products in ROS formation [80]. Using specific ROS indicators, we recently demonstrated the formation of apoplastic ROS at the site of −Pi induced Fe deposition [19]. The underlying mechanism remains elusive but CIII Prxs may constitute a missing link between Pi dependent ROS formation and CW remodeling.

Comparative transcriptome and proteome analysis allows in-depth dissection of gene expression

Our comparative transcriptome and proteome analysis revealed a highly significant but relatively low positive correlation for the abundance of PSR proteins and their cognate transcripts in all three genotypes tested (Fig. 3a, e). The majority of mRNA/protein pairs in our dataset showed discordant changes, which has been previously observed and discussed in *Arabidopsis* and other organisms like mice and humans and which is likely explained by (post-) translational regulation and/or a temporal delay between the regulation of transcript and protein abundance. In addition, technical limitations in the efficiency of protein identification (e.g., low abundant proteins and transmembrane proteins) may restrict the detection of proteins relative to their cognate transcripts [25, 81–84].

Correlation values significantly increased when gene activity was subcategorized. For example, we observed a strong positive correlation between protein and mRNA abundance when we focused on proteins that were Pi-responsive in all genotypes (Table 4). Similarly, we found an enhanced positive correlation when we compared only significantly regulated genes with their cognate proteins (Fig. 3b, c, d, e). Moreover, our observations suggest that the integration of transcriptome and proteome datasets can be used as a valuable complementary approach. For example, we identified 28 and 18 CIII Prx, regulated on the transcript and/or protein level, respectively. Only 5 of those showed correlative expression changes in both datasets (Additional file 7; Table S6, Additional file 10: Table S9). However, the integration of both approaches revealed regulation of 41 CIII Prx, suggesting that the majority of CIII Prx are involved in the response to Pi deprivation.

We demonstrate that spCCA is a useful tool to integrate all experimental factors in our investigation, including the proteome and transcriptome data, Pi-status and genotype in order to elucidate unknown correlations in this multidimensional dataset. Interestingly, the first two CVs of our spCCA indicated a prominent role of genes and proteins that were differentially regulated in Pi-replete *pdr2* seedlings (Fig. 3f, g). Indeed, detailed analysis of our datasets revealed that the majority of Pi-responsive genes was not significantly deregulated in *pdr2*, compared to the wild type (Additional file 2: Figure S1C, Additional file 4: Table S3). On the other hand, several Fe-related genes, CIII Prx and pectin modifying enzymes were differentially regulated in Pi-replete *pdr2* plants (Table 2, Table 3, Additional file 7: Table S6), indicating that conditional hypersensitivity in *pdr2* might be a cause of constitutive de-repression or sensitization of these genes/proteins. P5-type ATPases are orphan, membrane localized ER proteins with unknown substrate specificity [85]. Mutant studies on yeast *SPF1* and *Arabidopsis MIA/PDR2* strongly suggest a function in ER quality control, protein folding and regulation of secretory processes [13, 86–88]. Hyperaccumulation of pectin and callose in the CW of Pi-depleted *pdr2* roots [19] (and this study) support a function of PDR2 in regulating ER-dependent secretion.

Conclusions

We performed complementary transcriptomics and proteomics approaches to monitor changes in steady-state transcript and protein levels upon Pi deprivation of *Arabidopsis* wild-type, *pdr2* and *lpr1lpr2* roots. Our analysis reveals a set of genes and proteins that are involved in the regulation of Fe homeostasis, cell wall remodeling and ROS formation. We observed increased *FER1* and decreased *IRT1* expression in all genotypes, which are consistent with intracellular Fe accumulation and feed-back inhibited Fe uptake in Pi-depleted roots, respectively. Analysis of *fer1-3-4*, *fer1-2-3-4* and *irt1* mutants demonstrates that cellular Fe uptake and Fe storage in ferritin are not involved in Pi-dependent modulation of root growth. We provide evidence for the importance of apoplastic Fe redistribution to maintain root growth upon Pi-depletion and for a role of FRD3 in this process. Our data further reveal Pi-dependent regulation of cell wall-modifying enzyme expression and changes in the deposition of pectins in Pi-deprived roots. The high correlation between sites of Fe deposition and enhanced pectin accumulation suggests that pectins might be involved in Fe binding and/or Pi mobilization from Fe-P complexes.

Methods

Plant material and growth conditions

Arabidopsis thaliana accession Columbia (Col-0) and Col lines *pdr2-1*, *lpr1-1lpr2-1, irt1-1, frd3-7, fer1-3-4* and *fer1-2-3-4* were previously described [11, 13, 58, 89, 90]. The *pdr2-1* mutant was identified and characterized by our group [12, 13, 19]. The *irt1-1* (SALK_024525) and *frd3-7* (SALK_122235) lines were obtained from the European Arabidopsis Stock Center (NASC). The *lpr1-1lpr2-1* double mutant and the ferritin mutants (*fer1-(2)-3-4*) were kindly provided by T. Desnos [11] and J.F. Briat [90], respectively. Seeds were surface-sterilized and germinated on 0.8 % (w/v) Phyto-Agar (Duchefa) containing 50 μM Fe-EDTA and 2.5 mM KH_2PO_4, pH 5.6 (high or + Pi medium) or no Pi supplement (low or −Pi medium) as reported [13, 19].

Root growth measurement

The position of the root tip was marked on the back of the agar plate directly after seedling transfer from + Pi to + Pi or −Pi medium. Images were taken on a stereomicroscope and total increment of primary root length was calculated at the according time point using ImageJ software. For daily growth rate measurements, the root tip position was marked every 24 h. The distance between two marker-points defines the daily root growth.

Histochemical staining

Accumulation and distribution of Fe and callose in roots was monitored as previously described [19]. De-methyl esterified pectins were stained for 5–10 min in 0.05 % (w/v) Ruthenium Red solution (Applichem). Hydroxylamine-ferric chloride staining was adapted from Hornatowska and Reeve [63, 64]. Seedlings were initially incubated for 5–10 min in freshly prepared hydroxylamine solution (0.7 % NaOH, 0.7 % hydroxylamine hydrochloride in 60 % EtOH), followed by the addition of an equal (or higher) volume of a solution containing concentrated HCl/EtOH 95 % (1:2 ratio). The solution was removed and ferric chloride was added (10 % $FeCl_3$ in 60 % EtOH containing 0.1 N HCl). Seedlings were cleared using chloral hydrate solution (7:7:1 chloral hydrate:ddH_2O:glycerol). Samples were analyzed using a multizoom stereomicroscope (Nikon AZ100) for overview images and a Zeiss AxioImager bright field microscope for detail images.

RNA preparation and microarray hybridization

Seedlings (4-days-old) were transferred from + Pi to either + Pi or −Pi medium and roots were harvested after 20 h. RNA was extracted using the RNeasy Plant Mini Kit from Qiagen followed by an on-column DNA digestion (40 min) using Qiagen RNase-free DNase Set. Quality control and hybridization to ATH1 *Arabidopsis* GeneChips was done by NASC's Affymetrix Service (http://affymetrix.arabidopsis.info/).

Statistical analysis of mRNA expression data

Data preprocessing, generation of Venn diagrams and heat maps was performed using Arraystar 4.1 software (DNASTAR). Arrays were normalized with robust multi-array analysis (RMA) and quantile background correction. Pairwise comparisons were performed using a fold-change cutoff value of ≥ 1.5 for increased and of ≤0.66 for decreased transcript levels ($p \leq 0.05$; Student's *t*-test, no multiple testing correction). Gene ontology analysis was done with the preassigned settings of the Arraystar software using a cutoff value $p \leq 0.05$ and FDR (Benjamini Hochberg) correction. Hierarchical clustering was performed with 4870 ANOVA-filtered genes using the *hclust* package of the R software v.3.0.0. [91]. The ATTED-II database (http://atted.jp) was used to generate a list of *IRT1* co-regulated genes based on ATTED's mutual ranking. All other calculations and graphics were prepared using Microsoft Excel 2010 software.

Preparation of protein samples and LC-MS analysis

Plants were grown as for mRNA analysis. Proteins were extracted from root tissue and digested with trypsin. Peptides were injected into an EASY-nLC II nano liquid chromatography system, equipped with a Nanospray Flex ion source (Thermo Fisher Scientific) and electrosprayed into an Orbitrap Velos Pro mass spectrometer (Thermo Fisher Scientific). Details are described in Additional file 17.

Protein identification and relative quantification

The raw data was imported into Proteome Discoverer v.1.4 (PD). Peak lists generated with a precursor signal to noise ratio of 1.5 with PD were used to search the TAIR10 database amended with common contaminants (35,394 sequences, 14,486,974 residues) with the Mascot algorithm v.2.5 on an in-house Mascot server. The enzyme specificity was set to trypsin and two missed cleavages were tolerated. Carbamidomethylation of cysteine was set as a fixed modification and oxidation of methionine as a variable modification. The precursor tolerance was set to 7 ppm and the product ion mass tolerance was set to 0.8 Da. A decoy database search was performed to determine the peptide false discovery rate (FDR). The search results were imported into the Scaffold Q+ software v.4.1.1 (Proteome Software, Inc.). Peptide and protein FDRs were calculated and the identity thresholds set to 0.01 and 1 % respectively to control the family wise error rate of peptide and protein identifications.

The raw data was imported into Progenesis LC-MS v.4.1 (Nonlinear Dynamics) for relative protein quantification between LC-MS analyses. The peptide ion signal peak landscapes of LC-MS analyses were aligned using the analysis as a reference that gave the highest minimum and maximum number of vectors in the aligned set of analyses when each analysis was used as a reference. Ratiometric normalization in log space to a selected reference analysis over all aligned peptide ion signals was performed. The summed intensities of peptide ion signal peak isotope envelopes over time were used as a measure of peptide abundance. A coefficient of variance (CV) of peptide abundance of less than 50 % for a peptide in all LC-MS analyses of a biological condition (three replicate analyses of each of three biological replicates for a total of 9) was required for a peptide to be quantified. Protein abundance was inferred by the sum of all unique peptides mapping to a given protein (non-conflicting peptides). Protein abundance fold changes and corresponding p-values between the biological conditions were calculated.

Multidimensional scaling (MDS) analysis

Multidimensional scaling was conducted using the isoMDS function from the MASS package version 7.3-29 [92]. Technical replicates of the proteome analysis were averaged, reducing the original dataset to 18 biological replicate samples. Missing values were either imputed by half of the minimum intensity or excluded from further analysis. The resulting matrix of 3849x18 proteins was subjected to ANOVA ($p < 0.05$) revealing 412 consistent proteins. Intensities were log-transformed.

Supervised penalized canonical correlation analysis (spCCA)

SpCCA analysis was done according to [53]. ANOVA filtered transcriptome and proteome data sets were reduced to signals with a variance of ≥0.12 and ≥0.4 resulting in 1143 transcripts and 47 proteins. The experimental design consisted of a binaric matrix of 18 samples x 8 experimental factors (three genotypes: Col, *pdr2*, *lpr1lpr2*; two growth media: +Pi, −Pi agar; and three replicates). SpCCA was conducted with 25 resampling runs (n.r = 25) and 25 random start vectors (max.counter.test = 25) to optimize sparsity parameters in a grid search between (0,0,0) and (0.6,0.5,1) with small step sizes (0.05,0.05,0.1) for transcriptomics, proteomics and design dataset.

Ethics (and consent to participate)

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Microarray data sets with the reference number NASCARRAYS-648 were deposited on the NASCArrays database (http://affymetrix.arabidopsis.info/). The proteomics data have been deposited to the ProteomeXchange Consortium [93] via the PRIDE partner repository with the dataset identifier PXD003449 and 10.6019/PXD003449 (http://www.ebi.ac.uk/pride/archive/).

Additional files

Additional file 1: Table S1. ATH1 dataset. Shown is the relative average expression value of all probe sets (B-G) and the linear fold change of all pairwise comparions (L-AC). (XLSX 9833 kb)

Additional file 2: Figure S1. Correlation and GO term analysis. (A) Heat map of a hierarchical cluster analysis of the group of 48 transcripts altered in all three genotypes upon Pi-depletion. Relative expression values are shown. (B) Scatter plots presenting pairwise correlation analysis (log_2 fold changes) of the 48 commonly regulated genes upon Pi-depletion. FC, fold change. (see also Additional file 1: Table S1). (C) Correlation analysis of a subset of 241 Pi-responsive genes that were differentially regulated in wild-type and *pdr2* but not in *lpr1lpr2* roots ($p \leq 0.05$, Student's *t*-test; $0.66 \geq FC \geq 1.5$). The upper image shows log_2 fold changes of all genes upon Pi-starvation. The lower heat map illustrates the same gene set and expressional changes using a color code. (D) GO term analysis of a subset of 1680 genes that were either Pi-responsive in wild-type, *pdr2* and/or *lpr1lpr2* roots or that were differentially regulated in Pi-replete *pdr2* and/or *lpr1lpr2* roots ($p \leq 0.05$, Student's *t*-test; $0.66 \geq FC \geq 1.5$). Each segment in a wheel represent one GO term. The top five GO terms are listed and significance values are shown. The complete list of genes and GO terms is shown in Additional file 4: Table S3. (see also Additional file 4: Table S3, Additional file 5: Table S4, Additional file 6: Table S5, Additional file 7: Table S6). (PDF 246 kb)

Additional file 3: Table S2. Shown is a list of references that described Pi- or Fe-responsiveness of the genes listed in Table 1. (XLSX 14 kb)

Additional file 4: Table S3. Pi-responsive genes exclusively regulated in wild-type and *pdr2* roots. Shown is the linear fold change of all 241 genes that showed Pi-dependent expressional changes ($p \leq 0.05$, Student's *t*-test; $0.66 \geq$

FC ≥ 1.5) in wild-type and *pdr2* only, but not in *lpr1lpr2* roots. Highlighted are genes that were differentially expressed in *pdr2* (at least 2-fold higher or lower) compared to the wild-type. (XLSX 82 kb)

Additional file 5: Table S4. Pi-responsive and deregulated genes in *pdr2* or *lpr1lpr2* roots and GO term analysis. Shown is a list of 1680 genes that were either regulated in one of the tested lines under Pi-depletion or differentially regulated in *pdr2* or *lpr1lpr2* in Pi-replete conditions, compared to the wild-type ($p \leq 0.05$, Student's *t*-test; $0.66 \geq FC \geq 1.5$). Additional tabs show results from Gene Ontology analysis using the list of 1680 genes. BP, biological processes; MF, molecular function; CC, cellular compartment. (XLSX 1048 kb)

Additional file 6: Table S5. Regulated genes of the GO term "extracellular region". Listed are 322 genes whose encoded proteins are annotated to be located in the extracellular region (GO: 0005576) and which were either regulated in one of the tested lines under Pi-depletion or which were differentially regulated in *pdr2* or *lpr1lpr2* in Pi-replete conditions. This table is based on the list of 1680 genes (see Additional file 4: Table S3). (XLSX 98 kb)

Additional file 7: Table S6. Regulation of extracellular peroxidases. Listed are 29 peroxidases that are annotated to be located in the extracellular region and found to be regulated either in one of the tested lines under Pi-depletion or which were differentially regulated in *pdr2* or *lpr1lpr2* in Pi-replete conditions. Green and red fields depict significantly induced or repressed genes, respectively ($p \leq 0.05$, Student's *t*-test; $0.66 \geq FC \geq 1.5$). This table is based on the list of 322 regulated genes of the GO term "extracellular region" (see Additional file 6: Table S5). (XLSX 19 kb)

Additional file 8: Table S7. Proteome data. Scaffold v4.4.1 was used to aggregate and visualize protein identifications from the Mascot search engine (v2.5.) run via Proteome Discoverer (V1.4) with X!Tandem searches integrated into Scaffold. LFDR scoring and protein cluster analysis for protein grouping were used to identify proteins. Total spectra (#PSMs) per protein normalized to the total spectra of all proteins recorded for each biological condition are shown. (XLSX 643 kb)

Additional file 9: Table S8. Differentially regulated proteins. Listed are 1304 proteins that were either Pi-responsive in at least one genotype (Col-0, *pdr2* and/or *lpr1lpr2*) or which were already deregulated in one of the mutant lines grown on Pi-replete conditions ($p \leq 0.05$, $0.769 \geq FC \geq 1.3$). Green and red boxes represent proteins that were significantly induced or repressed, respectively. Blue boxes represent proteins that were significantly regulated ($p \leq 0.05$) but did not reach the preassigned cut-off fold change value. (XLSX 230 kb)

Additional file 10: Table S9. Regulation of peroxidases. (A) Listed are 23 peroxidases that were either Pi-responsive in at least one genotype (wild-type, *pdr2* and/or *lpr1lpr2*) or were already deregulated in one of the mutant lines grown on Pi-replete conditions ($p \leq 0.05$, $0.769 \geq FC \geq 1.3$). (B) Listed are 5 peroxidases that were regulated on transcript and protein level in at least one pairwise comparison. TC, transcript; PO, protein. Green and red boxes represent proteins which were significantly induced or repressed, respectively. (XLSX 60 kb)

Additional file 11: Table S10. Regulation of mRNA/protein pairs. Listed are mRNA/protein pairs that showed correlative expression upon Pi-deficiency. Shown is a list of 26 pairs for wild-type, 211 pairs for *pdr2* and 22 pairs for *lpr1lpr2*. Green and red boxes represent proteins that were significantly induced or repressed, respectively ($p \leq 0.05$, $0.769 \geq FC \geq 1.3$). Blue boxes represent proteins which were significantly regulated ($p \leq 0.05$) but did not reach the preassigned cut-off fold change value. (XLSX 73 kb)

Additional file 12: Figure S2. Fe staining and root growth assay. Perls/DAB Fe staining on 4-days-old seedlings that were transferred from + Pi to + Pi or –Pi medium for 20 h. Upper panels show mature root segments of wild-type, *pdr2* and *lpr1lpr2* seedlings, lower panels depict the root meristem and EZ, which shows early differentiation of root hairs under-Pi. Scale bar, 200 μm. (PDF 45 kb)

Additional file 13: Table S11. Protein/transcript list of spCCA analysis. Shown is the list of mRNAs/proteins that are highly relevant (high weight) within the three canonical variables found in the spCCA analysis. Values in tables illustrate the relative weight of each mRNA/protein. Negative values indicate that these mRNAs/proteins are anti-correlated to the pattern of the respective canonical variable as shown in Fig. 3 and Additional file 14: Figure S3. (XLSX 26 kb)

Additional file 14: Figure S3. spCCA analysis. (A) Shown are the experimental design factors used for the supervised correlation analysis. (B, C, and D) Canonical variables (CV) of the spCCA analysis representing a subset of transcripts/proteins that showed maximum correlation with the illustrated patterns generated by the spCCA algorithm. CVs in B and C are also shown in Fig. 4 (see also Additional file 13: Table S11). (PDF 961 kb)

Additional file 15: Figure S4. Fe distribution in *fer* and *irt1* mutant plants. (A) Semi-thin (1 μm) longitudinal sections of Perls/DAB stained root tips of wild-type seedlings after transfer from + Pi to + Pi or –Pi (20 h). Shown are overview (scale bar 100 μm) and detail (scale bar 25 μm) images of the root tip. Arrows indicate punctate Fe storages. (B) Perls/DAB Fe staining of wild-type and *irt1* seedlings. Upper and middle panels show mature and young differentiated root segments, respectively. Lower panels show the root meristem. Scale bar 100 μm. (PDF 1598 kb)

Additional file 16: Figure S5. Aniline blue staining on *frd3* roots and citrate application. (A) 4-days-old wild-type and *frd3-7* seedlings were transferred from + Pi to + Pi or –Pi medium for 2 days. Left: Aniline blue (callose) staining. Right: photographs. Scale bar, 200 μm. (B, and C) 4-days-old wild-type seedlings were transferred from + Pi to + Pi or –Pi medium supplemented with increasing concentrations of citrate. (B) Daily increase in primary root growth was measured over 3 days and illustrated in segmented boxes within the bar graph. (±SE, $n \geq 15$). Standard error was calculated from the average total root growth within 3 days. (C) Photograph of wild-type plants that were transferred for 5 days to –Pi medium, supplemented with different citrate concentrations. Each colored spot indicates the position of the root tip after the indicated time point. Scale bar, 1000 μm. (PDF 1948 kb)

Additional file 17: Detailed description of protein extraction and LC-MS analysis. (PDF 76 kb)

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
W.H. and P.M. performed LC-MS analysis, protein identification, quantification and statistical analysis. S.M. and S.N. performed spCCA and cluster analysis. S.M. provided support for all statistical analysis. S.A. designed the study and co-wrote the manuscript. J.M. designed the study, performed mRNA data processing and all principle data analysis, conducted all biological experiments, and wrote the manuscript. All authors have read and approved the final version of the manuscript.

Acknowledgements
We thank J.F. Briat for *ferritin* mutant seeds. Research at the Leibniz Institute of Plant Biochemistry was supported by institutional core funding provided by the state of Saxony-Anhalt and the Federal Republic of Germany. The publication of this article was funded by the Open Access fund of the Leibniz Association.

Author details
[1]Proteome Analytics Research Group, Leibniz Institute of Plant Biochemistry, D-06120 Halle (Saale), Germany. [2]Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, D-06120 Halle (Saale), Germany. [3]Department of Molecular Signal Processing, Leibniz Institute of Plant Biochemistry, D-06120 Halle (Saale), Germany. [4]Institute of Biochemistry and Biotechnology, Martin Luther University Halle-Wittenberg, D-06120 Halle (Saale), Germany. [5]Department of Plant Sciences, University of California-Davis, Davis, CA 95616, USA.

Received: 19 January 2016 Accepted: 18 April 2016
Published online: 28 April 2016

References
1. Shen J, Yuan L, Zhang J, Li H, Bai Z, Chen X, Zhang W, Zhang F. Phosphorus dynamics: from soil to plant. Plant Physiol. 2011;156(3):997–1005.

2. Lin WY, Huang TK, Leong SJ, Chiou TJ. Long-distance call from phosphate: systemic regulation of phosphate starvation responses. J Exp Bot. 2014;65(7): 1817–27.
3. Rubio V, Linhares F, Solano R, Martin AC, Iglesias J, Leyva A, Paz-Ares J. A conserved MYB transcription factor involved in phosphate starvation signaling both in vascular plants and in unicellular algae. Genes Dev. 2001; 15(16):2122–33.
4. Bari R, Datt Pant B, Stitt M, Scheible WR. PHO2, microRNA399, and PHR1 define a phosphate-signaling pathway in plants. Plant Physiol. 2006;141(3): 988–99.
5. Bustos R, Castrillo G, Linhares F, Puga MI, Rubio V, Perez-Perez J, Solano R, Leyva A, Paz-Ares J. A central regulatory system largely controls transcriptional activation and repression responses to phosphate starvation in Arabidopsis. PLoS Genet. 2010;6:9.
6. Thibaud MC, Arrighi JF, Bayle V, Chiarenza S, Creff A, Bustos R, Paz-Ares J, Poirier Y, Nussaume L. Dissection of local and systemic transcriptional responses to phosphate starvation in Arabidopsis. Plant J. 2010;64(5):775–89.
7. Peret B, Clement M, Nussaume L, Desnos T. Root developmental adaptation to phosphate starvation: better safe than sorry. Trends Plant Sci. 2011;16(8):442–50.
8. Abel S. Phosphate sensing in root development. Curr Opin Plant Biol. 2011; 14(3):303–9.
9. Hinsinger P, Herrmann L, Lesueur D, Robin A, Trap J, Waithaisong K, Plassard C. Impact of roots, microorganisms and microfauna on the fate of soil phosphorus in the rhizosphere. In: Annual Plant Reviews. Volume 48. John Chichester, UK: Wiley & Sons, Inc.; 2015: 375–407.
10. Reymond M, Svistoonoff S, Loudet O, Nussaume L, Desnos T. Identification of QTL controlling root growth response to phosphate starvation in Arabidopsis thaliana. Plant Cell Environ. 2006;29(1):115–25.
11. Svistoonoff S, Creff A, Reymond M, Sigoillot-Claude C, Ricaud L, Blanchet A, Nussaume L, Desnos T. Root tip contact with low-phosphate media reprograms plant root architecture. Nat Genet. 2007;39(6):792–6.
12. Ticconi CA, Delatorre CA, Lahner B, Salt DE, Abel S. Arabidopsis pdr2 reveals a phosphate-sensitive checkpoint in root development. Plant J. 2004;37(6): 801–14.
13. Ticconi CA, Lucero RD, Sakhonwasee S, Adamson AW, Creff A, Nussaume L, Desnos T, Abel S. ER-resident proteins PDR2 and LPR1 mediate the developmental response of root meristems to phosphate availability. Proc Natl Acad Sci U S A. 2009;106(33):14174–9.
14. Sanchez-Calderon L, Lopez-Bucio J, Chacon-Lopez A, Gutierrez-Ortega A, Hernandez-Abreu E, Herrera-Estrella L. Characterization of low phosphorus insensitive mutants reveals a crosstalk between low phosphorus-induced determinate root development and the activation of genes involved in the adaptation of Arabidopsis to phosphorus deficiency. Plant Physiol. 2006; 140(3):879–89.
15. Camacho-Cristobal JJ, Rexach J, Conejero G, Al-Ghazi Y, Nacry P, Doumas P. PRD, an Arabidopsis AINTEGUMENTA-like gene, is involved in root architectural changes in response to phosphate starvation. Planta. 2008; 228(3):511–22.
16. Yu H, Luo N, Sun L, Liu D. HPS4/SABRE regulates plant responses to phosphate starvation through antagonistic interaction with ethylene signalling. J Exp Bot. 2012;63(12):4527–38.
17. Gonzalez-Mendoza V, Zurita-Silva A, Sanchez-Calderon L, Sanchez-Sandoval ME, Oropeza-Aburto A, Gutierrez-Alanis D, Alatorre-Cobos F, Herrera-Estrella L. APSR1, a novel gene required for meristem maintenance, is negatively regulated by low phosphate availability. Plant Sci. 2013;205–206:2–12.
18. Karthikeyan AS, Jain A, Nagarajan VK, Sinilal B, Sahi SV, Raghothama KG. Arabidopsis thaliana mutant lpsi reveals impairment in the root responses to local phosphate availability. Plant Physiol Biochem. 2014;77:60–72.
19. Müller J, Toev T, Heisters M, Teller J, Moore KL, Hause G, Dinesh DC, Bürstenbinder K, Abel S. Iron-dependent callose deposition adjusts root meristem maintenance to phosphate availability. Dev Cell. 2015;33(2):216–30.
20. Ward JT, Lahner B, Yakubova E, Salt DE, Raghothama KG. The effect of iron on the primary root elongation of Arabidopsis during phosphate deficiency. Plant Physiol. 2008;147(3):1181–91.
21. Misson J, Raghothama KG, Jain A, Jouhet J, Block MA, Bligny R, Ortet P, Creff A, Somerville S, Rolland N et al. A genome-wide transcriptional analysis using Arabidopsis thaliana Affymetrix gene chips determined plant responses to phosphate deprivation. Proc Natl Acad Sci U S A. 2005;102(33):11934–9.
22. Hirsch J, Marin E, Floriani M, Chiarenza S, Richaud P, Nussaume L, Thibaud MC. Phosphate deficiency promotes modification of iron distribution in Arabidopsis plants. Biochimie. 2006;88(11):1767–71.
23. Zheng L, Huang F, Narsai R, Wu J, Giraud E, He F, Cheng L, Wang F, Wu P, Whelan J et al. Physiological and transcriptome analysis of iron and phosphorus interaction in rice seedlings. Plant Physiol. 2009;151(1):262–74.
24. Morcuende R, Bari R, Gibon Y, Zheng W, Pant BD, Blasing O, Usadel B, Czechowski T, Udvardi MK, Stitt M et al. Genome-wide reprogramming of metabolism and regulatory networks of Arabidopsis in response to phosphorus. Plant Cell Environ. 2007;30(1):85–112.
25. Lan P, Li W, Schmidt W. Complementary proteome and transcriptome profiling in phosphate-deficient arabidopsis roots reveals multiple levels of gene regulation. Mol Cell Proteomics. 2012;11(11):1156–66.
26. Woo J, Macpherson CR, Liu J, Wang H, Kiba T, Hannah MA, Wang XJ, Bajic VB, Chua NH. The response and recovery of the Arabidopsis thaliana transcriptome to phosphate starvation. BMC Plant Biol. 2012;12:62.
27. Wang J, Lan P, Gao H, Zheng L, Li W, Schmidt W. Expression changes of ribosomal proteins in phosphate- and iron-deficient Arabidopsis roots predict stress-specific alterations in ribosome composition. BMC Genomics. 2013;14:783.
28. Lin WD, Liao YY, Yang TJ, Pan CY, Buckhout TJ, Schmidt W. Coexpression-based clustering of Arabidopsis root genes predicts functional modules in early phosphate deficiency signaling. Plant Physiol. 2011;155(3):1383–402.
29. del Pozo JC, Allona I, Rubio V, Leyva A, de la Pena A, Aragoncillo C, Paz-Ares. A type 5 acid phosphatase gene from Arabidopsis thaliana is induced by phosphate starvation and by some other types of phosphate mobilising/ oxidative stress conditions. Plant J. 1999;19(5):579–89.
30. Cheng Y, Zhou W, El Sheery NI, Peters C, Li M, Wang X, Huang J. Characterization of the Arabidopsis glycerophosphodiester phosphodiesterase (GDPD) family reveals a role of the plastid-localized AtGDPD1 in maintaining cellular phosphate homeostasis under phosphate starvation. Plant J. 2011;66(5):781–95.
31. Liu TY, Aung K, Tseng CY, Chang TY, Chen YS, Chiou TJ. Vacuolar Ca2+/H+ transport activity is required for systemic phosphate homeostasis involving shoot-to-root signaling in arabidopsis. Plant Physiol. 2011;156(3):1176–89.
32. Vert G, Grotz N, Dedaldechamp F, Gaymard F, Guerinot ML, Briat JF, et al. IRT1, an Arabidopsis transporter essential for iron uptake from the soil and for plant growth. Plant Cell. 2002;14(6):1223–33.
33. Thomine S, Vert G. Iron transport in plants: better be safe than sorry. Curr Opin Plant Biol. 2013;16(3):322–7.
34. Colangelo EP, Guerinot ML. The essential basic helix-loop-helix protein FIT1 is required for the iron deficiency response. Plant Cell. 2004;16(12):3400–12.
35. Buckhout TJ, Yang TJ, Schmidt W. Early iron-deficiency-induced transcriptional changes in Arabidopsis roots as revealed by microarray analyses. BMC Genomics. 2009;10:147.
36. Yang TJ, Lin WD, Schmidt W. Transcriptional profiling of the Arabidopsis iron deficiency response reveals conserved transition metal homeostasis networks. Plant Physiol. 2010;152(4):2130–41.
37. Palmer CM, Hindt MN, Schmidt H, Clemens S, Guerinot ML. MYB10 and MYB72 are required for growth under iron-limiting conditions. PLoS Genet. 2013;9(11):e1003953.
38. Yuan Y, Wu H, Wang N, Li J, Zhao W, Du J, Wang D, Ling HQ. FIT interacts with AtbHLH38 and AtbHLH39 in regulating iron uptake gene expression for iron homeostasis in Arabidopsis. Cell Res. 2008;18(3):385–97.
39. Wang N, Cui Y, Liu Y, Fan H, Du J, Huang Z, Yuan Y, Wu H, Ling HQ. Requirement and functional redundancy of Ib subgroup bHLH proteins for iron deficiency responses and uptake in Arabidopsis thaliana. Mol Plant. 2013;6(2):503–13.
40. Cosgrove DJ. Growth of the plant cell wall. Nat Rev Mol Cell Biol. 2005;6(11): 850–61.
41. Wolf S, Greiner S. Growth control by cell wall pectins. Protoplasma. 2012; 249(2):169–75.
42. Francoz E, Ranocha P, Nguyen-Kim H, Jamet E, Burlat V, Dunand C. Roles of cell wall peroxidases in plant development. Phytochemistry. 2014;112:15–21.
43. Ferrari S, Vairo D, Ausubel FM, Cervone F, De Lorenzo G. Tandemly duplicated Arabidopsis genes that encode polygalacturonase-inhibiting proteins are regulated coordinately by different signal transduction pathways in response to fungal infection. Plant Cell. 2003;15(1):93–106.
44. Kanai M, Nishimura M, Hayashi M. A peroxisomal ABC transporter promotes seed germination by inducing pectin degradation under the control of ABI5. Plant J. 2010;62(6):936–47.
45. Chen ZH, Nimmo GA, Jenkins GI, Nimmo HG. BHLH32 modulates several biochemical and morphological processes that respond to Pi starvation in Arabidopsis. Biochem J. 2007;405(1):191–8.

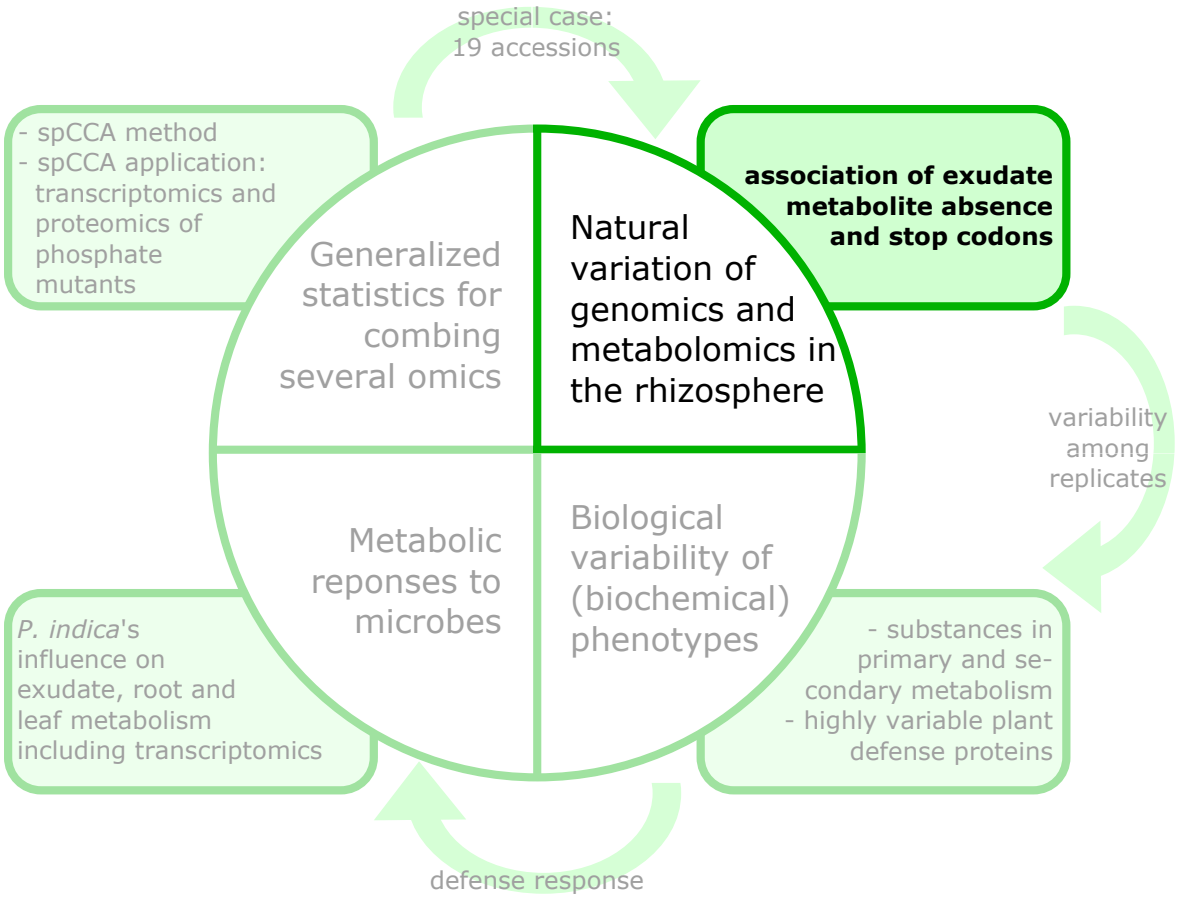
46. Miura K, Lee J, Miura T, Hasegawa PM. SIZ1 controls cell growth and plant development in Arabidopsis through salicylic acid. Plant Cell Physiol. 2010; 51(1):103–13.
47. Miura K, Rus A, Sharkhuu A, Yokoi S, Karthikeyan AS, Raghothama KG, Baek D, Koo YD, Jin JB, Bressan RA et al. The Arabidopsis SUMO E3 ligase SIZ1 controls phosphate deficiency responses. Proc Natl Acad Sci U S A. 2005;102(21):7760–5.
48. Miura K, Lee J, Gong Q, Ma S, Jin JB, Yoo CY, Miura T, Sato A, Bohnert HJ, Hasegawa PM. SIZ1 regulation of phosphate starvation-induced root architecture remodeling involves the control of auxin accumulation. Plant Physiol. 2011;155(2):1000–12.
49. Schmid NB, Giehl RF, Doll S, Mock HP, Strehmel N, Scheel D, Kong X, Hider RC, von Wiren N. Feruloyl-CoA 6′-Hydroxylase1-dependent coumarins mediate iron acquisition from alkaline substrates in arabidopsis. Plant Physiol. 2014;164(1):160–72.
50. Rodriguez-Celma J, Lin WD, Fu GM, Abadia J, Lopez-Millan AF, Schmidt W. Mutually exclusive alterations in secondary metabolism are critical for the uptake of insoluble iron compounds by Arabidopsis and medicago truncatula. Plant Physiol. 2013;162(3):1473–85.
51. Fourcroy P, Siso-Terraza P, Sudre D, Saviron M, Reyt G, Gaymard F, Abadia A, Abadia J, Alvarez-Fernandez A, Briat JF. Involvement of the ABCG37 transporter in secretion of scopoletin and derivatives by Arabidopsis roots in response to iron deficiency. New Phytol. 2014;201(1):155–67.
52. Kai K, Mizutani M, Kawamura N, Yamamoto R, Tamai M, Yamaguchi H, Sakata K, Shimizu B. Scopoletin is biosynthesized via ortho-hydroxylation of feruloyl CoA by a 2-oxoglutarate-dependent dioxygenase in Arabidopsis thaliana. Plant J. 2008;55(6):989–99.
53. Thum A, Mönchgesang S, Westphal L, Lübken T, Rosahl S, Neumann S, Posch S. Supervised Penalized Canonical Correlation Analysis. arXiv. 2014; preprint arXiv:1405.1534.
54. Jain A, Wilson GT, Connolly EL. The diverse roles of FRO family metalloreductases in iron and copper homeostasis. Front Plant Sci. 2014;5:100.
55. Curie C, Cassin G, Couch D, Divol F, Higuchi K, Le Jean M, Misson J, Schikora A, Czernic P, Mari S. Metal movement within the plant: contribution of nicotianamine and yellow stripe 1-like transporters. Ann Bot. 2009;103(1):1–11.
56. Reyt G, Boudouf S, Boucherez J, Gaymard F, Briat JF. Iron and ferritin dependent ROS distribution impact Arabidopsis root system architecture. Mol Plant. 2014;8(3):439–53.
57. Durrett TP, Gassmann W, Rogers EE. The FRD3-mediated efflux of citrate into the root vasculature is necessary for efficient iron translocation. Plant Physiol. 2007;144(1):197–205.
58. Roschzttardtz H, Seguela-Arnaud M, Briat JF, Vert G, Curie C. The FRD3 citrate effluxer promotes iron nutrition between symplastically disconnected tissues throughout Arabidopsis development. Plant Cell. 2011;23(7):2725–37.
59. Green LS, Rogers EE. FRD3 controls iron localization in Arabidopsis. Plant Physiol. 2004;136(1):2523–31.
60. Rogers EE, Guerinot ML. FRD3, a member of the multidrug and toxin efflux family, controls iron deficiency responses in Arabidopsis. Plant Cell. 2002; 14(8):1787–99.
61. Sterling C. Crystal-structure of ruthenium red and stereochemistry of its pectic stain. Am J Bot. 1970;57(2):172–5.
62. Krishnamurthy K. Methods in cell wall cytochemistry. Boca Raton, Florida, USA: CRC press; 1999.
63. Reeve RM. A specific hydroxylamine-ferric chloride reaction for histochemical localization of pectin. Stain Technol. 1959;34(4):209–11.
64. Hornatowska J. Visualisation of pectins and proteins by microscopy. 2005. STFI-Packforsk report.
65. Ravet K, Touraine B, Boucherez J, Briat JF, Gaymard F, Cellier F. Ferritins control interaction between iron homeostasis and oxidative stress in Arabidopsis. Plant J. 2009;57(3):400–12.
66. Bournier M, Tissot N, Mari S, Boucherez J, Lacombe E, Briat JF, Gaymard F. Arabidopsis ferritin 1 (AtFer1) gene regulation by the phosphate starvation response 1 (AtPHR1) transcription factor reveals a direct molecular link between iron and phosphate homeostasis. J Biol Chem. 2013;288(31):22670–80.
67. Meyer S, De Angeli A, Fernie AR, Martinoia E. Intra- and extra-cellular excretion of carboxylates. Trends Plant Sci. 2010;15(1):40–7.
68. Kobayashi Y, Hoekenga OA, Itoh H, Nakashima M, Saito S, Shaff JE, Maron LG, Pineros MA, Kochian LV, Koyama H. Characterization of AtALMT1 expression in aluminum-inducible malate release and its role for rhizotoxic stress tolerance in Arabidopsis. Plant Physiol. 2007;145(3):843–52.
69. Pant BD, Pant P, Erban A, Huhman D, Kopka J, Scheible WR. Identification of primary and secondary metabolites with phosphorus status-dependent abundance in Arabidopsis, and of the transcription factor PHR1 as a major regulator of metabolic changes during phosphorus limitation. Plant Cell Environ. 2015;38(1):172–87.
70. Ziegler J, Schmidt S, Chutia R, Muller J, Bottcher C, Strehmel N, Scheel D, Abel S. Non-targeted profiling of semi-polar metabolites in Arabidopsis root exudates uncovers a role for coumarin secretion and lignification during the local response to phosphate limitation. J Exp Bot. 2016;67(5):1421–32.
71. Krzeslowska M. The cell wall in plant cell response to trace metals: polysaccharide remodeling and its role in defense strategy. Acta Physiol Plant. 2011;33(1):35–51.
72. Gessa C, Deiana S, Premoli A, Ciurli A. Redox activity of caffeic acid towards iron(III) complexed in a polygalacturonate network. Plant Soil. 1997;190(2):289–99.
73. Nagarajah S, Posner AM, Quirk JP. Competitive adsorption of phosphate with polygalacturonate and other organic anions on kaolinite and oxide surfaces. Nature. 1970;228(5266):83–5.
74. Zhu XF, Wang ZW, Wan JX, Sun Y, Wu YR, Li GX, Shen RF, Zheng SJ. Pectin enhances rice (Oryza sativa) root phosphorus remobilization. J Exp Bot. 2015;66(3):1017–24.
75. Sudre D, Gutierrez-Carbonell E, Lattanzio G, Rellan-Alvarez R, Gaymard F, Wohlgemuth G, Fiehn O, Alvarez-Fernandez A, Zamarreno AM, Bacaicoa E et al. Iron-dependent modifications of the flower transcriptome, proteome, metabolome, and hormonal content in an Arabidopsis ferritin mutant. J Exp Bot. 2013;64(10):2665–88.
76. Chen SX, Schopfer P. Hydroxyl-radical production in physiological reactions. A novel function of peroxidase. Eur J Biochem. 1999;260(3):726–35.
77. Passardi F, Penel C, Dunand C. Performing the paradoxical: how plant peroxidases modify the cell wall. Trends Plant Sci. 2004;9(11):534–40.
78. Marjamaa K, Kukkola EM, Fagerstedt KV. The role of xylem class III peroxidases in lignification. J Exp Bot. 2009;60(2):367–76.
79. Passardi F, Tognolli M, De Meyer M, Penel C, Dunand C. Two cell wall associated peroxidases from Arabidopsis influence root elongation. Planta. 2006;223(5):965–74.
80. Daudi A, Cheng Z, O'Brien JA, Mammarella N, Khan S, Ausubel FM, Bolwell GP. The apoplastic oxidative burst peroxidase in Arabidopsis is a major component of pattern-triggered immunity. Plant Cell. 2012;24(1):275–87.
81. Robles MS, Cox J, Mann M. In-vivo quantitative proteomics reveals a key contribution of post-transcriptional mechanisms to the circadian regulation of liver metabolism. PLoS Genet. 2014;10(1):e1004047.
82. Baerenfaller K, Massonnet C, Walsh S, Baginsky S, Buhlmann P, Hennig L, Hirsch-Hoffmann M, Howell KA, Kahlau S, Radziejwoski A et al. Systems-based analysis of Arabidopsis leaf growth reveals adaptation to water deficit. Mol Syst Biol. 2012;8:606.
83. Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H et al. Mass-spectrometry-based draft of the human proteome. Nature. 2014;509(7502): 582–7.
84. Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. Global quantification of mammalian gene expression control. Nature. 2011;473(7347):337–42.
85. Palmgren MG, Nissen P. P-type ATPases. Annu Rev Biophys. 2011;40:243–66.
86. Cronin SR, Rao R, Hampton RY. Cod1p/Spf1p is a P-type ATPase involved in ER function and Ca2+ homeostasis. J Cell Biol. 2002;157(6):1017–28.
87. Vashist S, Frank CG, Jakob CA, Ng DT. Two distinctly localized p-type ATPases collaborate to maintain organelle homeostasis required for glycoprotein processing and quality control. Mol Biol Cell. 2002;13(11):3955–66.
88. Jakobsen MK, Poulsen LR, Schulz A, Fleurat-Lessard P, Moller A, Husted S, Schiott M, Amtmann A, Palmgren MG. Pollen development and fertilization in Arabidopsis is dependent on the MALE GAMETOGENESIS IMPAIRED ANTHERS gene encoding a type V P-type ATPase. Genes Dev. 2005;19(22): 2757–69.
89. Nishida S, Tsuzuki C, Kato A, Aisu A, Yoshida J, Mizuno T. AtIRT1, the primary iron uptake transporter in the root, mediates excess nickel accumulation in Arabidopsis thaliana. Plant Cell Physiol. 2011;52(8):1433–42.
90. Ravet K, Touraine B, Kim SA, Cellier F, Thomine S, Guerinot ML, Briat JF, Gaymard F. Post-translational regulation of AtFER2 ferritin in response to intracellular iron trafficking during fruit development in Arabidopsis. Mol Plant. 2009;2(5):1095–106.
91. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5(10):R80.

92. Venables W, Ripley BD. Modern applied statistics with S. New York, USA: Springer-Verlag; 2002.
93. Vizcaíno JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Ríos D, Dianes JA, Sun Z, Farrah T, Bandeira N, Binz PA,Xenarios I, Eisenacher M, Mayer G, Gatto L, Campos A, Chalkley RJ, Kraus HJ, Albar JP, Martinez-Bartolomé S, Apweiler R,Omenn GS, Martens L, Jones AR, Hermjakob H. ProteomeXchange provides globally co-ordinated proteomics data submission and dissemination. Nat Biotechnol. 2014;30(3):223–6.

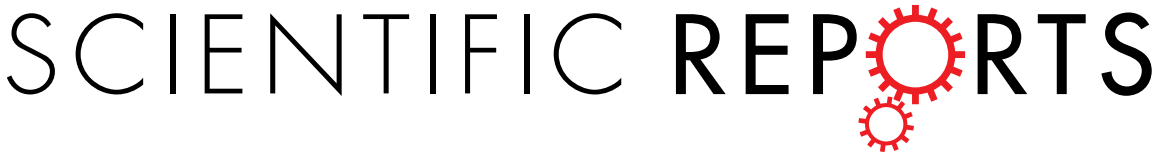
2.2 Natural variation of root exudates in *Arabidopsis thaliana* – linking metabolomic and genomic data

Mönchgesang, S.; Strehmel, N.; Schmidt, S.; Westphal, L.; Taruttis, F.; Müller, E.; Herklotz, S.; Neumann, S.; Scheel, D. Natural variation of root exudates in *Arabidopsis thaliana* – linking metabolomic and genomic data. *Sci Rep* **2016**, *6*.

equal contributions



www.nature.com/scientificreports



OPEN

Natural variation of root exudates in *Arabidopsis thaliana*-linking metabolomic and genomic data

Received: 14 February 2016
Accepted: 14 June 2016
Published: 01 July 2016

Susann Mönchgesang*, Nadine Strehmel*, Stephan Schmidt*, Lore Westphal, Franziska Taruttis†, Erik Müller, Siska Herklotz, Steffen Neumann & Dierk Scheel

Many metabolomics studies focus on aboveground parts of the plant, while metabolism within roots and the chemical composition of the rhizosphere, as influenced by exudation, are not deeply investigated. In this study, we analysed exudate metabolic patterns of *Arabidopsis thaliana* and their variation in genetically diverse accessions. For this project, we used the 19 parental accessions of the Arabidopsis MAGIC collection. Plants were grown in a hydroponic system, their exudates were harvested before bolting and subjected to UPLC/ESI-QTOF-MS analysis. Metabolite profiles were analysed together with the genome sequence information. Our study uncovered distinct metabolite profiles for root exudates of the 19 accessions. Hierarchical clustering revealed similarities in the exudate metabolite profiles, which were partly reflected by the genetic distances. An association of metabolite absence with nonsense mutations was detected for the biosynthetic pathways of an indolic glucosinolate hydrolysis product, a hydroxycinnamic acid amine and a flavonoid triglycoside. Consequently, a direct link between metabolic phenotype and genotype was detected without using segregating populations. Moreover, genomics can help to identify biosynthetic enzymes in metabolomics experiments. Our study elucidates the chemical composition of the rhizosphere and its natural variation in *A. thaliana*, which is important for the attraction and shaping of microbial communities.

In *Arabidopsis thaliana (A. thaliana)*, natural genetic variation has been intensively exploited to study a variety of traits related to plant development, stress response and nutrient content (for review, see Weigel[1]). Several publications have demonstrated that natural variation is a suitable basis for dissecting secondary metabolite pathways by using genetic mapping analyses. The genetics of glucosinolates and its link to pathogen and herbivore resistance have been investigated thoroughly[2–5]. A large variation of glucosinolates in leaves and seeds was observed for 39 genetically diverse Arabidopsis accessions[6]. Houshyani *et al.*[7] found that natural variation of the general metabolic response to different environmental conditions is not necessarily associated with the genetic similarity between nine accessions.

Many metabolomics studies focus on aboveground plant tissues. As a result, only limited information is available with regard to the metabolism of belowground parts of the plant.

Roots are crucial for the uptake of water and nutrients. For example, Agrawal *et al.*[8] utilized natural variation of *A. thaliana* to identify malic acid as a key mediator for nickel tolerance. To communicate with the belowground environment, plant roots also exude metabolites such as flavonoids, phenylpropanoids and glucosinolates[9], which can attract microorganisms or increase the resistance against pathogens[9–11]. These interactions take place in the rhizosphere, which is regarded as the space adjacent to roots[12]. As the properties of the rhizosphere differ strongly from the bulk soil in terms of microorganism abundance[13], as well as the qualitative and quantitative metabolic composition[14,15], investigations on root exudates are needed to assess the role of this microenvironment. Micallef *et al.*[16] demonstrated that the rhizobacterial community composition is influenced by varying exudation profiles.

Non-targeted metabolite profiling of secondary metabolites by liquid chromatography coupled to mass spectrometry (LC/MS) is an ideal analytical platform to link natural metabolite variation to biosynthetic pathways. It allows for the detection and quantification of semipolar compounds[17], when the resulting three-dimensional

Leibniz Institute of Plant Biochemistry, Department of Stress and Developmental Biology, Weinberg 3, 06120 Halle (Saale), Germany. †Present address: University of Regensburg, Josef-Engert-Str. 9, 93053 Regensburg, Germany. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to D.S. (email: dierk.scheel@ipb-halle.de)

signals with a specific mass-to-charge (*m/z*) ratio, retention time (RT) and intensity, so-called features, can be annotated. Depending on the nature of the compound, they are more likely to be detected upon electrospray ionization in the positive (ESI(+)) or negative mode (ESI(−)).

Our approach to investigate natural genetic variation of secondary metabolism in root exudates focuses on 19 *A. thaliana* accessions, which show a large degree of geographic and phenotypic diversity (Supplementary Table S1) and were used to generate the Multiparent Advanced Generation Inter-Cross (MAGIC) lines[18]. Whole genome sequencing revealed that the parental accessions and the MAGIC lines represent most of genetic variability of *A. thaliana* and therefore provide a valuable resource for genetic and metabolic studies[19,20].

The aim of this study is to find out if the root exudate composition in *A. thaliana* is genetically determined. For this purpose, we analysed which metabolites show natural variation, if similar metabolic phenotypes share a genetic base, in particular, if certain characteristics can be traced back to single nucleotide polymorphisms and hence, directly link phenotype and genotype.

Results

Non-targeted metabolite profiling of root exudates reveals distinct metabolic phenotypes for 19 Arabidopsis accessions. A clustering analysis was performed to find similarities between the metabolic profiles and sequence polymorphisms of the 19 founder accessions of the MAGIC population of *A. thaliana*. The dendrograms calculated from the metabolic features show a clear separation of accessions in Fig. 1a for exudates measured in ESI(−) and Fig. 1b in ESI(+). At a correlation threshold of 0.95 (dashed line), seven and five clusters, respectively, were observed.

No-0 and Po-0 (blue) were found in the same cluster (cluster 1, ESI(−); cluster 5 ESI(+)) in both ion modes. Ct-1 and Edi-0 (purple) also displayed high similarity in their metabolic profiles. Sf-2 and Kn-0 (green) were in close proximity and would have been in the same clade when cutting the ESI(+) dendrogram at a different threshold. Similar metabolic phenotypes were also detected in the exudation patterns of Wu-0 and Tsu-0, and additionally Mt-0 (orange). These three accessions either clustered in dendrogram branch 2 (ESI(−)) or 3 (ESI(+)).

In both metabolic dendrograms, one Oy-0 sample was observed as an outlier, which did not cluster with the other replicates of Oy-0. For Hi-0 and Ws-0, mixed clusters were observed. The positive ion mode generally harboured more outliers. As obvious from the quality control plots in Supplementary Fig. S1, the outlying samples did not show any extreme deviations on the technical side and were therefore not excluded from further analysis[21].

For the analysis of genetic diversity, sequence polymorphisms in coding sequences (CDS) extracted from the 19 genomes project[22] were used for a genetic clustering (Fig. 1c). One large dendrogram branch (L*er*-0, Kn-0, Wil-2; Ws-0, Ct-1, No-0; Hi-0, Tsu-0, Mt-0, Wu-0, Col-0, Rsch-4) had less than 825,000 mismatches (dashed line) while the outliers Bur-0, Sf-2, and Can-0 had increasing numbers of polymorphisms. Oy-0 and Po-0 formed a small cluster and were found in proximity to Edi-0, Zu-0 and the large dendrogram branch.

The metabolic analysis was based on a non-targeted metabolite profiling approach considering metabolic features characterised only by their *m/z* ratios, RTs and intensities. These characteristics are not sufficient to investigate the underlying molecules, its biosynthetic pathway and its potential in plant signaling. Annotations and identifications of metabolites, as shown in the next paragraph, are required to interpret non-targeted metabolic profiles in the biological context.

Semipolar secondary metabolites are the major components of the exudation patterns. Only 25 and 22 of the metabolic signals (455 (ESI(−)), 475 (ESI(+), respectively) could be assigned to metabolites which have been previously described as exudate-characteristic for Col-0[15]. Differential metabolites were detected by a generalized Welch-test between the 19 accessions; their colour-coded intensity map is shown in Fig. 2. Chemically related compounds were placed in groups separated by horizontal spacing.

Among the differential metabolites, there were several compounds with an aromatic moiety, such as the nucleoside thymidine and the amino acids Phe and Tyr. The amino acid derivative hexahomo-Met *S*-oxide had low abundance in the exudates of Sf-2 and was enriched in Mt-0.

A range of glucosinolate degradation products was characteristic for the exudates of some accessions. Edi-0 had rather low levels of indolic compounds and the isothiocyanate hydrolysis product of 8-MeSO-Octyl glucosinolate. Wu-0 showed a clear absence of the neoglucobrassicin (1-MeO-I3M) hydrolysis product 1-methoxy-indole-3-ylmethylamine (1-MeO-I3CH_2NH_2), while Sf-2 was missing the malonyl-glucoside of 6-hydroxyindole-3-carboxylic acid (6-(Malonyl-GlcO)-I3CH_2CO_2H). An unknown indole derivative ($C_{10}H_9NO_3$) was highly abundant in the exudates of Ct-1 and Wil-2, and lowly abundant in Sf-2. Generally, large amounts of the glucosinolate precursor and hydrolysis products were detected in the exudates of L*er*-0, Mt-0 and Wil-2.

Plant hormone-derived metabolites also differed between the 19 accessions. Two salicylic acid (SA) catabolites, 2,3 and 2,5-dihydroxybenzoic acid (DHBA) pentosides, were highly abundant in Col-0, Kn-0, L*er*-0, Mt-0, Wil-2, Ws-0 and Wu-0. No preference for the 3′ or 5′ hydroxylated variant of DHBA was noticed, and both isomers correlated positively with a Pearson correlation of 0.91. 9,10-dihydrohydroxy jasmonic acid (JA) *O*-sulfate was another differential plant hormone catabolite in *A. thaliana* exudates with low levels in Bur-0, Can-0 and Zu-0 and high levels in Col-0, Kn-0, Po-0, Rsch-4 and Wu-0.

Among the phenylpropanoids, the coumarin scopoletin and its glycosides differed in the exudates of the 19 accessions. A hexose-pentose conjugate of scopoletin as well as three other glycosides ($C_4H_{10}O$ Hex-DeoxyHex, $C_{12}H_{16}O_5$ Hex, $C_7H_{14}O_4$ Malonyl-Hex) were among the differentially abundant metabolites which were described for Col-0 exudates[15].

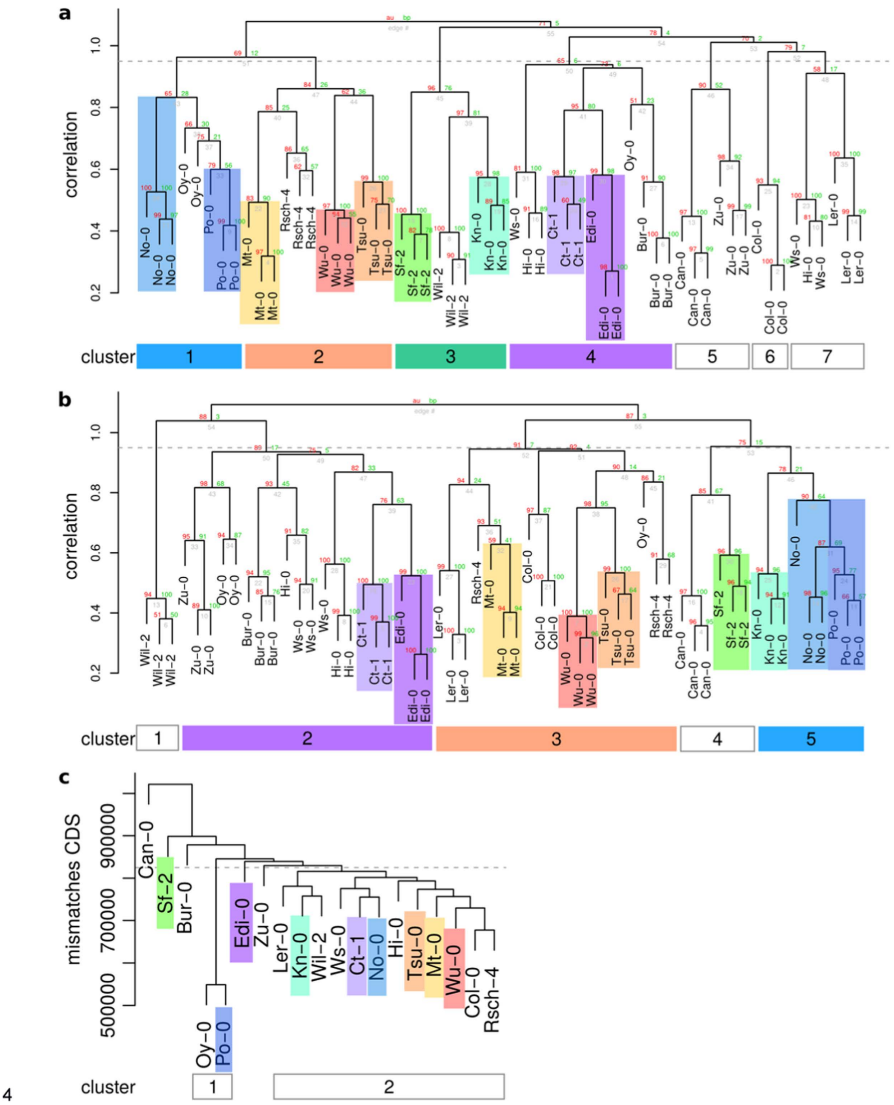


Figure 1. Hierarchical clustering of metabolic features from (**a**) exudates ESI(−), (**b**) ESI(+) and of (**c**) genetic distances. (**a**+**b**) Features were obtained by UPLC/ESI(−)-QTOF-MS (a) or UPLC/ESI(+)-QTOF-MS (**b**) from exudate samples and differed from the blank (Welch test, $p < 0.05$). Intensities were corrected for batch effects using SVA and subjected to average linkage clustering with correlation as a distance measure. (**c**) Variant tables of the 19 genomes project were reduced to coding regions, as annotated by TAIR. The sum of all mismatches was used as a distance matrix for average linkage clustering. Dendrograms were cut at a correlation threshold of 0.95 (dashed line). As cluster numbers were not comparable, consistent clusters were coloured across ion modes as a visual guidance.

Other differential phenylpropanoids include the monolignol glucoside syringin as well as both isomers of the sulfated dilignol G(8-*O*-4)FA *O*-sulfate consisting of coniferyl alcohol (G) and ferulic acid (FA): it was present at high levels in Kn-0 and Wil-2 exudates. Two hydroxylated fatty acids also showed natural variation and were highly abundant in Mt-0.

Several isoforms of known glycosylated metabolites (e.g. kaempferol triglycosides with *m/z* 739.21) were detected at different RTs indicating differences in sugar conjugation. The investigation of these putatively annotated metabolites can be facilitated by exploring polymorphisms in genes encoding their biosynthetic enzymes.

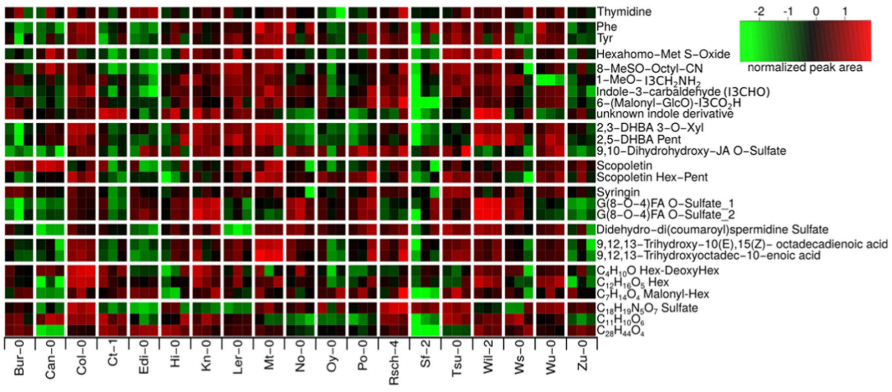


Figure 2. Colour-coded intensity matrix of differential metabolites occurring in exudates. Integrated peak areas were log-transformed and scaled to zero mean and standard variance. A Welch-test was used to find differentially abundant metabolites between the 19 accessions.

The absence of an indolic glucosinolate hydrolysis product and a hydroxycinnamic acid conjugate is genetically determined. Wiesner *et al.*[23] reported that the accession Wu-0 lacks the 1′-methoxylated indolic glucosinolate due to a premature stop codon in the *CYP81F4* gene[24]. Its frameshift mutation leads to a loss of function and subsequently to the absence of 1-MeO-I3M in roots and leaves[23], and also its amine, 1-MeO-I3CH$_2$NH$_2$, in the exudates of our hydroponic system.

To elucidate if further metabolite absences in the exudates like 1-MeO-I3CH$_2$NH$_2$ in Wu-0 can be traced back to a single gene, we developed a workflow to link genomic and metabolic patterns (Fig. 3). Features with the same absence pattern could be different molecular species of the same compound (adducts, isotopes, fragment or cluster ions). Alternatively, they may be different isomers from the same biosynthetic pathway with a common precursor.

Among the seven metabolic features with absence in two accessions, three were characteristic for Can-0 and L*er*-0. The hydroxycinnamic acid polyamine derivative cyclic didehydro-di(coumaroyl)spermidine sulfate previously identified in Col-0[15] and also detected in other accessions was clearly absent in Can-0 and L*er*-0 (Fig. 2). This compound with RT = 3.6 min was absent in the negative ion mode as $[M\text{-}H]^-$ adduct with $m/z = 514.17$ and $[M\text{-}2H + Na + CH_2O_2]^-$ adduct with $m/z = 582.15$. Another compound with $m/z = 514.17$ eluting at 4.2 min was also absent in Can-0 and L*er*-0. Tandem mass spectrometry (MS/MS) analysis revealed a sulfur trioxide loss in the fragmentation pattern similar to the sulfated cyclic didehydro-di(coumaroyl)spermidine conjugate. Can-0 carries a premature stop codon in the gene AT2G25150 encoding spermidine dicoumaroyl transferase (SCT), whereas in L*er*-0, a large deletion is present in the CDS of this gene[22]. Both accessions have no detectable levels of SCT transcript in their roots (Fig. 4a).

Thus, neither Can-0 nor L*er*-0 possess SCT activity to most likely produce cyclic didehydro-di(coumaroyl) spermidine sulfate and its isomer. To further support the data observed with these two accessions, we analysed the exudates of the homozygous knockout line SALK_098927C (Col-0 background), which indeed did not display any peaks with m/z 514.17 ESI(−) at 3.6 min, as shown in Fig. 4b, and thus confirm our hypothesis.

The above results for the Wu-0 and Can-0/L*er*-0 pattern showed the feasibility of such an association analysis to link compounds to their biosynthetic pathways. In specific cases, there is a direct connection between metabolic phenotype and genotype. Therein, metabolite variation among Arabidopsis accessions can be traced back to individual SNPs without trait segregation and QTL mapping.

Matching metabolic and genetic patterns can indicate compound class. Genetic alterations may be exploited to characterise so far unknown compounds which are part of related biosynthetic pathways[25]. MS/MS fragmentation facilitates the annotation of chemical substructures, which are often characteristic for a certain class of compounds. Knowledge about biosynthetic pathways can further support the assignment of unknown features to compound classes.

For the annotation of metabolites, collision-induced dissociation (CID-) MS was performed for 17 selected MS1 ESI(−) features obtained by the above described screening.

With the help of MS/MS spectra, nine out of 17 features were annotated and for five further features, the elemental composition was determined. An overview of compounds, fragment spectra and matching enzymes is given in Supplementary Table S5.

A compound (m/z 739.21, RT = 4.3 min) that was not found in the exudates of Wu-0 (Fig. 5a) was identified as a flavonoid with the same elemental composition ($C_{33}H_{40}H_{19}$) and fragment spectrum as kaempferol 3-*O*-Rha(1→2)Glc 7-*O*-Rha[15]. The RT shift indicates different glycosidic conjugation. This compound was identified as robinin (kaempferol 3-*O*-Rha-Gal 7-*O*-Rha) by an authentic standard having a galactose moiety instead of glucose in the diglycoside at the 3′ position (Fig. 5b). One out of the 16 premature stop codons characteristic for Wu-0 was present in AT2G22590.1, which encodes the UDP-glycosyltransferase (UGT) superfamily protein

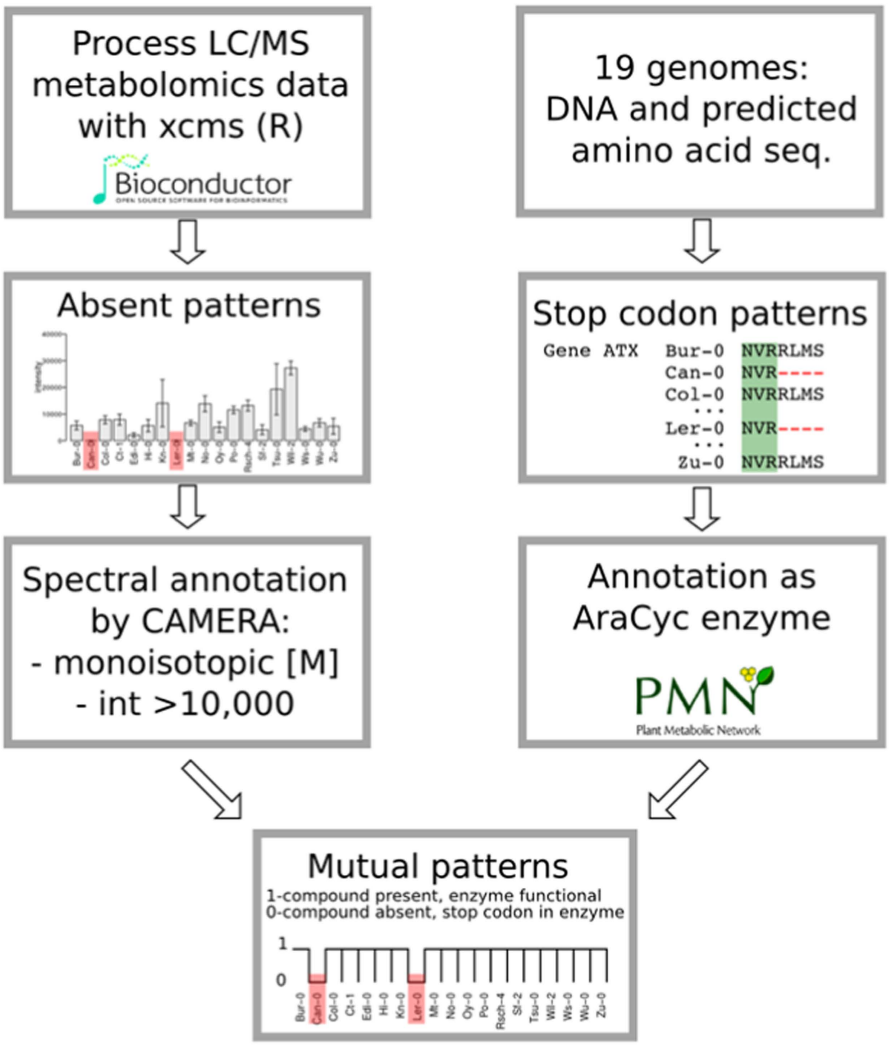


Figure 3. Workflow for matching metabolic patterns of absence with stop codons in genes annotated as AraCyc enzymes. For the metabolic data, 384 out of 455 metabolic features from the ESI(−) data set were absent in at least one accession. 38 of them were annotated as monoisotopic peak [M] by CAMERA. Approximately 32,000 stop codons were detected. 1,588 of AraCyc enzyme-encoding genes displayed a prematurely ended amino acid sequence possibly representing non-functional enzymes that can be causative for metabolite absence.

UGT91A1. This gene is coexpressed with the flavonol synthase 1 (FLS1, AT5G08640) and chalcone flavanone isomerase (TT5, AT3G55120) encoding genes that are annotated with the "flavonoid biosynthetic process" by Gene Ontology[26]. The exudates of the homozygous knockout line SALK_088702C (Col-0 background) were missing robinin and its UGT91A1 transcript levels in roots were diminished (Fig. 5c–e).

The hydroxylated fatty acid 9,12,13-trihydroxyoctadec-10-enoic acid (9,12,13-TriHOME, KEGG C14833) was not present in the exudates of Edi-0 and Zu-0 (Fig. 2). Its lack corresponds to a SNP pattern introducing a stop codon into a long-chain-alcohol *O*-fatty-acyltransferase gene (AT5G55360.1). The unsaturated variant 9,12,13-trihydroxyoctadec-10(E),15(Z)-enoic acid, however, could be detected in Edi-0 and Zu-0 exudates, but not in the Ct-1 accession, and accordingly, pointed to different polymorphism patterns. Besides, similar intensity distributions of both hydroxylated fatty acids were found across the exudates of the 19 accessions (Fig. 2).

These examples show that the direct search for a metabolite-enzyme-connection can provide valuable insights into biosynthetic pathways but require careful examination of the resulting candidate genes.

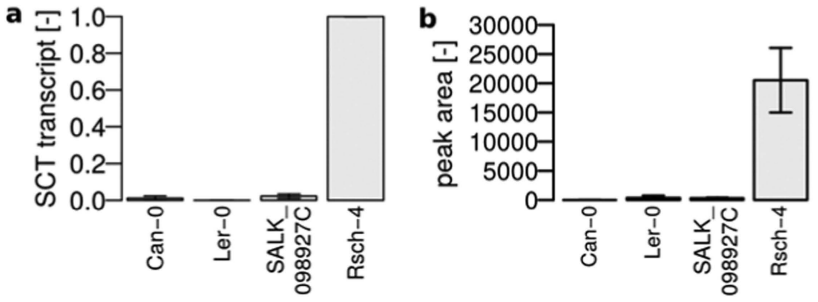


Figure 4. Natural and T-DNA insertion knockouts of SCT. (**a**) Relative transcript levels of SCT in root tissue as determined by qPCR, PP2A as reference, normalized to Rsch-4, mean ± s.e.m., n = 3. (**b**) Peak area counts of cyclic didehydro-di(coumaroyl)spermidine sulfate in exudates, mean ± s.e.m., n = 3.

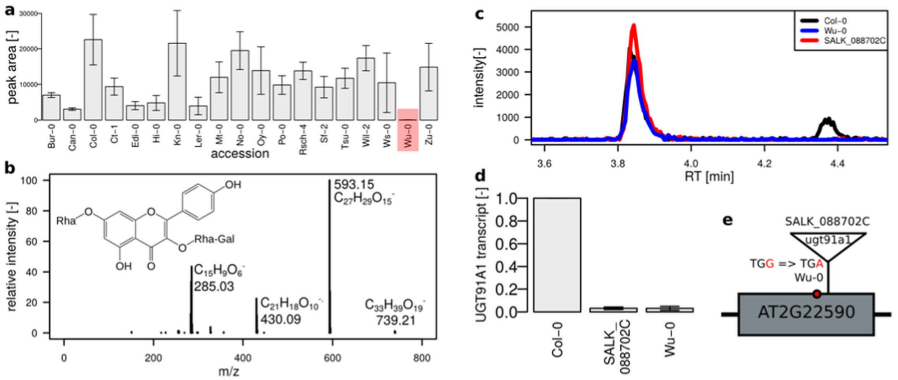


Figure 5. Robinin absence is linked to a stop codon in the UGT91A1 encoding gene. (**a**) Peak area counts, mean ± s.e.m. (n = 3) with absence in Wu-0 (highlighted in red) (**b**) MS/MS spectrum of robinin, 30 eV, (**c**) extracted ion chromatogram at *m/z* 739.21 with kaempferol 3-*O*-Rha(1→2)Glc 7-*O*-Rha eluting at 3.9 min and the galactose-conjugated robinin eluting at 4.3 min not detected in the natural knockout Wu-0 and T-DNA insertion line SALK_088702C, (**d**) relative transcript levels of UGT91A1 in roots as determined by qPCR, PP2A as reference, normalized to Col-0, mean ± s.e.m., n = 4, (**e**) schematic representation of the UGT91A1 gene (one exon) and the loss-of-function mutations in Wu-0 and SALK_088702C.

Discussion

This study showed how the exudation pattern of *A. thaliana* accessions is reflected by a genetic clustering of polymorphisms in their CDS. The previously reported similarity of the German and Norwegian accession Po-0 and Oy-0[22] was only observable at metabolic level in the ESI(−) dendrogram. The close relation was confirmed by the genetic clustering. However, we also observed closely related metabolic profiles of Po-0 with No-0 (blue), which has not been described before. Neither the metabolic proximity of Sf-2 and Kn-0 (green) nor of Ct-1 and Edi-0 (purple) were reflected by small genetic distances.

The similarity of the Wu-0, Tsu-0 and Mt-0 was present in both ESI dendrograms of the exudate analysis and seems to be genetically determined. The close genetic relation between the Japanese accession Tsu-0 and Mt-0 from Libya has already been reported by Nordborg *et al.*[19] as well as by De Pessemier *et al.*[27], and was confirmed for metabolic exudate and the CDS profiles (orange).

The clustering of metabolic profiles demonstrated that genetic variation between the 19 founder accessions of the Arabidopsis MAGIC population is indeed reflected in the exudate metabolome. This is in contrast to the previously reported only minor correlation between shoot metabolic and genetic similarity[7] of nine accessions, partially overlapping with the MAGIC founder lines. Compared to 149 SNPs that were used to estimate a genetic distance by Houshyani *et al.*[7], our analysis included 640,066 polymorphisms that were exclusively within CDS. The usage of SNPs in CDS ensures a comprehensive, but most direct genotype-phenotype-association, disregarding regulatory sequences. From hierarchical clustering, we can conclude that the three dendrograms reflect the genetic determination of the exudation profile of several Arabidopsis accessions. Both, the genetic and thus the metabolic profiles, may have been affected by selection processes at the collection sites[25]. Information on

environmental conditions, especially characteristic rhizosphere data of the original locations, would be of great interest, but unfortunately, these are not well documented[28].

In our study, a variety of glycosylated and sulfated compounds are the key metabolites that underlie natural variation in the exudates of the MAGIC parental lines. Scopoletin was found both as an aglycone and hexose-pentose conjugate. However, glucosinolates were only detected as degradation products (amines, carbaldehydes, isothiocyanates). Currently, we cannot elucidate whether glucosinolate exudation is initiated by myrosinase activation or if hydrolysis was caused by the sample preparation procedure.

Previously, hormones were described as constituents of root exudates[29]. Despite that, plant hormones were difficult to detect with the analytical method due to their low abundance. Plant hormone-derived metabolites were detected as glycosylated and sulfated in case of SA and JA, respectively. Natural variation is reflected by a great spectrum of glycosidic conjugation. This was shown for SA catabolites. SA was present in the exudates of Col-0 in the study of Strehmel *et al.*[15] but did not pass their stringent filtering criteria to be included in their exudate compound collection, while SA derivatives with 2,3 or 2,5- dihydroxy-substituted benzoic acid pentose conjugates passed the filter. As shown in Supplementary Fig. S2, high amounts of SA were found in Kn-0, Wil-2 and Wu-0, the lowest amount was present in Sf-2 exudates, one of the accessions with also low DHBA pentoside levels. Interestingly, solely pentosides but no hexosides of DHBA were detected in the root exudates of Col-0[15]. Li *et al.*[30] investigated the discrimination of hexose and pentose conjugation in 96 *A. thaliana* accessions. Combined QTL and association mapping pointed to a locus on chromosome 5 within proximity of a gene encoding a putative UGT with pentose specificity. The findings of this study support the previously reported low ratio of pentose-hexose conjugates for Edi-0[30]. Sf-2 was the accession with the lowest DHBA pentoside-hexoside ratio, which may be caused by a non-functional pentose-conjugating UGT and a background hexose-transferase activity that leads to a DHBA hexoside phenotype.

Chemically related compounds often derive from the same biosynthetic pathway. The characterisation of these metabolites might be facilitated by combining metabolic patterns with genomic data. Thus, an analysis workflow was developed which compares metabolite and sequence polymorphism patterns. In order to reduce the complexity, qualitative metabolic patterns were extracted and compared with the presence of premature stop codons in enzyme-encoding genes. The absence of a sulfated cyclic di(dehydrocoumaroyl)-spermidine was traced back to a single genomic alteration diminishing SCT activity in Can-0 and L*er*-0. These data support the hypothesis postulated by Strehmel *et al.*[15] that the cyclic conjugate is derived from di(coumaroyl)spermidine synthesized from spermidine and coumaroyl-CoA by SCT as illustrated in Fig. 6. A subsequent oxidative ring formation and sulfonylation led to sulfated cyclic di(dehydrocoumaryol)-spermidine[31]. Nevertheless, the coumaroyl spermidine transferase activity can hardly be inferred from the gene annotation as "HXXD-type acyl transferase family protein". This workflow furthermore pointed towards the substrate specificity of UGT91A1. Previous studies have shown that UGT91A1 is regulated by MYB transcription factors and speculated about its involvement in glycosylation of flavonols or flavonol glycosides[32]. We could show that in the absence of UGT91A1 enzymatic activity no galactose transfer to kaempferol 3-*O*-Rha 7-*O*-Rha (kaempferitrin) is catalysed to produce robinin. However, the presence of the glucose-substituted isomer kaempferol 3-*O*-Rha(1→2)Glc 7-*O*-Rha implies the involvement of a different UGT not accepting galactose but rather glucose as a substrate. We hereby found that UGT91A1 might have similar flavonoid substrate specificity as UGT73C6 and UGT78D1[33]. However, the patterns of two closely related hydroxylated fatty acids did not show mutual absences. Their intensity distributions were similar and point out the threshold issue in the absence definition. The SNP in AT5G55360 is likely to be a false positive candidate that needs to be excluded by a careful interpretation.

Future investigations will focus on the refinement of our approach by addressing the following points: i) When is a peak defined as absent? We relied on the decision of the peak-picking method centWave[34] in the xcms package[35]. If the algorithm found a peak at a particular *m/z* and RT in one accession but could erroneously not match its peak criterion in any replicates of another accession, the peak was defined as absent. ii) For a proof of concept, our workflow only included nonsense mutations in CDS of single genes. More complex studies would include amino acid exchanges in CDS, alterations in promoter regions as well as cases of gene function redundancies.

Linking stop codons with metabolite absences helps with the elucidation of secondary metabolite pathways but still requires fragment spectra to be interpreted manually and gene annotations have to be carefully checked for a possible involvement within the biosynthetic pathway of the metabolite. The connection has to be validated by knockout lines of the respective candidate genes.

Our study revealed natural variation in the root exudate composition of 19 genetically diverse accessions of *A. thaliana*. Combining nonsense mutations with metabolic patterns of the exudates facilitated to determine the genetic base of specific metabolite absences. Furthermore, the integration of sequence data can help to identify compound classes in metabolomics experiments. Our study can aid to further unravel biochemical and molecular processes in the rhizosphere by providing a metabolomics resource of root exudates (MetaboLights, accession number MTBLS160, http://www.ebi.ac.uk/metabolights/MTBLS160). Future investigations should aim at correlating metagenomics with exudation profiles in order to deduce characteristics that can be exploited to circumvent limiting abiotic factors and decrease the susceptibility towards biotic stresses.

Methods

Plant material. Seeds of the accessions Bur-0, Col-0, Can-0, Ct-1, Edi-0, Hi-0, Kn-0, L*er*-0, Mt-0, No-0, Oy-0, Po-0, Rsch-4, Sf-2, Tsu-0, Wil-2, Ws-0, Wu-0, and Zu-0 of *A. thaliana* (Supplementary Table S1) were obtained from the European Arabidopsis Stock Centre. The T-DNA insertion lines SALK_098927C and SALK_088702C were obtained from the SALK institute and Dr. Ralf Stracke (Bielefeld), respectively, and characterised as elaborated in the Supplementary Methods.

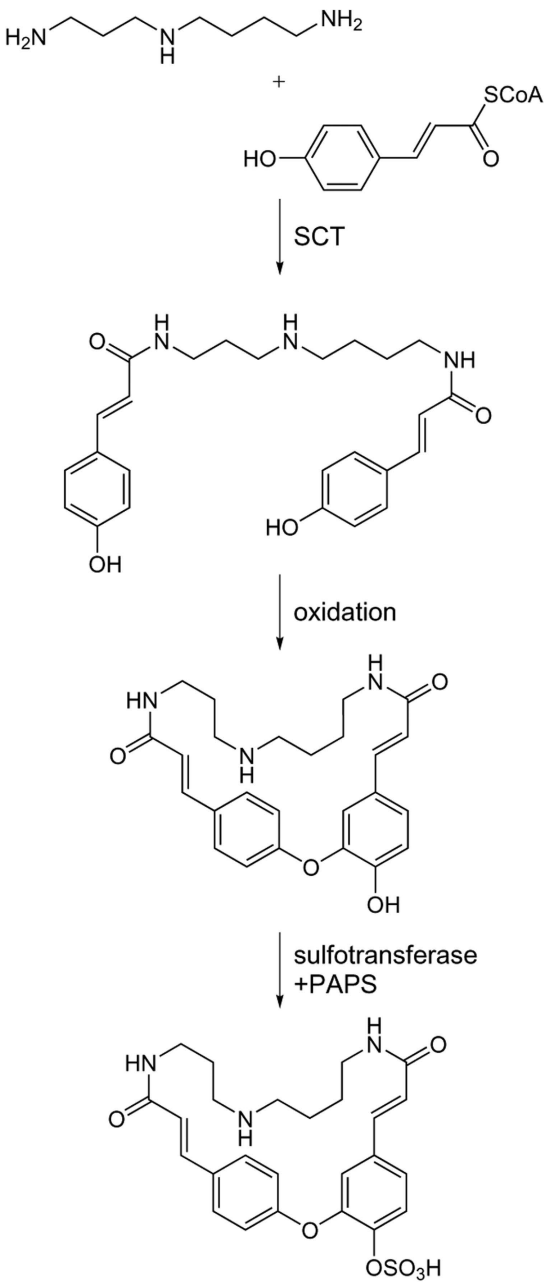


Figure 6. Biosynthetic pathway of cyclic didehydro-di(coumaroyl) spermidine sulfate. Di(coumaroyl) spermidine is synthesized by SCT[47] and subsequent oxidative ring closure and sulfonylation leads to cyclic didehydro-di(coumaroyl) spermidine sulfate, PAPS = 3′-phosphoadenosine-5′-phosphosulfate.

Plant cultivation. All seeds were surface-sterilized prior to plant cultivation. Then, all lines were cultivated in a hydroponic system with three independent biological experiments as previously described[15] and in the Supplement. Culture medium was used as a blank. Medium was collected after one-week-exudation (week 5–6) and resulted in 57 pooled root exudates (of four plants each).

Sample preparation. Root exudates were prepared according to Strehmel *et al.*[15] and as described in Supplementary Methods.

Non-targeted metabolite profiling analysis. Changes in metabolism were analysed by ultra-performance liquid chromatography coupled to electrospray ionization quadrupole time–of–flight mass spectrometry (UPLC/ESI-QTOF-MS) according to Böttcher *et al.*[36].

All mass spectra were acquired in centroid mode and recalibrated on the basis of lithium formate cluster ions.

A detailed description of plant cultivation, sample preparation and metabolite profiling can be found in Supplementary Methods.

Data analysis. Raw data files were converted to mzData using CompassXPort version 1.3.10 (Bruker Daltonics 4.0). Subsequently, the R package xcms version 1.41.0[35] was used for feature detection, alignment and filling of missing values. On this account, features were detected with the help of the centWave algorithm according to Tautenhahn *et al.*[34] (snthr = 5, scanrange = c(1,3060), ppm = 20, peak width = c(5,12)), matched across samples (xcms function group, minfrac = 0.75, bw = 2, mzwid = 0.05, max = 50), corrected for retention time shifts (method = "loess") and grouped again. Missing values were imputed with the xcms function fillPeaks which integrates raw chromatographic data. The data matrix was extracted using the diffreport function.

DataAnalysis 4.0 (Bruker Daltonics) was used for generation of extracted ion chromatograms, deconvolution of compound mass spectra and calculation of elemental compositions. For relative quantification of compounds extracted ion chromatograms from the non-targeted analysis were integrated with QuantAnalysis 2.0 (Bruker Daltonics) using the quantifier ions as listed in Supplementary Table S3. Peak areas were log-transformed and z-scaled to achieve normal distribution. Differential metabolites were detected by a generalized Welch-test between the 19 accessions (unequal variances, one-way layout, $p < 0.05$, corrected for multiple testing by Benjamini-Hochberg's method[37]).

All statistical procedures were performed with the R statistical language version 3.0.0[38] and the Bioconductor environment[39]. All data are available from the MetaboLights repository under the accession number MTBLS160 (see Supplementary Methods).

Hierarchical clustering. Before hierarchical clustering, remaining missing values were replaced with half of the minimum feature intensity. Feature intensities were logarithmized, z-transformed and checked for normality with a Kolmogorov-Smirnow test. Non-biological sources of variation were removed by surrogate variable analysis from the SVA package version 3.8.0[40]. In order to discriminate between experimental artifacts and metabolic features in the non-targeted analysis, a generalized Welch test (unequal variances, one-way layout) was applied to find differential features ($p < 0.05$, corrected for multiple testing by Benjamini-Hochberg's method[37]) between the 19 accessions and blank. As a post-hoc test, 2-sample Welch tests were used to find features that were differential ($p < 0.05$) from the blank in at least one accession. This resulted in 455 out of 1950 ESI(−) and 475 out of 3738 ESI(+) metabolic features used for hierarchical clustering. Hierarchical clustering was performed via multiscale bootstrap resampling with the R package pvclust version 1.2–2[41], which improves robustness by providing an approximately unbiased p-value (AU, red number in Fig. 1). Pearson correlation was used as distance measure and average linkage as a linkage method. Since the combination of both ion modes results in redundancy by compounds giving rise to several features, each mode was processed separately. Consistent clusters between the ESI(−) and ESI(+) mode were coloured.

Unspecific signals were more pronounced (87% vs. 75%) in ESI(+) vs. ESI(−). This had led to us to focus on ESI(−) in subsequent analyses.

Sequence analysis. Genetic distances were estimated from the variant tables available from the 19 genomes project[22]. Loci were reduced to CDS as annotated by the R packages Bsgenome.Athaliana.TAIR.TAIR9[42] and Genomic Ranges version 1.14.4[43]. For each variant locus, 19×19 comparisons were conducted. In order to construct a distance matrix, mismatches were penalized by increasing the distance by 1. The sum of matrices over all 6,400,466 loci was used as a distance matrix (Supplementary Table S2) for hierarchical clustering via the hclust package with average linkage.

Predicted amino acid sequences were processed with BioPerl (Bio::Tools::Run::Alignment::Clustalw, Bio::SeqIO, Bio::Seq, and Bio::AlignIO) and aligned with the Clustalw algorithm with ktuple = 2 and a BLOSUM scoring matrix. Multiple sequence alignments were evaluated for premature ending with the R packages Biostrings version 2.30.1 and plyr version 1.8.1.

Combination of metabolic and genetic patterns. A metabolic feature was defined as absent when below the limit of detection in all replicates of an accession. Applying this stringent definition, the peak list created from aligning all spectra from ESI(−) was screened for metabolic features with absence, thus reducing the number of features by 25% for exudates ESI(−). The distribution of absence across the 19 accessions is referred to as a pattern. The length of a pattern is the number of accessions that lack the same feature, i.e. a feature absent in Can-0 und Zu-0 is a pattern of length two. Out of the 455 metabolic features in the exudate data set (ESI(−)), 384 were missing in at least one accession. 46 were missing in exactly one accession (length = 1), 52 were absent in two accessions (length = 2) (see Supplementary Table S4). The R package CAMERA version 1.23.2[44] was used for annotation of adduct species and isotope information. In order to find an association between metabolic patterns of absence and its genetic background, features with a pattern of absence, a monoisotopic annotation by CAMERA and a minimal median intensity of 10,000 were evaluated. 31 features that passed the intensity threshold were matched with stop codon patterns resulting in 9/7/1 features of absence with length 1/2/3.

These matching features or their corresponding quasi-molecular ion were subjected to fragmentation by MS/MS with 10, 20 and 30 eV. Stop codon patterns were derived from multiple sequence alignments of AraCyc enzyme genes[45] (ftp.plantcyc.org/Pathways/BLAST_sets/aracyc_enzymes.fasta, Dec 2015) as annotated by TAIR10_functional annotations from TAIR.org[46].

References

1. Weigel, D. Natural variation in Arabidopsis: from molecular genetics to ecological genomics. *Plant Physiol.* **158,** 2–22 (2012).
2. Mithen, R., Clarke, J. H., Lister, C. & Dean, C. Genetics of aliphatic glucosinolates. III. Side chain structure of aliphatic glucosinolates in Arabidopsis thaliana. *Heredity (Edinb).* **74,** 210–215 (1995).
3. Mithen, R. & Campos, H. Genetic variation of aliphatic glucosinolates in Arabidopsis thaliana and prospects for map based gene cloning. *Entomologica Experimentalis et Applicanta*. **53,** 202–205 (1996).
4. Magrath, R. *et al.* Genetics of aliphatic glucosinolates. I. Side chain elongation in Brassica napus and Arabidopsis thaliana. *Heredity (Edinb)*. **72,** 290–299 (1994).
5. Mitchell-Olds, T. & Pedersen, D. The molecular basis of quantitative genetic variation in central and secondary metabolism in Arabidopsis. *Genetics*. **149,** 739–747 (1998).
6. Kliebenstein, D. J. *et al.* Genetic control of natural variation in Arabidopsis glucosinolate accumulation. *Plant Physiol.* **126,** 811–825 (2001).
7. Houshyani, B. *et al.* Characterization of the natural variation in Arabidopsis thaliana metabolome by the analysis of metabolic distance. *Metabolomics*. **8,** 131–145 (2012).
8. Agrawal, B., Lakshmanan, V., Kaushik, S. & Bais, H. P. Natural variation among Arabidopsis accessions reveals malic acid as a key mediator of Nickel (Ni) tolerance. *Planta*. **236,** 477–489 (2012).
9. Bais, H. P., Weir, T. L., Perry, L. G., Gilroy, S. & Vivanco, J. M. The role of root exudates in rhizosphere interactions with plants and other organisms. *Annu Rev Plant Biol.* **57,** 233–266 (2006).
10. Schmid, N. B. *et al.* Feruloyl-CoA 6′-Hydroxylase1-dependent coumarins mediate iron acquisition from alkaline substrates in Arabidopsis. *Plant Physiol.* **164,** 160–172 (2014).
11. van de Mortel, J. E. *et al.* Metabolic and transcriptomic changes induced in Arabidopsis by the rhizobacterium Pseudomonas fluorescens SS101. *Plant Physiol.* **160,** 2173–2188 (2012).
12. Hiltner, L. Über neue Erfahrungen und Probleme auf dem Gebiete der Bodenbakteriologie. *Arbeiten der Deutschen Landwirtschaft Gesellschaft*. **98,** 59–78 (1904).
13. Bulgarelli, D., Schlaeppi, K., Spaepen, S., Ver Loren van Themaat, E. & Schulze-Lefert, P. Structure and functions of the bacterial microbiota of plants. *Annu Rev Plant Biol.* **64,** 807–838 (2013).
14. Bressan, M. *et al.* Exogenous glucosinolate produced by Arabidopsis thaliana has an impact on microbes in the rhizosphere and plant roots. *ISME J.* **3,** 1243–1257 (2009).
15. Strehmel, N., Böttcher, C., Schmidt, S. & Scheel, D. Profiling of secondary metabolites in root exudates of Arabidopsis thaliana. *Phytochemistry*. **108C,** 35–46 (2014).
16. Micallef, S. A., Shiaris, M. P. & Colon-Carmona, A. Influence of Arabidopsis thaliana accessions on rhizobacterial communities and natural variation in root exudates. *J Exp Bot*. **60,** 1729–1742 (2009).
17. Böttcher, C., von Roepenack-Lahaye, E. & Scheel, D. Resources for metabolomics in *Genetics and Genomics of the Brassicaceae* Vol. 9 *Plant Genetics and Genomics: Crops and Models* (eds Bancroft, I. & Schmidt, R.) 469–503 (Springer, 2011).
18. Kover, P. X. *et al.* A Multiparent Advanced Generation Inter-Cross to fine-map quantitative traits in Arabidopsis thaliana. *PLoS Genet*. **5,** e1000551 (2009).
19. Nordborg, M. *et al.* The pattern of polymorphism in Arabidopsis thaliana. *PLoS Biol*. **3,** e196 (2005).
20. Clark, R. M. *et al.* Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. *Science*. **317,** 338–342 (2007).
21. Editorial. How robust are your data? *Nature Cell Biology*. **11** (2009).
22. Gan, X. *et al.* Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature*. **477,** 419–423 (2011).
23. Wiesner, M., Schreiner, M. & Zrenner, R. Functional identification of genes responsible for the biosynthesis of 1-methoxy-indol-3-ylmethyl-glucosinolate in Brassica rapa ssp. chinensis. *BMC Plant Biol*. **14,** 124 (2014).
24. Pfalz, M. *et al.* Metabolic engineering in Nicotiana benthamiana reveals key enzyme functions in Arabidopsis indole glucosinolate modification. *Plant Cell*. **23,** 716–729 (2011).
25. Keurentjes, J. J. *et al.* The genetics of plant metabolism. *Nat Genet*. **38,** 842–849 (2006).
26. Obayashi, T. *et al.* ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis. *Nucleic Acids Res*. **35,** D863–869 (2007).
27. De Pessemier, J., Chardon, F., Juraniec, M., Delaplace, P. & Hermans, C. Natural variation of the root morphological response to nitrate supply in Arabidopsis thaliana. *Mech Dev*. **130,** 45–53 (2013).
28. Trontin, C., Tisne, S., Bach, L. & Loudet, O. What does Arabidopsis natural variation teach us (and does not teach us) about adaptation in plants? *Curr Opin Plant Biol*. **14,** 225–231 (2011).
29. Ziegler, J. *et al.* Simultaneous analysis of apolar phytohormones and 1-aminocyclopropan-1-carboxylic acid by high performance liquid chromatography/electrospray negative ion tandem mass spectrometry via 9-fluorenylmethoxycarbonyl chloride derivatization. *J Chromatogr A*. **1362,** 102–109 (2014).
30. Li, X. *et al.* Exploiting natural variation of secondary metabolism identifies a gene controlling the glycosylation diversity of dihydroxybenzoic acids in Arabidopsis thaliana. *Genetics*. **198,** 1267–1276 (2014).
31. Negishi, M. *et al.* Structure and function of sulfotransferases. *Arch Biochem Biophys*. **390,** 149–157 (2001).
32. Stracke, R. *et al.* Differential regulation of closely related R2R3-MYB transcription factors controls flavonol accumulation in different parts of the Arabidopsis thaliana seedling. *Plant J*. **50,** 660–677 (2007).
33. Jones, P., Messner, B., Nakajima, J., Schaffner, A. R. & Saito, K. UGT73C6 and UGT78D1, glycosyltransferases involved in flavonol glycoside biosynthesis in Arabidopsis thaliana. *J Biol Chem*. **278,** 43910–43918 (2003).
34. Tautenhahn, R., Böttcher, C. & Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics*. **9,** 504 (2008).
35. Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem*. **78,** 779–787 (2006).
36. Böttcher, C. *et al.* The multifunctional enzyme CYP71B15 (PHYTOALEXIN DEFICIENT3) converts cysteine-indole-3-acetonitrile to camalexin in the indole-3-acetonitrile metabolic network of Arabidopsis thaliana. *Plant Cell*. **21,** 1830–1845 (2009).
37. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. B* **57,** 289–300 (1995).
38. R: A Language and Environment for Statistical Computing (2014).
39. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. **5,** R80 (2004).

40. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*. **3,** 1724–1735 (2007).
41. Suzuki, R. & Shimodaira, H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*. **22,** 1540–1542 (2006).
42. BSgenome.Athaliana.TAIR.TAIR9: Full genome sequences for Arabidopsis thaliana (TAIR9).
43. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput Biol*. **9,** e1003118 (2013).
44. Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R. & Neumann, S. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal Chem*. **84,** 283–289 (2012).
45. Mueller, L. A., Zhang, P. & Rhee, S. Y. AraCyc: a biochemical pathway database for Arabidopsis. *Plant Physiol*. **132,** 453–460 (2003).
46. Huala, E. *et al.* The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res*. **29,** 102–105 (2001).
47. Luo, J. *et al.* A novel polyamine acyltransferase responsible for the accumulation of spermidine conjugates in Arabidopsis seed. *Plant Cell*. **21,** 318–333 (2009).

Acknowledgements

We thank Dr. Ralf Stracke for providing seeds. This work was supported by the Joint Initiative for Research and Innovation of the Leibniz Association (SAW-2011-IPB-3 97, "Chemical Communication in the Rhizosphere"). Metabolomics expertise by Dr. Christoph Böttcher und expert technical assistance by Sylvia Krüger, Jessica Thomas and Susanne Kirsten are gratefully acknowledged. The publication of this article was funded by the Open Access fund of the Leibniz Association.

Author Contributions

D.S. designed the project. The hydroponic system was designed by D.S., N.S. and S.S. The genetic model was designed by S.M., E.M., S.S., L.W. and S.N. N.S., S.S. and S.M. performed measurement and general data analysis. Non-targeted profiling was analysed by N.S., F.T. and S.M.; the targeted analysis was performed by N.S., S.H. and S.M. S.N. submitted the data to MetaboLights. Expertise and proofreading was provided by L.W., D.S., N.S., S.S. and S.N. S.M. and N.S. structured and wrote the manuscript. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

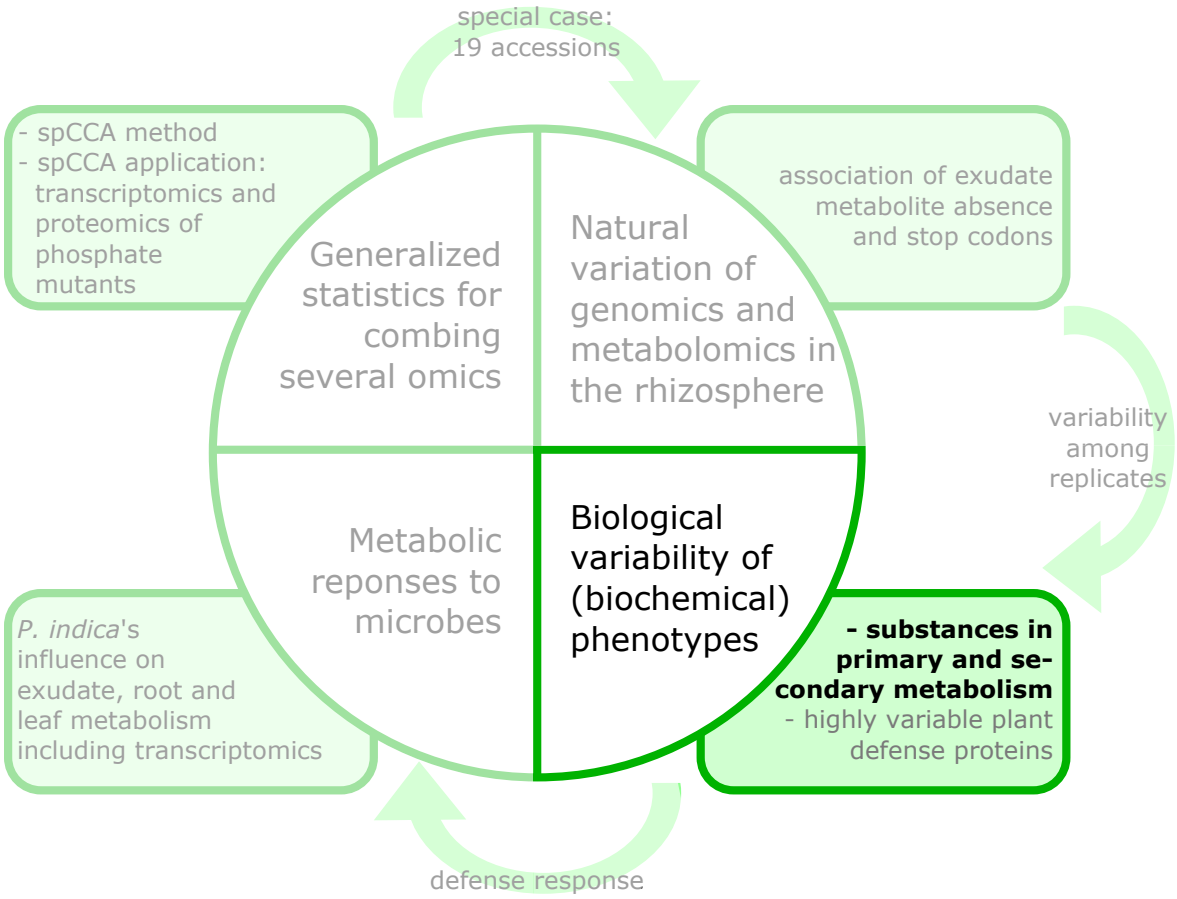
How to cite this article: Mönchgesang, S. *et al.* Natural variation of root exudates in *Arabidopsis thaliana*-linking metabolomic and genomic data. *Sci. Rep.* **6**, 29033; doi: 10.1038/srep29033 (2016).

2.3 Biological variability of biochemical phenotypes

2.3.1 Plant-to-plant variability in root metabolite profiles of 19 *Arabidopsis thaliana* accessions is substance-class-dependent

Mönchgesang, S.; Strehmel, N.; Trutschel, D.; Westphal, L.; Neumann, S.; Scheel, D. Plant-to-Plant Variability in Root Metabolite Profiles of 19 *Arabidopsis thaliana* Accessions Is Substance-Class-Dependent. *Int J Mol Sci* **2016**, *17*.

equal contributions



International Journal of
Molecular Sciences

MDPI

Communication

Plant-to-Plant Variability in Root Metabolite Profiles of 19 *Arabidopsis thaliana* Accessions Is Substance-Class-Dependent

Susann Mönchgesang [1,*,†], Nadine Strehmel [1,†], Diana Trutschel [1,2,3], Lore Westphal [1], Steffen Neumann [1] and Dierk Scheel [1,*]

1 Leibniz Institute of Plant Biochemistry, Stress and Developmental Biology, Weinberg 3, 06120 Halle (Saale), Germany; nadine.strehmel@ipb-halle.de (N.S.); diana.trutschel@ipb-halle.de (D.T.); lore.westphal@ipb-halle.de (L.W.); steffen.neumann@ipb-halle.de (S.N.)
2 Institute of Computer Science, Martin-Luther-University Halle-Wittenberg, Von-Seckendorff-Platz 1, 06120 Halle (Saale), Germany
3 German Center for Neurodegenerative Disesaes, Stockumer Straße 12, 58453 Witten, Germany
* Correspondence: susann.moenchgesang@ipb-halle.de (S.M.); dierk.scheel@ipb-halle.de (D.S.); Tel.: +49-345-5582-1475 (S.M.); +49-345-5582-1400 (D.S.)
† These authors contributed equally to this work.

Academic Editor: Ute Roessner
Received: 30 June 2016; Accepted: 12 September 2016; Published: 16 September 2016

Abstract: Natural variation of secondary metabolism between different accessions of *Arabidopsis thaliana* (*A. thaliana*) has been studied extensively. In this study, we extended the natural variation approach by including biological variability (plant-to-plant variability) and analysed root metabolic patterns as well as their variability between plants and naturally occurring accessions. To screen 19 accessions of *A. thaliana*, comprehensive non-targeted metabolite profiling of single plant root extracts was performed using ultra performance liquid chromatography/electrospray ionization quadrupole time-of-flight mass spectrometry (UPLC/ESI-QTOF-MS) and gas chromatography/electron ionization quadrupole mass spectrometry (GC/EI-QMS). Linear mixed models were applied to dissect the total observed variance. All metabolic profiles pointed towards a larger plant-to-plant variability than natural variation between accessions and variance of experimental batches. Ratios of plant-to-plant to total variability were high and distinct for certain secondary metabolites. None of the investigated accessions displayed a specifically high or low biological variability for these substance classes. This study provides recommendations for future natural variation analyses of glucosinolates, flavonoids, and phenylpropanoids and also reference data for additional substance classes.

Keywords: LC/MS; GC/MS; *Arabidopsis*; secondary metabolism; natural variation; individual variability; metabolite profiling

1. Introduction

Metabolomics is one of the "-omics" disciplines in plant science. With the help of hyphenated techniques such as gas chromatography coupled to mass spectrometry (GC/MS) or liquid chromatography-coupled mass spectrometry (LC/MS), a large spectrum of small molecules within a plant can be analysed. *Arabidopsis thaliana* (*A. thaliana*) is a model species to investigate secondary metabolic pathways. Naturally occurring accessions and their distinct phenotypes have evolved in different habitats and full genome sequencing revealed a substantial number of single nucleotide polymorphisms [1]. Compared to seeds and shoots, root metabolism is not as well investigated, but in plants it is crucial in order to provide the molecular building blocks for physical anchorage in the ground and to regulate all belowground processes. By root exudation, plants also communicate with

their surrounding rhizosphere and soil microorganisms. In general, due to the relatively low biomass of *Arabidopsis*, especially in roots, material of several plants is pooled before sample preparation. With increasing sensitivity and decreasing costs of analytical techniques, pooling does not seem to be technically necessary anymore. Indeed, in some cases it is interesting to focus on individual variability to investigate which mechanisms determine plant metabolism without stress exposure. Once the plant material is pooled, the information on individual plants is irreversibly lost. Vice versa, smart experimental design allows for both—investigating variances on different levels (replicates) and detecting differences between accessions.

Several metabolomics studies examined the contribution of different variance sources to the total observed variance [2,3]. For nuclear magnetic resonance (NMR) metabolomics, Lewisetal et al. [2] found that extraction and instrumental deviations accounted for less than 10% and 1%, respectively, of the total variance in leaves of the accession L*er*-0. The substantial plant-to-plant variability of 52% in L*er*-0 could be reduced by pooling several plants to facilitate the separation of L*er*-0 from Col-0 samples. Reducing biological variability by pooling might allow for the fast detection of the effect of interest but nevertheless, it might miss subtle between-plant effects. Similar trends for extraction and instrumental variance were observed in comprehensive LC/MS-based metabolomics studies of Col-0 shoots [3]. Trutschel et al. [3] also provide a solution for how to incorporate different kinds of replicates into a powerful experimental design without the need for sample pooling.

Previous studies have investigated plant-to-plant variability during leaf development. The area of leaf six varied substantially between plants of the isogenic accession Col-0 at the same developmental stage, and this variability seems to converge in mature leaves [4]. Li et al. [5] determined there was 33%–40% plant-to-plant variability between the oil content of Col-0 seeds, and pointed out that this fact needs to be considered to draw statistically valid conclusions.

Plant-to-plant variability has neither been investigated in root metabolism nor have previous studies incorporated more than two *A. thaliana* accessions into a comprehensive root metabolic profiling analysis. Here, we analysed root metabolic profiles of 19 accessions, which were the founders of the multiparent advanced generation inter-cross (MAGIC) collection of *A. thaliana* [1,6], using a single-plant setup in a hydroponic system.

The aim of this study was to decompose the total variance of root metabolite profiles observed in untreated plants into the components attributable to (1) natural variation between accessions; (2) experimental batch; and (3) individual variability between plants. Furthermore, we investigated the relative biological variability of three important substance classes: glucosinolates (GSLs), flavonoids, and phenylpropanoids including oligolignols which seem to play a vital role in root (but not shoot) metabolism. Following the analysis of 19 accessions in their entirety, the variability of each accession was analysed to identify any particular highly or lowly variable accessions.

2. Results

2.1. Variability between Plants Is a Greater Source of Variance than Natural Variation between Accessions

Many studies on natural variation are primarily interested in differences between the accessions, and reduce plant-to-plant variability by pooling material to obtain fast results. However, to obtain a comprehensive picture of variability, the variance at each level of the experimental design should be incorporated.

The experimental setup of our study, shown in Figure 1, resulted in 222 single-plant LC/MS measurements in each electrospray ionization (ESI) mode. The alignment of chromatograms and spectra over 222 samples was performed, deviations in retention time (RT) and mass-to-charge ratio (m/z) were small across all samples (Figure S1) reflecting a sufficient quality of the measurements to analyse the effects of accession, experimental batch, and individual plant. Linear mixed models with all experimental levels as random effects were applied to decompose the total metabolic variance.

Int. J. Mol. Sci. **2016**, *17*, 1565

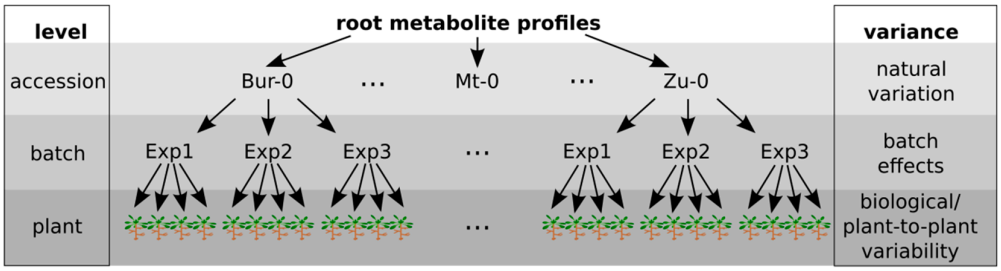


Figure 1. Nested experimental design with three levels. Each variance level had multiple replicates—to assess natural variation, 19 accessions of *Arabidopsis thaliana* (*A. thaliana*) were grown. Three independent biological experiments were performed to estimate non-biological variance derived from the experimental batch. To assess individual variability, four plants were harvested in each biological experiment for each accession. Single-plant root extracts were subjected to liquid chromatography-coupled mass spectrometry (LC/MS) and gas chromatography-coupled mass spectrometry (GC/MS) analysis.

The non-targeted metabolic profiles of the 19 accessions indicated that the between-accession variance is smaller than the plant-to-plant-variability over all features. The results for ESI(−) are shown in Figure 2a and for ESI(+) in Supplementary Figure S2.

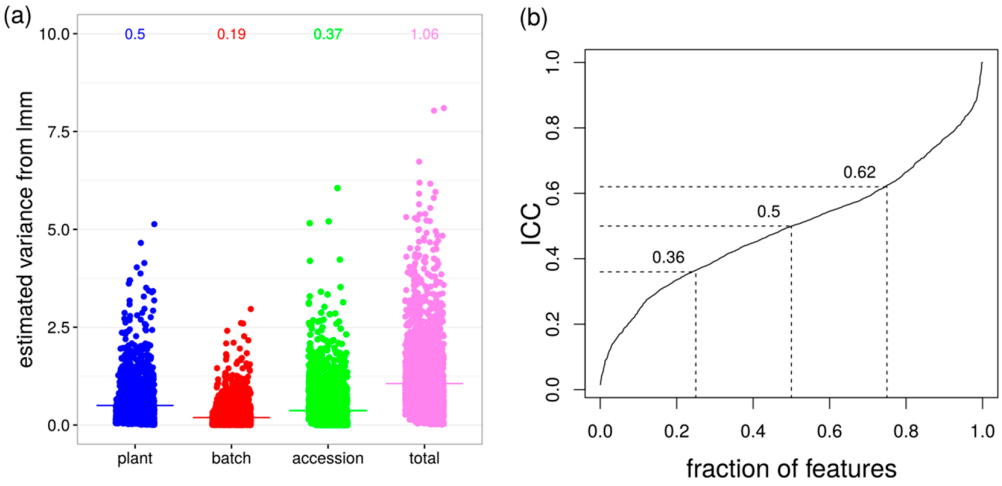


Figure 2. Variance decomposition of LC/electrospray ionization (ESI)(−) MS data set. (**a**) Variances for plant, batch and accession were estimated with a linear mixed model (lmm), dot—variance of one feature, bar and number—mean variance over 2730 features; (**b**) cumulative intraclass correlation (ICC) distribution for all features ($\sigma^2_{plant}/\sigma^2_{total}$), dotted lines indicate 25%, 50% and 75% quantiles.

The mean between-plant variance $\sigma^2_{plant} = 0.50$ is 20% larger than the between-accession variance $\sigma^2_{accession} = 0.37$. The estimated mean between-experiment variation $\sigma^2_{batch} = 0.19$ is less than 40% of σ^2_{plant}. On average, plant-to-plant variability contributes to approximately half of the total variance ($\sigma^2_{plant}/\sigma^2_{total} = 0.47$). However, this biological variance has to be interpreted in the context of the total variance for comparisons across features and platforms, i.e., knowing whether the feature with the highest σ^2_{plant} also exhibits large σ^2_{total}. It may also occur that a feature with high σ^2_{plant} has low σ^2_{total}, which determines the experimental design to include more replicates on the plant level in a potential validation study.

The intraclass correlation (ICC) according to Sampson et al. [7], here $\sigma^2_{plant}/\sigma^2_{total}$, reflects which fraction of total variance is attributable to the single plant and thus, a relative biological variability.

The mean ICC ≈ 0.5 of a data set could either be representative for the majority of features (narrow interquartile range) or only for a few features if the interquartile range is broad. Figure 2b shows the cumulative ICC distribution over all features, with the fraction of features (*x*-axis) in increasing ICC (*y*-axis) order. The distribution revealed that 25%, 50%, and 75% of all these features had an ICC up to 0.36, 0.50, and 0.62. This implies that for half of the features, the plant-to-plant variability contributes to less than 50% to the total variance, and for the other half this variance level explains more than 50% of the total variance. In summary, in our non-targeted analysis of root metabolic natural variation, plant-to-plant variability seems to be larger than between-accession variance. If a broad range of metabolites are of interest, it is important to know the biological variability that is exhibited by most metabolites. If only a small subset of the non-targeted analysis is in research focus, it will be sufficient to deal with the biological variability of a certain substance class.

2.2. Plant-to-Plant Variability in Secondary Metabolism Is Substance-Class-Dependent, but Not Accession-Specific

A difficulty in non-targeted metabolomics is the assignment of the measured features to metabolites and their potential role in pathways in a living system. To facilitate the interpretation of plant-to-plant variability, three sets of annotatable compounds were quantified by integrating peak areas of the extracted ion chromatograms and analysed for their variances at each level (Table S1). In Figure 3, GSLs, flavonoids, and phenylpropanoids are indicated by circles, triangles, and squares, respectively. GSLs were the substance class with the highest plant-to-plant variability (σ^2_{plant} = 3.16, Figure 3a left, circles) compared to flavonoids and phenylpropanoids. They also showed a large deviation of the single metabolite plant variance from the mean of the substance class. Similarly, σ^2_{total} = 5.03 was highest for GSLs in the comparison to flavonoids (σ^2_{plant} = 1.63, σ^2_{total} = 2.60) and phenylpropanoids (σ^2_{plant} = 1.24, σ^2_{total} = 2.88).

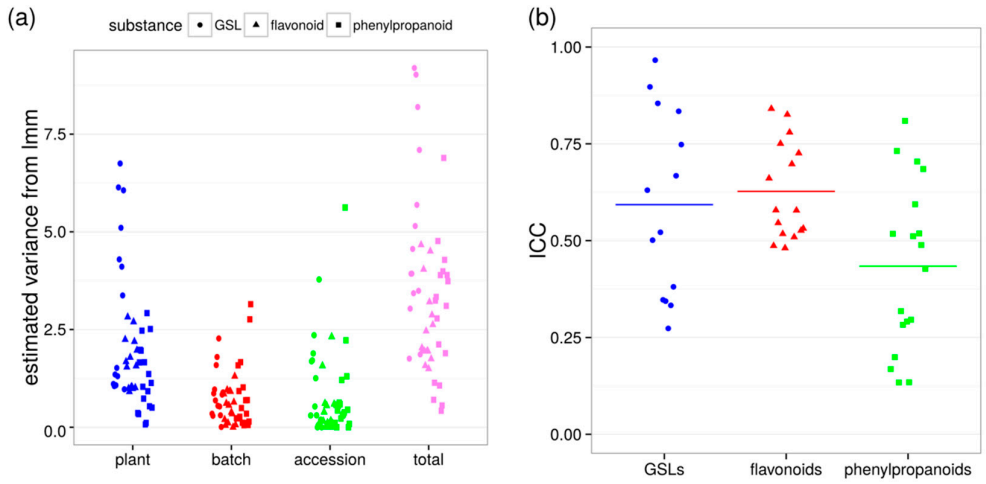


Figure 3. Biological variability of annotated secondary metabolites. (**a**) Variances for plant, batch and accession were estimated with a linear mixed model (lmm), dot—variance of one metabolite; (**b**) ICCs for glucosinolates (GSLs), flavonoids, and phenylpropanoids, dot—ICC of one metabolite, bar—mean ICC for substance class.

With the current experimental setup of four plants in three batches for a total of 12 plants per accession, the minimal detectable log fold-change to distinguish between two accessions is 3.94, 2.97 and 3.24 for glucosinolates, flavonoids, and phenylpropanoids, respectively, with a power of 0.8 and a significance level of 0.05. However, plant-to-plant variability needs to be interpreted in the context of total variance to find out at which experimental level the main observation is made.

If $\sigma^2_{plant} \approx \sigma^2_{total}$, nearly all of the total variance would be caused by plant-to-plant variability and a large number of plants would be required to analyse effects beyond this experimental level, i.e., between accessions. If $\sigma^2_{plant}/\sigma^2_{total} \approx 0$, it would be sufficient to use one plant per accession. Glucosinolates and phenylpropanoids show a large range of ICCs. For flavonoid metabolites, the ICCs are rather high but similar for all analysed members of the substance class (Figure 3b). Hence, calculations with the mean ICCs like above will provide sufficient power for analyses of flavonoids, but not for all metabolites of the classes glucosinolates and phenylpropanoids.

A set of primary metabolites was also analysed for their plant-to-plant variability (Table S2) but, in comparison to secondary metabolism, the ICC distributions of carbohydrates, organic acids, amino acids, and phosphates covered a large range (Figure S3). As expected, the primary metabolism is more stable than secondary metabolism, the latter showing substance-class specific ICC distributions.

Until here, we assumed all accessions to have equal variances at the plant and batch level. In addition, we analysed if the accessions differ with regard to their plant-to-plant variability. For this purpose, linear mixed models were applied to estimate the variances of secondary metabolites for each accession separately. As shown in Figure S4, there are no clear highly and lowly variable accessions across the measured substance classes. However, Edi-0 showed relatively low ICCs for GSLs and flavonoids. Hi-0 and Sf-2 showed higher ICCs for all three compound classes.

In our analysis, taking the ICCs of secondary metabolite classes into consideration seems to be more important than the selection of accessions.

3. Discussion

Our study investigated natural variation and plant-to-plant variability of 19 key accessions in a comprehensive metabolite profiling approach. Measuring single plant extracts prevented the irreversible information loss resulting from pooling plant material and allows to distinguish between accessions and still analyse plant-to-plant variability. Environmental variation was kept to a minimum by a randomized growth regimen and selecting plants with approximately the same vigor for analyses. Both non-targeted LC/MS ionization modes indicated a higher plant-to-plant variability than natural variation between accessions and variance due to experimental batches. Plant-to-plant variability contributed to 47%–50% of the total variance, which is higher than previously reported for one particular compound class in seeds of one accession [5]. As our total variance was the sum of plant, batch and accession variance, the ICCs referring to the sum of plant and batch variance, like in the oil seed study [5], would have been larger.

Furthermore, we chose a range of secondary and primary metabolite classes for more specific analyses. Both data sets indicated that the plant-to-plant variability had the greatest contribution to the total variance of these metabolite classes. For GSLs, flavonoids and phenylpropanoids, the means of σ^2_{batch} and $\sigma^2_{accession}$ were in the same order of magnitude, whereas for primary metabolite sets $\sigma^2_{accession}$ was less pronounced with values one order of magnitude below σ^2_{batch}. The minimal detectable effects were quite large and impractical with the given experimental setup of three experiments with four plants each. Possible combinations of biological and technical replicates to reliably detect a smaller effect can be calculated with the implementation provided by Trutschel et al. [3]. All annotated substance classes displayed higher mean ICCs than the non-targeted data sets they were derived from. The higher the fraction of features with high ICCs, the higher the number of plants that is required to maintain the power in a statistical analysis. This should be taken into consideration for future experimental designs. Flavonoid metabolites have similar ICCs within their substance class and therefore, calculation with mean ICC of the substance class will be sufficient to obtain reliable results for most metabolites in this class. Contrarily, GSLs and phenylpropanoids displayed a large ICC spread and require a substance-specific estimation of variance prior to future analyses. A previous study of root exudates has demonstrated that there are substance-specific differences in some metabolite classes due to alterations in the biosynthetic pathways [8]. Since some metabolites are specifically induced during stress response, they might not have been expressed in

the unperturbed physiological state that was the focus of this study. The analysis of plant-to-plant variability in each accession revealed that ICC distributions are not distinct for any of the 19 accessions with the few exceptions of Edi-0, Hi-0, and Sf-2. However, our set of 19 accessions is too small to draw a general conclusion about accession-specific plant-to-plant variability and more accessions have to be analysed in future.

There are hints that biological variability converges after development [4] and upon exposure to stress factors [9,10]. A study of *Arabidopsis* plants exposed to a biotic stress factor, namely the endophytic fungus *Piriformospora indica*, showed substantial metabolic variability in untreated control samples and only a small spread of co-cultivated samples in principal component analyses. These samples were no single plant measurements but the batch variances in both sample classes were identical and thus, the observed deviation is expected to result from plant-to-plant variability [9]. Töpfer et al. [10] found that upon abiotic stress treatment, certain metabolites were robust in their abundance from plant to plant and displayed low coefficients of variation, whereas other metabolites showed larger plant-to-plant variability.

For future natural variation studies, it might be worth considering measuring single plants and make the data available for further analyses answering research questions on a different experimental level. We have provided estimated variances for selected substances in Supplementary Tables S1 and S2. Furthermore, we provide exemplary data and the functions in an R script for variance estimation in the Supplementary Folder S1 as well as data for additional substance classes in the targeted analysis in MTBLS338 in the MetaboLights repository. This knowledge can be exploited to appropriately design an experiment prior to its conduction because it may differ between a non-targeted screen and the analysis of specific substance classes.

4. Materials and Methods

4.1. Plant Cultivation

The *A. thaliana* accessions Bur-0, Can-0, Col-0, Ct-1, Edi-0, Hi-0, Kn-0, L*er*-0, Mt-0, No-0, Oy-0, Po-0, Rsch-4, Sf-2, Tsu-0, Wil-2, Ws-0, Wu-0, and Zu-0 were obtained as seeds from the European Arabidopsis Stock Centre (Nottingham, UK) and surface sterilized prior to plant cultivation. All accessions were cultivated in a hydroponic system under 8 h light and 22 °C as described previously [11] and in the protocol section of MTBLS338 with four plants in each of the three independent biological experiments. All samples were rotated in the growth chamber to minimize position effects. Primary root length and root fresh weight are given in MTBLS338. Out of 228 root samples, 210 and 222 from individual plants could be used for the GC/MS and LC/MS analysis, respectively.

4.2. Liquid Chromatography/Mass Spectrometry (LC/MS)

For LC/MS analysis, 40 mg root material were extracted in 200 µL 80% methanol/water (v/v) twice according to Böttcher et al. [12] and reconstituted in 30% methanol (v/v) containing 5 µM 2,4-dichlorophenoxyacetic acid as an internal standard. Upon full loop injection into an Acquitiy UPLC system (Waters, Eschborn/Germany) mounted with a HSS T3 column (100 × 1.0 mm, 1.8 µM particle size), samples were separated at a flow rate of 150 µL/min with mixtures of A (water/0.1% formic acid) and B (acetonitrile/0.1% formic acid) with a 20 min gradient: 0–1 min isocratic 95% A, 5% B; 1–16 min linear 5%–95% B; 16–18 min isocratic 95% B; 18–18.01 min linear 95%–5% B; 18.01–20 min isocratic 5% B. Eluates were ionized using an Apollo II source (Bruker Daltonics, Billerica, MA, USA) into a MicroTOF-Q I hybrid quadrupole time-of-flight mass analyzer (Bruker Daltonics) in both ionization modes with a mass range m/z 80–1000. Mass spectrometry settings were applied as previously described [11] and in the protocol section of MTBLS338.

All LC/MS runs were acquired as centroid spectra and recalibrated with lithium formate cluster ions for each measurement. Vendor .d file formats were converted into the open standard mzData with CompassXPort (Bruker Daltonics, Billerica, MA, USA).

4.3. Gas Chromatography/Mass Spectrometry (GC/MS)

For GC/MS analysis, 40 µL of the root extract were vacuum-evaporated and subjected to a derivatization with (1) methoxyamine hydrochloride and (2) *N*,*O*-bis(trimethylsilyl)-trifluoroacetamide as previously described [13]. Derivatized samples were injected in a splitless manner into a split/splitless inlet of an Agilent 6890N GC and a ZB-5 column (30 m × 0.25 mm, 0.25 m 95% dimethyl/5% diphenyl polysiloxane film, 10 m integrated guard column, Phenomenex, Aschaffenburg, Germany) at 230 °C. An Agilent 5975 Series Mass Selective Detector (Agilent Technologies, Waldbronn, Germany) was used to detect eluting compounds from m/z 70 to 600. Vendor file format conversion and baseline correction was performed by MetAlign [14].

4.4. Data Analysis

Statistical analysis was performed using R version 3.2.0 and the Bioconductor environment [15,16]. Functions are available as an R script in the Supplementary Folder S1.

4.4.1. Raw Data Processing

All LC/MS data analysis was performed with the R packages XCMS and CAMERA [17–19]. Features were extracted with centWave (snthr = 10, ppm = 20, peakwidth = c(5,12), scanrange = c(1,3600)) and grouped (minfrac = 0.75, bw = 5, mzwid = 0.05), corrected for retention shifts and re-grouped with smaller bandwidth (bw = 2). Missing values were imputed by integration of raw data (fillPeaks) and with random numbers around the minimal intensity value across the samples.

Baseline-corrected GC/MS tags with intensities above 500 peak height were subsequently processed with TagFinder [20] and mass spectral features were grouped according to their common retention time. Clusters with at least 3 correlating tags were extracted and identified according to matching the Golm Metabolome Database [21]. In GC/MS, 15,539 tags were detected and 98 metabolites were annotated (Table S3).

All data were log-transformed to approximate a normal distribution for further statistics.

4.4.2. Targeted LC/MS Analysis

For the targeted analysis, DataAnalysis 4.2 (Bruker Daltonics, Billerica, MA, USA) was used to extract ion chromatograms, deconvolute mass spectra and determine the elemental composition. Peak areas (minimum peak area = 500) of extracted ion chromatograms were integrated with QuantAnalysis 2.0 (Bruker Daltonics, Billerica, MA, USA) to quantify compound abundances with quasi-molecular ions as listed in Table S4 [11,22]. In the LC/MS measurements, 3305 peaks ESI(+) and 2730 peaks ESI(−) were detected and all together 139 compounds could be annotated.

4.4.3. Variance Estimation with Linear Mixed Models

A linear mixed model (R package lme4, version 1.1-11, [23]) with accession, batch and plant as random effects was applied to log-transformed metabolite abundances to estimate variance contribution of each experimental level assuming equal variances for each accession. Linear mixed models with batch and plant as random effects were applied separately to each accession to examine accession-specific variances. Intraclass correlations (ICCs) were calculated as the ratio of σ^2_{plant} and σ^2_{total} according to Sampson et al. [7] and plotted as a cumulative distribution. Further analysis was constrained to known metabolites to allow for a better interpretation. The minimal detectable effect sizes were estimated with the power calculations for multilevel experiments [3].

4.5. Data Availability

All data sets including the targeted analyses are available from the MetaboLights repository under the accession number MTBLS338 [24].

5. Conclusions

This study investigated the variability in root metabolite profiles of 19 *A. thaliana* accessions. It revealed that plant-to-plant variability can be a substantial component of the overall variability in a natural variation analysis. Additionally, several selected substance classes were characterized by differing intraclass correlations. To exploit the full potential of a non-targeted metabolite profiling, single-plant measurements should be acquired and correctly integrated into the analysis. Hence, different substance classes of interest might require a customised experimental set-up.

Supplementary Materials: Supplementary materials can be found at www.mdpi.com/1422-0067/17/9/1565/s1.

Acknowledgments: The authors thank Christoph Böttcher and Stephan Schmidt for metabolomics advice. Expert technical assistance by Sylvia Krüger, Julia Göhricke, Siska Herklotz and Jessica Thomas is gratefully acknowledged.

Author Contributions: Dierk Scheel conceived the study. Dierk Scheel, Steffen Neumann, Lore Westphal and Nadine Strehmel supervised the study. Nadine Strehmel and Susann Mönchgesang designed and performed the experiments and analysed the data, Diana Trutschel performed statistical analysis. Susann Mönchgesang wrote the manuscript. Lore Westphal provided advice on natural variation and critical manuscript feedback. Steffen Neumann submitted the data to MetaboLights. All authors read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

EI	electron ionization
ESI	electrospray ionization
GC/MS	gas chromatography/mass spectrometry
GSL	glucosinolate
ICC	intraclass correlation
LC/MS	liquid chromatography/mass spectrometry

References

1. Gan, X.; Stegle, O.; Behr, J.; Steffen, J.G.; Drewe, P.; Hildebrand, K.L.; Lyngsoe, R.; Schultheiss, S.J.; Osborne, E.J.; Sreedharan, V.T.; et al. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* **2011**, *477*, 419–423. [CrossRef] [PubMed]
2. Lewis, J.; Baker, J.M.; Beale, M.H.; Ward, J.L. Metabolite Profiling of GM Plants—The importance of robust experimental design and execution. In *Genomics for Biosafety in Plant Biotechnology*; Nap, J.-P., Atanassov, A., Stiekema, W.J., Eds.; IOS Press: Amsterdam, The Netherlands, 2004.
3. Trutschel, D.; Schmidt, S.; Grosse, I.; Neumann, S. Experiment design beyond gut feeling: Statistical tests and power to detect differential metabolites in mass spectrometry data. *Metabolomics* **2015**, *11*, 851–860. [CrossRef]
4. Granier, C.; Massonnet, C.; Turc, O.; Muller, B.; Chenu, K.; Tardieu, F. Individual leaf development in *Arabidopsis thaliana*: A stable thermal-time-based programme. *Ann. Bot.* **2002**, *89*, 595–604. [CrossRef] [PubMed]
5. Li, Y.; Beisson, F.; Pollard, M.; Ohlrogge, J. Oil content of *Arabidopsis* seeds: The influence of seed anatomy, light and plant-to-plant variation. *Phytochemistry* **2006**, *67*, 904–915. [CrossRef] [PubMed]
6. Kover, P.X.; Valdar, W.; Trakalo, J.; Scarcelli, N.; Ehrenreich, I.M.; Purugganan, M.D.; Durrant, C.; Mott, R. A Multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. *PLoS Genet.* **2009**, *5*, e1000551. [CrossRef] [PubMed]
7. Sampson, J.N.; Boca, S.M.; Shu, X.O.; Stolzenberg-Solomon, R.Z.; Matthews, C.E.; Hsing, A.W.; Tan, Y.T.; Ji, B.T.; Chow, W.H.; Cai, Q.; et al. Metabolomics in epidemiology: Sources of variability in metabolite measurements and implications. *Cancer Epidemiol. Biomark. Prev.* **2013**, *22*, 631–640. [CrossRef] [PubMed]
8. Mönchgesang, S.; Strehmel, N.; Schmidt, S.; Westphal, L.; Taruttis, F.; Muller, E.; Herklotz, S.; Neumann, S.; Scheel, D. Natural variation of root exudates in *Arabidopsis thaliana*—Linking metabolomic and genomic data. *Sci. Rep.* **2016**, *6*, 29033. [CrossRef] [PubMed]
9. Strehmel, N.; Mönchgesang, S.; Herklotz, S.; Kruger, S.; Ziegler, J.; Scheel, D. *Piriformospora indica* Stimulates Root Metabolism of *Arabidopsis thaliana*. *Int. J. Mol. Sci.* **2016**, *17*. [CrossRef] [PubMed]

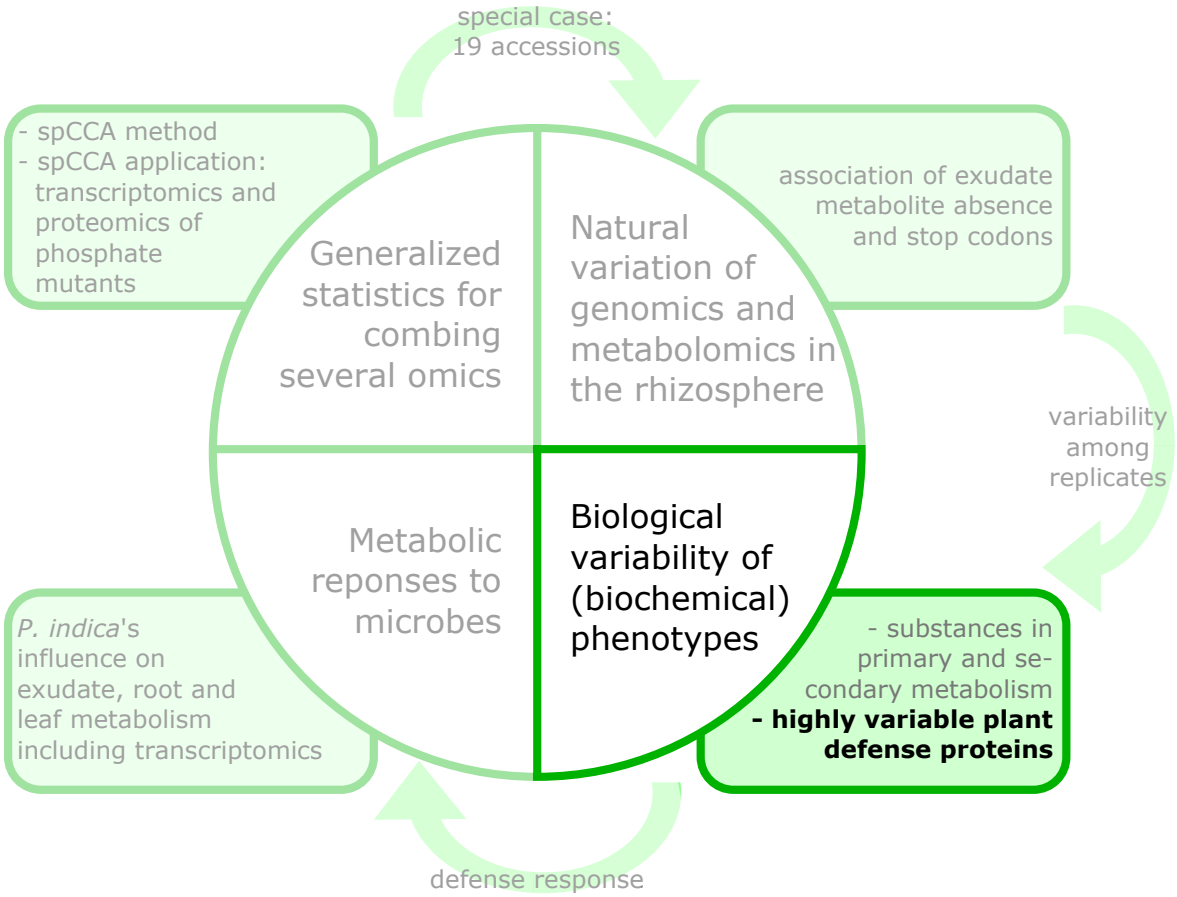
10. Töpfer, N.; Scossa, F.; Fernie, A.; Nikoloski, Z. Variability of metabolite levels is linked to differential metabolic pathways in *Arabidopsis*'s responses to abiotic stresses. *PLoS Comput. Biol.* **2014**, *10*, e1003656. [CrossRef] [PubMed]
11. Strehmel, N.; Böttcher, C.; Schmidt, S.; Scheel, D. Profiling of secondary metabolites in root exudates of *Arabidopsis thaliana*. *Phytochemistry* **2014**, *108*, 35–46. [CrossRef] [PubMed]
12. Böttcher, C.; Westphal, L.; Schmotz, C.; Prade, E.; Scheel, D.; Glawischnig, E. The multifunctional enzyme CYP71B15 (PHYTOALEXIN DEFICIENT3) converts cysteine-indole-3-acetonitrile to camalexin in the indole-3-acetonitrile metabolic network of *Arabidopsis thaliana*. *Plant Cell* **2009**, *21*, 1830–1845. [CrossRef] [PubMed]
13. Buhtz, A.; Witzel, K.; Strehmel, N.; Ziegler, J.; Abel, S.; Grosch, R. Perturbations in the Primary Metabolism of Tomato and *Arabidopsis thaliana* Plants Infected with the Soil-Borne Fungus *Verticillium dahliae*. *PLoS ONE* **2015**, *10*, e0138242. [CrossRef] [PubMed]
14. Lommen, A. MetAlign: Interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. *Anal. Chem.* **2009**, *81*, 3079–3086. [CrossRef] [PubMed]
15. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2015.
16. Gentleman, R.C.; Carey, V.J.; Bates, D.M.; Bolstad, B.; Dettling, M.; Dudoit, S.; Ellis, B.; Gautier, L.; Ge, Y.; Gentry, J.; et al. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.* **2004**, *5*, R80. [CrossRef] [PubMed]
17. Tautenhahn, R.; Böttcher, C.; Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinform.* **2008**, *9*, 504. [CrossRef] [PubMed]
18. Smith, C.A.; Want, E.J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* **2006**, *78*, 779–787. [CrossRef] [PubMed]
19. Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T.R.; Neumann, S. CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal. Chem.* **2012**, *84*, 283–289. [CrossRef] [PubMed]
20. Luedemann, A.; Strassburg, K.; Erban, A.; Kopka, J. TagFinder for the quantitative analysis of gas chromatography—Mass spectrometry (GC-MS)-based metabolite profiling experiments. *Bioinformatics* **2008**, *24*, 732–737. [CrossRef] [PubMed]
21. Kopka, J.; Schauer, N.; Krueger, S.; Birkemeyer, C.; Usadel, B.; Bergmuller, E.; Dormann, P.; Weckwerth, W.; Gibon, Y.; Stitt, M.; et al. GMD@CSB.DB: The Golm Metabolome Database. *Bioinformatics* **2005**, *21*, 1635–1638. [CrossRef] [PubMed]
22. Lassowskat, I.; Böttcher, C.; Eschen-Lippold, L.; Scheel, D.; Lee, J. Sustained mitogen-activated protein kinase activation reprograms defense metabolism and phosphoprotein profile in *Arabidopsis thaliana*. *Front. Plant Sci.* **2014**, *5*, 554. [CrossRef] [PubMed]
23. Bates, D.; Mächler, M.; Bolker, B.; Walker, S. Fitting Linear Mixed-Effects Models Using LME4. *J. Stat. Softw.* **2015**, *67*, 1–48. [CrossRef]
24. Plant-to-Plant Variability in Root Metabolite Profiles of 19 Arabidopsis thaliana Accessions Is Substance-Class Dependent. Available online: http://www.ebi.ac.uk/metabolights/MTBLS338 (accessed on 13 September 2016).

2.3.2 Assessment of label-free quantification in discovery proteomics and impact of technological factors and natural variability of protein abundance

Al Shweiki, M. R.; Mönchgesang, S.; Majovsky, P.; Thieme, D.; Trutschel, D.; Hoehenwarter, W. Assessment of Label-free Quantification in Discovery Proteomics and Impact of Technological Factors and Natural Variability of Protein Abundance. *J Proteome Res* (revised version submitted on 21/10/2016).

equal contributions

now available in the final version:
DOI: 10.1021/acs.jproteome.6b00645
http://pubs.acs.org/doi/full/10.1021/acs.jproteome.6b00645



Assessment of Label-free Quantification in Discovery Proteomics and impact of Technological Factors and Natural Variability of Protein Abundance

Journal:	*Journal of Proteome Research*
Manuscript ID	pr-2016-006452.R1
Manuscript Type:	Article
Date Submitted by the Author:	21-Oct-2016
Complete List of Authors:	Al Shweiki, MHD Rami; Universitatsklinikum Ulm, Neurology Mönchgesang, Susann; Leibniz Institute of Plant Biochemistry, Majovsky, Petra; Leibniz-Institut fur Pflanzenbiochemie Thieme, Domenika; Leibniz-Institut fur Pflanzenbiochemie Trutschel, Diana; Deutsches Zentrum fur Neurodegenerative Erkrankungen Standort Witten Hoehenwarter, Wolfgang; Leibniz Institute of Plant Biochemistry (IPB), Proteome Analytics Research Group

Assessment of Label-free Quantification in Discovery Proteomics and impact of Technological Factors and Natural Variability of Protein Abundance

MHD Rami Al Shweiki1,§, Susann Mönchgesang2,§, Petra Majovsky1,§, Domenika Thieme1, Diana Trutschel2,3,4 and Wolfgang Hoehenwarter1*

1 Research Group Proteome Analytics, Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120, Halle (Saale), Germany

2 Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120, Halle (Saale), Germany

3 Deutsches Zentrum für Neurodegenerative Erkrankungen, Stockumer Straße. 12, 58453, Witten, Germany

4 Martin-Luther-University Halle-Wittenberg, Von-Seckendorff-Platz 1, 06120, Halle (Saale), Germany

§ These authors contributed equally to this work

*Address correspondence to:

Dr. Wolfgang Hoehenwarter

Email: Wolfgang.Hoehenwarter@ipb-halle.de

Tel: +49 345 5582 1411

Fax: +49 345 5582 1409

Running Title: Assessment of Variability in Label-free Quantification

Abstract

We evaluated the state of label-free discovery proteomics focusing especially on technological contributions and contributions of naturally occurring differences in protein abundance to the inter-sample variability in protein abundance estimates in this highly peptide-centric technology. First, the performance of popular quantitative proteomics software, Proteome Discoverer, Scaffold, MaxQuant and Progenesis QIP was benchmarked using their default parameters and some modified settings. Beyond this the inter- sample variability in protein abundance estimates was decomposed into variability introduced by the entire technology itself and variable protein amounts inherent to individual plants of the *Arabidopsis thaliana* Col-0 accession. The technical component was considerably higher than the biological inter-sample variability suggesting an effect on the degree and validity of reported biological changes in protein abundance. Surprisingly, the biological variability, protein abundance estimates and protein fold changes were recorded differently by the software used to quantify the proteins, warranting caution in the comparison of discovery proteomics results. As expected, around 99 % of the proteome was invariant in the isogenic plants in the absence of environmental factors; however few proteins showed substantial quantitative variability. This naturally occurring variation between individual organisms can have an impact on the causality of reported protein fold changes.

Keywords: natural variability, shotgun proteomics, label-free quantification, MaxQuant, Progenesis, biological variability, experimental variability, protein abundance

Introduction

The past five years have seen great advances in the field of proteomics. Particularly the maturation of shotgun or discovery proteomics has been profound, delivering on the –omics promise of quantitative measurement of all of the proteins of higher eukaryotes [1]. The most significant developments may have been in terms of sensitivity and quantification. Both can be attributed to improvements in liquid chromatography (LC) and mass spectrometry (MS) hardware, particularly nano-UPLC and fast scanning high resolution accurate mass MS and proteomics data analysis software.

The incompatibility of metabolic labeling and shotgun proteomics in clinical trials and agronomics field studies has lately led to the popularization of label-free quantitative proteomics approaches. Research into the mass spectrometric measurement of the abundance of different proteins in a biological sample by direct quantification of the signal response of their derivative peptides goes back more than ten years. Bondarenko and Wang and co-workers [2, 3] showed a linear relationship between protein abundance and peptide peak areas. Liu and others [4, 5] showed the same for the number of total MS/MS spectra recorded for a given protein. Numerous strategies and algorithms implementing these two themes, peptide ion signal peak intensity and peptide spectral match (PSM) (spectral counting) based quantification have since been published (reviewed in [6-8] and extensively cited in [9]). All of them are reported to quantify protein abundance with reasonable accuracy over a dynamic range of two to four orders of magnitude. Nevertheless, few have been independently validated or have found general consensus in the proteomics community.

In this study we set out to perform a rigorous evaluation of the label-free shotgun proteomics technology as a whole. To begin, we comparatively assessed the performance of four of the currently most popular professional software for label-free quantification of proteins in MS- based discovery proteomics experiments returned in a search of the Thomson Reuters Web of Science (Supplementary Table 1). They are available commercially or free of charge, promise to be robust,

accurate and facile even for a non-expert user and are supported by comprehensive instructions and /or on-line user communities and help forums as well as direct support from the vendors where applicable.

Proteome Discoverer (henceforth termed PD) is a MS data analysis platform provided by Thermo Fisher Scientific for its mass spectrometers (http://www.thermoscientific.com/content/tfs/en/product/proteome-discoverer-software.html). Its main focus is protein identification. MaxQuant LFQ (MQ) is freely available from the MPI of Biochemistry in Martinsried, Germany [9, 10] and quantifies proteins across samples using the maximum (pair-wise) peptide ratio information from extracted peptide ion signal intensities. These are normalized by minimizing the overall fold changes of all peptides across all fractions prior to normalization. Progenesis QIP (QIP) marketed by Waters (http://www.nonlinear.com/progenesis/qi/) also quantifies proteins based on peptide ion signal peak intensity (the quantification algorithms are not published). It allows full operator control over every processing step including alignment of peptide ion signal landscapes and indeed individual peptide ion signal peaks. Scaffold is a software suite from Proteome Software (http://www.proteomesoftware.com/products/scaffold/) that serves mainly as a quantitative data analysis and integration platform. This software also features advanced statistical procedures for refinement of quantitative search engine results and various spectral counting or peak ion intensity based protein quantification indices (PQIs).

We then went beyond the software benchmark validation which has been the focus of several recent studies [11-17] to an analysis of the variability in protein abundance estimates in repeated sample analysis (inter-sample variability) introduced by the entire shotgun proteomics technology from protein extraction to software supported protein quantification. To this end, individual, essentially isogenic plants of the *Arabidopsis thaliana* Col-0 accession, cultivated and harvested under controlled conditions were sampled to focus on the technological contribution to the inter-sample variability. We found it to be substantial and to have an impact on the degree of protein abundance fold changes that can be reported as biological in nature at different statistical confidence levels.

Importantly, we also found well known issues such as protein inference from measured peptides persist in discovery proteomics and confound protein quantification, leading to vastly different abundance estimates of the same proteins by MQ and QIP.

In addition, we observed marked variability in the abundance of around 25 proteins in the individual *Arabidopsis* plants. This naturally occurring variability found even in an isogenic background under controlled cultivation and harvest conditions can also affect protein abundance fold changes and their reported causality in response to experimental conditions, especially in studies wherein individual organisms are pooled. This underscores the importance of *a priori* consideration and development of suitable experimental design.

Experimental Procedures

Standard Proteins and Plant Material

α-Lactoalbumin (αLA), Apotransferin (APO), ß-Lactoglobulin A and B (ßLA and ßLB), Bovine Serum Albumin (BSA), Carbonic Anhydrase (CAH), Cytochrome C (CYTC), Fetuin (FET) and Myoglobin (MYO) were purchased from Sigma Aldrich GmbH. Ovalbumin (OVA) was purchased from Protea Biosciences (West Virginia, USA).

Arabidopsis thaliana ecotype Columbia (Col-0) seeds were cold-stratified and sterilized with chlorine gas. Seeds were sown in steam sterilized soil. Germinated seedlings were transplanted into individual pots at the two cotyledon stage. Plants were grown in a growth chamber in one flat consisting of 60 pots under short day conditions (8 hour photo period) at 22°C for six weeks with weekly watering. Plant shoots (leaf rosette) were harvested by cutting slightly above soil level and immediately frozen in liquid nitrogen.

Standard Protein Dilution Series

Standard proteins were dissolved in ddH_2O. SDS-PAGE of standard proteins was performed according to Laemmli [18]. Disulfide bonds were reduced with dithiothreitol (DTT) and alkylated with an excess of iodoacetamide (IAA). Proteins were digested with trypsin at an enzyme to protein ratio of 1:50 (w/w) at 37°C over night. Tryptic peptides were desalted with in-house made STAGE Tips containing 6 layers of 3M Empore C18 solid phase extraction matrix (3M, Minneapolis, USA) in a 100 µl pipette tip as described [19].

Six isobaric mixtures of the ten digested standard proteins were prepared to give a six point dilution series (10 fmol/µl, 30 fmol/µl, 100 fmol/µl, 300 fmol/µl, 600 fmol/µl, 1000 fmol/µl) of all of them. The molar composition in fmol/µl and the weight amount of total protein in µg/µl of each of the mixtures is given in Supplementary Table 2. One µl of each mixture was injected into an LC-MS system to measure the dilution series under naked conditions. The six mixtures were spiked into one µg of trypsin digested *Arabidopsis thaliana* total protein extract and injected to give a measurement of the dilution series under matrix conditions. The ratio of standard to matrix protein weight amount was approximately 11% (Supplementary Table 2).

Arabidopsis thaliana Total Protein Extraction

Frozen leaf rosettes were ground to a fine, light green powder under liquid nitrogen with a mortar and pestle. Proteins were extracted from the plant tissue with a phenol based procedure described in detail previously [19].

Liquid Chromatography and Mass Spectrometry

Peptides were injected into an EASY-nLC 1000 nano liquid chromatography system (Thermo Fisher Scientific). Peptides were separated using C18 reverse phase chemistry using an Acclaim PepMap 100 pre-column (length 2cm, inner diameter 75 µm, particle diameter 3 µm) in-line with an EASY-Spray ES803 column (length 50 cm, inner diameter 75 µm, particle diameter 2 µm) (both from Thermo

Fisher Scientific). Peptides were eluted with a 180 min gradient increasing from 5% to 40% ACN in ddH_2O (60 min gradient for QC measurements of the standard proteins individually) and a flow rate of 300 nl/min and electrosprayed into an Orbitrap Velos Pro mass spectrometer (Thermo Fisher Scientific) with an EASY-Spray ion source (Thermo Fisher Scientific). The source voltage was set to 2 kV, the S-Lens RF level to 50%. The delta multipole offset was -7.00. The instrument method consisted of one survey (full) scan of the entire ion population in the Orbitrap mass analyzer followed by up to 20 data dependent CID product ion scans of selected precursor ions in the linear quadrupole ion trap (LTQ). A single micro scan per mass spectrum was acquired in both mass analyzers. The AGC target value was set to 1e06 and the maximum injection time (max IT) to 500 ms in the Orbitrap. The parameters were set to 1e04 and 100 ms in the LTQ with an isolation width of 2 Da and normalized collision energy of 35 for precursor isolation and MS/MS scanning. Dynamic exclusion was enabled with a repeat count of 1, a repeat duration of 30 s an exclusion duration of 60s and a relative exclusion width of 10 ppm. Full scan mass spectra were internally calibrated on the fly using the lock mass option with the m/z 445.120024. Four blank injections were run following every sample injection to reduce carryover of the most abundant proteins to approximately 5%. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [20] partner repository with the dataset identifier PXD004025 and 10.6019/PXD004025

Standard Protein Dilution Series Quantification under Naked Conditions

Thermo .raw files were imported into Proteome Discoverer v1.4. (PD). Peak lists were generated with a precursor signal to noise ratio of 1.5 and default settings were used to search a custom made database containing the sequences of the protein standards taken from NCBI (STD, 10 sequences, 3040 residues) with the Mascot algorithm v.2.5.1 on an in-house Mascot server. The enzyme specificity was set to trypsin and two missed cleavages were tolerated. Carbamidomethylation of cysteine was set as a fixed modification and oxidation of methionine as a variable modification. The precursor tolerance was set to 7 ppm and the product ion mass tolerance was set to 0.8 Da. A decoy

database search was performed to determine the peptide false discovery rate (FDR) with the Target Decoy PSM Validator module. A 1% peptide FDR threshold was applied.

Mascot .mgf files were imported into the Skyline software v.3.1.0.7382 [21] to produce a spectral library of the standard proteins. The STD standard proteins database in .FASTA format was imported into Skyline to populate a spectral tree. The .raw files were imported into Skyline and XICs of the standard peptides in the respective measurements were produced using the settings described previously [19].

Validation of Label-free Quantitative Proteomics Software

Thermo. Raw files were imported into all tested software without any file conversion. The software versions, settings and employed PQIs are summarized in Table 1. All database searches were performed using a concatenated database that combined TAIR10 amended with common contaminates and STD (TAIR10STD, 35,414 sequences, 14,493,054 residues).

For PD peptides and proteins were identified with the Mascot software as described above. A protein quantitation matrix containing the #PSMs for each identified protein in each measurement was produced by opening all of the .msf files of the measurements in a single report. The NSAF PQI was calculated manually according to [22].

Additionally, peptide precursor ion intensities were extracted from the .raw files using the Precursor Ions Area Detector module. The .msf files were imported into the Scaffold Q+ (Scaffold) software. A 1% peptide and 1% protein FDR threshold was applied. Two protein quantitation matrices were produced, one containing the NSAF, the other the Top3 PQI by selecting the respective PQIs in the Quantitative Analysis setup.

MaxQuant used its integrated Andromeda search engine to identify peptides and proteins. Carbamidomethylation of cysteine was set as a fixed modification and oxidation of methionine and

acetylation of protein N-termini as variable modifications. For all other settings and parameters deviating from software defaults see Table 1.

Progenesis QIP settings were software defaults with the exceptions listed in Table 1. Mascot .mgf files for database search with Mascot were created in QIP excluding MS/MS spectra ranking greater than 5. Peptides and proteins were identified as described above for PD. The peptide FDR threshold was adjusted to 1% and the Mascot search result was imported into Progenesis as an .xml file. Peak picking and alignment of Ovalbumin peptide ion signal peaks were manually adjusted using the Review Peak Picking Window and the Select and Edit Buttons in the Run tab. All peptide and protein identifications and estimated abundance values for all software settings and PQIs are provided as supplemental data.

Table 1. Software Settings and PQIs evaluated in the study. All listed settings are deviations from the software defaults which are given in the text. Settings for database search are also given in the text. The functionality and algorithms behind the MQ settings are explained in detail in [10]. LFQ is the PQI implemented in MaxLFQ and also described in detail in [9]. iBAQ refers to intensity based absolute quantification, more information can be found in [11, 23]. Top3 uses the peak areas of a protein's 3 peptides with the most intense mass spectrometric signal response for quantification [24]. The QIP PQI Non-conflicting Peptides means only peptides unique to a protein were used for quantification. NSAF refers to the normalized spectral abundance factor, details can be found in [22].

Software Name	Version	Software Parameter Set Name	Settings Deviating from Default	PQI
MaxQuant	1.5.0.0	MQ-Def-LFQ	Default	LFQ
		MQ-MBR-LFQ	Default; Match Between Runs (MBR) enabled	LFQ
		MQ-Mod-LFQ	Default; Match Between Runs (MBR) enabled; Minimum Ratio set to 1; Precursor mass tolerance 7 ppm	LFQ
		MQ-MBR-iBAQ	Default, Match Between Runs (MBR) enabled, iBAQ enabled	iBAQ
Progenesis QIP	4.1.	QIP-Def-Top3	Default; Mascot search see text	Top3
		QIP-Def-NoC	Default; Mascot search see text	Non-conflicting Peptides
		QIP-Ses-Top3	Peak Picking Parameters Sensitivity set to 5 (Maximum); Mascot search see text	Top3
		QIP-Ses-NoC	Peak Picking Parameters Sensitivity set to 5 (Maximum); Mascot search see text	Non-conflicting Peptides
Proteome Discoverer	1.4.	PD-NSAF	see text	NSAF
Scaffold Q+	4.4.3.	Scaffold-NSAF	Peptide and Protein ID from PD; Protein FDR 1%	NSAF
		Scaffold-Top3	Peptide and Protein ID, Peak Intensity from PD; Protein FDR 1%	Top3

Experimental Design and Statistical Rationale

For validation of the label-free quantitative proteomics software, the same weight amount of each of the six isobaric mixtures of the ten digested standard proteins was spiked into 1 µg each of the same tryptic digest of *Arabidopsis thaliana* total protein extract. Each of these six mixtures was measured four times in a block design wherein all of the mixtures were measured once in random order before proceeding to the second block wherein all of the mixtures were measured again in random order and so forth. This design eliminated biological and experimental variability in *Arabidopsis* protein abundance and minimized variability potentially introduced by sample handling and the LC-MS system to focus solely on any variability in protein abundance estimates introduced by the tested software and thereby on their individual merits.

In order to analyze the correlation between measured and expected protein abundance, all possible fold changes between measured protein abundance s at all molar amounts were calculated for each protein and compared to their expected fold-changes by a linear model. Permutation analysis was performed for testing the significance of the linear models. Eighty percent of the measured $\log_{10}$ fold changes were resampled 10 times. Each resampled replicate was permuted. Linear models and their slopes were calculated for each of the 10 permuted sets with the expected $\log_{10}$ fold changes. The null hypothesis of no linear correlation between measured and expected $\log_{10}$ fold changes (H_0 : linear model slope a = 0) was tested using a Student's t-test and the slopes of ten unpermuted, resampled sets.

The overall performance of all tested software parameter sets was scored. Formally the score is expressed in equation (1) with *aFC* being the expected $\log_{10}$ fold change, *mFC* the measured $\log_{10}$ fold change, *n* the number of $\log_{10}$ fold changes quantified for all standard proteins, 168 being the maximum number, i.e. all possible fold changes for all standards and *k* the number of *Arabidopsis* background proteins quantified in at least 12 measurements.

(1):

$$Score := \left(\frac{\sum_{i=1}^{n} (aFC_i - mFC_i)^2 \cdot 168}{n^2 \cdot k^2} \cdot 1e + 06 \right)^{-1}$$

This term takes into account the accuracy and precision of protein quantification by way of the mean deviation of measured from expected standard protein abundance over the entire range of fold changes and molar amounts. It also takes into account the sensitivity of protein quantification by rewarding proportionally larger numbers of quantified standard protein $\log_{10}$ fold changes as well as larger absolute numbers of quantified background matrix proteins.

To assess the variability in protein abundance estimates from analysis to analysis introduced by the entire shotgun proteomics technology (intra-sample or technical variability) four six-week-old *Arabidopsis thaliana* Col-0 plants with essentially no genetic polymorphism grown under the same conditions were extracted independently three times and measured resulting in a multilevel (hierarchical) experimental design. The measured proteins were quantified using three of the software and parameter sets independently, QIP-Def-Top3, MQ-MBR-LFQ and MQ-MBR-iBAQ. This design allowed assessment of the variability of protein abundance estimates from plant to plant (inter-sample or biological variability), in this case virtually excluding genetic and environmental factors.

To decompose the total measured variance in protein abundance into biological and technical components, a linear mixed model with no fixed effects and biological and technical replicates as random effects was fit onto every protein quantified by each of the three software parameter sets to adjust for dependencies in multilevel structure. A second analysis was performed with software as a fixed and biological and technical replicates as random effects on the mutual set of proteins quantified by all three software parameter sets. PQI values output from the software and parameter sets were $\log_2$-transformed to approximate a normal distribution for further statistical testing. The

expected values for each software and parameter set were estimated using model-based least square means. The intra-class correlation (ICC) is defined in [25] as the ratio of inter-sample to total variance. The statistical analysis was performed using the statistical software language R, version 3.1.3, the lme4 package, version 1.1-11 [26] and lmerTest, version 2.0 [27]. Further details and R code are provided as a supplement.

Results

Benchmark Set of Protein Standards

We selected a set of ten standard proteins spanning the molecular weight (Mw) and isoelectric point (pI) range of the majority of cellular protein monomers (Table 2). We assessed their purity with SDS-PAGE (Supplementary Figure 1) and LC-MS (data not shown) and found minimal contamination in line with high purity chemicals.

We digested the proteins with trypsin and measured 100 fmol on column of each protein digest with DDA UPLC-HR/AM MS using a 50 cm LC column. We identified a substantial number of unique peptides resulting in high sequence coverage for all of the proteins (Table 2). We made six mixtures of the ten digested standard proteins with approximately equal weight (coefficient of variance 0.33%) but different molar amounts. These were used to produce a six point dilution series of all standard proteins comprising 10, 30, 100, 300, 600 and 1000 fmol and so numerous small and large fold changes over the full two orders of magnitude. The dilution series was measured in triplicate. We extracted the #PSMs and the peak areas of the three most abundant peptides per protein (Top 3) using a 1% PSM FDR threshold. These PQIs were plotted as the dependent variables against the protein molar amounts. As expected, strong correlation to a linear model was observed for both PQIs for all proteins (Supplementary Figure 2). The highest R^2 was observed for proteins with the highest #PSM to peptides ratio and not necessarily the largest proteins, underscoring the impact of peptide specific properties on protein quantification. As also may be expected, the slopes of the lines were different

for the individual proteins as well as deviating from unity, implying every protein quantified in a discovery proteomics experiment requires its own proportionality coefficient to correctly infer its abundance.

Table 2. Proteins used as standards in the study. The sequence coverage, number of PSMs and number of peptides recorded in a measurement of 100 fmol on column are given. The number of peptides unique to a protein is given in parentheses.

Standard Name	Token	MW (Da)	% Coverage	# PSMs	# Peptides	pI
Cytochrome C	CYTC	12327	42.86	17	7 (7)	9.50
α-Lactoalbumin	αLA	14178	40.58	28	7 (7)	5.14
Myoglobin	MYO	17000	63.69	161	9 (9)	7.81
ß-Lactoglobulin B	BLB	18276	48.15	74	10 (1)	4.92
ß-Lactoglobulin A	BLA	18363	40.12	40	13 (2)	4.86
Ovalbumin	OVA	42750	15.32	28	5 (5)	5.29
BSA	BSA	66000	80.40	494	68 (68)	6.18
Fetuin	FET	48400	22.60	73	8 (8)	5.50
Carbonic anhydrase	CAH	29000	31.50	30	9 (9)	6.92
Transferrin	TRA	80000	70.17	610	60 (60)	7.08

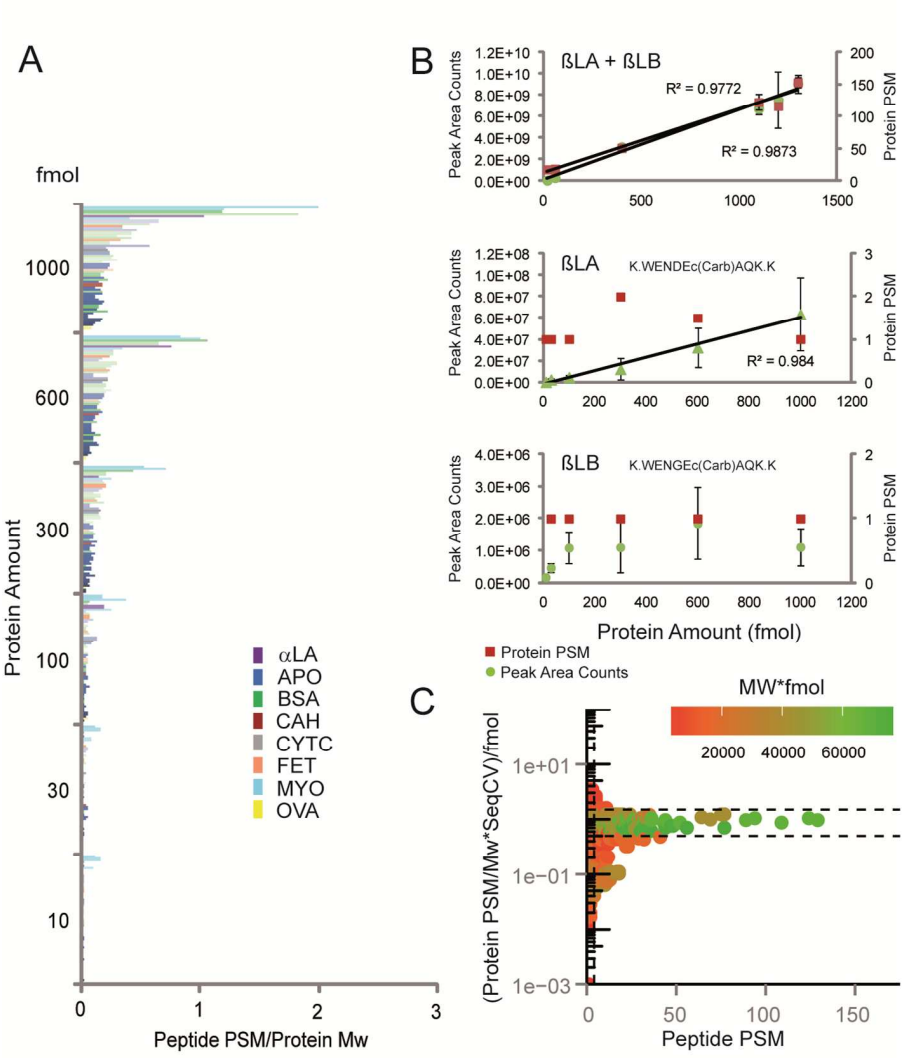
Inference of Standard Protein Abundance from Measured Peptide Ion Signal Response

Currently discovery proteomics experiments are inherently peptide-centric so we were interested in quantification of the standard proteins on the peptide level. We plotted the protein molar amounts against the #PSMs of each peptide normalized to the Mw of each peptides respective progenitor protein as a proxy for the number of quantifiable peptides per protein. We mostly observed a linear relationship between peptide signal response and protein molar amounts with this PQI (Figure 1A). Although this may seem intuitive from the protein results described above this need not be so because the proteins are quantified using the sum of peptide signal responses. Indeed, when using raw #PSMs, peptide responses were often far from linear over the range of protein molar amounts.

Two of the proteins in our standard set (ßLA and ßLB) were genetic isoforms distinguished at only two positions in their primary structure and hence by two unique peptides each. Both #PSMs and Top 3 could not quantify them because it was impossible to fractionally assign the PQIs to their respective molar amounts. The combined molar amounts however showed excellent linearity to both

PQIs (Figure 1B, top panel). The two discriminating peptides that were identified (one for each isoform) showed a very weak signal response (Figure 1B, center and bottom panels). Only the peptide WENDEc(Carb)AQK unique to ß-LA showed a strong linear relationship between its peak area and the protein's molar amount and so could be potentially suitable for quantification.

Figure 1



To further investigate the relationship between inferred protein and measured peptide abundance we normalized each protein's #PSMs at each molar amount to its Mw and multiplied this term by the respective protein sequence coverages. The sequence coverage is greatly affected by the mass spectrometric peptide ion signal response even at low protein abundance and therefore serves as a proxy to incorporate the factor of peptides physico-chemical properties on protein quantification.

The PQI was then set in relation to the molar amount for each protein. Note that this term corresponds to the proportionality between measured and actual protein abundance (slope of the linear model). We plotted it as a function of the #PSMs of each identified peptide of each protein (Figure 1C). Two things become apparent: First, with increasing number of peptide PSMs which conceivably can be generalized to any measure of peptide signal response, the relationship between measured and actual protein abundance converges on direct proportionality. Second, this is primarily the case for larger and more abundant proteins (plot coloring). While this is not surprising, it does imply there may be substantial error associated with quantification of smaller and low abundant proteins (red points) as observed previously [11].

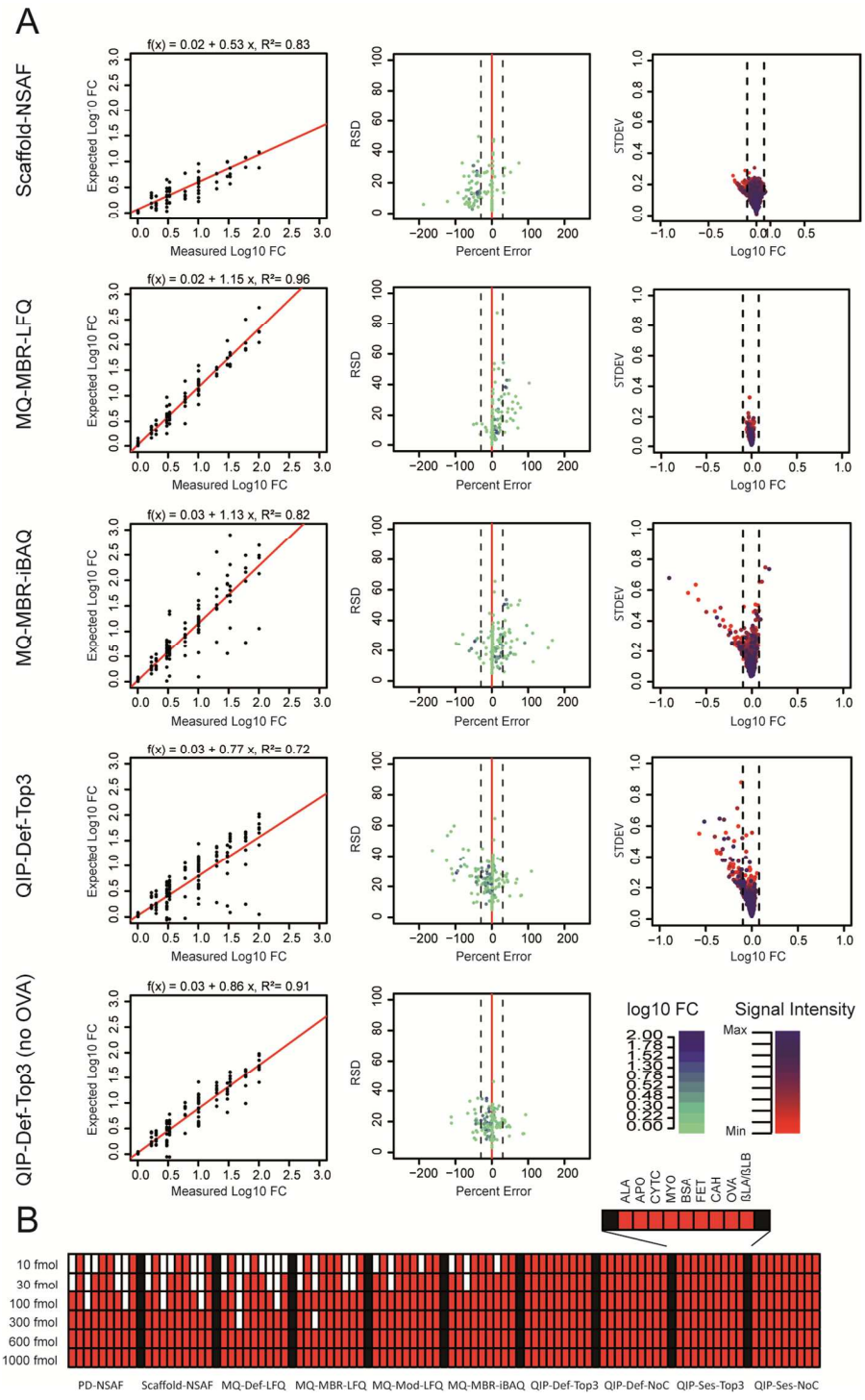
Accuracy and Precision of Label-Free Quantitative Proteomics Software

Peptides and proteins were identified and quantified in quadruplicate measurements of the six isobaric standard mixtures spiked into the *Arabidopsis* peptide matrix with PD, Scaffold (Scaffold used the PD output files so linked to PD), MQ and QIP. We tested each software using the software defaults (i.e. ready to use) as well as with sets of modified settings and different PQIs; for a detailed description see Experimental Procedures and Table 1. Importantly, only Scaffold and MQ had the option of applying a protein FDR calculation in addition to the commonly employed peptide FDR. All FDR thresholds were set to 1% for peptide and protein identification in all analyses.

We investigated the relationship between the measured and expected fold changes for each of the standard proteins detected in all quadruplicate measurements to assess the accuracy and precision of protein quantification. Figure 2 shows the results for the settings that performed most favorably for each software, the results for all other settings are shown in Supplementary Figure 3. Table 3 shows the slopes of all linear models. As expected, all software and all settings achieved reasonable agreement with a linear model as evidenced by a coefficient of determination (R^2) of > 0.7 with the exception of Scaffold employing the Top3 PQI (Table 3, Supplementary Figure 3 2nd column). Both Scaffold and PD alone produced much improved R^2 employing the NSAF PQI (Figure 2A top panel,

Supplementary Figure 3 1st column) suggesting precursor ion intensity extraction from the raw data using PD was sub-optimal. MQ and QIP both showed more precise quantification of standard protein log_{10} fold changes with all settings and PQIs over the entire range of molar abundance using R^2 as a measure of global precision (Table 3). MQ, however, consistently delivered higher R^2 than QIP.

Figure 2



The slope of the linear model reflects the accuracy of the measurement of the $\log_{10}$ fold changes, a slope of one meaning direct proportionality between measured and expected protein abundance. PD and Scaffold employing NSAF both showed a 0.5 fold underestimation of $\log_{10}$ fold changes indicating this PQI is semi-quantitative at best. Strikingly, a slope deviating only slightly from unity was achieved by MQ for all settings and PQIs including the software defaults, so quantification of the set of standard proteins over the entire dynamic range of 10 to 1000 fmol was very accurate. QIP was substantially less accurate employing all settings and PQIs; however upon closer inspection it became clear, that one of the standard proteins, OVA had been incorrectly quantified over the entire range of protein molar amounts. We made use of the extensive QIP interface to manually correct the peak picking and alignment of all OVA derived peptide ion signal peaks using the default settings and Top3 PQI which proved to be the most advantageous for QIP (QIP-Def-Top3 (mod OVA), Supplementary Figure 3 1st column). This improved both the slope and R^2 slightly. We then removed OVA from the QIP results entirely (QIP-Def-Top3 (no OVA), Figure 2A bottom panel) which lead to a slope close to unity and an $R^2 > 0.9$, in agreement with the MQ results.

To visualize the accuracy and precision of protein quantification more directly the relative error in percent of the measured $\log_{10}$ fold changes and their relative standard deviation (RSD) in the quadruplicate measurements were plotted (Figure 2A center panel and Supplementary Figure 3). PD and Scaffold employing the NSAF PQI greatly underestimated protein abundance $\log_{10}$ fold changes by 30 to 100%; especially larger ones (note for example a 50 % error of a $\log_{10}$ fold change of 2 means a 50 fold error). MQ showed very conservative percent errors of < 30% using the default settings and the match between runs (MBR) option enabled with the LFQ PQI. Small and larger $\log_{10}$ fold changes were quantified similarly accurately with these parameters. The accuracy decreased somewhat using modified settings with a minimum ratio count of 1 as opposed to the default of 2 and when using the IBAQ PQI with MBR settings. Both of these parameter sets were designed to increase the number of standard proteins quantified at low abundance (10 to 30 fmol), i.e. the sensitivity. The percent error was still mostly below 30 %. However, some $\log_{10}$ fold changes, and particularly the largest ones (1.78

and 2), were equally over and under- estimated. QIP also quantified most of the proteins correctly with an error of <30 % although a substantial number of log_{10} fold changes were more severely underestimated with all parameter sets. However, this could be reconciled when OVA quantification was manually corrected and more so when the protein was removed entirely. In this case all remaining proteins were quantified very accurately over the entire range of log_{10} fold changes including the largest ones. The RSD of the quadruplicate measurements of log_{10} fold changes was generally below 40 % for all software and all parameter sets.

Table 3. Ranked list of the evaluated software and parameter sets and of some of the results of the evaluation. Coefficient of determination and slope of the linear model, number of quantified log_{10} fold changes for all protein standards (168 maximum meaning every possible pair wise fold change for every protein) and p-value of significance testing of the linear models are given. The number of *Arabidopsis thaliana* background matrix proteins quantified in at least 1, 12 and all 24 measurements as well as the software/parameter set score are also given.

Software Parameter Set	R^2	Slope	Quantified Log_{10} Fold Changes (n)	P-value	*Arabidopsis* Matrix Proteins Quantified in ≥ Measurements			Score
					1	12	24	
QIP-Def-Top3 (no OVA)	0.909	0.86	147	3.985-13	2178	2178	2170	100.85
MQ-MBR-LFQ	0.958	1.15	120	1.058E-09	2366	1671	1294	53.42
QIP-Def-Top3 (mod OVA)	0.797	0.79	168	1.022E-13	2178	2178	2170	50.86
QIP-Def-Top3	0.731	0.77	168	1.107E-09	2178	2178	2170	38.26
QIP-Def-NoC	0.74	0.74	168	2.605E-12	2178	2177	2166	37.88
QIP-Ses-NoC	0.733	0.74	168	4.489E-12	2182	2182	2169	37.46
QIP-Ses-Top3	0.723	0.77	168	7.730E-12	2182	2182	2169	37.46
MQ-MBR -iBAQ	0.815	1.13	145	2.473E-12	2319	2221	1811	35.44
MQ-Mod-LFQ	0.773	1.1	145	6.307E-10	2314	2199	1793	30.12
MQ-Def-LFQ	0.961	1.18	101	6.308E-09	2362	1275	881	27.73
PD-NSAF	0.858	0.54	110	3.593E-11	2605	1894	1391	17.58
Scaffold-NSAF	0.828	0.53	110	9.134E-10	2543	1878	1307	16.12
Scaffold-Top3	0.391	2.15	106	2.200E-09	2768	2005	1462	0.61

Sensitivity and LOQ

We found it important to determine the absolute limit of quantification (LOQ) to measure the sensitivity as an additional performance metric of the software. To do so, we examined if a quantitative value was available at each molar amount of the dilution series for each standard protein in all quadruplicate measurements for each software and parameter set (Figure 2B). PD and

Scaffold using the NSAF PQI were both unable to quantify four of the standard proteins at 10 and 30 fmol as well as some others even at higher molar amounts. The same held true for MQ using the default and MBR settings with the LFQ PQI. MQ sensitivity improved markedly with the modified settings and the IBAQ PQI with MBR settings. This explains the better accuracy and precision of the former two parameter sets, because a lack of quantitative data points at low molar amounts translates directly into a lack of measured large fold changes which inherently tend to be associated with higher errors. Remarkably, QIP performed perfectly for all parameter sets, quantifying all proteins at all molar amounts in all four measurements. The relationship between experimentally observed and expected protein amounts was highly significant for all software and parameter sets.

Quantification of *Arabidopsis thaliana* Background Matrix Proteins

It was also interesting to judge the softwares' performance in regard to the *Arabidopsis thaliana* background matrix. The number of *Arabidopsis* proteins that gave a quantitative value in one, 12 and all 24 of the measurements was counted (Table 3). All software quantified more than 2000 proteins with all parameter sets in at least one of the 24 measurements. PD and Scaffold with the NSAF PQI (and incidentally also with the Top3 PQI) produced the highest number of quantified proteins in a single measurement. It decreased sharply, however, in 12 and 24 measurements when the repeatability of protein quantification became a requirement. Note that the number of quantified proteins was very similar between PD and Scaffold with the NSAF PQI suggesting the 1% peptide FDR in this instance delivers sufficiently stringent protein identification compared with the 1% protein FDR applied additionally in Scaffold.

The number of quantified proteins decreased most greatly in 12 and 24 measurements using MQ with the default settings and the LFQ PQI. The default settings link peptide identification and protein quantification. A peptide ion signal peak must be annotated with an MS/MS identification in each measurement wherein it is to be quantified. Presumably the MQ feature extraction (peak picking) algorithm gives a higher quality signal than a simple PSM count, which forms the basis of the NSAF

PQI. This raises the LOQ explaining the substantially lower number of quantified proteins especially in 24 measurements compared to PD and Scaffold. The number of proteins quantified in all measurements was increased when the MBR settings were applied. The MBR option allows the calibration of peptide retention times over all LC-MS measurements in a set and thereby alignment of peptide ion signal peaks. Peak and protein identification can thus be transferred from one LC-MS measurement to all measurements in the set.

MQ with the modified settings as well as with the IBAQ PQI with MBR settings mostly negated the decline in quantified proteins. The minimum ratio count of 1 in the modified settings reduces the number of peptide ratios required for protein quantification to 1, mitigating the constraints on protein quantification and increasing sensitivity. QIP, much as MQ, applies peptide ions signal peak picking, normalization and feature alignment prior to identification. Notably, QIP did not suffer from any decline in the number of proteins quantified in any number of measurements for all parameter sets. This may suggest QIP enforces quantification of peptide ion signals in all aligned measurements, presumably if a high quality signal peak is encountered in at least one measurement in the aligned set.

To assess the accuracy and precision of quantification of the matrix proteins, the median $\log_{10}$ fold change of the PQI values of each measurement to the mean of all measurements was plotted on the abscissa and the standard deviation of the $\log_{10}$ fold changes on the ordinate for each protein for the most favorable parameter sets for each software (Figure 2A right panel). Perhaps surprisingly in light of the significant error in quantifying the standard proteins, Scaffold with the NSAF PQI exhibited high accuracy and precision likely because there was apparently no variability in the abundance of the *Arabidopsis* proteins. As expected from the results for the standard proteins, MQ with the MBR settings and the LFQ PQI was the most accurate and precise. MQ with the iBAQ PQI showed slightly higher standard deviations as well as some underestimation of proteins with a weaker PQI response however this was only a small fraction of all quantified proteins. QIP with the default parameters and the Top3 PQI showed similar behavior but with the $\log_{10}$ fold change of somewhat more proteins

being underestimated. However QIP also quantified a substantially higher number of *Arabidopsis* proteins in all.

Software Ranking

We chose to aggregate the tested metrics into a score to rank the software according to their performance in label free protein quantification (Table 3). QIP ranked highest when OVA was removed and third when OVA quantification was manually corrected, reflecting the high accuracy and precision under these circumstances as well as QIP's generally superior sensitivity. However, these circumstances represent a handicap. Under unbiased conditions MQ with the MBR settings and the LFQ PQI outperformed QIP regardless of the employed QIP parameter set, highlighting MQ's superior accuracy and precision albeit with somewhat modest sensitivity. Moreover, MQ-with the MBR settings and the iBAQ PQI was only slightly outscored by QIP. The iBAQ PQI which substantially increased the number of quantified standard and matrix proteins especially in 24 measurements still delivered high R^2 and a slope close to unity, suggesting MQ-iBAQ is generally *en par* with QIP. The MQ modified LFQ parameter set performed similarly to the iBAQ parameter set. The MQ default settings with the LFQ PQI had the lowest score for all MQ parameter sets because of poor sensitivity despite delivering the highest R^2. As expected, PD and Scaffold ranked well below MQ and QIP due to the very poor approximation of expected standard protein $\log_{10}$ fold changes and their overall modest sensitivity.

Variability in Protein Abundance Estimates in the Shotgun Proteomics Technology

We were interested in the repeatability of large scale protein quantification from sample to sample. The shotgun proteomics protocol comprises at the least protein extraction, digestion, LC-MS and software-supported protein identification and quantification. All of these steps are possible avenues of variability and so far we had only investigated the latter. The variability vested in the entire

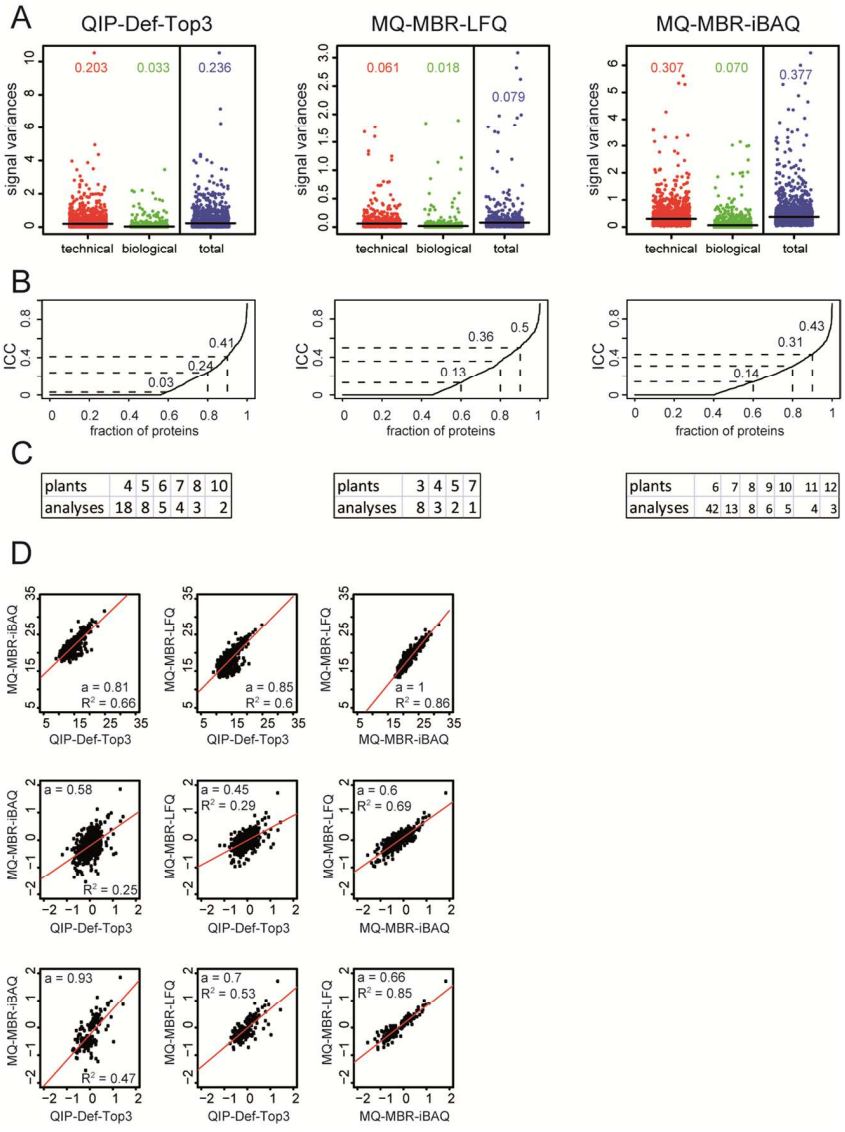
technology strongly impacts the accuracy of protein abundance estimates and the difference in fold changes that can be confidently inferred between samples and ultimately determines its utility.

We made use of a nested experimental design described by Trutschel and co-workers [28] wherein repeated experimental analysis of multiple biological samples allows decomposition of the total variance of measured protein abundance in the data set into additive technical and biological components. The measurements were analyzed with the three highest scoring software/parameter sets, QIP using the default settings and the Top3 PQI (QIP-Def-Top3) and MQ with the MBR settings and the LFQ PQI (MQ-MBR-LFQ) as well as with the iBAQ PQI (MQ-MBR-iBAQ). 2382, 1414 and 1996 proteins were respectively quantified in all 12 analyses (3 of each plant) reflecting the increased sensitivity of QIP and MQ with the iBAQ PQI as already reported.

A linear mixed model (LMM) was used to estimate the contributions of the shotgun proteomics technology (technical variance) and the plants themselves (biological variance) to the total variance of the abundance of each protein inferred by each of the software (Figure 3A). The respective mean estimated variances for all quantified proteins were also calculated. As expected, the mean estimated biological variance was very small regardless of which software and parameter sets were used to quantify the proteins. Perhaps not unexpectedly in light of its complexity particularly in plant samples, the mean estimated variance associated with the technology was 6.15 fold higher for QIP-Def-Top3, 3.39 fold for MQ-MBR-LFQ and 4.39 fold for MQ-MBR-iBAQ than the respective mean estimated biological variances. Interestingly, none of the mean biological variances estimated for the quantitative results from the three software that by definition should be unaffected by the software and PQIs used for protein quantification were the same. This can be explained by the different number of proteins quantified by each of the software, the variance of every one of them having a technical and biological component. The ratio of the estimated mean biological to total variance, known as the intra-class correlation (ICC) according to [25] was also determined for every protein for the three software (Figure 3B). Forty to 50% of the proteins exhibited ICCs below 0.03 so only

negligible biological variability in their abundance and 10 % had approximately equal technical and biological variance components (ICC of around 0.5).

Figure 3



Considering the high mean technical to biological variance ratios we were interested in the minimal effect (fold change in protein abundance) that can be inferred with confidence (α = 0.05, power = 80%) with the two-level hierarchical experimental design used here (3 analyses of four plants each) and the estimated technical and biological variances. The minimal detectable fold change was 1.68 for QIP using the default settings and the Top3 PQI, 1.39 for MQ with the MBR settings and the LFQ

PQI and 1.99 with the iBAQ PQI (calculated according to [28]). The possible combinations of the number of individual plants and the number of independent analyses of each necessary to detect a 1.41 fold change (0.5 $\log_2$ fold change) at the same confidence was also calculated (Figure 3C). These results suggest some caution is warranted in the interpretation of shotgun proteomics results in plants, particularly for small fold changes in protein abundance.

To further address the issue of different biological variance estimates by the three software, we used the set of 1028 proteins quantified by all of them as an input for the LMM. These proteins had exactly the same protein group accessions for all three software, i.e. for both MQ parameters sets and QIP. The least square means ofPQIs for each protein produced by the three software were plotted as independent variables (Figure 3D, top panel). Moderate correlation was observed for QIP with any of the MQ parameters sets whereas a strong positive correlation was observed between MQ using the LFQ and the iBAQ PQIs. It is clear that different PQIs cannot be directly compared. Nevertheless, modest positive correlation particularly for less intense signals suggests that relative protein abundance estimates by QIP and MQ were also quite different.

To investigate the possible implications of different relative estimates of protein abundance by the software on biological conclusions, we made use of a discovery proteomics dataset quantifying changes in protein abundance in response to phosphate deprivation in *Arabidopsis* roots that we published previously [29]. We analyzed three measurements each of three pools of wild type Col-0 seedling roots grown on plates containing or lacking phosphate in the growth medium with all three software and parameter sets. We plotted the least square means of PQIs for both phosphate replete and deprived samples and observed similar correlations between MQ and QIP and both MQ parameter sets as above (Supplementary Figure 4A, top left and right panels). Upon plotting the $\log_2$ fold changes in protein abundance incurred by a lack of phosphate (minus/plus phosphate) we found correlation as well as direct proportionality (R^2 and slope of the respective linear model) deteriorated markedly between all software and parameter sets (Figure 3D, center panel). Indeed, there was essentially a lack of positive correlation between QIP and MQ. To shed a better light on *de facto*

changes in protein abundance we then plotted only the statistically significant ($p<0.05$ in any of the software's estimates) $\log_2$ fold changes between the two nutrient regimes (Figure 3D bottom panel). Positive correlation and approximate direct proportionality was evident with R^2 and slope values similar to the plots of the PQIs themselves. This indicates that trends and directionalities of changes in protein abundance are conserved between MQ and QIP. Nevertheless, there are still some substantial discrepancies, particularly for the relatively small changes in protein abundance upon phosphate deprivation reported here and in [29].

Naturally Occurring Variability in Isogenic Plants

Finally, we looked at the proteins with a particularity high biological component of the total variance of their abundance. We selected the proteins quantified by MQ with the MBR settings and the LFQ PQI because these parameters allowed the most confident inference of small fold changes in protein abundance and produced the 90 % quantile of quantified proteins with the highest ICC values of all three software parameters sets. Together this permitted the most confident and sensitive detection of biologically variant proteins.

Twenty-five proteins showed substantial variance in their abundance that was of a biological nature (Supplementary Table 3). This was in spite of the essentially isogenic background and controlled cultivation of the sampled plants that eliminates major genetic and environmental variation as a source of quantitative changes in protein abundance. As expected under such conditions the relative biological variance for nearly all other proteins was minimal as reflected by the ICCs of the 80% quantiles (80% of proteins had ICCs ≤ 0.36, Figure 3B). We performed gene ontology (GO) analysis of the twenty five proteins and found that set of proteins was highly enriched for proteins responsive to abiotic and biotic stimuli as well as to stress and proteins involved in cellular metabolic processes (Supplementary Figure 4B).

The protein whose abundance varied the most in the four plants was PATHOGENESIS-RELATED PROTEIN 5 (PR5). Incidentally, the technical component of its total variance in protein abundance

was negligible so the protein had a very high ICC of 0.96. The abundance of other proteins involved in the plant response to biotic pathogens such as MITOCHONDRIAL HSO70 2 (MTHSC70-2) and ANKYRIN REPEAT CONTAINING PROTEIN 22 (AKR2) was also highly variable in the sampled plants (Supplementary Figure 4C).

Discussion

We conducted a study to assess the quantitative performance of label-free shotgun proteomics in particular focusing on the sample-to-sample variability in protein abundance estimates and its implications by the technology as a whole. The performance of label-free quantification software has received renewed attention in the recent literature [11-17]. However, much of this work was directed towards the workings of individual PQIs. Ahrné and co-workers in particular report extensively on the merits of 5 of the most well-known label-free PQIs for absolute protein quantification, the conclusions of the work of course being equally applicable to relative protein quantification [11].

A set of ten standard proteins in a total plant extract was used to test the software. We decided to use a relatively small set of proteins because this allowed us to measure each of the physico-chemically distinct proteins and the more than 150 derived tryptic peptides at six different molar amounts and so to examine quantitative software performance for each of them over a dynamic range of two orders of magnitude for both small and large fold changes. The range was selected because 1 fmol of the individual protein digests of BSA, MYO, ßLA and ßLB and 3 fmol of each protein in a combined dilution series was detected in preliminary measurements. We hypothesized that 10 fmol would be reasonably close to the absolute sensitivity limit of our LC-MS system using a data dependent (DDA) scan strategy common in shotgun proteomics experiments yet still be detectable in a complex peptide matrix. On the other hand 1000 fmol is the amount of more abundant cellular proteins and so we did not think it necessary or feasible to extend the dynamic range further upwards because the ratio of standards to matrix would have become high.

In their important paper Ahrné et al. direct considerable attention to analysis and discussion of not only linearity but, more meaningfully, proportionality between actual and PQI inferred protein abundance [11]. This question addresses one of the quintessential yet seldom treated issues in quantitative proteomics: the relationship between protein abundance and MS peptide ion signal response. In a landmark paper, Silva and co-workers showed direct proportionality between these two variables meaning signal response is a linear function of protein abundance with a slope of one when the signal intensity of the three most abundant peptides ions of a protein is used for quantification [24]. Direct proportionality between the total signal and total protein abundance has also been reported [4], however, not all of the thousands of proteins measured in a discovery proteomics experiment exhibit direct, or even the same proportionality. This was already evident from our measurements of the standards under naked conditions (Supplementary Figure 2 and Figure 1B, top panel). In this case results will be semi-quantitative at best, because the measured fold changes cannot be equated with actual changes in protein abundance. This illustrates the central challenge faced by quantitative proteomics software solutions.

Reassuringly, both of the tested software that were primarily conceived for label-free protein quantification, MQ and QIP, demonstrated accurate and precise proportionality estimates over two orders of magnitude of protein abundance for small and large fold changes. They also showed accurate and precise quantification of proteins on a large scale as exemplified by quantification of *Arabidopsis* background matrix proteins. This is a tribute to their feasibility for shotgun proteomics studies by a person with limited expertise. We have verified this independently of previous work by the software developers employing a mixture of two proteomes [9].

Which of the software is preferable may depend on the study type. Based on our experimental approach and LC-MS system we recommend MQ with the MBR settings and the LFQ PQI for datasets where quantification of up to one or two thousand proteins is desired such as in comparative shotgun proteomics analyses of SDS PAGE bands, organelle proteomes, immuno-precipitates or prokaryotic proteomes because of it superior accuracy and precision. QIP may be the better choice

when maximum quantitative coverage of complex proteomes is the goal because of its increased sensitivity. One drawback is that QIP does not allow the upfront calculation of a protein FDR which should be the norm in today's proteomics studies. MQ with the iBAQ PQI could be a good compromise which has probably led to its recent popularity [23, 30]. The decision for the right software also bears some financial aspects. MQ is available free of charge whereas QIP is commercial software. Scaffold and PD which are geared more towards protein identification, data integration, statistical analysis and results visualization did not perform as well as MQ and QIP which were designed explicitly for protein quantification. However, this may also be attributed to the spectral counting based NSAF PQI and the relatively simple peak intensity extraction algorithm integrated in PD v1.4.

Our analysis of the variability of protein abundance measurements by the shotgun proteomics technology as a whole yielded some unexpected insights. Clearly variability introduced by the technology will outweigh the biological component of the total variability in our experiment. It is also no surprise that the biological component was generally very small as seen from the ICC plots. We expected the relatively large technical component to be especially prominent in plant proteomics studies because of the involved procedure to extract proteins for in-solution digestion from these tissues [31, 32]. Conceivably this could also hold true for many human and animal tissues. It may well be that the technical variance is much smaller for other types of samples such as single cell organisms or cell cultures. Nevertheless, the calculations of statistical confidence and power for the detection of fold changes in protein abundance imply that some care must be taken in the interpretation of quantitative shotgun proteomics results particularly in plants and in up-front experimental design in general. A hierarchical experimental design wherein proteins from several biologically identical samples are repeatedly extracted and measured is advantageous, in some cases necessitating large numbers of samples and repetitions, particularly when the variability between biological conditions is expected to be high [28].

The difference in the biological variance estimates from the results of the three software parameter sets was astonishing. Biological variance should by definition not be influenced by experimental factors and we expected its estimate to be the same for all three software results. Initially we explained this discrepancy by the different number of proteins, each having a biological variance component quantified by the software. However, we saw another, even more astonishing discrepancy when the intersection of proteins quantified by all three software parameter sets was analyzed for each variance source, particularly so when protein fold changes were considered The modest and respective absence of positive correlation between MQ and QIP (Figure 3D) showed, that the relative quantification of the same proteins was quite different by these two. We attribute this result to different sets of peptides derived from the individual proteins being identified and quantified by MQ and QIP and used to infer protein abundance. Incidentally it seems not so much to be an issue of the employed PQI, i.e. the actual method of protein quantification, because relative protein abundance estimates were consistent for two PQIs within MQ. What does this mean and how can it be reconciled with the accurate and precise quantification of the standard proteins by both software discussed above?

The peptides used to quantify the standard proteins were all unique to the respective standards (Table 2). Therefore they could be unequivocally assigned to their progenitors and used to correctly infer protein abundance. Many of the *Arabidopsis* proteins have some degree of primary structure homology, resulting in a large number of peptides that are shared between them upon enzymatic digestion. By extension these peptides all have their own biological variance components. Three scenarios are readily conceivable and presumably all three contribute to the observed discrepancies. (1) MQ and QIP quantified the same peptides relatively differently. This is underscored by the erroneous quantification of OVA in the set of standard proteins and implied by the seemingly enforced quantification of peptide ion signals by QIP. (2) Different sets of peptides are identified, quantified and used to estimate the abundance of the same protein (same protein group accession) by QIP and MQ. This is undoubtedly the case but may be secondary as shown by the generally high

correlation between the two employed MQ PQIs. (3) MQ and QIP grouped and assigned peptides differently to different sets of prospective progenitors, thereby leading to distinct reconstructions of biological variance and indeed distinct estimates of the abundance of the same proteins. We investigated this issue of protein grouping further by determining and plotting both least square mean intensities and log_2 fold changes of sets of mutually quantified protein groups not assigned the same protein group accession by the two software, but sharing at least one common protein accession number (Supplementary Figure 4A, bottom left and right panels). The correlation and direct proportionality of the linear model were diminished somewhat compared to the sets of proteins with exactly the same protein group accessions assigned by both software. This indicates differential grouping has an effect on protein quantification.

Although trends and directionalities were conserved when statistically significant changes in protein abundance were quantified by both MQ and QIP, it was interesting to see, that R^2 values were considerably lower when log_2 fold changes rather than PQIs were plotted. We explain this by way of a cumulative error potential in fold changes which comprise two abundance estimates of the same protein as opposed to a single estimate of PQI values (expressed in equations 2 and 3 wherein I refers to the measured intensity, P_n to protein n and α to a deviation from the measured intensity introduced by the respective quantification software):

(2): $PQI_{P_1|PD,MQ} : \left(I_{P_1} + \alpha_{P_1|PD}\right) = a\left(I_{P_1} + \alpha_{P_1|MQ}\right) + b.$

(3): $log_2 FC_{P_2,P_1|PD,MQ} : log_2\left(I_{P_2} + \alpha_{P_2|PD}\right) - log_2\left(I_{P_1} + \alpha_{P_1|PD}\right) = a\left(log_2\left(I_{P_2} + \alpha_{P_2|MQ}\right) - log_2\left(I_{P_1} + \alpha_{P_1|MQ}\right)\right) + b.$

These points illustrate perhaps the most significant issue in peptide centric proteomics which is however largely ignored *in lieu* of a solution: the problem of protein inference [33, 34]. It is further exemplified by the inability of the tested software to correctly quantify the two ß-lactoglobulin isoforms in the set of standards because the distinguishing peptides could not be quantified (Figure 1

B). These findings highlight the importance of considering protein quantification at the peptide level in discovery proteomics experiments and further advocate prudence in the interpretation and comparison of shotgun proteomics results from different software and experimental platforms. The major issues elucidated in this study as well as some possible actions to take to mitigate them are summarized below in Table 4.

Table 4. Major caveats identified in this study, their effect on protein quantification and possible solutions.

Caveat	Problem	Possible Solution
High variability in protein abundance estimates introduced by discovery proteomics technology.	Small fold changes (<2) in abundance are difficult to quantify accurately.	Hierarchical experimental design and *a priori* calculation of required number of replicates.
Variability in protein abundance estimates introduced by genetic factors present.	Inference of causality may be hampered.	Hierarchical experimental design and *a priori* calculation of required number of replicates.
Different peptide quantification, PQIs and protein inference by proteomics software.	Different abundance estimates of proteins with shared sets of peptides.	Cautiously compare discovery proteomics results. Quantify proteins with two proteomics software.
Peptides distinguishing homologous proteins escape detection.	Highly homologous proteins are difficult to quantify.	Perform targeted proteomics experiment with the distinguishing peptides.
Proportionality between PQI value and protein abundance not unitary.	Protein abundance estimates are semi-quantitative at best.	Use peak area based software and PQI. Calibrate measurements with spike-in standards.
Small proteins generate a limited set of peptides for quantification.	Small proteins are difficult to quantify.	Perform targeted proteomics experiment with suitable peptide or peptides.
Low abundant proteins generate a limited set of peptides with ion signal for quantification	Low abundant proteins are difficult to quantify	Perform targeted proteomics experiment with peptides with the best ESI response.

In our experiments we determined the variability of protein abundance in plants of the commonly used *Arabidopsis thaliana* Col-0 accession with essentially no genome wide genetic polymorphism. *Arabidopsis thaliana* is the most important angiosperm model plant. It is an annual, self pollinating species that is naturalized worldwide. As expected, the mean variance of the abundance of all proteins in the practically isogenic plants grown and harvested under controlled conditions was very small. Incidentally, biological variance in metabolite abundance was found to be much higher in this accession [28]. This is no surprise as the metabolome is further downstream from the genome than the proteome.

A handful of proteins showed some sizable variability in their abundance in the sampled plants. This naturally occurring variability in biological samples in the absence of genetic or environmental factors warrants consideration in the experimental design of comparative molecular studies and when

reporting the causality in response to experimental conditions. Incidentally, it was a surprise to see that these proteins are nearly all involved in stress response. It is documented that epigenetic mechanisms underlie most types of stress response in plants [35-37] and there is specific evidence for this at the PR5 locus [38], one of the most well known markers for induction of salicylic acid (SA) mediated resistance to biotrophic pathogens [39] and systemic acquired resistance (SAR) [40, 41].

References

1. Mann, M.; Kulak, N. A.; Nagaraj, N.; Cox, J., The coming age of complete, accurate, and ubiquitous proteomes. *Molecular cell* **2013,** 49, (4), 583-90.

2. Bondarenko, P. V.; Chelius, D.; Shaler, T. A., Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. *Analytical chemistry* **2002,** 74, (18), 4741-9.

3. Wang, W.; Zhou, H.; Lin, H.; Roy, S.; Shaler, T. A.; Hill, L. R.; Norton, S.; Kumar, P.; Anderle, M.; Becker, C. H., Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Analytical chemistry* **2003,** 75, (18), 4818-26.

4. Liu, H.; Sadygov, R. G.; Yates, J. R., 3rd, A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Analytical chemistry* **2004,** 76, (14), 4193-201.

5. Rappsilber, J.; Ryder, U.; Lamond, A. I.; Mann, M., Large-scale proteomic analysis of the human spliceosome. *Genome research* **2002,** 12, (8), 1231-45.

6. Bantscheff, M.; Lemeer, S.; Savitski, M. M.; Kuster, B., Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Analytical and bioanalytical chemistry* **2012,** 404, (4), 939-65.

7. Megger, D. A.; Bracht, T.; Meyer, H. E.; Sitek, B., Label-free quantification in clinical proteomics. *Biochimica et biophysica acta* **2013,** 1834, (8), 1581-90.

8. Neilson, K. A.; Ali, N. A.; Muralidharan, S.; Mirzaei, M.; Mariani, M.; Assadourian, G.; Lee, A.; van Sluyter, S. C.; Haynes, P. A., Less label, more free: approaches in label-free quantitative mass spectrometry. *Proteomics* **2011,** 11, (4), 535-53.

9. Cox, J.; Hein, M. Y.; Luber, C. A.; Paron, I.; Nagaraj, N.; Mann, M., Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Molecular & cellular proteomics : MCP* **2014,** 13, (9), 2513-26.

10. Cox, J.; Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology* **2008,** 26, (12), 1367-72.

11. Ahrne, E.; Molzahn, L.; Glatter, T.; Schmidt, A., Critical assessment of proteome-wide label-free absolute abundance estimation strategies. *Proteomics* **2013,** 13, (17), 2567-78.

12. Mueller, L. N.; Brusniak, M. Y.; Mani, D. R.; Aebersold, R., An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. *Journal of proteome research* **2008,** 7, (1), 51-61.

13. Nahnsen, S.; Bielow, C.; Reinert, K.; Kohlbacher, O., Tools for label-free peptide quantification. *Molecular & cellular proteomics : MCP* **2013,** 12, (3), 549-56.

14. Ramus, C.; Hovasse, A.; Marcellin, M.; Hesse, A. M.; Mouton-Barbosa, E.; Bouyssie, D.; Vaca, S.; Carapito, C.; Chaoui, K.; Bruley, C.; Garin, J.; Cianferani, S.; Ferro, M.; Van Dorssaeler, A.; Burlet-Schiltz, O.; Schaeffer, C.; Coute, Y.; Gonzalez de Peredo, A., Benchmarking quantitative label-free LC-MS data processing workflows using a complex spiked proteomic standard dataset. *Journal of proteomics* **2016,** 132, 51-62.

15. Sjodin, M. O.; Wetterhall, M.; Kultima, K.; Artemenko, K., Comparative study of label and label-free techniques using shotgun proteomics for relative protein quantification. *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences* **2013,** 928, 83-92.

16. Cappadona, S.; Baker, P. R.; Cutillas, P. R.; Heck, A. J.; van Breukelen, B., Current challenges in software solutions for mass spectrometry-based quantitative proteomics. *Amino acids* **2012,** 43, (3), 1087-108.

17. Matzke, M. M.; Brown, J. N.; Gritsenko, M. A.; Metz, T. O.; Pounds, J. G.; Rodland, K. D.; Shukla, A. K.; Smith, R. D.; Waters, K. M.; McDermott, J. E.; Webb-Robertson, B. J., A comparative analysis of computational approaches to relative protein quantification using peptide peak intensities in label-free LC-MS proteomics experiments. *Proteomics* **2013,** 13, (3-4), 493-503.

18. Laemmli, U. K., Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **1970,** 227, (5259), 680-5.

19. Majovsky, P.; Naumann, C.; Lee, C. W.; Lassowskat, I.; Trujillo, M.; Dissmeyer, N.; Hoehenwarter, W., Targeted proteomics analysis of protein degradation in plant signaling on an LTQ-Orbitrap mass spectrometer. *Journal of proteome research* **2014,** 13, (10), 4246-58.

20. Vizcaino, J. A.; Csordas, A.; Del-Toro, N.; Dianes, J. A.; Griss, J.; Lavidas, I.; Mayer, G.; Perez-Riverol, Y.; Reisinger, F.; Ternent, T.; Xu, Q. W.; Wang, R.; Hermjakob, H., 2016 update of the PRIDE database and its related tools. *Nucleic acids research* **2016,** 44, (D1), D447-56.

21. MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J., Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **2010,** 26, (7), 966-8.

22. Zybailov, B.; Mosley, A. L.; Sardiu, M. E.; Coleman, M. K.; Florens, L.; Washburn, M. P., Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. *Journal of proteome research* **2006,** 5, (9), 2339-47.

23. Schwanhausser, B.; Busse, D.; Li, N.; Dittmar, G.; Schuchhardt, J.; Wolf, J.; Chen, W.; Selbach, M., Global quantification of mammalian gene expression control. *Nature* **2011,** 473, (7347), 337-42.

24. Silva, J. C.; Gorenstein, M. V.; Li, G. Z.; Vissers, J. P.; Geromanos, S. J., Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Molecular & cellular proteomics : MCP* **2006,** 5, (1), 144-56.

25. Sampson, J. N.; Boca, S. M.; Shu, X. O.; Stolzenberg-Solomon, R. Z.; Matthews, C. E.; Hsing, A. W.; Tan, Y. T.; Ji, B. T.; Chow, W. H.; Cai, Q.; Liu da, K.; Yang, G.; Xiang, Y. B.; Zheng, W.; Sinha, R.; Cross, A. J.; Moore, S. C., Metabolomics in epidemiology: sources of variability in metabolite measurements and implications. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **2013,** 22, (4), 631-40.

26. R Core Team. , R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. **2014,** ISBN 3-900051-07-0, URL http://www.R-project.org.

27. Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. , *Package 'lmerTest'*, 2016.

28. Trutschel, D., Schmidt, S., Grosse, I., Neumann, S. , Experiment design beyond gut feeling: statistical tests and power to detect differential metabolites in mass spectrometry data. *Metabolomics* **2015,** 11, 851-860.

29. Hoehenwarter, W.; Monchgesang, S.; Neumann, S.; Majovsky, P.; Abel, S.; Muller, J., Comparative expression profiling reveals a role of the root apoplast in local phosphate response. *BMC plant biology* **2016,** 16, 106.

30. Wilhelm, M.; Schlegl, J.; Hahne, H.; Moghaddas Gholami, A.; Lieberenz, M.; Savitski, M. M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; Mathieson, T.; Lemeer, S.; Schnatbaum, K.; Reimer, U.; Wenschuh, H.; Mollenhauer, M.; Slotta-Huspenina, J.; Boese, J. H.; Bantscheff, M.; Gerstmair, A.; Faerber, F.; Kuster, B., Mass-spectrometry-based draft of the human proteome. *Nature* **2014,** 509, (7502), 582-7.

31. Isaacson, T.; Damasceno, C. M.; Saravanan, R. S.; He, Y.; Catala, C.; Saladie, M.; Rose, J. K., Sample extraction techniques for enhanced proteomic analysis of plant tissues. *Nature protocols* **2006,** 1, (2), 769-74.

32. Rose, J. K.; Bashir, S.; Giovannoni, J. J.; Jahn, M. M.; Saravanan, R. S., Tackling the plant proteome: practical approaches, hurdles and experimental tools. *The Plant journal : for cell and molecular biology* **2004,** 39, (5), 715-33.

33. Jungblut, P. R.; Holzhutter, H. G.; Apweiler, R.; Schluter, H., The speciation of the proteome. *Chemistry Central journal* **2008,** 2, 16.

34. Nesvizhskii, A. I.; Aebersold, R., Interpretation of shotgun proteomic data: the protein inference problem. *Molecular & cellular proteomics : MCP* **2005,** 4, (10), 1419-40.

35. Luna, E.; Bruce, T. J.; Roberts, M. R.; Flors, V.; Ton, J., Next-generation systemic acquired resistance. *Plant physiology* **2012,** 158, (2), 844-53.

36. Luna, E.; Ton, J., The epigenetic machinery controlling transgenerational systemic acquired resistance. *Plant signaling & behavior* **2012,** 7, (6), 615-8.

37. Molinier, J.; Ries, G.; Zipfel, C.; Hohn, B., Transgeneration memory of stress in plants. *Nature* **2006,** 442, (7106), 1046-9.

38. Slaughter, A.; Daniel, X.; Flors, V.; Luna, E.; Hohn, B.; Mauch-Mani, B., Descendants of primed Arabidopsis plants exhibit resistance to biotic stress. *Plant physiology* **2012,** 158, (2), 835-43.

39. Nawrath, C.; Metraux, J. P., Salicylic acid induction-deficient mutants of Arabidopsis express PR-2 and PR-5 and accumulate high levels of camalexin after pathogen inoculation. *The Plant cell* **1999,** 11, (8), 1393-404.

40. Uknes, S.; Mauch-Mani, B.; Moyer, M.; Potter, S.; Williams, S.; Dincher, S.; Chandler, D.; Slusarenko, A.; Ward, E.; Ryals, J., Acquired resistance in Arabidopsis. *The Plant cell* **1992,** 4, (6), 645-56.

41. Ward, E. R.; Uknes, S. J.; Williams, S. C.; Dincher, S. S.; Wiederhold, D. L.; Alexander, D. C.; Ahl-Goy, P.; Metraux, J. P.; Ryals, J. A., Coordinate Gene Activity in Response to Agents That Induce Systemic Acquired Resistance. *The Plant cell* **1991,** 3, (10), 1085-1094.

Acknowledgments

MHD Rami Al Shweiki was sponsored by the DAAD funding program “Study Scholarships for Graduates of All Disciplines 2015/2016” No. 57139980. We thank the Leibniz Association for support and funding. We thank Dr. Jan Grau for insightful discussion of statistical procedures. We thank Dr. Steffen Neumann for critical discussion of the manuscript. The publication of this article was funded by the Open Access fund of the Leibniz Association.

The authors declare no competing financial interest.

Supporting Information. The following files are available free of charge.

Supplementary Table 1. Search of Web of Science with proteomics software names.

Supplementary Table 2. Composition of the protein standard mixtures.

Supplementary Table 3. Biologically variant proteins.

Supplemental Figure 1. SDS-PAGE of the selected standards.

Supplemental Figure 2. Quantification of the standards under naked conditions.

Supplemental Figure 3. Quantification of the standards in *Arabidopsis* matrix

Supplemental Figure 4. Variance of protein abundance estimates

Figure Legends

Figure 1. Inference of Standard Protein Abundance from Measured Peptide Ion Signal Response. A. The mean number of PSMs in the three measurements of each molar amount of each standard protein normalized to the respective standard protein's molecular weight is plotted. Only peptides with a sum of 24 or greater mean PSMs in measurements of all six molar amounts were plotted. This cut-off was chosen to eliminate peptides with less than 4 PSMs at any one point of the dilution series (6*4 = 24) as PSM based quantification with values below 4 was shown to be highly inaccurate [citation]. B. Quantification of ßLA and ßLB isoforms. Top panel; mean number of PSMs and mean Top3 PQI (ion signal peak area of three most intense peptide ion signals) in the three measurements of each molar amount of common peptides between ßLA and ßLB. Error bars denote standard errors (SE). Center panel; as top panel but PQIs for the tryptic peptide WENDEc(Carb)AQK from AA 61 to 69 on ßLA discriminating the isoforms. Bottom panel; as center panel for the homologous peptide on ßLB WENGEc(Carb)AQK. C. The proportionality between measured and actual sample protein abundance expressed as a function of peptide ion signal response. The number of PSMs multiplied by the sequence coverage normalized to the molecular weight of every standard protein at each molar amount is set in relation to the respective molar amounts on the $\log_{10}$ scaled ordinate. Note this corresponds to the slope of a linear model of proportionality between measured and actual protein abundance. The number of PSMs of each peptide of each standard protein at each molar amount is on the abscissa. There is a general trend towards direct proportionality especially for larger, more abundant proteins. Dashed lines parallel to the x-axis show 0.5 and 1.5 ordinate values, respectively.

Figure 2. Validation of Label-free Quantitative Proteomics Software. A. Left panel; proportionality between measured and expected standard protein $\log_{10}$ fold changes. Log_{10} fold changes were calculated exclusively for mean PQI values of four measurements of standard protein molar amounts. Center panel; percent error and variability of measured standard protein $\log_{10}$ fold changes. Percent

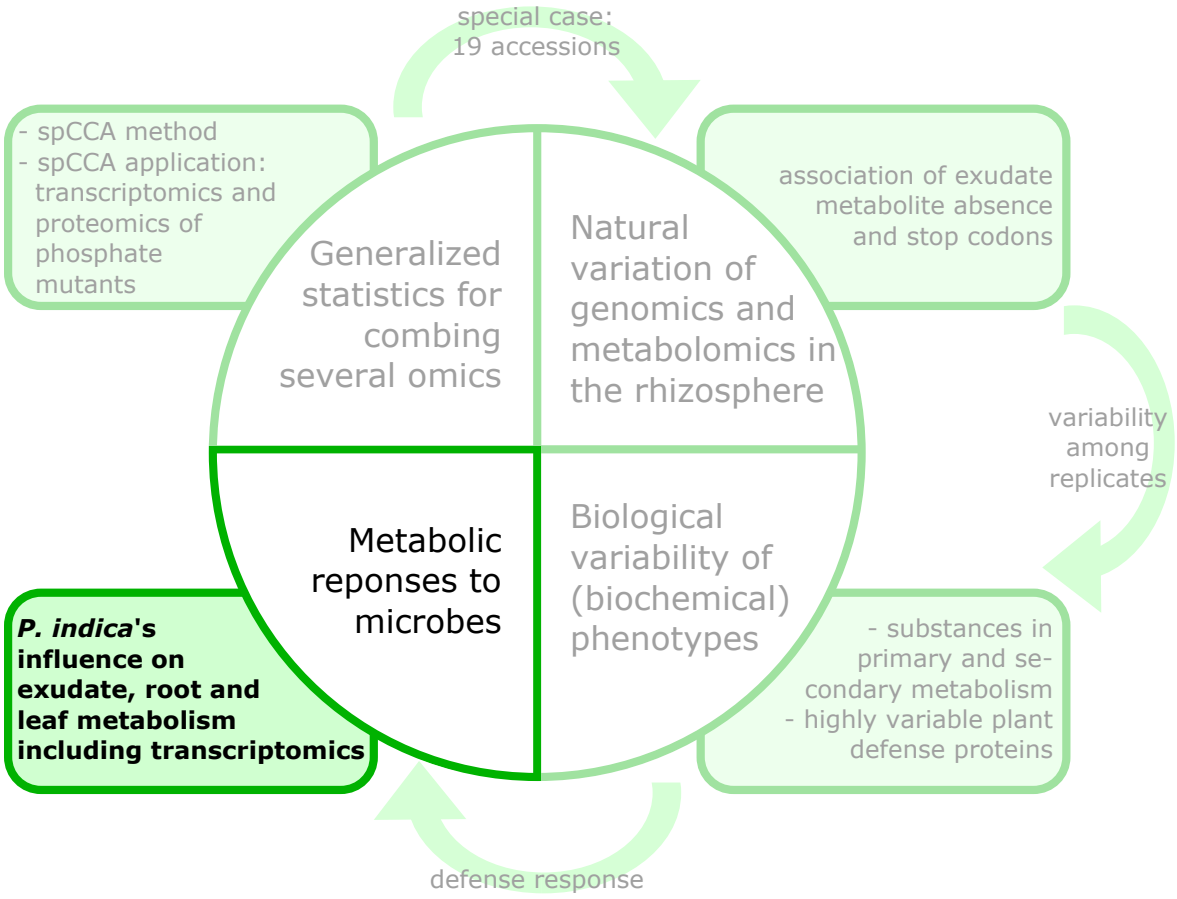
error was calculated as the difference between measured and expected $\log_{10}$ fold changes normalized to the respective expected $\log_{10}$ fold change multiplied by 100. When the $\log_{10}$ expected fold change was zero, only the difference multiplied by 100 was used. As above, only $\log_{10}$ measured fold changes calculated for mean PQI values of four measurements of standard protein molar amounts were considered. The RSD of measured $\log_{10}$ fold changes was calculated from all possible pair wise fold changes of four measurements of standard protein molar amounts (4*4 = 16). Dashed horizontal lines denote -30% and 30% error respectively. Coloring is according to $\log_{10}$ fold change (from green to blue). Right panel; Volcano plot showing the median $\log_{10}$ fold change of measured to mean PQI of 12 or more measurements of *Arabidopsis thaliana* background matrix proteins and the standard deviation of the respective $\log_{10}$ fold changes. Dashed horizontal lines denote 0.097 and 0.079 $\log_{10}$ fold changes (0.8 and 1.2 fold changes) respectively. Coloring is according to PQI signal intensity (from red to blue). FC denotes Fold change. B. Heat map showing standard proteins quantified at molar amounts with each of the evaluated software parameter sets. Red indicates standard proteins were quantified in all four measurements of the respective molar amount, white indicates a failure to do so. Note the molar amounts for ßLA/ßLB differ from the six point dilution series of the other standards because peptides shared between the two could be quantified. Therefore ßLA/ßLB molar amounts are the sum of the two in each mixture (in ascending order: 20, 60, 400, 1100, 1200, 1300 fmol).

Figure 3. Analysis of the contributions of the shotgun proteomics technology (technical variance) and the plants (biological variance) to the total variance of protein abundance. A. The total variance of the abundance of every protein measured in all 3 repeated analyses of every plant in the nested experimental design and quantified by three software parameter sets (QIP-def-Top3, MQ-MBR-LFQ and MQ-MBR-iBAQ) was decomposed into technical and biological variance using a LMM for each of the software parameter sets. Mean variances for all proteins are indicated by horizontal bars and by numbers. B. The ratio of biological to total variance (ICC) is shown for the accumulating fraction of all quantified proteins. The 60, 80 and 90% quantiles are indicated by dashed lines. C. The possible

combinations of the numbers of plants and repeated analyses of each plant to detect fold changes in protein abundance of 1.41 with 95% confidence and 80% power are shown. D. Top panel: The expected protein abundance values (PQI) of each quantified protein were estimated for the three software and parameter sets using model-based least square means and plotted as independent variables for pair wise comparisons of software and parameter sets. Center panel: $\log_2$ fold changes of protein abundance values estimated as above for measurements of root proteomes deprived of and replete with phosphate (minus/plus phosphate). Bottom panel: As center panel showing only statistically significant fold changes (Student's two sample T-test, $n = 9$, $\alpha = 0.05$).

2.4 *Piriformospora indica* stimulates root metabolism of *Arabidopsis thaliana*

Strehmel, N.; Mönchgesang, S.; Herklotz, S.; Krüger, S.; Ziegler, J.; Scheel, D. *Piriformospora indica* Stimulates Root Metabolism of *Arabidopsis thaliana*. *Int J Mol Sci* **2016**, *17*.



International Journal of
Molecular Sciences

MDPI

Article

Piriformospora indica Stimulates Root Metabolism of *Arabidopsis thaliana*

Nadine Strehmel [1,*], Susann Mönchgesang [1], Siska Herklotz [1], Sylvia Krüger [1], Jörg Ziegler [2] and Dierk Scheel [1,*]

1 Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle (Saale), Germany; susann.moenchgesang@ipb-halle.de (S.M.); siska.herklotz@pflanzenphys.uni-halle.de (S.H.); sylvia.krueger@ipb-halle.de (S.K.)
2 Department of Molecular Signal Processing, Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle (Saale), Germany; joerg.ziegler@ipb-halle.de
* Correspondence: nstrehme@ipb-halle.de (N.S.); dierk.scheel@ipb-halle.de (D.S.); Tel.: +49(0)-345-5582-1400 (D.S.)

Academic Editors: Ute Roessner and Jianhua Zhu
Received: 14 May 2016; Accepted: 28 June 2016; Published: 8 July 2016

Abstract: *Piriformospora indica* is a root-colonizing fungus, which interacts with a variety of plants including *Arabidopsis thaliana*. This interaction has been considered as mutualistic leading to growth promotion of the host. So far, only indolic glucosinolates and phytohormones have been identified as key players. In a comprehensive non-targeted metabolite profiling study, we analyzed *Arabidopsis thaliana*'s roots, root exudates, and leaves of inoculated and non-inoculated plants by ultra performance liquid chromatography/electrospray ionization quadrupole-time-of-flight mass spectrometry (UPLC/(ESI)-QTOFMS) and gas chromatography/electron ionization quadrupole mass spectrometry (GC/EI-QMS), and identified further biomarkers. Among them, the concentration of nucleosides, dipeptides, oligolignols, and glucosinolate degradation products was affected in the exudates. In the root profiles, nearly all metabolite levels increased upon co-cultivation, like carbohydrates, organic acids, amino acids, glucosinolates, oligolignols, and flavonoids. In the leaf profiles, we detected by far less significant changes. We only observed an increased concentration of organic acids, carbohydrates, ascorbate, glucosinolates and hydroxycinnamic acids, and a decreased concentration of nitrogen-rich amino acids in inoculated plants. These findings contribute to the understanding of symbiotic interactions between plant roots and fungi of the order of Sebacinales and are a valid source for follow-up mechanistic studies, because these symbioses are particular and clearly different from interactions of roots with mycorrhizal fungi or dark septate endophytes

Keywords: plant; fungus; interaction; exudates; roots; leaves; metabolite profiling; liquid chromatography/mass spectrometry (LC/MS); gas chromatography/mass spectrometry (GC/MS)

1. Introduction

Piriformospora indica is a root-interacting endophytic fungus and has been found in the Indian Thar-Dessert [1]. It belongs to the order of *Sebacinaceous* (Basidiomycota) [2] and yields a growth promotion effect with various crop plants such as barley, tobacco, maize, and tomato, but also with the model plant *Arabidopsis thaliana* [3]. Previous studies showed that *P. indica* promotes nutrient uptake and helps plants to survive under biotic (pathogenic organisms) [4,5] and abiotic (water, temperature, salt, toxins, heavy metal ions) stress conditions [6,7]. Furthermore, it stimulates plant growth, biomass, and seed production [8,9]. The fungus colonizes the epidermal and rhizodermal part of the roots and forms pearshaped spores, which accumulate within the roots and on the root surface. *P. indica* grows inter- and intracellularly [10] but does not invade the endodermis and aerial parts of the plant.

This endosymbiotic interaction has been considered as mutualistic, as it leads to an improved nutrient state in the host [11]. After establishment of this endosymbiotic interaction, the plant obtains more phosphorous and water through extracellular hyphae of the fungus, whereas the fungus is supplied with nitrogen and hydrocarbons in form of plant amino acids [11–15].

P. indica can be cultivated with the model plant *A. thaliana*. In general, *P. indica* colonization is host-dependent and occurs in two phases: Early interactions are biotrophic in barley and *A. thaliana*, but can switch to saprotrophy or maintain biotrophy in later stages, respectively [15]. Host metabolism determines the availability of nitrogen, and the subsequent induction of nitrogen transporters and a possible nutritional switch of *P. indica* from biotrophy to saprotrophy. *A. thaliana* had been shown to provide sufficient nitrogen sources in form of increased levels of amino acids like Gln and Asn at 14 dpi.

During the initial phase of this interaction, defense genes are activated and reactive oxygen species (ROS) produced by the host against *P. indica* [16]. However, *P. indica* can rescue plants with elevated ROS levels by providing antioxidants [17]. After recognition of *A. thaliana*, *P. indica* releases effectors into the rhizosphere, which induce a response in the host [18]. Moreover, an increase in the intracellular calcium concentration in the host's root cells is provoked, which triggers an intracellular signaling cascade (mitogen-activated protein kinase signaling pathways) [19,20], whereupon ethylene signaling components and ethylene-response transcription factors are required [21,22]. Furthermore, cytokinins and auxins play a crucial role in the maintenance of this symbiotic interaction [23]. Lahrmann et al. [24] and others showed that the colonization of *A. thaliana* with *P. indica* correlates with the induction of salicylic acid catabolites and jasmonate as well as glucosinolate metabolism [25,26]. Indolics were identified as key players in the maintenance of this mutualistic interaction. Especially indolic glucosinolates and reaction products are required to restrict the growth of *P. indica*.

Since the genomes of both organisms have been completely sequenced, both partners offer an ideal opportunity to study mutualistic interactions of plants and root endophytes in the rhizosphere [27,28]. Thus, we investigated the metabolic response of *A. thaliana* to *P. indica* under hydroponic conditions by non-targeted liquid chromatography/mass spectrometry (LC/MS) and gas chromatography/mass spectrometry (GC/MS)-based metabolite profiling. For this purpose, we chose ultra performance liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry (UPLC/ESI-QTOFMS) for the profiling of secondary metabolites and gas chromatography coupled to electron ionization quadrupole mass spectrometry (GC/EI-QMS) for the profiling of primary metabolites. Both platforms gain a snapshot of biochemical processes within a cell. Whereas reversed-phase LC/MS allows for the profiling of semipolar compounds [29], namely indolics, flavonoids, phenylpropanoids, glucosinolates and their degradation products, GC/EI-QMS covers main parts of central carbon metabolism [30]. Regardless of the choice of analysis platform, all samples can be grouped according to their common metabolic fingerprint. For this purpose we set up a standardized co-cultivation system, which supports the growth of both partners in close association to each other and the consequent profiling of roots and their exudates as well as leaves.

2. Results and Discussion

To study the interaction of *P. indica* with *A. thaliana*, a sterile hydroponic cultivation system was developed, which allows for the simultaneous profiling of roots and their exudates (Supplementary Figure S1). For this purpose, *P. indica* was precultivated on agar plates and *A. thaliana* on agar-filled, bottom-cut PCR tubes protruding into a liquid culture medium. After two weeks, both organisms were brought together in half-strength Murashige-Skoog (MS) medium supplemented with 0.5% sucrose (*w*/*v*) and Gamborg B5 vitamins such as myo-inositol, nicotinic acid, pyridoxin, and thiamine. According to our preliminary studies both components are essential for this symbiosis and hence the growth promotion effect of *A. thaliana*.

2.1. P. indica Promotes Shoot Growth of A. thaliana under Specific Culturing Conditions in a Hydroponic System after Root Colonization

If both components (sucrose and Gamborg B5 vitamins) were supplied for the co-cultivation studies, the shoot biomass increased by 22% ($p = 4.2 \times 10^{-5}$) in inoculated samples compared to the control confirming the previously reported growth promotion effect in soil [31]. Although *P. indica* colonizes the roots, the root biomass did not change after two weeks of co-cultivation (Figure 1) leading to the assumption that *P. indica* might provoke a systemic effect in *A. thaliana*. Although previous studies have shown a growth promotion effect in roots [23,31], we anticipated slight deviations in a hydroponic system compared to soil due to different physicochemical properties.

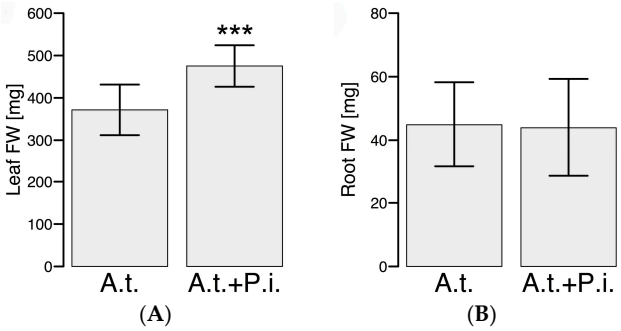


Figure 1. Leaf and root fresh weight of *A. thaliana* (A.t.) after co-cultivation with *P. indica* (P.i.) in a hydroponic system. *A. thaliana* was co-cultivated for two weeks with an agar plug containing mycelia of *P. indica*. For control *A. thaliana* was solely cultivated with an agar plug in 0.5× Murashige & Skoog (MS) medium supplemented with 0.5% sucrose (w/v) and vitamins: (**A**) shoot fresh weight (FW); (**B**) root fresh weight (FW). Values represent the mean ± SD (standard deviation) of three independent experiments (control samples: $n = 3 \times (3 - 5)$ and co-cultivated samples: $n = 3 \times 5$). Each replicate n comprises a pool of 24 plants. Significance analysis of differences between control and co-cultivated samples was performed by *t*-test: ***, $p \leqslant 0.001$.

To investigate how *P. indica* interacts with the host in a hydroponic system, fluorescence microscopy images were recorded using green fluorescent protein (GFP)-labeled *P. indica* and the interaction monitored at 14 dpi.

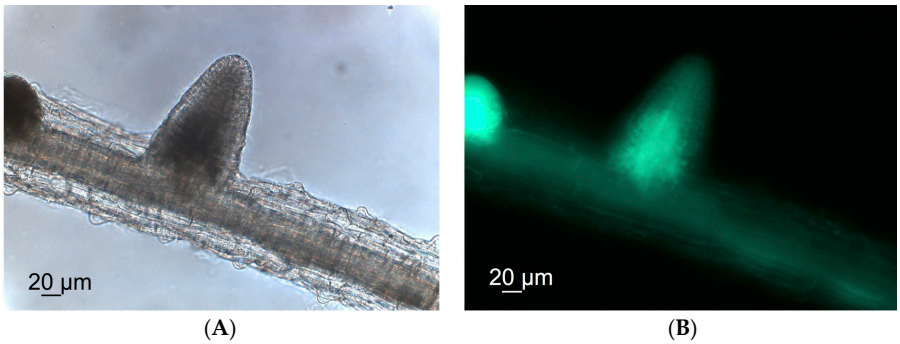


Figure 2. Microscopic image of an inoculated root with a GFP-labeled *P. indica* strain. (**A**) brightfield image; (**B**) fluorescence image.

P. indica grows inter- and intracellularly in root cells of *A. thaliana* when co-cultivated in soil [31]. In order to test if *P. indica* still forms fungal hyphae at the root surface under hydroponic conditions, a GFP-labeled *P. indica* strain was used to visualize colonization. Only weak autofluorescence signals

were detected in the non-inoculated roots and roots inoculated with the non-labeled *P. indica* strain (Supplementary Figure S2). In contrast, roots inoculated with the GFP-labeled strain exhibited a very strong fluorescence already after a 3 s exposure time showing that *P. indica* colonizes the root surface and penetrates the root of *A. thaliana* (Figure 2). Interestingly, *P. indica* was predominantly detected in lateral roots. According to these observations, we concluded that *P. indica* colonizes roots of *A. thaliana* and as a consequence likely leads to changes in root and shoot metabolism. So far, only indolic glucosinolates and hormones have been discussed in depth [24,26].

2.2. *P. indica* Alters the Exudation of Secondary Metabolites by *A. thaliana* Roots

Hormonal regulation has been described to accompany the colonization of *P. indica* on *A. thaliana* roots [22–24,32–35]. An enrichment analysis (Table 1 and Supplementary Table S1) of the upregulated root transcripts 14 dpi as published in Lahrmann et al. [24] revealed an overrepresentation of genes involved in the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway "Biosynthesis of plant hormones" (ath01070).

Table 1. Overrepresented KEGG pathways among upregulated *A. thaliana* root transcripts 14 dpi [15].

Term	Fold Enrichment	*p*-Value *
ath00966: Glucosinolate biosynthesis	10.4	8.89×10^{-8}
ath00940: Phenylpropanoid biosynthesis	3.8	5.84×10^{-7}
ath00360: Phenylalanine metabolism	3.7	6.21×10^{-5}
ath00903: Limonene and pinene degradation	3.8	1.12×10^{-4}
ath00680: Methane metabolism	3.5	1.70×10^{-4}
ath00945: Stilbenoid, diarylheptanoid and gingerol biosynthesis	3.7	2.20×10^{-4}
ath00910: Nitrogen metabolism	3.9	5.56×10^{-3}
ath00260: Glycine, serine and threonine metabolism	3.7	7.07×10^{-3}
ath00460: Cyanoamino acid metabolism	5.0	1.25×10^{-2}
ath00960: Tropane, piperidine and pyridine alkaloid biosynthesis	5.5	2.05×10^{-2}
ath01070: Biosynthesis of plant hormones	1.6	3.54×10^{-2}
ath00400: Phenylalanine, tyrosine and tryptophan biosynthesis	3.3	4.44×10^{-2}

* *p*-value was corrected according to Benjamini-Hochberg.

As shown in Supplementary Figure S3, *P. indica* significantly affects phytohormone levels in root exudates and roots, respectively. A higher concentration of hormones was found in exudates of co-cultivated samples as compared to control samples. This effect was in particular pronounced for jasmonate (JA), and jasmonyl-isoleucine (JA-Ile), both showing a more than 10-fold increase in the exudates and its potential role was discussed in reference [24]. In roots, the hormone content was also increased, but to a lower extent for JA, and JA-Ile, for which only a two- to four-fold increase was observed. 12-oxo-phytodienoic acid (OPDA), the precursor of JA and JA-Ile, also accumulated in roots but could not be detected in exudates irrespective of the conditions.

Besides the transcriptional regulation of hormone biosynthesis, hormone responses were overrepresented biological processes in Gene Ontology (GO:0009753 response to jasmonic acid stimulus, GO:0009751 response to salicylic acid stimulus). The analysis of Gene Ontology (GO) terms (Supplementary Table S1) further pointed to the involvement of secondary metabolic processes as the top two enriched processes (GO:0019748) ranked after defense response (GO:0006952). Consequently, roots and their exudates were comprehensively profiled for changes in primary and secondary metabolism upon *P. indica* colonization. Root exudates were only profiled for changes in semipolar metabolism, as in a screen for primary metabolites (GC/MS) all blank samples (samples without plant and/or fungus) already exhibited a considerable number of primary metabolites. Representative base peak chromatograms are depicted in Figure 3 and reveal a unique metabolic fingerprint for both ionization modes, ESI(+) and ESI(−). A principal component analysis (PCA) could verify this

assumption. For both ionization modes 89% of the total variance was explained by the first principal component (PC1) and 3% ESI(+) as well as 4% ESI(−) by PC2 (Supplementary Figure S4).

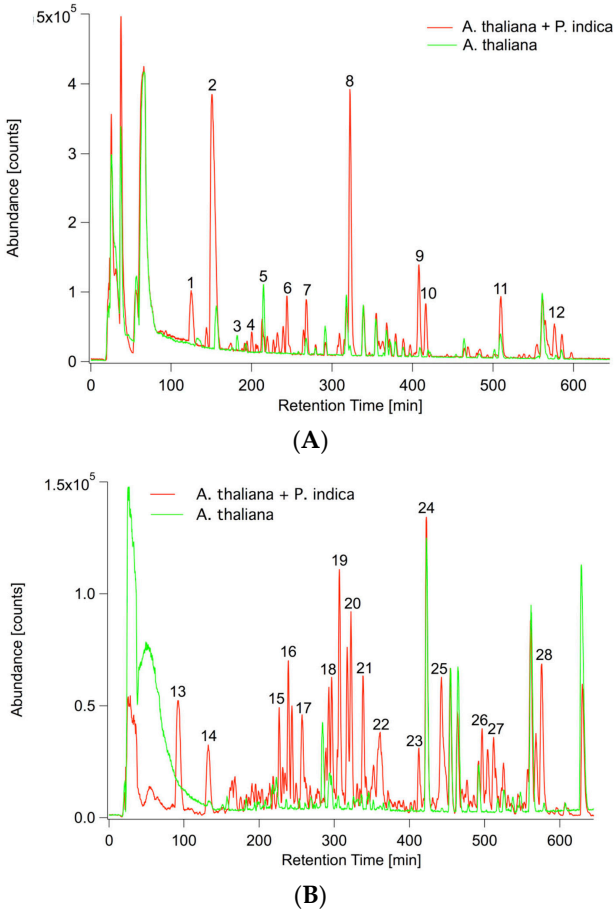


Figure 3. Overlay of representative UPLC/ESI(+/−)-QTOFMS base peak chromatograms (*m*/*z* 100–1000) of inoculated (red) and non-inoculated (green) *A. thaliana* exudates. (**A**) ESI(+): positive ionization mode; (**B**) ESI(−): negative ionization mode. **1**: 8-MeSO–Octyl–NH_2; **2**: $C_{10}H_{15}N_3$; **3**: H–Ile–Ile–OH; **4**: 1-MeO–I3CH_2NH_2; **5**: $C_9H_7N_3O_3$; **6**: $C_{17}H_{34}NO_9P$; **7**: Scopoletin; **8**: 8-MeSO–Octyl–CN; **9**: $C_{16}H_{29}NO_8$; **10**: $C_{12}H_{20}O_4$; **11**: $C_{14}H_{28}O_5$; **12**: $C_{28}H_{42}O_6$; **13**: Pantothenic acid; **14**: $C_{16}H_{26}O_8$; **15**: $C_{16}H_{23}N_3O_8$; **16**: $C_{16}H_{23}N_3O_8$; **17**: $C_{13}H_{24}O_6$; **18**: $C_{12}H_{22}O_5$; **19**: $C_9H_{18}O_4$; **20**: $C_{13}H_{22}O_5$; **21**: $C_{12}H_{22}O_5$; **22**: $C_{25}H_{41}N_3O_9$; **23**: $C_{14}H_{26}O_5$; **24**: Internal standard 2,4-Dichlorophenoxyacetic acid; **25**: 9,12,13-Trihydroxyoctadec-10-enoic acid; **26**: $C_{12}H_{18}O_4$; **27**: $C_{28}H_{44}O_6$; **28**: $C_{28}H_{42}O_6$.

Non-targeted UPLC/ESI(+/−)-QTOFMS-based metabolite profiling revealed that the concentration of 200 out of 341 detected ESI(+) components as well as 271 out of 377 ESI(−) components was significantly affected ($p < 0.01$) due to the presence of *P. indica*. A total of 28 (ESI(+)) as well as 24 (ESI(−)) components were down- and 172 (ESI(+)) as well as 247 (ESI(−)) components were upregulated due to the inoculation implying that *P. indica* stimulates root exudation of *A. thaliana*.

As already observed by Lahrmann et al. [24], the amount of compounds associated with nucleoside and aromatic amino acid metabolism was reduced in concentration by the inoculation, while that of aliphatic and indolic glucosinolate metabolism (except for 4-hydroxy-indole-3-carbaldehyde), dihydroxybenzoic acid (DHBA) conjugates, JA metabolism as well as fatty acid and pantothenic acid metabolism was increased. A number of phenylpropanoids including coumarins and oligolignols

(except for scopoletin and G(8-*O*-4)FA sulfate) showed reduced levels in the exudates of inoculated samples (Figure 4) leading to the assumption that these oligomers are further metabolized inside the cell and not exuded, very likely to oligolignols or to lignin [36], a main constituent of the cell wall. Both, glucosinolate (ath00966) and phenylpropanoid biosynthesis (ath00940), were among the overrepresented KEGG pathways of root transcripts (Table 1).

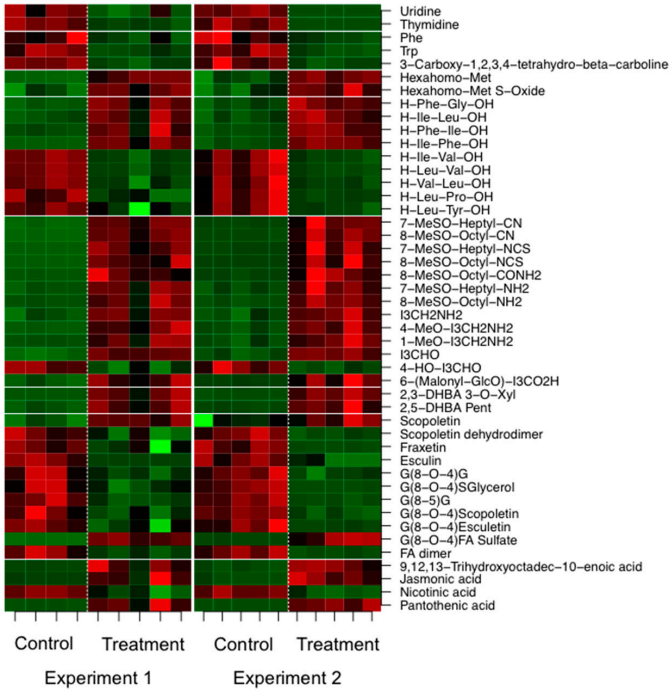


Figure 4. Differentially expressed metabolites ($p \leqslant 0.01$) in root exudates of *A. thaliana* after co-cultivation with the fungus *P. indica* for two weeks across two independent biological experiments. Intensity values were log-transformed and z-scored row-wise. Red: maximal intensity; Green: minimal intensity.

Nicotinic acid, an important precursor for vitamin B6, and thus, key player in the photoprotection of plants [37], also decreased in concentration upon co-cultivation in the root exudates (Figure 4). Obviously, nicotinic acid is required by *P. indica*. If this compound was not supplemented, no growth-promoting effect was observed of the host.

In the exudates we also detected differences in the dipeptide pool, namely the concentration of Phe-Gly, Ile-Leu, Phe-Ile and Ile-Phe was enhanced, while that of Ile-Val, Leu-Val, Val-Leu, Leu-Pro and Leu-Tyr was reduced in the co-cultivated samples (Figure 4). These differences might originate from different functionalities of the respective dipeptides and require further investigation. So far, Komarova et al. [38] showed that peptide transporters (*AtPTR5* and *AtPTR5*) facilitate the uptake of nitrogen from the rhizosphere.

2.3. Changes in the Root Metabolism of A. thaliana

The secondary metabolic changes detected in root exudates, especially that of glucosinolate biosynthesis, phenylpropanoid biosynthesis, and phenylalanine metabolism should also be reflected in root metabolism. In addition, transcriptionally enriched KEGG pathways of primary metabolism (Table 1), such as nitrogen metabolism (ath00910), glycine, serine and threonine metabolism (ath00260), and cyanoamino acid metabolism (ath00460) were expected in the GC/MS-based metabolite profiles.

2.3.1. Non-Targeted GC/MS Based Metabolite Profiling Reveals Perturbations in the Primary Root Metabolism

Figure 5 shows two representative extracted ion chromatograms of m/z 73 (equals $C_3H_9Si^+$ and is a typical fragment for trimethylsilylated compounds) obtained from a pool of inoculated and non-inoculated roots. Again, the inoculated profile is distinct compared to the non-inoculated root metabolic profile. Forty-eight percent of the total variance was explained by PC1 and 13% by PC2 and are plotted in Supplementary Figure S5.

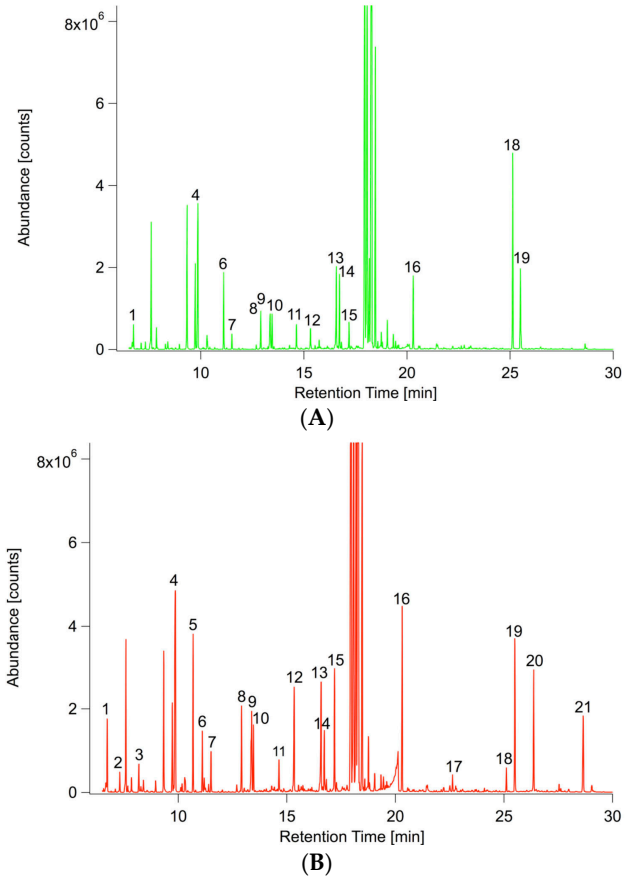


Figure 5. Representative extracted ion chromatograms (m/z 73) of inoculated and non-inoculated *A. thaliana* root extracts. (**A**) Non-inoculated root; (**B**) with *P. indica* inoculated root. **1**: Lactic acid (2TMS); **2**: Alanine (2TMS); **3**: Sulfuric acid (2TMS); **4**: Phosphoric acid (3TMS); **5**: Glyceric acid (3TMS); **6**: Serine (3TMS); **7**: Threonine (3TMS); **8**: Malic acid (3TMS); **9**: Pyroglutamic acid (2TMS); **10**: GABA (3TMS); **11**: Glutamic acid (3TMS); **12**: Asparagine (3TMS); **13**: Glutamic acid (3TMS); **14**: Glutamine (3TMS); **15**: Citric acid (4TMS); **16**: Myo-Inositol (6TMS); **17**: Glucose-6-phosphate (1MeOX, 6TMS); **18**: Thiamine hexoside; **19**: Sucrose (8TMS); **20**: Unknown; **21**: Unknown.

Non-targeted GC/EI-Q-MS based metabolite profiling revealed 287 out of 801 differentially accumulated components. Among them, we detected amino acids (e.g., Asn, Thr, Leu, 3-Cyano-Ala, beta-Ala, Val, Ala, Gln, ornithine, Pro, pyro-Glu, and GABA), organic acids (e.g., citrate, 2-oxoglutarate, fumarate, malate, oxalate, glycerate, fumarate, and 3-hydroxy-3-methylglutaric acid), carbohydrates (e.g., 1-*O*-methylglucopyranoside, 1-*O*-methylgalactopyraoside, maltose, raffinose, trehalose, xylose, ribose), polyols (erythritol, myo-inositol), phosphates (e.g., glycerol-3-phosphate, phosphate, glycerophosphoglycerol), and sulfates (e.g., sulfate, thiamine, thiamine-hex) belonging to the starch

and sucrose metabolism, glycolysis, tricarboxylic acid (TCA) cycle, amino acid metabolism, and urea metabolism. All compounds showed increased levels in the inoculated roots (Figure 6) except for pyruvate, erythritol, allantoin, and 4-methyl-5-thiazoleethanolglycopyranoside (for spectrum see Supplementary Figure S6) indicating that the initially applied amount of sucrose and thiamine is metabolized by *P. indica*. We observed an increase in the concentration of Asn, Gln, Ser, Thr, and Ala at 14 dpi. Serine was also increased in its levels as described by Lahrmann et al. [15], but did not pass the defined significance level ($p = 0.051$). In general, the data collected are in good accordance with the transcriptional changes and lead us to the hypothesis that *A. thaliana* provides nitrogen to the fungus so that *P. indica* can maintain a biotrophic nutritional state [15]. In the leaf profiles, the N-rich amino acids (Gln, Arg, Asn, 3-Cyano-Ala, and ornithine) were among the few differentially accumulated compounds decreasing in concentration upon colonization and consequently showed the opposite trend (Supplementary Figure S7) compared to the roots. This raises the question if these amino acids are transported to the root to feed *P. indica*. Most likely, these amino acids are required to balance the nutritional state of *P. indica*. To trace the flow of nutrients, further investigations are required. The change in the concentration of organic acids and carbohydrates was comparable for roots and leaves except that less differential changes were observed in the leaf profiles. These results again show that *P. indica* activates primary root metabolism of *A. thaliana*. According to our data, both partners appear to offer each other nutrients to maintain a mutualistic interaction, since an enhanced amount of P, S, and N (in the form of amino acids) was observed in roots of *A. thaliana* colonized with *P. indica*.

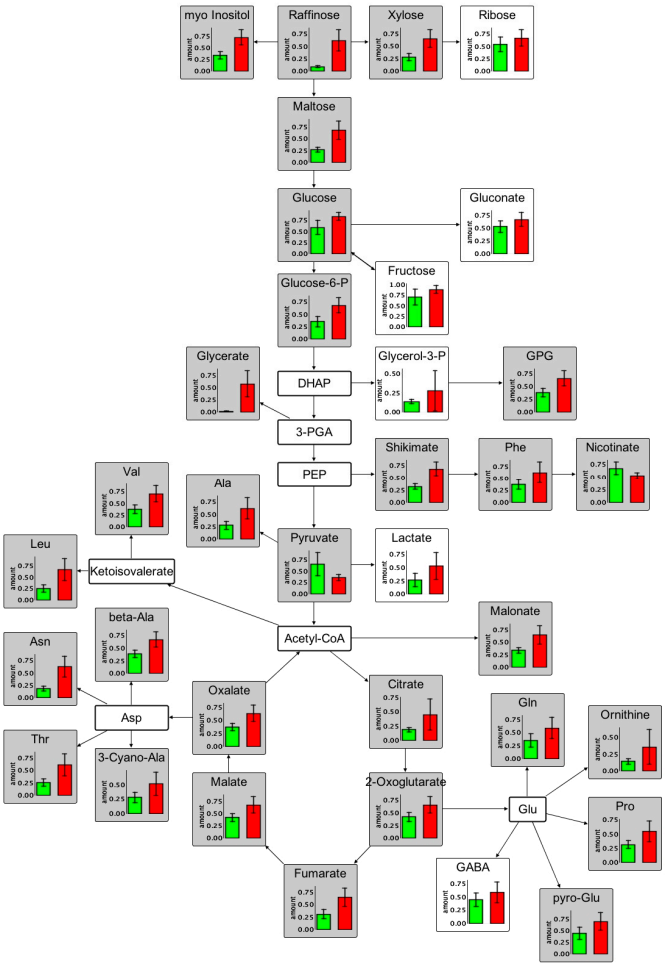


Figure 6. *Cont.*

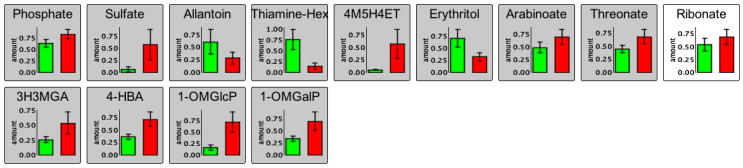


Figure 6. Differentially expressed primary metabolites occurring in root extracts of *A. thaliana*. Control and treatments are color-coded; control: *A. thaliana* (green) and treatment: *A. thaliana* + *P. indica* (red). Compounds with $p < 0.01$ are specifically marked by grey color or left blank for $0.01 \leqslant p \leqslant 0.05$. GPG: glycerophosphoglycerol; 4M5HET: 4-methyl-5-hydroxyethylthiazole; 3H3MGA: 3-hydroxy-3-methylglutaric acid; 4-HBA: 4-hydroxybenzoic acid; 1-OMGclP: 1-*O*-methyl-glucopyranoside; 1-OMGalP: 1-*O*-methylgalactopyranoside.

2.3.2. LC/MS Based Non-Targeted Metabolite Profiling Shows an Induction of Aliphatic and Indolic Glucosinolate Metabolism, Flavonoids, and Oligolignols in Roots

Besides primary metabolism, secondary root metabolism was investigated, since one category "secondary metabolic process" was a highly ranked candidate in the GO enrichment analysis. A unique fingerprint was observed in the root LC/MS profiles (Figure 7). According to Supplementary Figure S5, 76% of the entire variance was explained by PC1 and 0.07% by PC2 for the positive mode. These values were similar for the negative mode (PC1: 74%; PC2: 0.08%) and indicate that secondary metabolism is perturbed to a greater extent than primary metabolism.

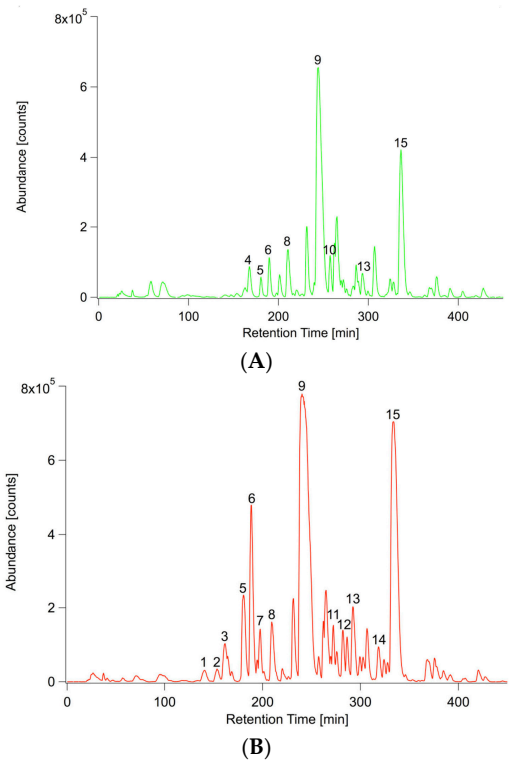


Figure 7. Representative UPLC/ESI(−)-QTOFMS base peak chromatograms (m/z 100–1000) of inoculated and non-inoculated *A. thaliana* root extracts. (**A**) Non-inoculated root; (**B**) with *P. indica* inoculated root. **1**: 7MeSO Heptyl GSL; **2**: 2,5 DHBA-Pent; **3**: I3M GSL; **4**: $C_{14}H_{18}O_{10}$; **5**: $C_{17}H_{24}O_{10}$; **6**: 8 MeSO Octyl GSL; **7**: Scopolin; **8**: 4MeO-I3M GSL **9**: 1MeO-I3M GSL; **10**: $C_{18}H_{32}O_{11}$; **11**: $C_{19}H_{18}O_3$; **12**: $C_{19}H1_8O_3$; **13**: 7MeS Heptyl GSL; **14**: $C_{38}H_{46}O_{18}$; **15**: 8MeS Octyl GSL.

Int. J. Mol. Sci. **2016**, *17*, 1091

In these profiles, 167 out of 329 detected compounds (ESI(+)) were altered in abundance and 188 out of 359 for the negative ionization mode due to the presence of *P. indica*. Similarly to the exudates, a higher number of compounds displayed upregulated abundance in the inoculated samples compared to the non-inoculated samples. From these numbers one can once more conclude that *P. indica* stimulates secondary root metabolism as well.

In accordance with Lahrmann et al. [24], aliphatic and indolic glucosinolates as well as their breakdown products, aromatic amino acids, coumarins, oligolignols, and flavonoids accumulated in inoculated roots (Figure 8) confirming the transcript data (KEGG, Table 1: glucosinolate ath00966 and phenylpropanoid biosynthesis ath00940). Although the plant seems to be in a defensive stage, no camalexin was detected in these profiles. In the leaf profiles an increased amount of aliphatic and indolic glucosinolates as well as their breakdown products, JA conjugates, oligolignols, and hydroxycinnamic acid amides was detected (Supplementary Figure S8). Several flavonoids (glycosylated kaempferol and quercetin) were only detected as differential in the root profiles and not in the leaf profiles, leading to the conclusion that this substance class plays an important role in the mutualistic interaction of *A. thaliana* and *P. indica*. Recently, Lahrmann et al. [24] stated that it remains to be clarified if flavonoids are accumulating in roots of *A. thaliana* upon interaction with *P. indica*. Indeed, we show that flavonoids accumulate in roots of *A. thaliana* upon co-cultivation with *P. indica*. Most likely, enhanced flavonoid biosynthesis, in addition to JA signaling [39], may also function as a signal for *P. indica*.

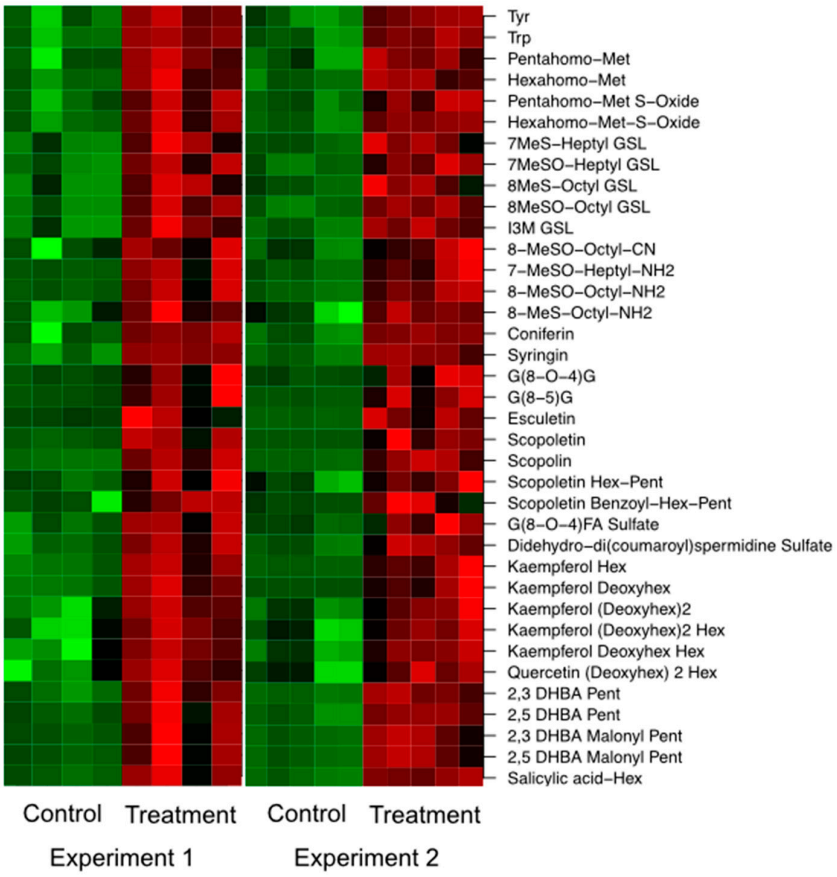


Figure 8. Differentially expressed secondary metabolites occurring in root extracts of *A. thaliana* across two independent biological experiments. Candidates were retrieved from a two-sided *t*-test ($p < 0.01$). For visualization, intensity values were log-transformed and z-scored row-wise. Red: maximal intensity; green: minimal intensity.

3. Materials and Methods

3.1. Chemicals and Standards

All chemicals were of highest analytical grade (>99%) and obtained from Carl Roth GmbH + Co. KG (Karlsruhe, Germany), Difco Microbiology (Lawrence, KS, USA), Duchefa Biochemie B.V. (Haarlem, The Netherlands), Merck KGaA (Darmstadt, Germany), and Sigma-Aldrich (Steinheim, Germany).

3.2. Pre-Cultivation of P. indica

P. indica was cultured on agar plates (1.5% (*w*/*v*) agar) for 3 weeks at 28 °C in the dark using Aspergillus minimal medium [29]. For this purpose, a punched out agar block with mycelia of *P. indica* was placed in the center of a culture plate.

3.3. Conduction of Co-Cultivation Studies and Production of Plant Material

Co-cultivation studies were performed as previously described [25]. In short, two-week old *A. thaliana* plantlets were co-cultivated for two weeks with *P. indica* in a hydroponic system under short day conditions (23 °C, 8 h light, 180 µE· m^{-2}· s^{-1} and 21 °C, 16 h dark). After two weeks of co-cultivation (four-week old plants), the medium containing the nutrient solution and the root exudates was filtered and stored at 4 °C in Schott flasks until further processing. At harvest, roots were cut below the bottom of the PCR tube and blotted dry with a paper towel before shock freezing in liquid nitrogen. Finally, they were stored at −80 °C until further processing. More technical details are visualized in Supplementary Figure S1. Media composition is summarized in Appendix A.

3.4. LC/MS-Based Metabolite Profiling

For LC/MS-based metabolite profiling (**UPLC**: Acquity, Waters, Eschborn, Germany; **MS**: MicrOTOF–Q I hybrid quadrupole time-of-flight mass spectrometer equipped with an Apollo II electrospray ion source, Bruker Daltonik GmbH, Bremen, Germany), the ground tissue material was processed by solid liquid extraction using methanol/water, 80/20 (*v*/*v*) (40 mg root fresh weight corresponds to 200 µL extraction solution and 50 mg leaf fresh weight corresponds to 400 µL extraction solution). Analytes of the nutrient solution were extracted by a reversed-phase solid phase extraction procedure (180 mL medium result in 120 µL analysis solution).

3.4.1. Preparation of Nutrient Solutions for LC/MS Analysis

All exudate samples were prepared and analyzed by UPLC/ESI-QTOFMS as presented in Lahrmann et al. [24]. In short, the nutrient solution was spiked with 20 µM 2-(2,4-dichlorophenoxy) acetic acid, evaporated until dryness, reconstituted in 9 mL water/methanol 95/5 (*v*/*v*) and subjected to a Bond Elut PPL cartridge (200 mg, 3 mL, Agilent Technologies, Böblingen, Germany). Finally, the eluate was subjected to a solid phase extraction workup and reconstituted in 120 µL water/methanol 70/30 (*v*/*v*) prior to LC/MS analysis. Technical details of the solid phase extraction workup can be found in the Appendix B.

3.4.2. Sample Preparation and Profiling of Tissue Material for LC/MS Analysis

The plant material was processed according to Böttcher et al. [29]. As already described, the frozen material was extracted twice with methanol/water, 80/20 (*v*/*v*) and reconstituted in methanol/water, 30/70 (*v*/*v*) prior to LC/MS-analysis. More details of the extraction procedure can be found in the Appendix B.

3.4.3. Non-Targeted LC/MS-Based Profiling and Data Analysis

Changes in the secondary plant metabolism were analyzed by UPLC/ESI-QTOFMS. Samples were injected onto an Acquity UPLC system (Waters, Eschborn, Germany), equipped with an HSS

T3 column (100 × 1.0 mm, particle size 1.8 µm, Waters), and separated using a binary gradient (A: water/0.1% (*v*/*v*) formic acid; B: acetonitrile/0.1% (*v*/*v*) formic acid). Eluting compounds were detected in positive and negative ionization mode from *m*/*z* 100–1000 using a MicroTOF–Q I hybrid quadrupole time-of-flight mass spectrometer equipped with an Apollo II electrospray ion source (Bruker Daltonics, Billerica, MA, USA). All instrument parameters and further settings can be found in the Appendix B.

Raw data files were converted to mzData using CompassXPort version 1.3.10 (Bruker Daltonics). For feature detection, alignment, and filling of missing values the R package XCMS version 1.41.0 [40] was used. Settings are summarized in Appendix B.

The intensities of the resulting features (*m*/*z*-retention time pairs) were $\log_2$ transformed and subjected to a two-sided Student's *t*-test. Relevant mass spectral features were extracted within a predefined range (isolation width: ±0.02 *m*/*z*) and elemental compositions were calculated applying a default error range (15 ppm). Putative elemental compositions were checked for consistency while analyzing elemental compositions of fragment ions and neutral losses of collision-induced dissociation (CID)-mass spectra. For acquisition of CID mass spectra quasi-molecular cluster ions were isolated at the Q1 (isolation width: ±3 *m*/*z*) and fragmented inside the collision cell using argon as collision gas. Product ions were detected as described above. All mass spectral data can be found in the MetaboLights repository (MTBLS341) [41].

3.5. GC/MS Based Metabolite Profiling

3.5.1. Sample Preparation of Tissue Material

One hundred µL extract of the remainder from the LC/MS-based metabolite profiling studies was spiked with 100 µM succinic acid-2,2,3,3-d_4, dried down in a vacuum concentrator, and stored at −20 °C until further processing.

3.5.2. Preparation of Samples for Non-Targeted Metabolite Profiling and Analysis of GC/MS Profiles

Dried down extracts were subjected to a two-step derivatization process using methoxyamine hydrochloride and *N*,*O*-bis(trimethylsilyl)trifluoroacetamide. Derivatized samples were injected splitless at 230 °C onto an Agilent 6890N GC equipped with a split/splitless inlet and a ZB-5 column (30 m × 0.25 mm, 0.25 µm 95% dimethyl/5% diphenyl polysiloxane film, 10 m integrated guard column, Phenomenex, Aschaffenburg, Germany). Eluting components were detected from *m*/*z* 70–600 by using an Agilent 5975 Series Mass Selective Detector (Agilent Technologies, Waldbronn, Germany). For the generation of the metabolite profiles, chromatograms were baseline-corrected using Metalign [42]. Peak intensities above 500 arbitrary ion current units were imported into the TagFinder software [43], aligned using the retention index model of van den Dool and grouped according to their common retention time and mass spectral features. For statistical analysis, peak intensities of cluster (cluster size > 3) were normalized to the internal standard (succinic acid-2,2,3,3-d_4). Then, all data were $\log_2$-transformed and submitted to a two-sided Students *t*-test. Finally, resulting mass spectral features were identified via best mass spectral and retention index match using the Golm Metabolome Database [44] and the NIST2012 software (May 2011, National Institute of Standards and Technology, Gaithersburg, MD, USA). Details of the derivatization protocol and instrument parameters can be found in the Appendix C.

All statistical analysis was either performed with the R statistical language, the Bioconductor environment, the package pcaMethods or Microsoft Excel.

3.6. Hormone Analysis

Hormone profiling was conducted as described in Ziegler et al. [45] (for further information see Appendix D). Root material was homogenized, extracted in methanol, and processed firstly using a hydrophobic solid phase extraction cartridge (Chromabond Sorbent HR-XC, Macherey-Nagel, Düren,

Germany) and secondly with an anion exchange solid phase extraction cartridge (Diethylaminoethyl Sephadex (DEAE-Sephadex)). For the root exudates the anion exchange step was omitted.

Analytes were separated by an Agilent 1290 Infinity HPLC system and detected on-line by ESI-MS/MS using an API 3200 triple-quadrupole LC-MS/MS system equipped with an ESI Turbo Ion Spray interface (AB Sciex, Darmstadt, Germany). Triple quadrupole scans were acquired in the multiple reaction monitoring mode (MRM) with Q1 and Q3 set at unit resolution. Scheduled MRM was performed with a window of 90 s and a target scan time of 0.1 s. Selected MRM transitions and compound specific parameters can be found in Ziegler et al. [45].

3.7. Microscope Images

Bright-field and fluorescence microscopic images were recorded with a Stemi 2000 Axio Imager stereomicroscope (Carl Zeiss MicroImaging GmbH, Göttingen, Germany). For bright-field images a Plan Apochromat 20×/0.75 objective with 20× magnification was used and for fluorescence images a Plan Apochromat 20×/0.75 objective with 20× magnification, a GFP-Filter 450–490 nm, filterset 9 and the Axio Imager camera.

3.8. Transcript Enrichment Analysis

Overrepresentation analysis of the overexpressed genes in *Arabidopsis* 14 dpi as published in Lahrmann et al. [24] was performed with DAVID [46,47] against the default background genes from TAIR using KEGG pathways [48] and Gene Ontology [49].

3.9. Data Availability

All data sets are available from the MetaboLights repository [41] under the accession number MTBLS341.

4. Conclusions

The mutualistic interaction of *P. indica* with *A. thaliana* resulted in an increased shoot biomass production, but not root biomass after a two-week co-cultivation. Interestingly, the presence of *P. indica* had an obvious effect on the root's primary and secondary metabolism and the exudation rate, but not on leaf metabolism of *A. thaliana*. Apparently, *P. indica* stimulates the belowground metabolism of *A. thaliana*, but not the shoot metabolism. The metabolic changes identified can be considered as potential biomarkers, which need to be tackled in the near future. Previous studies and this study have shown that indolic glucosinolates and hormones are important for the interaction. The induction of the defense response might indicate that the plant tries to balance fungal growth and maintain its mutualism. This assumption could be confirmed by the analysis of appropriate mutants. In the future, new mutants, especially of the flavonoid metabolism, need to be obtained to investigate the mutualistic interaction in more depth. It is possible that plant-growth promoting microorganisms can be valuable tools for crop improvement [7,50], as they promote the plant growth and help the plant to cope with abiotic and biotic stress factors.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/17/7/1091/s1.

Acknowledgments: This work was supported by the German Leibniz association (PAKT project 'Chemical Communication in the Rhizosphere'). The authors thank Steffen Neumann for upload of the data into MetaboLights and Alga Zuccaro for proofreading and the GFP-labeled *P. indica* strain. The publication of this article was funded by the Open Access fund of the Leibniz Association.

Author Contributions: Dierk Scheel and Nadine Strehmel conceived and designed the experiments; Siska Herklotz and Sylvia Krüger performed the experiments; Nadine Strehmel and Susann Mönchgesang analyzed the LC/MS and GC/MS data; Nadine Strehmel wrote the paper; Susann Mönchgesang and Dierk Scheel edited parts of the manuscript; Jörg Ziegler conducted the hormone profiling and proofread the manuscript. All authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

A. thaliana	*Arabidopsis thaliana*
P. indica	*Piriformospora indica*
GC	Gas chromatography
UPLC	Ultraperformance liquid chromatography
ESI	Electrospray ionisation
QTOF	Quadupole time of flight mass spectrometer
SD	Standard deviation
ET	Ethylene
JA	Jasmonic acid
GSL	Glucosinolate

Appendix A (Media for Co-Cultivation Studies)

Cultivation of *P. indica*: For the cultivation of *P. indica* Complete Medium was used and prepared as follows: *stock solution 1*: 12% (*w*/*v*) $NaNO_3$, 1.04% (*w*/*v*) KCl, 1.04% (*w*/*v*) $MgSO_4 \cdot 7H_2O$, 3.03% (*w*/*v*) KH_2PO_4 and *stock solution 2*: 0.6% (*w*/*v*) $MnCl_2 \cdot 4H_2O$, 0.265% (*w*/*v*) $ZnSO_4 \cdot 7H_2O$, 0.15% (*w*/*v*) H_3BO_3, 0.075% (*w*/*v*) KI, 0.025% (*w*/*v*) $Na_2MO_4 \cdot 2H_2O$, 0.013% (*w*/*v*) $CuSO_4 \cdot 5H_2O$. The final medium consisted of a mix of 5% (*v*/*v*) *stock solution 1*, 2% (*w*/*v*) Glucose, 0.2% (*w*/*v*) Bacto-Pepton, 0.1% (*w*/*v*) yeast extract, 0.1% (*w*/*v*) Casamino acids and 0.1% (*v*/*v*) *stock solution 2*.

Cultivation of *A. thaliana* and co-cultivation of *A. thaliana* with *P. indica*: For the pre- and co-cultivation stage 0.221% (*w*/*v*) Premix (M0231; Duchefa Biochemie B.V.) and 0.5% (*w*/*v*) sucrose in water were used. The pH was adjusted to 5.9 with 1 M KOH prior to autoclaving.

Appendix B (UPLC/ESI-QTOFMS)

C18-SPE: Bond Elut PPL cartridges were washed with 1 mL methanol, conditioned with 1 mL water/formic acid, 98/2 (*v*/*v*), loaded with 4 mL sample solution, washed with 1 mL water, and eluted with 2 mL methanol/formic acid, 99/2 (*v*/*v*); eluates were evaporated in a vacuum centrifuge and the residue were reconstituted in 120 µL water/methanol, 70/30 (*v*/*v*).

Extraction of root material: 200 µL methanol/water, 80/20 (*v*/*v*), pre-cooled at −28 °C, were added to the tissue; the mixture was allowed to reach room temperature within 15 min with occasionally vortexing; after sonication for 15 min at 20 °C and centrifugation for 10 min at 16,000× *g* the supernatant was transferred to a new 2 mL polypropylene tube; the remaining plant material was extracted a second time with 200 µL methanol/water, 80/20 (*v*/*v*); both extracts were combined and evaporated to dryness at 40 °C using a vacuum centrifuge; the residue was redissolved in methanol/water, 30/70 (*v*/*v*) according to fresh weight (40 mg = 200 µL), sonicated for 10 min at 20 °C, centrifuged for 5 min at 16,000× *g*, and the supernatant subjected to UPLC/ESI-QTOFMS analysis

UPLC settings: Full loop (loop volume: 2.5 µL); gradient: (flow rate: 150 µL· min^{-1}) 0–1 min, isocratic 95% A, 5% B; 1–16 min, linear from 5% to 95% B; 16–18 min, isocratic 95% B; 18–18.01 min, linear from 95% to 5% B; 18.01–20 min, isocratic 5% B.

ESI(+) settings: Nebulizer gas, nitrogen, 1.6 bar; dry gas, nitrogen, 6 L· min^{-1}, 190 °C; capillary, −5000 V; end plate offset, −500 V; funnel 1 RF, 200 Vpp; funnel 2 RF, 200 Vpp; in-source CID energy, 0 V; hexapole RF, 100 Vpp; quadrupole ion energy, 3 eV; collision gas, argon; collision energy, 3 eV; collision RF 200 Vpp; transfer time, 70 µs; pre pulse storage, 5 µs; spectra rate, 3 Hz.

ESI(−) settings: All parameters were maintained except for the nebulizer gas (1.4 bar), capillary (4000 V), quadrupole ion energy (5 eV), collision energy (7 eV), and collision RF (150 Vpp).

Data acquisition: centroid mode; recalibration on the basis of lithium formate cluster ions after injecting 20 µL 10 mM lithium hydroxide 49.9/49.9/0.2 (dissolved in isopropanol/water/formic acid; *v*/*v*/*v*).

XCMS settings: Feature detection with the help of the centWave algorithm (sntresh: 3, prefilter: (3.100), ppm: 25, peak width: (5.12); feature alignment with the help of the XCMS function group.density (minfrac: 0.75, bw: 2, mzwid: 0.05); missing values replacement by the XCMS function fillPeaks.

Analysis of raw data: DataAnalysis 4.2 software (Bruker Daltonics) for deconvolution and generation of extracted ion chromatograms

Appendix C (GC/EI-QMS)

Derivatization: Residues were reconstituted in 40 µL methoxyaminehydrochloride (20 mg/mL in pyridine, Sigma-Aldrich), the solution thoroughly vortexed and incubated at 40 °C for 1.5 h. An 80 µL mix comprising 70 µL *N*,*O*-bis(trimethylsilyl)trifluoroacetamide (BSTFA, Macherey-Nagel) and 10 µL alkane reference mixture (dodecane, pentadecane, nonadecane, docosane, octacosane and dotriacontane each to a final concentration of 80 $\mu g \cdot mL^{-1}$ dissolved in pyridine) were added and incubation at 40 °C proceeded for an additional 30 min; the solution was centrifuged and the supernatant transferred to a GC vial.

GC settings: Carrier gas helium, constant flow: 1 $mL \cdot min^{-1}$; temperature program: 70 °C for 1 min, gradient 9 $K \cdot min^{-1}$ to 300 °C, 5 min at 300 °C.

EI settings: Transfer line 300 °C; ion source temperature 230 °C; scan rate 3 Hz

Metalign settings: Maximum amplitude: 6,000,000, peak slope factor: 0.5, peak threshold factor: 1, average peak width at half height: 5.

TagFinder settings: Peak Finder (Smooth Width Apex Finder: 3; Low Intensity Threshold: 500; Max: Merging Time Width: 0.3); Time Scanner (Time Scan Width: 1; Min Fragment Intensity: 500); Tag Gen Filter (Tag Mass: 76, 146, 150–600; Sample Counts > 5); Intensity Calculator (Simple: MAX_INTENSITY); Tag Correlation (Correlation Method: PearsonCor; Significance Level: SIG_005); Tag Clustering (Core Adjacency Option: SAME_CORE; Min Core Option: INPUT_VALUE); Tag Output (Min Cluster Size: 3)

Appendix D (Hormone Analysis)

Profiling of Root Tissue

Homogenization and extraction: Root material was homogenized in bead beater and extracted with; 200 µL methanol (supplemented with 2 ng abscisic acid-d_6 (ABA-d_6), 5 ng indole-3-acetic acid-$^{13}C_6$ (IAA-$^{13}C_6$), 5 ng jasmonic acid-d_6 (JA-d_6), 0.74 ng jasmonyl isoleucine-d_2 (JA-Ile-d_2), 30 ng 12-oxo phytodienoic acid (OPDA-d_5), 1.5 ng salicylic acid-d_4 (SA-d_4), 5 ng zeatin (Z-d_5), 5 ng trans-zeatin-riboside-d_5 (tZ9R-d_5), 5 ng dihydrozeatin riboside-d_5 (DHZR-d_5). After vortexing for 20 min the supernatant was clarified by two rounds of centrifugation at 10,000 rpm for 5 min. Before loading on the HR-XC SPE 1 mL water/acetic acid, 98/2 (v/v) was added.

HR-XC: The resin was conditioned with 1 mL methanol followed by 1 mL water (the liquid was passed through SPE 96 well plate (50 mg HR-XC resin per well) by centrifugation at 250× g for 5 min using a JS5.3 bucket rotor in an Avanti J-26XP centrifuge (Beckman). Samples were transferred to the SPE 96 well plate, the resin washed with 1 mL H_2O. Analytes were eluted successively by adding 1 mL MeOH (for acidic hormones) and 1 mL methanolic ammonia (0.35 M) for zeatins.

DEAE-Sephadex: The resin was washed with 1 mL methanol. The methanolic eluates from the HR-XC plates were loaded onto DEAE-Sephadex (acetate form, 50 $mg \cdot well^{-1}$) filled. After washing with 1 mL methanol, the analytes were eluted with 1 mL of 3 M acetic acid in methanol.

Further processing: eluates were transferred to 2 mL Eppendorf tubes and evaporated to dryness; residues were dissolved in 40 µL of 20% (v/v) methanol, diluted with 40 µL of water and centrifuged at 10,000× g for 10 min.

LC: Agilent 1290 Infinity HPLC; Nucleoshell RP18 column (50 × 3 mm, particle size 2.7 µm; Macherey-Nagel, Düren, Germany) at 30 °C; eluent (A: water/0.2% (*v*/*v*) acetic acid; B: acetonitrile/0.2% (*v*/*v*) acetic acid); flow rate: 0.5 mL· min^{-1}; gradient for cytokinins: 2% B for 0.5 min, followed by a linear increase to 28% B within 3 min; increase to 98% in 0.5 min followed by an isocratic period of 1.5 min at 98% B, starting conditions restored within the next 0.5 min, and the column was allowed to re-equilibrate for 1 min at 2% B; gradient for acidic phytohormones: B increased from 10% to 80% within 9 min after an initial hold at 10% B for 0.5 min; further increase to 98% B within 0.5 min; isocratic period at 98% B for 1.5 min; column re-equilibrated at 10% B for 1 min.

ESI(+) for cytokinins: curtain gas 50 psi, ion spray voltage 3500 V, ion source temperature 650 °C, nebulizing and drying gas 70 psi and 50 psi.

ESI(−) for acidic phytohormones: negative ion mode curtain gas 50 psi, ion spray voltage −4500 V, ion source temperature 350 °C, nebulizing and drying gas 70 psi and 50 psi.

Data evaluation: Peak areas were calculated automatically using the IntelliQuant algorithm of the Analyst 1.6.2 software (AB Sciex, Darmstadt, Germany) and manually supervised. All other calculations were performed with Excel (Microsoft Office Professional Plus 2010).

Profiling of Root Exudates

Sample preparation: Exudates were processed according to LC/MS-based metabolite profiling protocol; the residues were reconstituted in 200 µL methanol (supplemented with 0.5 ng ABA-d_6, 2.5 ng IAA-$^{13}C_6$, 1 ng JA-d_6, 0.1 ng JA-Ile-d_2, 4 ng OPDA-d_5, 0.4 ng SA, 2.5 ng Z-d_5, 2.5 ng tZ9R-d_5, 2.5 ng DHZR-d_5); incubated for 15 min at room temperature; after centrifugation, the supernatant was processed as described for root extracts, except for the omission of the DEAE-Sephadex SPE.

References

1. Verma, S.; Varma, A.; Rexer, K.H.; Hassel, A.; Kost, G.; Sarbhoy, A.; Bisen, P.; Bütehorn, B.; Franken, P. *Piriformospora indica*, gen. et sp. nov., a new root-colonizing fungus. *Mycologia* **1998**, *90*, 896–903. [CrossRef]
2. Weiss, M.; Selosse, M.A.; Rexer, K.H.; Urban, A.; Oberwinkler, F. Sebacinales: A hitherto overlooked cosm of heterobasidiomycetes with a broad mycorrhizal potential. *Mycol. Res.* **2004**, *108*, 1003–1010. [CrossRef] [PubMed]
3. Varma, A.; Bakshi, M.; Lou, B.; Hartmann, A.; Oelmueller, R. *Piriformospora indica*: A Novel Plant Growth-Promoting Mycorrhizal Fungus. *Agric. Res.* **2012**, *1*, 117–131. [CrossRef]
4. Sun, C.; Shao, Y.; Vahabi, K.; Lu, J.; Bhattacharya, S.; Dong, S.; Yeh, K.-W.; Sherameti, I.; Lou, B.; Baldwin, I.T.; et al. The beneficial fungus *Piriformospora indica* protects Arabidopsis from *Verticillium dahliae* infection by downregulation plant defense responses. *BMC Plant Biol.* **2014**, *14*, 268. [CrossRef] [PubMed]
5. Daneshkhah, R.; Cabello, S.; Rozanska, E.; Sobczak, M.; Grundler, F.M.W.; Wieczorek, K.; Hofmann, J. *Piriformospora indica* antagonizes cyst nematode infection and development in Arabidopsis roots. *J. Exp. Bot.* **2013**, *64*, 3763–3774. [CrossRef] [PubMed]
6. Camehl, I.; Sherameti, I.; Seebald, E.; Johnson, J.M.; Oelmüller, R. Role of Defense Compounds in the Beneficial Interaction Between *Arabidopsis thalina* and *Piriformospora indica*. In *Piriformospora indica: Sebacinales and Their Biotechnological Applications*; Varma, A., Kost, G., Oelmüller, R., Eds.; Springer-Verlag Berlin Heidelberg: New York, NY, USA, 2013; Volume 33, pp. 239–250.
7. Gill, S.S.; Gill, R.; Trivedi, D.K.; Anjum, N.A.; Sharma, K.K.; Ansari, M.W.; Ansari, A.A.; Johri, A.K.; Prasad, R.; Pereira, E.; et al. *Piriformospora indica*: Potential and Significance in Plant Stress Tolerance. *Front. Microbiol.* **2016**, *7*, 332. [CrossRef] [PubMed]
8. Das, A.; Kamal, S.; Shakil, N.A.; Sherameti, I.; Oelmueller, R.; Dua, M.; Tuteja, N.; Johri, A.K.; Varma, A. The root endophyte fungus *Piriformospora indica* leads to early flowering, higher biomass and altered secondary metabolites of the medicinal plant, *Coleus forskohlii*. *Plant Signal. Behav.* **2012**, *7*, 103–112. [CrossRef] [PubMed]
9. Prasad, R.; Kamal, S.; Sharma, P.K.; Oelmueller, R.; Varma, A. Root endophyte *Piriformospora indica* DSM 11827 alters plant morphology, enhances biomass and antioxidant activity of medicinal plant Bacopa monniera. *J. Basic Microbiol.* **2013**, *53*, 1016–1024. [CrossRef] [PubMed]

10. Varma, A.; Verma, S.; Sudha; Sahay, N.; Butehorn, B.; Franken, P. *Piriformospora indica*, a cultivable plant-growth-promoting root endophyte. *Appl. Environ. Microbiol.* **1999**, *65*, 2741–2744. [PubMed]
11. Oelmueller, R.; Sherameti, I.; Tripathi, S.; Varma, A. *Piriformospora indica*, a cultivable root endophyte with multiple biotechnological applications. *Symbiosis* **2009**, *49*, 1–17. [CrossRef]
12. Yadav, V.; Kumar, M.; Deep, D.K.; Kumar, H.; Sharma, R.; Tripathi, T.; Tuteja, N.; Saxena, A.K.; Johri, A.K. A phosphate transporter from the root endophytic fungus *Piriformospora indica* plays a role in phosphate transport to the host plant. *J. Biol. Chem.* **2010**, *285*, 26532–26544. [CrossRef] [PubMed]
13. Bakshi, M.; Vahabi, K.; Sherameti, I.; Oelmuller, R.; Bhattacharya, S.; Baldwin, I.; Varma, A.; Yeh, K.-W.; Johri, A.K. WRKY6 restricts *Piriformospora indica*-stimulated and phosphate-induced root development in Arabidopsis. *BMC Plant Biol.* **2015**, *15*, 305. [CrossRef] [PubMed]
14. Kumar, M.; Yadav, V.; Kumar, H.; Sharma, R.; Singh, A.; Tuteja, N.; Johri, A.K. *Piriformospora indica* enhances plant growth by transferring phosphate. *Plant Signal. Behav.* **2011**, *6*, 723–725. [CrossRef] [PubMed]
15. Lahrmann, U.; Ding, Y.; Banhara, A.; Rath, M.; Hajirezaei, M.R.; Doehlemann, S.; von Wiren, N.; Parniske, M.; Zuccaro, A. Host-related metabolic cues affect colonization strategies of a root endophyte. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 13965–13970. [CrossRef] [PubMed]
16. Vahabi, K.; Sherameti, I.; Bakshi, M.; Mrozinska, A.; Ludwig, A.; Reichelt, M.; Oelmuller, R. The interaction of Arabidopsis with *Piriformospora indica* shifts from initial transient stress induced by fungus-released chemical mediators to a mutualistic interaction after physical contact of the two symbionts. *BMC Plant Biol.* **2015**, *15*, 58. [CrossRef] [PubMed]
17. Matsuo, M.; Johnson, J.M.; Sherameti, I.; Hieno, A.; Tokizawa, M.; Yamamoto, Y.Y.; Nomoto, M.; Tada, Y.; Godfrey, R.; Obokata, J.; et al. High REDOX RESPONSIVE TRANSCRIPTION FACTOR1 Levels Result in Accumulation of Reactive Oxygen Species in *Arabidopsis thaliana* Shoots and Roots. *Mol. Plant* **2015**, *8*, 1253–1273. [CrossRef] [PubMed]
18. Rafiqi, M.; Jelonek, L.; Akum, N.F.; Zhang, F.; Kogel, K.-H. Effector candidates in the secretome of *Piriformospora indica*, a ubiquitous plant-associated fungus. *Front. Plant Sci.* **2013**, *4*, 228. [CrossRef] [PubMed]
19. Vadassery, J.; Oelmuller, R. Calcium signaling in pathogenic and beneficial plant microbe interactions what can we learn from the interaction between *Piriformospora indica* and *Arabidopsis thaliana*. *Plant Signal. Behav.* **2009**, *4*, 1024–1027. [CrossRef] [PubMed]
20. Johnson, J.M.; Oelmueller, R. Agony to harmony—what decides? calcium signaling in beneficial and pathogenic plant—fungus interactions—what we can learn from the Arabidopsis/*Piriformospora indica* symbiosis. *Mol. Microb. Ecol. Rhizosphere* **2013**, *2*, 833–850.
21. Camehl, I.; Sherameti, I.; Venus, Y.; Bethke, G.; Varma, A.; Lee, J.; Oelmueller, R. Ethylene signalling and ethylene-targeted transcription factors are required to balance beneficial and nonbeneficial traits in the symbiosis between the endophytic fungus *Piriformospora indica* and *Arabidopsis thaliana*. *New Phytol.* **2010**, *185*, 1062–1073. [CrossRef] [PubMed]
22. Camehl, I.; Oelmueller, R. Do ethylene response factors-9 and -14 repress PR gene expression in the interaction between *Piriformospora indica* and Arabidopsis? *Plant Signal. Behav.* **2010**, *5*, 932–936. [CrossRef] [PubMed]
23. Vadassery, J.; Ritter, C.; Venus, Y.; Camehl, I.; Varma, A.; Shahollari, B.; Novak, O.; Strnad, M.; Ludwig-Mueller, J.; Oelmueller, R. The role of auxins and cytokinins in the mutualistic interaction between Arabidopsis and *Piriformospora indica*. *Mol. Plant Microbe Interact.* **2008**, *21*, 1371–1383. [CrossRef] [PubMed]
24. Lahrmann, U.; Strehmel, N.; Langen, G.; Frerigmann, H.; Leson, L.; Ding, Y.; Scheel, D.; Herklotz, S.; Hilbert, M.; Zuccaro, A. Mutualistic root endophytism is not associated with the reduction of saprotrophic traits and requires a noncompromised plant innate immunity. *New Phytol.* **2015**, *207*, 841–857. [CrossRef] [PubMed]
25. Vahabi, K.; Camehl, I.; Sherameti, I.; Oelmueller, R. Growth of Arabidopsis seedlings on high fungal doses of *Piriformospora indica* has little effect on plant performance, stress, and defense gene expression in spite of elevated jasmonic acid and jasmonic acid-isoleucine levels in the roots. *Plant Signal. Behav.* **2013**, *8*, e26301. [CrossRef] [PubMed]
26. Nongbri, P.L.; Johnson, J.M.; Sherameti, I.; Glawischnig, E.; Halkier, B.A.; Oelmueller, R. Indole-3-acetaldoxime-derived compounds restrict root colonization in the beneficial interaction between Arabidopsis roots and the endophyte *Piriformospora indica*. *Mol. Plant Microbe Interact.* **2012**, *25*, 1186–1197. [CrossRef] [PubMed]

27. Zuccaro, A.; Lahrmann, U.; Gueldener, U.; Langen, G.; Pfiffi, S.; Biedenkopf, D.; Wong, P.; Samans, B.; Grimm, C.; Basiewicz, M.; et al. Endophytic life strategies decoded by genome and transcriptome analyses of the mutualistic root symbiont *Piriformospora indica*. *PLoS Pathog.* **2011**, *7*, e1002290. [CrossRef] [PubMed]
28. Lahrmann, U.; Zuccaro, A. Opprimo ergo sum-evasion and suppression in the root endophytic fungus *Piriformospora indica*. *Mol. Plant Microbe Interact.* **2012**, *25*, 727–737. [CrossRef] [PubMed]
29. Böttcher, C.; Chapman, A.; Fellermeier, F.; Choudhary, M.; Scheel, D.; Glawischnig, E. The Biosynthetic Pathway of Indole-3-Carbaldehyde and Indole-3-Carboxylic Acid Derivatives in Arabidopsis. *Plant Physiol.* **2014**, *165*, 841–853. [CrossRef] [PubMed]
30. Lisec, J.; Schauer, N.; Kopka, J.; Willmitzer, L.; Fernie, A.R. Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nat. Protoc.* **2006**, *1*, 387–396. [CrossRef] [PubMed]
31. Peskan-Berghoefer, T.; Shahollari, B.; Giong, P.H.; Hehl, S.; Markert, C.; Blanke, V.; Kost, G.; Varma, A.; Oelmueller, R. Association of *Piriformospora indica* with *Arabidopsis thaliana* roots represents a novel system to study beneficial plant-microbe interactions and involves early plant protein modifications in the endoplasmic reticulum and at the plasma membrane. *Physiol. Plant.* **2004**, *122*, 465–477. [CrossRef]
32. Peskan-Berghoefer, T.; Vilches-Barro, A.; Mueller, T.M.; Glawischnig, E.; Reichelt, M.; Gershenzon, J.; Rausch, T. Sustained exposure to abscisic acid enhances the colonization potential of the mutualist fungus *Piriformospora indica* on *Arabidopsis thaliana* roots. *New Phytol.* **2015**, *208*, 873–886. [CrossRef] [PubMed]
33. Cosme, M.; Wurst, S.; Cosme, M.; Franken, P.; Cosme, M.; Lu, J.; Erb, M.; Lu, J.; Erb, M.; Stout, M.J.; et al. A fungal endophyte helps plants to tolerate root herbivory through changes in gibberellin and jasmonate signaling. *New Phytol.* **2016**. [CrossRef] [PubMed]
34. Khatabi, B.; Schafer, P. Ethylene in mutualistic symbioses. *Plant Signal. Behav.* **2012**, *7*, 1634–1638. [CrossRef] [PubMed]
35. Schaefer, P.; Pfiffi, S.; Voll, L.M.; Zajic, D.; Chandler, P.M.; Waller, F.; Scholz, U.; Pons-Kuehnemann, J.; Sonnewald, S.; Sonnewald, U.; et al. Manipulation of plant innate immunity and gibberellin as factor of compatibility in the mutualistic association of barley roots with *Piriformospora indica*. *Plant J.* **2009**, *59*, 461–474. [CrossRef] [PubMed]
36. Vanholme, R.; Demedts, B.; Morreel, K.; Ralph, J.; Boerjan, W. Lignin biosynthesis and structure. *Plant Physiol.* **2010**, *153*, 895–905. [CrossRef] [PubMed]
37. Bilski, P.; Li, M.Y.; Ehrenshaft, M.; Daub, M.E.; Chignell, C.F. Vitamin B6 (pyridoxine) and its derivatives are efficient singlet oxygen quenchers and potential fungal antioxidants. *Photochem. Photobiol.* **2000**, *71*, 129–134. [CrossRef]
38. Komarova, N.Y.; Thor, K.; Gubler, A.; Meier, S.; Dietrich, D.; Weichert, A.; Suter Grotemeyer, M.; Tegeder, M.; Rentsch, D. AtPTR1 and AtPTR5 transport dipeptides in planta. *Plant Physiol.* **2008**, *148*, 856–869. [CrossRef] [PubMed]
39. Hause, B.; Schaarschmidt, S. The role of jasmonates in mutualistic symbioses between plants and soil-born microorganisms. *Phytochemistry* **2009**, *70*, 1589–1599. [CrossRef] [PubMed]
40. Smith, C.A.; Want, E.J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* **2006**, *78*, 779–787. [CrossRef] [PubMed]
41. MetaboLights. Available online: http://www.ebi.ac.uk/metabolights/MTBLS341 (accessed on 3 July 2016).
42. Lommen, A. MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. *Anal. Chem.* **2009**, *81*, 3079–3086. [CrossRef] [PubMed]
43. Luedemann, A.; Strassburg, K.; Erban, A.; Kopka, J. TagFinder for the quantitative analysis of gas chromatography-mass spectrometry (GC-MS)-based metabolite profiling experiments. *Bioinformatics* **2008**, *24*, 732–737. [CrossRef] [PubMed]
44. Golm Metabolome Database. Available online: http://gmd.mpimp-golm.mpg.de/.
45. Ziegler, J.; Qwegwer, J.; Schubert, M.; Erickson, J.L.; Schattat, M.; Burstenbinder, K.; Grubb, C.D.; Abel, S. Simultaneous analysis of apolar phytohormones and 1-aminocyclopropan-1-carboxylic acid by high performance liquid chromatography/electrospray negative ion tandem mass spectrometry via 9-fluorenylmethoxycarbonyl chloride derivatization. *J. Chromatogr. A* **2014**, *1362*, 102–109. [CrossRef] [PubMed]
46. DAVID Bioinformatics Resources 6.7. Available online: https://david.ncifcrf.gov/.

47. Huang, D.W.; Sherman, B.T.; Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **2008**, *4*, 44–57. [CrossRef] [PubMed]
48. Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30. [CrossRef] [PubMed]
49. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef] [PubMed]
50. Ansari, M.W.; Gill, S.S.; Tuteja, N. *Piriformospora Indica* a Powerful Tool for Crop Improvement. *Proc. Indian Natl. Sci. Acad.* **2014**, *80*, 317–324. [CrossRef]

2.5 Contributions to publications

1. Thum, A.; Mönchgesang, S.; Westphal, L.; Lübken, T.; Rosahl, S.; Neumann, S.; Posch, S. Supervised Penalized Canonical Correlation Analysis. *arXiv* **2014**, *1405.1534*.

   - design: 0 %
   - experimentation: 0 %
   - data analysis: SM tested the R package and discussed the project (10 %)
   - manuscript writing: SM gave critical feedback on manuscript (10 %)

2. Hoehenwarter, W.; Mönchgesang, S.; Neumann, S.; Majovsky, P.; Abel, S.; Müller, J. Comparative expression profiling reveals a role of the root apoplast in local phosphate response. *BMC Plant Biol* **2016**, *16*, 106.

   - design: 10 %
   - experimentation: 0 %
   - data analysis: SM analyzed and integrated high throughput omics data (40 %)
   - manuscript writing: SM wrote the methods and results part including figures for spCCA and critically read the manuscript (10 %)

3. Mönchgesang, S.; Strehmel, N.; Schmidt, S.; Westphal, L.; Taruttis, F.; Müller, E.; Herklotz, S.; Neumann, S.; Scheel, D. Natural variation of root exudates in *Arabidopsis thaliana* – linking metabolomic and genomic data. *Sci Rep* **2016**, *6*.

   - design: SM extended the initial SNP model by EM with restriction to nonsense mutations and chose validation candidates (40 %)
   - experimentation: NS and SS performed initial experiments, NS and SM performed studies with T-DNA insertion lines (15 %)
   - data analysis: 80 %
   - manuscript writing: SM designed and prepared all figures and wrote the manuscript (95 %)

equal contributions

Susann Mönchgesang	Steffen Neumann	Wolfgang Hoehenwarter	Dierk Scheel
Place, Date	*Place, Date*	*Place, Date*	*Place, Date*
Signature	*Signature*	*Signature*	*Signature*

4. Mönchgesang, S.; Strehmel, N.; Trutschel, D.; Westphal, L.; Neumann, S.; Scheel, D. Plant-to-Plant Variability in Root Metabolite Profiles of 19 *Arabidopsis thaliana* Accessions Is Substance-Class-Dependent. *Int J Mol Sci* **2016**, *17*.

    - design: 40 %
    - experimentation: 0 %
    - data analysis: 70 %
    - manuscript writing: SM designed and prepared all figures with DT, SM wrote the manuscript (90 %)

5. Al Shweiki, M. R.; Mönchgesang, S.; Majovsky, P.; Thieme, D.; Trutschel, D.; Hoehenwarter, W. Assessment of Label-free Quantification in Discovery Proteomics and Impact of Technological Factors and Natural Variability of Protein Abundance. *J Proteome Res*, (revised version submitted on 21/10/2016).

    - design: 20 %
    - experimentation: 0 %
    - data analysis: SM performed all statistical analyses (40 %)
    - manuscript writing: SM contributed to most figures, wrote the methods part and critical feedback (15 %)

6. Strehmel, N.; Mönchgesang, S.; Herklotz, S.; Krüger, S.; Ziegler, J.; Scheel, D. *Piriformospora indica* Stimulates Root Metabolism of *Arabidopsis thaliana*. *Int J Mol Sci* **2016**, *17*.

    - design: 5 %
    - experimentation: 0 %
    - data analysis: 40 %
    - manuscript writing: SM prepared some figures and extended introduction, methods and results part (10 %)

equal contributions

Susann Mönchgesang	Dierk Scheel	Wolfgang Hoehenwarter
Place, Date	*Place, Date*	*Place, Date*
Signature	*Signature*	*Signature*

3 Discussion and perspectives

This thesis focused on integrative analysis methods for high throughput omics data. Thereby, emphasis was laid on customization of the analysis pipeline according to the experimental design of the individual study. Several experimental design factors were investigated in each study ranging from different omics levels to stress conditions. Four of the publications investigated root and exudate metabolomics of *A. thaliana* to shed light on belowground interactions. It was shown that integrating labor-intense omics with better annotated omics (genomics, transcriptomics) can rapidly advance research.
However, these customized data integration strategies shown in this thesis can be applicable to other experimental designs and should inspire to develop individual data analysis workflows to address further research questions with a different combinations of experimental design factors as shown in Figure 5.

3.1 SpCCA is a versatile tool to connect multiple datasets

In section 2.1, spCCA was applied in two manuscripts as a supervised method to combine two or more datasets. The major advantage of spCCA is the biological interpretation of resulting canonical variables. It is a universal approach that looks for linear combinations of up- and downregulated features involved in one biological process. The spCCA approach was initially tested on transcriptomics and metabolomics datasets measuring the responses of several *A. thaliana* genotypes to *Phytophthora infestans*. The comparison of the supervised against the non-supervised approach showed that only two out of first ten canonical variables were easily interpretable in the non-supervised pCCA. Contrarily, spCCA was not able to detect the canonical variable associated with two mutants that was revealed by non-supervised pCCA. However, in this multifactorial experimental design, spCCA was useful to interpret the resulting canonical variables with biological processes.
In a comparative transcriptomics-proteomics study, the phosphate deficiency response was dissected in the wild type Col-0, the hypersensitive mutant *pdr2* and the insensitive double-mutant *lpr1/lpr2*. Both individual datasets revealed modified metal homeostasis, cell wall remodeling and oxidative stress as a result of phosphate depletion. SpCCA identified a larger number of regulated CIII peroxidases than could be inferred from the individual datasets. In combination with wet lab experiments, the results of the integrative analysis of high throughput omics data provide a valid base for future investigations.

3.2 Linking metabolite absences with stop codons is a functional association analysis

For experimental designs considering one factor with many levels, e.g. the factor accession with 19 possible levels for the parental lines of the MAGIC collection, classic statistics like spCCA that rely on matrix operations would not succeed. Instead, customized workflows to find associations between different omics levels are required. In the manuscript in section 2.2, two factors were examined: one of them had 19 levels, namely accessions of *A. thaliana*, that were investigated on two omics levels. For selected exudate compounds, a connection between metabolic patterns and genetic variation was detected. This knowledge was extended to an unbiased approach to quickly investigate metabolite absences and potentially linked biosynthetic enzymes. A user-friendly web-application was developed to facilitate the collaboration between computational biologists and analytical chemists. This graphical user interface, as illustrated in Figure 7, conveniently allowed for matching of metabolic patterns with genomics alterations alongside structural and functional annotations.

The study showed that root exudation phenotypes are genetically determined. More specifically, metabolite absences are linked to nonsense mutations, which could be successfully validated in wet lab experiments with the corresponding knockout lines for representatives of three substance classes. Glycosylation constitutes an important modification underlying natural variation, as demonstrated for robinin in Wu-0 and a unique DHBA hexoside fingerprint in Sf-2 among the 19 accessions. This approach was also used to identify further root metabolites that show qualitative variation [50].

Our exudate dataset provides a valuable resource to further elucidate the complex interplay between a plant and its rhizosphere. To deepen the mechanistic understanding of exudation, transporters could be monitored for genetic alterations. AraMemnon, the *Arabidopsis* membrane proteine database, provides a group of drug/metabolite transporters and could be screened for further proteins with at least four transmembrane domains [51]. However, query options are rather limited and do not allow for automatic searches of the database. A major drawback is the predictive and not experimentally validated nature of the transmembrane domains. For a functional coherence, metagenomics data from soil bacteria and fungi were acquired for these 19 accessions within the SAW project "Chemical Communication in the Rhizosphere" and are ought to be correlated with the exuded metabolites.

These exudation profiles, obtained from the pooled nutrient solution of four plants, revealed high variability and therefore, we measured single plant root extracts to analyze biological variability in the subsequent metabolomics study of plant roots.

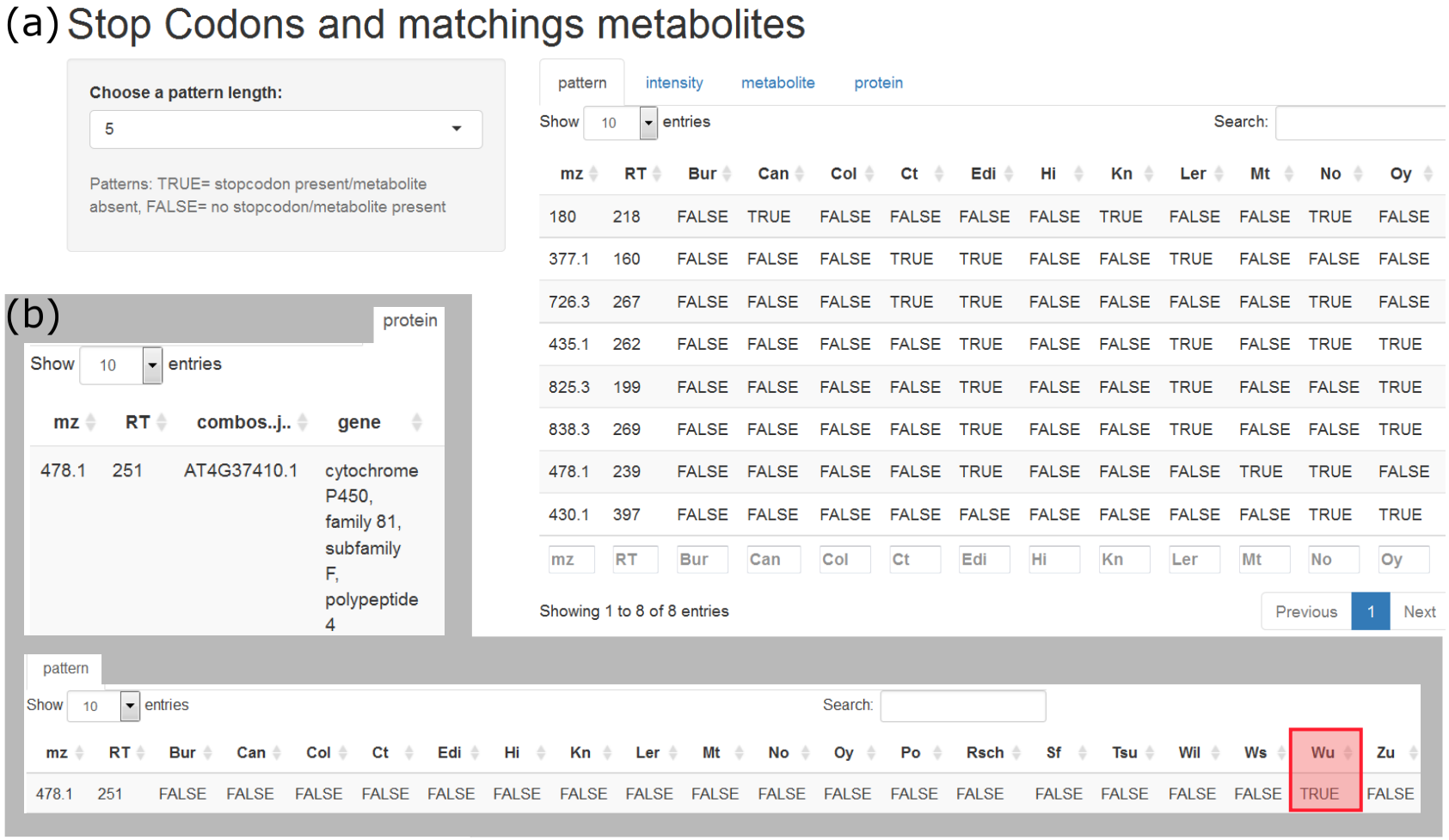


Figure 7: Graphical representation of root metabolite patterns in a Shiny Web Application. (a) In the side bar, the desired pattern length can be chosen from a dropdown menu. In the main panel, the tabs "pattern", "intensity" and "metabolite" contain primarily metabolic information, with mean intensities for each accession and metabolite annotations into isotopes and adducts [52]. The panel "protein" displays the matching enzyme-encoding genes with nonsense mutations and the top 5 enriched Gene Ontology terms among these genes if applicable. Filtering and sorting can be performed by each column. (b) The grey part illustrates the absence of neoglucobrassicin (m/z 478.1, RT 251 s, ESI(-)) in Wu-0 and AT34G37410 (CYP81F4) as one of the genes with a premature stop codon.

3.3 Plant-to-plant variability increases along the omics hierarchy

In the studies in section 2.3, multi-level experimental designs were utilized to address the central question of biological vs. technical replicates and revealed a substantial plant-to-plant variability for certain proteins and metabolites. To assess plant-to-plant variability, the intra-class correlation (ICC) according to Sampson *et al.* [53] is a useful measure. This measure is defined as the ratio of plant to total variability and thus, the proportion to which the individual plant contributes to the overall variance of a feature can be identified.
The proteomics study in section 2.3.2 evaluated the performance of commonly used proteomics softwares. The variability introduced by shotgun proteomics technology is higher than sample preparation. Moreover, the quantiles of the cumulative ICC distribution revealed more than 40 % of all proteins with no plant-to-plant-variability. In the metabolomics study in section 2.3.1, substantial plant-to-plant variability was observed for a set of glucosinolates, phenylpropanoids and flavonoids (75 % of metabolites with ICC$>$0.36).
In a similar metabolomics experiment with *Arabidopsis* leaves, the plant-to-plant variability was greater than the variance attributable to the factors instrumentation and sample preparation [54]. The publication in section 2.3.1 demonstrated that plant-to-plant variability of *Arabidopsis* roots was even greater than natural variation between accessions and non-biological variation between experimental batches. ICCs were useful to compare feature variances across platforms and different methods for variance estimation because they allow the interpreation of biological in context of the total variance. The numerical values for plant-to-plant variability were indeed not comparable between the leaf and root study: Whereas Trutschel *et al.* [54] applied a hierarchical *t*-test to estimate the contribution of each factor in the leaf metabolic profiles, a linear mixed model was applied to decompose the total observed root variances resulting in different numerical values of each contributing factor. Nevertheless, the ICC allows the comparison and the proportion of biological variability to the total variability was similar in the metabolomics studies of roots and leaves. Both cumulative ICC distribution followed a polynomial trend with a median ICC of 0.50 and 0.58, in contrast to the proteomics manuscript in section 2.3. Thus, the variability of biochemical entities increases along the omics hierarchy.

3.4 Studying *P. indica* reveals metabolic insights into a mutualistic interaction

Microorganisms have a great influence on host metabolism and its exudation. The root endophytic fungus *P. indica* has a growth-promoting effect on many hosts. In a comprehensive metabolomics study it could be shown that *P. indica* stimulates metabolism of *A. thaliana* at the site of colonization and exudation processes, but not leaf metabolism (section 2.4).

Transcriptomics data pointed towards secondary metabolism and other metabolic processes. The previously reported induced biosynthesis of glucosinolates of the indolic type was confirmed [55]. This leads to the hypothesis that even though the interaction is considered as mutualistic the plant balances fungal growth with a moderate defense response. Nitrogen-rich amino acids and flavonoids accumulated differentially in roots and leaves and proved as interesting candidates for follow-up studies. Current invesigations in our laboratory include mutants of the flavonoid biosynthesis pathway, transparent testa (tt) knockout lines *tt4* and *tt5*, to explore the compound-class specific effects. As an integrative method spCCA (section 2.1) could be applied onto the metabolomics data of three tissues of three genotypes under two growth conditions.

3.5 Implications for experimental design

This series of studies demonstrated how valuable a custom-made analysis pipeline is. My thesis aimed at the combination of reductionistic and holistic approaches.
The right choice of integrative methods obeys experimental limitations. SpCCA was shown to be versatile. Nevertheless, if the number of levels of one factor far exceeds the others, it is not the best data fusion approach. In this case, a pattern search and careful manual interpretation can elucidate connections between genotype and biochemical phenotype.
Another key contribution of this thesis is identifying the substantial variability of the omics levels downstream of nucleic acids. This is partially caused by more laborious sample preparation procedures as well as natural variability and both factors should be taken into consideration for future analyses. Pooling material from several plants is a commonly accepted way to deal with high plant-to-plant variability. Decreased analytical costs no longer require pooling of plant material, and statistical advances facilitate the correct integration of replicate types in the analysis. I would like to encourage researchers to measure comprehensive data without an irreversible loss of information, make them publicly available and allow fellow researchers to utilize these data for studies with different research questions in mind.
Both studies in section 2.3 underline the importance of *a priori* estimation of variances and advise how to handle them in experimental designs to obtain reproducible results and statistically valid conclusions.

3.6 Outlook – systematic approaches on the rise

Metabolism is highly dynamic and specific within an organism. With LC/MS and GC/MS, large number of samples can be handled to record metabolic snapshots for further use in systems biology approaches. Herein, different omics information is integrated to investigate the regulation of metabolism at the systems level. Metabolic networks may be reconstructed by utilizing

genomics data about enzymes for metabolic pathways and integrating quantitative metabolic information about the compounds therein. Software tools like VANTED and Cytoscape that are freely available to visualize complex network information [56, 57] and the creation of a systems biology markup language (SBML) as a standard format for computational modeling accelerate metabolic modeling studies [58].

The association of biosynthetic enzymes with metabolites in section 2.2 is related to a systems biology approach primarily aimed at finding patterns among known metabolites and unknown *m/z* & RT features. MS-based omics techniques deliver masses, RTs and fragment spectra. Structure elucidation and identification are important to take analyses beyond the nontargeted level and allow for biological interpretation of the measured biological entities. Due to the limited number of building blocks and connection sites, proteomics results in regular mass patterns that can be easily annotated with the available software tools. In metabolomics, there are more possible substructures and combinations thereof. To-date, no workflow can guarantee metabolite "identification". Metabolite identification is like collecting clues about the compound structure from orthogonal techniques like MS and possibly NMR. The MS spectrum points towards the accurate mass and the isotope pattern towards the elements. A sum formula from an MS spectrum needs to be validated by MS/MS spectra, which also allow for the annotation of fragments and hence, substructure annotation. The whole procedure requires expertise for spectral interpretation and are limited by the availability of a reference standard. Identification remains a bottleneck in metabolomics and could be tackled by integrating computational methods into integrative analysis workflows. Current approaches also aim at the identification of substructures to annotate either compound classes or to find sets of metabolites with common modifications. MetFamily uses a combination of MS and MS/MS measurements to find chemical substructures that are distinct for the levels of an experimental factor [59]. Since the fragmentation patterns in LC/MS metabolomics are dependent on the applied analytical conditions, an inhouse library with fragments and their corresponding substructures occurring in the commonly applied analytical set-up may be used to classify metabolites by either simple "has a substructure" assignments or overrepresentation of fragments that are characteristic for a substance class. A traditional tool in genomics like enrichment analysis, e.g. of Gene Ontology terms, has already been successfully utilized in metabolomics to annotate sets of metabolites, like implemented in Metabolite Set Enrichment Analysis and BiNChe – given that the compounds are listed in either simple metabolite set or chemical ontology databases, respectively [60, 61]. In most cases, the knowledge about functionally relevant substructures or compound classes would already be sufficient to generate hypotheses. Further genotype-phenotype studies could integrate the information about shared characteristic fragments between the MS/MS spectra of unknown metabolites, which would strengthen the hypothesis that these metabolites derive from the same biosynthetic pathway or underwent the same biochemcial transformation and that their abundance is likely to be associated with a SNP. These predictions need to be validated in experiments, possibly also with a targeted

workflow.

Future studies should address the flux of metabolites and the dynamics of metabolic transport within the plant. Several methods have been established to track metabolites in the living system like radioactive and isotopic labeling by either glucose substrates or a CO_2 environment with ^{13}C that can be analyzed by constraint-based modeling in a flux balance analysis (FBA) [62]. This technique could elucidate the biosynthesis dynamics and flow of metabolites from carbon assimilation in leaves to exudation in the rhizosphere.

This thesis followed a data-driven approach and thus, generating hypotheses for validation in biochemical studies. Future studies may combine the power of high throughput data acquired in omics experiments with prior hypotheses to initially allow for both hypothesis- and data-driven approaches. To do so, experimental designs in omics studies should be thoughtfully conceived.

References

[1] Koornneef, M.; Meinke, D. The development of Arabidopsis as a model plant. *Plant J* **2010**, *61*, 909–921.

[2] Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **2000**, *408*, 796–815.

[3] Weigel, D. Natural variation in Arabidopsis: from molecular genetics to ecological genomics. *Plant physiology* **2012**, *158*, 2–22.

[4] Huala, E.; Dickerman, A.W.; Garcia-Hernandez, M.; Weems, D.; Reiser, L.; LaFond, F.; Hanley, D.; Kiphart, D.; Zhuang, M.; Huang, W.; Mueller, L.A.; Bhattacharyya, D.; Bhaya, D.; Sobral, B.W.; Beavis, W.; Meinke, D.W.; Town, C.D.; Somerville, C.; Rhee, S.Y. The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res* **2001**, *29*, 102–105.

[5] Garcia-Hernandez, M.; Berardini, T.Z.; Chen, G.; Crist, D.; Doyle, A.; Huala, E.; Knee, E.; Lambrecht, M.; Miller, N.; Mueller, L.A.; Mundodi, S.; Reiser, L.; Rhee, S.Y.; Scholl, R.; Tacklind, J.; Weems, D.C.; Wu, Y.; Xu, I.; Yoo, D.; Yoon, J.; Zhang, P. TAIR: a resource for integrated Arabidopsis data. *Funct Integr Genomics* **2002**, *2*, 239–253.

[6] Rhee, S.Y.; Beavis, W.; Berardini, T.Z.; Chen, G.; Dixon, D.; Doyle, A.; Garcia-Hernandez, M.; Huala, E.; Lander, G.; Montoya, M.; Miller, N.; Mueller, L.A.; Mundodi, S.; Reiser, L.; Tacklind, J.; Weems, D.C.; Wu, Y.; Xu, I.; Yoo, D.; Yoon, J.; Zhang, P. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res* **2003**, *31*, 224–228.

[7] Witzel, K.; Neugart, S.; Ruppel, S.; Schreiner, M.; Wiesner, M.; Baldermann, S. Recent progress in the use of 'omics technologies in brassicaceous vegetables. *Front Plant Sci* **2015**, *6*, 244.

[8] Hatem, A.; Bozdag, D.; Toland, A.E.; Catalyurek, U.V. Benchmarking short sequence mapping tools. *BMC Bioinformatics* **2013**, *14*, 184.

[9] Baginsky, S.; Gruissem, W. Arabidopsis thaliana proteomics: from proteome to genome. *J Exp Bot* **2006**, *57*, 1485–1491.

[10] Baerenfaller, K.; Grossmann, J.; Grobei, M.A.; Hull, R.; Hirsch-Hoffmann, M.; Yalovsky, S.; Zimmermann, P.; Grossniklaus, U.; Gruissem, W.; Baginsky, S. Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. *Science* **2008**, *320*, 938–941.

[11] Salek, R.M.; Beisken, S.; Emery, L.; Grandison, R. Metabolomics: An introduction. *Train online* **2016**, *http://www.ebi.ac.uk/training/online/course/introduction–metabolomics/ [accessed on: 31/10/2016]*.

[12] Patti, G.J.; Yanes, O.; Siuzdak, G. Innovation: Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol* **2012**, *13*, 263–269.

[13] Schauer, N.; Fernie, A.R. Plant metabolomics: towards biological function and mechanism. *Trends Plant Sci* **2006**, *11*, 508–516.

[14] Worley, B.; Powers, R. Multivariate Analysis in Metabolomics. *Curr Metabolomics* **2013**, *1*, 92–107.

[15] Varmuza, K.; Filzmoser, P. *Introduction to multivariate statistical analysis in chemometrics*; CRC Press: Boca Raton, 2009; pp. xiii, 321 p.

[16] Korte, A.; Farlow, A. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* **2013**, *9*, 29.

[17] Mitteroecker, P.; Cheverud, J.M.; Pavlicev, M. Multivariate Analysis of Genotype-Phenotype Association. *Genetics* **2016**, *202*, 1345–1363.

[18] van der Sluis, S.; Posthuma, D.; Dolan, C.V. TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genet* **2013**, *9*, e1003235.

[19] Keller, M.A.; Piedrafita, G.; Ralser, M. The widespread role of non-enzymatic reactions in cellular metabolism. *Curr Opin Biotechnol* **2015**, *34*, 153–161.

[20] Beisken, S.; Eiden, M.; Salek, R.M. Getting the right answers: understanding metabolomics challenges. *Expert Rev Mol Diagn* **2015**, *15*, 97–109.

[21] Weckwerth, W. *The handbook of plant metabolomics*; Wiley-Blackwell: Weinheim, 2013; p. 424.

[22] Weckwerth, W. Metabolomics in systems biology. *Annu Rev Plant Biol* **2003**, *54*, 669–689.

[23] Fernie, A.R. The future of metabolic phytochemistry: larger numbers of metabolites, higher resolution, greater understanding. *Phytochemistry* **2007**, *68*, 2861–2880.

[24] D'Auria, J.C.; Gershenzon, J. The secondary metabolism of Arabidopsis thaliana: growing like a weed. *Current opinion in plant biology* **2005**, *8*, 308–16.

[25] Oksman-Caldentey, K.M.; Inze, D. Plant cell factories in the post-genomic era: new ways to produce designer secondary metabolites. *Trends Plant Sci* **2004**, *9*, 433–440.

[26] Dempsey, D.A.; Vlot, A.C.; Wildermuth, M.C.; Klessig, D.F. Salicylic Acid biosynthesis and metabolism. *Arabidopsis Book* **2011**, *9*, e0156.

[27] Lei, Z.; Huhman, D.V.; Sumner, L.W. Mass spectrometry strategies in metabolomics. *J Biol Chem* **2011**, *286*, 25435–25442.

[28] Kopka, J.; Schauer, N.; Krueger, S.; Birkemeyer, C.; Usadel, B.; Bergmuller, E.; Dormann, P.; Weckwerth, W.; Gibon, Y.; Stitt, M.; Willmitzer, L.; Fernie, A.R.; Steinhauser, D. GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics* **2005**, *21*, 1635–1638.

[29] Sumner, L.W.; Amberg, A.; Barrett, D.; Beale, M.H.; Beger, R.; Daykin, C.A.; Fan, T.W.; Fiehn, O.; Goodacre, R.; Griffin, J.L.; Hankemeier, T.; Hardy, N.; Harnly, J.; Higashi, R.; Kopka, J.; Lane, A.N.; Lindon, J.C.; Marriott, P.; Nicholls, A.W.; Reily, M.D.; Thaden, J.J.; Viant, M.R. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* **2007**, *3*, 211–221.

[30] Hiltner, L. Über neue Erfahrungen und Probleme auf dem Gebiete der Bodenbakteriologie. *Arbeiten der Deutschen Landwirtschaft Gesellschaft* **1904**, *98*, 59–78.

[31] Fourcroy, P.; Siso-Terraza, P.; Sudre, D.; Saviron, M.; Reyt, G.; Gaymard, F.; Abadia, A.; Abadia, J.; Alvarez-Fernandez, A.; Briat, J.F. Involvement of the ABCG37 transporter in secretion of scopoletin and derivatives by Arabidopsis roots in response to iron deficiency. *New Phytol* **2014**, *201*, 155–167.

[32] Yazaki, K. Transporters of secondary metabolites. *Curr Opin Plant Biol* **2005**, *8*, 301–307.

[33] Schmid, N.B.; Giehl, R.F.; Doll, S.; Mock, H.P.; Strehmel, N.; Scheel, D.; Kong, X.; Hider, R.C.; von Wiren, N. Feruloyl-CoA 6'-Hydroxylase1-dependent coumarins mediate iron acquisition from alkaline substrates in Arabidopsis. *Plant Physiol* **2014**, *164*, 160–172.

[34] Schmidt, H.; Gunther, C.; Weber, M.; Sporlein, C.; Loscher, S.; Böttcher, C.; Schobert, R.; Clemens, S. Metabolome analysis of Arabidopsis thaliana roots identifies a key metabolic pathway for iron acquisition. *PLoS One* **2014**, *9*, e102444.

[35] Ziegler, J.; Schmidt, S.; Chutia, R.; Müller, J.; Böttcher, C.; Strehmel, N.; Scheel, D.; Abel, S. Non-targeted profiling of semi-polar metabolites in Arabidopsis root exudates uncovers a role for coumarin secretion and lignification during the local response to phosphate limitation. *J Exp Bot* **2016**, *67*, 1421–1432.

[36] Oburger, E.; Kirk, G.D.J.; Wenzel, W.W.; Puschenreiter, M.; Jones, D.L. Interactive effects of organic acids in the rhizosphere. *Soil. Biol. Biochem.* **2009**, *41*, 449–457.

[37] Jones, P.; Messner, B.; Nakajima, J.; Schaffner, A.R.; Saito, K. UGT73C6 and UGT78D1, glycosyltransferases involved in flavonol glycoside biosynthesis in Arabidopsis thaliana. *J Biol Chem* **2003**, *278*, 43910–43918.

[38] Agrawal, B.; Lakshmanan, V.; Kaushik, S.; Bais, H.P. Natural variation among Arabidopsis accessions reveals malic acid as a key mediator of Nickel (Ni) tolerance. *Planta* **2012**, *236*, 477–489.

[39] Fomina, M.; Hillier, S.; Charnock, J.M.; Melville, K.; Alexander, I.J.; Gadd, G.M. Role of Oxalic Acid Over excretion in Transformations of Toxic Metal Minerals by Beauveria caledonica. *Appl Environ Microbiol* **2005**, *71*, 371–381.

[40] Strehmel, N.; Böttcher, C.; Schmidt, S.; Scheel, D. Profiling of secondary metabolites in root exudates of Arabidopsis thaliana. *Phytochemistry* **2014**, *108C*, 35–46.

[41] Komarova, N.Y.; Thor, K.; Gubler, A.; Meier, S.; Dietrich, D.; Weichert, A.; Suter Grotemeyer, M.; Tegeder, M.; Rentsch, D. AtPTR1 and AtPTR5 transport dipeptides in planta. *Plant Physiol* **2008**, *148*, 856–869.

[42] Oburger, E.; Dell'mour, M.; Hann, S.; Wieshammer, G.; Puschenreiter, M.; Wenzel, W.W. Evaluation of a novel tool for sampling root exudates from soil-grown plants compared to conventional techniques. *Environmental and Experimental Botany* **2013**, *87*, 235–247.

[43] Mathieu, L.; Lobet, G.; Tocquin, P.; Perilleux, C. "Rhizoponics": a novel hydroponic rhizotron for root system analyses on mature Arabidopsis thaliana plants. *Plant Methods* **2015**, *11*, 3.

[44] van Dam, N.M.; Bouwmeester, H.J. Metabolomics in the Rhizosphere: Tapping into Belowground Chemical Communication. *Trends Plant Sci* **2016**, *21*, 256–265.

[45] Kuijken, R.C.P.; Snel, J.F.H.; Heddes, M.M.; Bouwmeester, H.J.; Marcelis, L.F.M. The importance of a sterile rhizosphere when phenotyping for root exudation. *Plant and Soil* **2015**, *387*, 131–142.

[46] Ramakrishna, A.; Ravishankar, G.A. Influence of abiotic stress signals on secondary metabolites in plants. *Plant Signal Behav* **2011**, *6*, 1720–1731.

[47] Arbona, V.; Manzi, M.; de Ollas, C.; Gómez-Cadenas, A. Metabolomics as a Tool to Investigate Abiotic Stress Tolerance in Plants. *Int J Mol Sci* **2013**, *14*, 4885–4911.

[48] Jorge, T.F.; Rodrigues, J.A.; Caldana, C.; Schmidt, R.; van Dongen, J.T.; Thomas-Oates, J.; Antonio, C. Mass spectrometry-based plant metabolomics: Metabolite responses to abiotic stress. *Mass Spectrom Rev* **2016**, *35*, 620–649.

[49] Tenenboim, H.; Brotman, Y. Omic Relief for the Biotically Stressed: Metabolomics of Plant Biotic Interactions. *Trends Plant Sci* **2016**, *21*, 781–791.

[50] Mönchgesang, S.; Strehmel, N.; Trutschel, D.; Westphal, L.; Neumann, S.; Scheel, D. Plant-to-Plant Variability in Root Metabolite Profiles of 19 Arabidopsis thaliana Accessions Is Substance-Class-Dependent. *Int J Mol Sci* **2016**, *17*.

[51] Schwacke, R.; Schneider, A.; van der Graaff, E.; Fischer, K.; Catoni, E.; Desimone, M.; Frommer, W.B.; Flugge, U.I.; Kunze, R. ARAMEMNON, a novel database for Arabidopsis integral membrane proteins. *Plant Physiol* **2003**, *131*, 16–26.

[52] Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T.R.; Neumann, S. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal Chem* **2012**, *84*, 283–289.

[53] Sampson, J.N.; Boca, S.M.; Shu, X.O.; Stolzenberg-Solomon, R.Z.; Matthews, C.E.; Hsing, A.W.; Tan, Y.T.; Ji, B.T.; Chow, W.H.; Cai, Q.; Liu da, K.; Yang, G.; Xiang, Y.B.; Zheng, W.; Sinha, R.; Cross, A.J.; Moore, S.C. Metabolomics in epidemiology: sources of variability in metabolite measurements and implications. *Cancer Epidemiol Biomarkers Prev* **2013**, *22*, 631–640.

[54] Trutschel, D.; Schmidt, S.; Grosse, I.; Neumann, S. Experiment design beyond gut feeling: statistical tests and power to detect differential metabolites in mass spectrometry data. *Metabolomics* **2015**, *11*, 851–860.

[55] Lahrmann, U.; Strehmel, N.; Langen, G.; Frerigmann, H.; Leson, L.; Ding, Y.; Scheel, D.; Herklotz, S.; Hilbert, M.; Zuccaro, A. Mutualistic root endophytism is not associated with the reduction of saprotrophic traits and requires a noncompromised plant innate immunity. *New Phytol* **2015**, *207*, 841–857.

[56] Junker, B.H.; Klukas, C.; Schreiber, F. VANTED: a system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics* **2006**, *7*, 109.

[57] Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **2003**, *13*, 2498–2504.

[58] Hucka, M.; Finney, A.; Sauro, H.M.; Bolouri, H.; Doyle, J.C.; Kitano, H.; Arkin, A.P.; Bornstein, B.J.; Bray, D.; Cornish-Bowden, A.; Cuellar, A.A.; Dronov, S.; Gilles, E.D.; Ginkel, M.; Gor, V.; Goryanin II.; Hedley, W.J.; Hodgman, T.C.; Hofmeyr, J.H.; Hunter, P.J.; Juty, N.S.; Kasberger, J.L.; Kremling, A.; Kummer, U.; Le Novere, N.; Loew, L.M.; Lucio, D.; Mendes, P.; Minch, E.; Mjolsness, E.D.; Nakayama, Y.; Nelson, M.R.; Nielsen, P.F.; Sakurada, T.; Schaff, J.C.; Shapiro, B.E.; Shimizu, T.S.; Spence, H.D.; Stelling, J.; Takahashi, K.; Tomita, M.; Wagner, J.; Wang, J.; Forum, S. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **2003**, *19*, 524–531.

[59] Treutler, H.; Tsugawa, H.; Porzel, A.; Gorzolka, K.; Tissier, A.; Neumann, S.; Balcke, G.U. Discovering Regulated Metabolite Families in Untargeted Metabolomics Studies. *Anal Chem* **2016**, *88*, 8082–8090.

[60] Xia, J.; Wishart, D.S. MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res* **2010**, *38*, W71–7.

[61] Moreno, P.; Beisken, S.; Harsha, B.; Muthukrishnan, V.; Tudose, I.; Dekker, A.; Dornfeldt, S.; Taruttis, F.; Grosse, I.; Hastings, J.; Neumann, S.; Steinbeck, C. BiNChE: a web tool and library for chemical enrichment analysis based on the ChEBI ontology. *BMC Bioinformatics* **2015**, *16*, 56.

[62] Chokkathukalam, A.; Kim, D.H.; Barrett, M.P.; Breitling, R.; Creek, D.J. Stable isotope-labeling studies in metabolomics: new insights into structure and dynamics of metabolic networks. *Bioanalysis* **2014**, *6*, 511–524.

Appendix

List of acronyms and abbreviations

ABC ATP-binding cassette

CCA canonical correlation analysis

EI electron ionization

ESI electrospray ionization

FBA flux balance analysis

GC gas chromatography

GSL glucosinolate

GWAS genome-wide association analysis

HCA hierarchical clustering analysis

INDEL insertions and deletion

LC liquid chromatgraphy

MAGIC Multiparent Advanced Generation Inter-Cross

MATE multidrug and toxic compound extrusion

MDS multi-dimensional scaling

MS mass spectrometry

m/z mass-to-charge ratio

PCA principal component analysis

PLS partial least squares

PTR peptide transporter

QTL quantitative trait loci

RT retention time

SNP single nucleotide polymorphism

spCCA supervised penalized canonical correlation analysis

TOF time-of-flight

List of publications

- Al Shweiki, M. R.; Mönchgesang, S.; Majovsky, P.; Thieme, D.; Trutschel, D.; Hoehenwarter, W. Assessment of Label-free Quantification in Discovery Proteomics and Impact of Technological Factors and Natural Variability of Protein Abundance. *J Proteome Res* (revised version submitted on 21/10/2016).

- Hoehenwarter, W.; Mönchgesang, S.; Neumann, S.; Majovsky, P.; Abel, S.; Müller, J. Comparative expression profiling reveals a role of the root apoplast in local phosphate response. *BMC Plant Biol* **2016**, *16*, 106.

- Mönchgesang, S.; Ruttkies, C.; Treutler, H.; Heisters, M. Meeting Report: Plant Science Student Conference (PSSC) 2015 – Young researchers in green biotechnology. *Biotechnol J* **2015**, *10*, 1666-1667.

- Mönchgesang, S.; Strehmel, N.; Schmidt, S.; Westphal, L.; Taruttis, F.; Müller, E.; Herklotz, S.; Neumann, S.; Scheel, D. Natural variation of root exudates in *Arabidopsis thaliana* – linking metabolomic and genomic data. *Sci Rep* **2016**, *6*.

- Mönchgesang, S.; Strehmel, N.; Trutschel, D.; Westphal, L.; Neumann, S.; Scheel, D., Plant-to-Plant Variability in Root Metabolite Profiles of 19 *Arabidopsis thaliana* Accessions Is Substance-Class-Dependent.*Int J Mol Sci* **2016**, *17*.

- Rubbiani, R.; Can, S.; Kitanovic, I.; Alborzinia, H.; Stefanopoulou, M.; Kokoschka, M.; Mönchgesang, S.; Sheldrick, W.S.; Wölfl, S.; Ott, I. Comparative in vitro evaluation of N-heterocyclic carbene gold(I) complexes of the benzimidazolylidene type. *J Med Chem* **2011**, *54*, 8646-8657.

- Sinha, D.; Chong, L.; George, J.; Schlüter, H.; Mönchgesang, S.; Mills, S.; Li, J.; Parish, C.; Bowtell, D.; Kaur, P.; Australian Ovarian Cancer Study Group. Pericytes Promote Malignant Ovarian Cancer Progression in Mice and Predict Poor Prognosis in Serous Ovarian Cancer Patients. *Clin Cancer Res* **2016**, *22*, 1813-1824.

- Strehmel, N.; Mönchgesang, S.; Herklotz, S.; Krüger, S.; Ziegler, J.; Scheel, D., *Piriformospora indica* Stimulates Root Metabolism of *Arabidopsis thaliana*. *Int J Mol Sci* **2016**, *17*.

- Thum, A.; Mönchgesang, S.; Westphal, L.; Lübken, T.; Rosahl, S.; Neumann, S.; Posch, S. Supervised Penalized Canonical Correlation Analysis. *arXiv* **2014**, *1405.1534*.

equal contributions

Curriculum vitae

Susann Mönchgesang

Curriculum Vitae

Personal Details

Birth 7 June 1988, Erfurt
Nationality German

Education

Oct 2013 - Sep 2016 **Biochemistry (PhD),** *University of Halle-Wittenberg*
Focus: Metabolomics

Sep 2011 - Jul 2013 **Molecular Biotechnology (MSc),** *University of Heidelberg*
Focus: Drug research, GPA: 1.0

Feb 2012 - Jul 2012 **Biotechnology,** *University of Melbourne, Australia*
Exchange semester in Professional Master

Sep 2008 - Jul 2011 **Molecular Biotechnology (BSc),** *University of Heidelberg*
Focus: Drug research, GPA: 1.6

Practical Experience

Oct 2013 - Sep 2016 **PhD student,** *Leibniz Institute of Plant Biochemistry Halle (S.),*
Thesis: Metabolomics and biochemical omics data – integrative approaches

Nov 2012 - Jul 2013 **Master student,** *German Cancer Research Center Heidelberg*
Thesis: Potential epigenetic modifications and their implications on novel human melanoma therapy

Jul 2012 - Sep 2012 **Research intern,** *MPI for Medical Research Heidelberg*
Protein binding and reaction kinetics of a chaperone

Apr 2012 - Jun 2012 **Guest scientist,** *Peter MacCallum Cancer Centre Melbourne*
Survival analysis of ovarian cancer patients

Feb 2012 - Apr 2012 **Visiting student,** *Bio21, University of Melbourne*
Protein modifications in Alzheimer's and Parkinson's Disease

Nov 2011 - Feb 2012 **Industrial intern,** *Roche Diagnostics Mannheim*
Analytics and preparation of pump insulins

Mar 2011 - Jun 2011 **Bachelor student,** *IPMB, University of Heidelberg*
Thesis: Effects of benzimidazol-2-ylidene gold(I) complexes on mitochondrial activity

Aug 2010 - Sep 2010 **DAAD-RISE student,** *University of Toledo (OH), USA*
Spatiotemporal imaging of salivary glands

Feb 2010 - Mar 2010 **Industrial intern,** *Jena Bioscience and JenaGen*
Development of a new sequencing method

Fritz-Reuter-Straße 8 – 06114 Halle (Saale)
+49 (0) 345 5582 1475
susann.moenchgesang@ipb-halle.de

Acknowledgement

With this dissertation an eventful journey through graduate school has reached its destination. There were many people along the way that inspired and encouraged me:

I wish to thank Dierk Scheel and Steffen Neumann for the opportunity of such an interdisciplinary project; Dierk for believing in me and the stimulating environment in SEB and within the SAW project; Steffen for getting me acquainted with many collaborators and interdisciplinary communication.

I appreciate the time and effort of the referees for examining my thesis.

I am deeply grateful to Nadine Strehmel for getting me started in the metabolomics field, our fruitful projects and the friendship- your "not a big effort, just set it up and measure/quantify/heatmap it" way helped me to find my way (not only to do research) – and I am glad that sometimes my enthusiasm would pay back to you.

I am thankful to my group Christoph Ruttkies, Michael Gerlich, Diana Trutschel, Hendrik Treutler, Kristian Peters, Sarah Scharfenberg and Daniel Schober for coffee, cake, ice-cream and discussions. I would like to sincerely say thank you to Lore Westphal for an open ear, proofreading and genetics expertise. The current metabolomics group with Karin Gorzolka and Sophie Dietz has adopted me and shared their knowledge (and other things) with me. Furthermore, I wish to thank the former members Stephan Schmidt and Christoph Böttcher. Without the help of Sylvia Krüger, Julia Göhricke, Susanne Kirsten and Lennard Eschen-Lippold, my plants would not have been raised that professionally, their metabolic extracts would not have been measured that accurately and those polymerases would not have amplified that precisely. My collaborations turned out to be very flourishing- thanks to Wolfgang Hoehenwarter and Jens Müller.

Ich möchte meiner Familie für ihre Unterstützung danken und für ihre Versuche, meine Forschung zu verstehen, auch wenn es böhmische Dörfer waren. Thank you, Michi, for lots of science and girl talk alongside sparkling wine and, Fanny & Chrissie, for cheering me up and PhD competition. Finally, I wish to acknowledge Micha – thank you for everything from A(ustralia) to Z(illertal).

Declaration in lieu of oath

I herewith declare that

1. I wrote this thesis independently and without external help;
2. I used no other sources and supporting materials than those indicated;
3. the adoption of quotations from the literature and the internet as well as thoughts from other authors were indicated as such in the thesis;
4. this thesis or parts of it have not been submitted in any form to another faculty as part of an examination;
5. I have not applied for a doctoral degree previously.

Susann Mönchgesang

Place, Date

Signature