Characterization of poplar metabotypes via mass difference enrichment analysis

Franco Moritz [1#], Moritz Kaling [1, 2#]

Jörg-Peter Schnitzler [2]*, Philippe Schmitt-Kopplin [1, 3]*

[1] Research Unit Analytical BioGeoChemistry, Helmholtz Zentrum München (HMGU), D-85764, Neuherberg, Germany

[2] Research Unit Environmental Simulation, Institute of Biochemical Plant Pathology, Helmholtz Zentrum München (HMGU), D-85764, Neuherberg, Germany

[3] Chair of Analytical Food Chemistry, Technische Universität München (TUM), D-85354, Freising, Germany

# These authors contributed equally to the present work and share first authorship

*correspondence to jp.schnitzler@helmholtz-muenchen.de and schmitt-kopplin@helmholtz-muenchen.de

Abstract

Instrumentation technology for metabolomics has advanced drastically in recent years in terms of sensitivity and specificity. Despite these technical advances, data analytical strategies are still in their infancy in comparison with other 'omics'. Plants are known to possess an immense diversity of secondary metabolites. Typically, more than 70% of metabolomics data are not amenable to systems biological interpretation due to poor database coverage. Here, we propose a new general strategy for mass spectrometry-based metabolomics that incorporates all exact mass features with known sum formulae into the evaluation and interpretation of metabolomics studies. We extend the use of mass differences, commonly used for feature annotation, by re-defining them as variables that reflect the remaining 'omic' domains. The strategy uses exact mass difference network analyses exemplified for the metabolomic description of two gray poplar (*Populus* x *canescens*) genotypes that differ in their capability to emit isoprene. This strategy established a direct connection between the metabotype and the non-isoprene emitting phenotype, as mass differences pertaining to prenylation reactions were over-represented in non-isoprene emitting poplars. The analysis of mass differences was not only able to grasp the known chemical biology of poplar but it also improved the interpretability of yet unknown biochemical relationships.

A major part of mass spectrometric data is not amenable to data interpretation as metabolite databases are far from being complete. This work presents the concept and rules on how Mass Difference Enrichment Analysis (MDEA) enables data driven analysis and interpretation of metabolomics data. This new metabolomics approach is presented vis-à-vis the biochemically well-characterized gray poplar isoprene emitting and non-emitting mutants, and yields results that are in perfect accordance with prior metabolite and physiological knowledge. MDEA is shown to extend prior knowledge supporting the formulation of new, testable biochemical working hypotheses.

Keywords: Populus x canescens – Systems chemical biology – metabolomics – networks – mass difference analysis

List of Abbreviations

Abbreviation	Meaning
Da	Dalton
DCA	Dicarboxylic acid
DHAP	Dihydroxyacetone Phosphate
DMAPP	Dimethylallyl Pyrophosphate
DXS	1-deoxy-D-xylulose 5-phosphate Synthase
FA	Fatty acid
FPP	Farnesyl Pyrophosphate
FT-ICR-MS	Fourier Transform Ion Cyclotron Resonance Mass Spectrometry
GPP	Geranyl Pyrophosphate
HMDB	Human Metabolome Database
HTG	Hemiterpene Glucosides
IE	Isoprene Emitting
KEGG	Kyoto Encyclopedia of Genes and Genomes
MDA	Malondialdehyde
MDEA	Mass Difference Enrichment Analysis
MDiA	Mass Difference Analysis
MDiN	Mass Difference Network
MEP	Methylerythritol Phosphate

MS	Mass Spectrometry/Mass Spectrometer
MS/MS	Tandem Mass Spectrometry
NE	Non-Isoprene Emitting
OPLS-DA	Orthogonal Partial Least Squares Discriminant Analysis
PcISPS	Isoprene Synthase
PEP	Phosphoenolpyruvate
POP	Poplar Metabolome
PP	Pyrophosphate
MDB	Mass-difference based building blocks
rpairs	KEGG reaction pairs
SIM	Selected Ion Monitoring
SIMMS/MS	Selected Ion Monitoring Tandem Mass Spectrometry
TCA	Tricarboxylic acid
WT	Wild Type

Introduction

From genes via transcripts to proteins and finally to metabolites, is the "classical" view of the cellular information cascade. These days this flow of information is interpreted as a cellular network in which the different layers (genes, transcripts, proteins and metabolites) interact with, and influence, each other. Nonetheless, the nature and direction of these interactions are under constant debate. The scientific discipline, which aims to understand these cellular bionetworks globally is called systems biology (Ideker *et al.* 2001). The integration of genomic, transcriptomic, proteomic, and metabolomic data is a major challenge as all of these domains have their own time scale and are measured using different analytical techniques.

Well-known methods for metabolic pathway/network analysis are (i) constraint-based modelling such as flux balance analysis (e.g. Boyle & Morgan 2009), (ii) stable isotope feeding (e.g. Masakapalli *et al.* 2010) or (iii) the reconstruction of the differential biochemical Jacobian from a predefined fluctuation matrix and the covariance matrix of metabolomics measurements (Steuer *et al.* 2003, Nägele *et al.* 2014).

Recently, a comprehensive study showed how more than 80% of 2435 *Arabidopsis thaliana* metabolic features were altered due to cytoplasmic genome variation (Joseph *et al.* 2013). Notably, 91.2% of these features were unknown, which means that the known partition that would have been available for prior-knowledge driven metabolic network analysis amounted to 8.8% of all features. If the proportion of unknown molecular features is reported at all, their proportion is consistently documented to vary between 70% and 90% (e.g. Walker *et al.* 2014a, Witting *et al.* 2015).

As simultaneous identification and quantification of the entire metabolome is not possible using state-of-the-art instrumentation, there is also no data-analytical method available that incorporates these 90% of unknowns into phenotypic descriptions and interpretations. There have been approaches that use Gaussian graphical models towards a stepwise incorporation of yet unidentified features into bioinformatic evaluations (Krumsiek *et al.* 2011).

In recent years, mass spectral analysis has been extended using mass-difference networks (MDiNs) whose reconstruction requires accurate m/z-features or molecular masses as nodes, which can be connected by mass-difference-based building blocks (MDBs) as edges.

The concept was first introduced by Breitling *et al.* (2006a, 2006b) where they proposed two scientific applications of the concept. Firstly, MDiNs enable *ab initio* pathway detection, therefore, reasonably, metabolic pathways must be subgraphs of MDiNs, given the detection of the corresponding metabolites. Secondly, the common biochemical ancestry of connected metabolites in conjunction to metabolic difference analysis can support feature identification if one of the metabolites is known.

This database-driven exploitation of MDiNs was refined by a series of papers (Gipson *et al.* 2008, Rogers *et al.* 2009, Weber & Viant 2010, Watrous *et al.* 2012, Morreel *et al.* 2014). However, ultimately all approaches leave features that can be neither identified by MS/MS nor they be mapped into metabolite databases apart. Doing so is perfectly appropriate when high confidence molecular formula assignments to unknown/unidentified features are not possible. Despite the fact that ultra-high resolution (UHR) mass spectrometers have become commonplace, the majority of assigned molecular formulas do not match metabolite databases, while their detected m/z peaks in direct infusion mode are amenable to neither immediate MS/MS identification nor isotopic pattern matching. Moreover, non-targeted LC-

MS methods have too low sensitivities and require duty cycles that are too fast for true LC-UHR-MS.

While the methods described above have a tremendous potential to consolidate and elaborate on presently known metabolomes, and to extend this knowledge (if m/z-peak abundances are sufficiently large), they are bound to miss the impact of enzymatic promiscuity (Pichersky & Lewinsohn 2011) and as of yet unassigned enzyme specificities. Particularly in plant secondary metabolism, enzymes can either utilize multiple substrates or produce a variety of products from one substrate, e.g. terpene synthases (Kampranis *et al.* 2007) and O-methyltransferases (Schwab 2003). Metabolites coexist in very concentrated solutions and it has even been hypothesized they form deep eutectic fluids in a cell (Choi *et al.* 2011). Moreover, chemical activity is a function of compound concentration, meaning spontaneous reactions between metabolites are likely to occur. Global descriptions of molecular (metabolite) pools are therefore unlikely to be reflected by measurements of free metabolites alone.

Focusing on the assignment of molecular formulas, rather than on metabolite identification, Tziotis *et al.* (2011) generalized molecular formula propagation through MDiNs as a means of database-independent molecular formula assignment for UHR-MS features. The given method was applied on a multitude of analytical matrices (Müller *et al.* 2013, Walker *et al.* 2014b, Zhang *et al.* 2014, Forcisi *et al.* 2015, Moritz *et al.* 2015, Witting *et al.* 2015) and extensions to low-resolution mass spectrometry have been described (Forcisi *et al.* 2015).

Previously analyzed *Populus x canescens* (gray poplar) lines knocked-down in isoprene synthase (PcISPS; EC 4.2.3.27) (Behnke *et al.* 2007) revealed large phenotypic changes. The resulting lack of isoprene (2-methyl-1,3-butadiene) emission in poplar results in large metabolic (Behnke *et al.* 2010a, Way *et al.* 2013, Velikova *et al.* 2015), transcriptomic

(Behnke *et al.* 2010b), proteomic (Velikova *et al.* 2014), and physiological modifications (Behnke *et al.* 2007, Behnke *et al.* 2012). However, as interpretations of KEGG-based m/z feature annotations related to merely 3% of all molecular formulas, we aimed to explore the remaining 97% of the dataset. In the spirit of gene set enrichment analysis (Subramanian *et al.* 2005), we used mass difference enrichment analysis (MDEA) as a tool for mass difference analysis (MDiA), which mines MDBs that are associated with statistically important m/z features with molecular formulas of both known and unknown metabolites (Zhang *et al.* 2014, Moritz *et al.* 2015).

We used a Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FT-ICR-MS) dataset on *Populus x canescens* isoprene-emitting (IE) and non-isoprene-emitting (NE) poplar lines published recently (Kaling *et al.* 2015). We show that – regardless of the proportion of knowledge on single metabolic features – MDiA (i.e. MDEA) mines MDBs that are in agreement with prior knowledge, and we confirm them to be the major building blocks of the given poplar metabolome using SIM-stitch MS/MS experiments. We show that the approaches of Breitling *et al.* (2006a, 2006b) and Weber and Viant (2010) can easily be integrated, and how an extension to Walker *et al.* (2014b) enables the formulation of hypotheses that can serve as a basis for hypothesis-driven research in future endeavors.

Material and Methods

Plant material

The dataset of a UV experiment described in Kaling *et al.* (2015) was used. In brief: the four genotypes consisted of (i) two isoprene non-emitting (NE) RNAi lines Ra2 and Rb7 and (ii) two isoprene emitting (IE) lines wild type (WT) and empty vector control (EV) gray poplar. Each group (NE and IE) had a total of 16 plants, where half of the pants were exposed to high

UV radiation for 13 days. Sample preparation was performed as described in Kaling *et al.* (2015).

FT-ICR-MS measurements

FT-ICR MS measurements were performed as described in Kaling *et al.* (2015).

Data annotation

The 211 FT-ICR-MS spectra were internally calibrated, aligned, exported as ascii files, and combined with an error of 1 ppm using an in-house written program (Lucio *et al.* 2011). Peaks, which were present in just one spectrum were removed from the matrix. M/z features whose mass defect cannot be realized on the basis of any valid combination of C, H, N, O, P, and S, given their m/z, were omitted. Isotopologues of one molecular species must correlate across samples if the samples' origin is comparable. De-isotoping was performed in two steps. Firstly, searching for commonly known mass differences (e.g. ^{13}C-^{12}C = 1.003355 Da) within an error window of 3 ppm, and secondly, omission of the heavy isotope candidate if its ion abundance correlates with the monoisotopic candidate peak ($r \geq 0.95$).

Molecular formula assignment was performed on the basis of Tziotis *et al.* (2011) with an additionally applied Senior filter (Senior 1951), and a box for search direction randomization. Both edges and annotations were allowed to be rejected by the overall MDiN context so as to maximize the overall consensus of all formulas and MDBs. Molecular formula assignment was performed using the manually curated MDB list described herein. The overall error tolerance was set to 5 ppm and the error tolerance for MDB assignment was set to 0.2 ppm. The error over m/z distribution was centered on 0.04 ppm with a standard deviation of 0.1 ppm at m/z = 200 and 0.25 ppm at m/z ≥ 300 (Fig. S1).

Each mass spectrum contained 2,276 ± 886 non-randomly occurring m/z peaks. On average, 1,579 ± 487 successful monoisotopic molecular formula assignments were obtained per spectrum, which amounted to 4335 annotated molecular formulas over all 211 mass spectra (Table S1). The central composition of molecular formulas was CHO (1,499), N (2,205), S (1,938) and P (892). Overall, low mass defects indicated a high proportion of desaturation and oxygenation (not shown). A comparison of all 4,335 empirical molecular formulas to KEGG yielded 129 matches. Database queries were performed at hand of molecular formulas as error-bound m/z-queries amount to substantial proportions of false hits. As exemplified in Fig. S1, the empirical error distribution was i) not continually centered on 0 ppm as it is not linear, and (ii) shows systematic oscillations of error values. At mass spectral regions that are systematically not centered on 0 ppm, false database matching occurs by default if minimal errors are considered as a criterion of annotation goodness. Corresponding database query strategies are especially prone to false positive annotations if the error distribution is not centered, and prone to false negative annotations if the correct formula is not listed in the database.

Generation of MDB lists

The Breitling *et al.* (2006a, 2006b) approach: Mass-difference-based building blocks were defined to cover i) the functional list as described in Tziotis *et al.* (2011), further exchanges of small functionalities, (ii) amino acids, (iii) their corresponding keto-acids, (iv) even-chained fatty acids (C2-C16), (v) dicarboxylic acids, (vi) phenylpropanoids, (vii) cofactors and nucleotides/nucleosides, as well as (viii) derivatives of pentoses, hexoses, and their disaccharides. Furthermore, multiple prenylation steps were added. Co-enzymes were implicitly considered in the formulation of the reaction classes that they catalyze. The transformation of the above compound classes into mass-difference-based building blocks

(MDBs) was performed by modifications according to the following reaction classes: condensation/hydrolysis (Type A, CoA; M-H_2O), decarboxylative condensation (Type B, CoA, Pyridoxal-Phosphate; M-H_2O-CO_2), and hydrogenation of carbonyls with consecutive condensation (Type C, CoA, NADH, Pyridoxal-P; M+H_2-H_2O). The full list of MDBs and reaction types is given in Table S2.

The Weber and Viant (2010) approach: Mining of KEGG rpairs from KEGGAPI. MDBs that span two rpairs because 13 shortest paths of a length of k = 2 were obtained between D-Glucose and β-D-Fructose-6-phosphate after removal of metabolite entries such as 'electron', 'proton' or H_2O. Such multiplicity is problematic if the aim is to describe an MDB class that is intended to serve as a class of variables. The full list of MDBs and reaction types is given in Table S2.

Statistical analysis

The molecular formula / intensity matrix of annotated features was used for principal component analysis (PCA) and orthogonal partial least squares discriminant analysis (OPLS-DA) using SIMCA-P (v13.0.0.0, Umetrics, Umeå, Sweden). Isoprene emission was utilized as the Y-variable and the molecular formulas and their respective intensities were defined as X-variables. Based on the principal component 1 loadings of the OPLS-DA model (R2Y(cum) = 0.957, Q2(cum) = 0.767, Figure 2 - Figure supplement 1), 10% of the most important molecular formulas for the characterization of NE and IE were extracted and used for mass-difference enrichment analysis (Fig. S4).

Mass-difference enrichment analysis

MDEA was performed by separately using the two MDB lists described herein. MDiNs were reconstructed using the theoretical neutral masses of the 4335 molecular formulas annotated above. Assuming them to be error free, MDiNs were reconstructed at a networking error of 0.01 ppm to accommodate for mass deviations that derive from rounding. Given CHNOPS, there are usually no isobaric annotations at errors < 0.05 ppm. However, exceptions do exist. The MDiN, which was built using the curated MDB list contained 63,608 edges, 15,537 and 18,143 of which were incident to IE nodes and NE nodes, respectively. The KEGG rpair MDiN comprised 65,180 edges, 16,072 and 19,102 of which were incident to IE nodes and NE nodes, respectively. The entire results are provided as '.xlsx' files in Table S2. An inference pertaining to whether an MDB is over/under-represented among nodes of interest requires testing as to whether it was observed more or less frequently than should be expected by chance given the entire model dataset. The Fisher exact test, which assumes a hypergeometric distribution, was used as follows: (1) reconstruct an MDiN from a population of nodes, *P*, and a set of edges, *E*, (2) define a sample set *S*, that contains nodes of interest, (3) count the set of edges, E_S, that are connected to *S*, (4) count the number of edges of one MDB type *R* in *E*; and (5) count the corresponding number of edges, R_S, that are connected to *S*. The probability for R_S edges to be connected to *S*, given *E*, *R* and E_S is calculated as follows (using the MATLAB function *hygecdf*):

$$p = F(R_S|E, R, E_S) = \sum_{i=0}^{R_S} \frac{\binom{R}{i}\binom{E-R}{E_S-i}}{\binom{E}{E_S}} \qquad \text{Eq. 1}$$

Using the MATLAB function *hygestat*, it was possible to calculate the values μ and σ, which are the expected amount of R_S and the expected standard deviation given *E*, *R*, $E_{S:}$

$$\mu = \frac{E_S \cdot R}{E} \qquad \text{Eq. 2}$$

and

$$\sigma = \sqrt{\frac{E_S \cdot R \cdot (E-R) \cdot (E-E_S)}{E^2 \cdot (E-1)}} \qquad \text{Eq. 3}$$

The values μ and σ were used to calculate a Z-score that is indicative for enrichment or depletion of an MDB in *S*:

$$Z = \frac{R_S - \mu}{\sigma} \qquad \text{Eq. 4.}$$

The equations above assume a hypergeometric distribution; a discrete probability distribution that is very commonly used for modelling discrete problems as the one stated above. Eq. 1 to 4 enable the calculation of the expected frequency, μ, of an *R*, given a number of randomly selected edges that equals the number of edges that was found in association to a given MDB (R_s).. The equations for the calculation of the hypergeometric standard deviation of μ, and the expression of the de-pauperation or enrichment of an MDB in terms of Z-scores (deviation of the MDB frequency from μ in multiples of σ). As an approximation, Z-scores of $z \approx 2$ and $z \approx 2.5$ associate to the p-values $p \approx 0.05$ and $p \approx 0.01$, respectively. Naturally, there is no guarantee for R_S to follow a hypergeometric distribution, and as E_S becomes smaller, both p-values and Z-scores become biased (overestimated). The Matlab code, which was used to calculate the MDEA statistics is available as Table S5. Cross-validation via bootstrapping of *S* was performed 10,000 times to obtain an additional frequentist measure of significance. Notably, the present concept is entirely data-driven, as it does not differentiate whether compounds/metabolites are known to metabolic databases or not. The general workflow towards MDEA can be viewed in Fig. S5.

Incidence matrix

Even if database coverage is poor, it is possible to describe empirical metabolomics datasets by MDB-based transformations of molecules from metabolomic databases. MDiNs that encompass database entries and empirical data contain an interface between both metabolome spaces. Molecular formulas that were assigned to empirical m/z features can be expressed as combinations of molecular formulas from databases and MDBs. To analyze and visualize whether specific MDBs are used to generate arbitrary groups of non-annotated empirical features, it is necessary to convert the given data into an appropriate structure.

The adjacency matrix (Harary 1962), the Laplacian (Merris 1994), and the incidence matrix (Fulkerson & Gross 1965) (IM) are structures that represent a graph. IM rows pertain to nodes *v* and IM columns pertain to edges *e*. Each edge is listed as a distinct variable. Non-zero entries imply the incidence of a node v_i and an edge e_j. This structure was chosen for the representation of a transformation map that explains the data as a function of KEGG database entries.

Firstly, the 854 KEGG nodes and the 4,206 POP nodes were co-networked. As the aim was the description of the production of POP metabolites that are not a subset of the KEGG metabolome, 129 POP molecular formulas found in KEGG were omitted. Both previously described MDB lists were combined into one list of 450 MDBs. The MDiN consisted of 117,693 edges (the 129 omitted features would have amassed 8,644 additional edges). 492 KEGG formulas (substrates) were directly connected to 2,316 POP formulas (products) by means of 16,109 edges. 2,138 and 1,988 edges were connected to IE nodes and NE nodes respectively (Table S3).

Secondly, MDEA was used to analyze which MDBs were specifically associated to IE and NE nodes. 20 and 31 MDBs obtained Z-scores major 2 for their association to IE and NE respectively.

Thirdly, 42 human metabolome database (HMDB, Wishart *et al.* 2007) compound classes were assigned to 486 out of 492 KEGG formulas that served as substrates. Multiple compound class assignments were considered as independent variables. Edges are replaced by MDBs and compound classes replace nodes in the KEGG-POP-IM. Non-zero values indicate the sum of incidences a compound class has to a given MDB class. Figure 5 was created using the clustergram function within MATLAB over the KEGG-POP-IM. The KEGG-POP-IM is a generic extension to Walker *et al.* (2014b), who annotated an unknown metabolite of major impact by *in silico* conjugation of empirical data.

Empirical NULL distributions

To accommodate for an inappropriate assumption of the hypergeometric NULL distribution, α-values were empirically determined by performing MDEAs over 10,000 bootstrapped marker sets of the same size as the original set of markers. A-values for over-representation and underrepresentation are provided as 'xlsx' files (Table S2).

MS/MS experiments

One IE leaf and one NE leaf were investigated by fragmentation experiments that were performed using the multiple adjacent selected ion monitoring (SIM) method (Southam *et al.* 2007). ESI parameters were set as described above. The spectra were acquired over a time domain transient of 4 Megawords and an Ion-accumulation time of 1.3 s. The SIM window size was set to 30 Da. The SIM window was first centered around 260 m/z and was then shifted towards 440 m/z values in 15 Da steps (13 windows). Fragmentation of each SIM window was performed with four different fragmentation energies: (i) 0 eV, (ii) 5 eV, (iii) 10 eV, and (iv) 15 eV. Each spectrum was acquired for 56 scans.

Data were calibrated and aligned following spectral overlap. Annotation was performed as described above. Each spectrum was then divided into parent and daughter sections. MDiNs were created using KEGG rpairs and the manually curated MDB list. For MDEA, only valid parent→daughter (P→D) pairs were considered. The MDEA variables were generated as follows. *E* represents the sum of all MDBs that were P→D pairs in IE and NE.; *R* is the sum of all MDBs that were P→D pairs with P being a marker for either NE or IE; E_S defines the frequency of P→D pairs for each MDB; and R_S: gives the frequency of P→D pairs with P being a marker for each MDB.

Results

Interpretation of m/z feature statistics

The gray poplar dataset contains 211 mass spectra, each of which encompasses 2276 ± 886 non-randomly occurring m/z peaks after calibration, alignment, and exclusion of exported noise peaks. A mass difference network (MDiN)-based annotation strategy according to Tziotis *et al.* (2011) resulted in an average of 1,579 ± 487 successful monoisotopic molecular formula assignments per spectrum, with a final amount of 4,335 annotated molecular formulas across all spectra. The quality of the formula assignment is displayed as error over m/z plots (Fig. S1), indicating a good spectral alignment, as well as a slightly non-linear systematic error distribution.

Prior to the application of any further statistics, it was necessary to drop all m/z features that could not be annotated successfully, as they are potential artefacts by nature. Furthermore, knowing that the 4,335 assigned molecular formulas pertain to m/z features, both terms will be used interchangeably across the remainder of this work. Multivariate statistical approaches are common practice in metabolomics, as they are appropriate in screening large datasets for features that are potentially associated to a given experimental intervention. Here, orthogonal partial least squares discriminant analysis (OPLS-DA)-modeling was used and the top 10% of most important features were used for the characterization of molecular formulas in NE and IE genotypes (Fig. 2, Fig S1 and Fig S2).

Comparing the assigned molecular-formulas of our gray poplar dataset with the KEGG database resulted in only 129 database hits (3%), which limits the interpretation of metabolomics results, and may result in false statements.

Initial observation of differences in the molecular formula annotations (compositional space) revealed that pure CHO, CHOP and CHOS compositions are by trend upregulated in the NE genotypes, while a majority of N-containing compositions were IE-specific (Fig. 1, Fig. S2). The average cyclomatic number *u* according to Senior (1951; later termed double bond equivalent or degree of unsaturation) was significantly lower in NE than in IE (uNE = 6.6, uIE = 9.1 | p = $2.2*10^{-19}$). Consequently, compounds in the IE genotypes are characterized by a higher amount of double bonds, which is supported by significantly lower H/C ratios and H/(C+N) ratios in IE (pH/C= $1.4*10^{-12}$,pH/(C+N)=$2.2*10^{-23}$). These general analyses imply there are a lower amount of C- and N-aromatics in NE, as well as a higher amount of nitrogen-free compounds.

Among the 129 KEGG molecular formula hits (242 KEGG compounds), 17 molecular formulas (55 KEGG entries) and 22 molecular formulas (47 KEGG compounds) were accumulated in NE and IE, respectively (Table S1). It is common to describe and interpret metabolomes in terms of statistics over the KEGG pathway map participation of designated markers (Kankainen *et al.* 2011). The present analysis shows that 97% of the assigned molecular formulas in leaf extracts of IE and NE poplars do not match with KEGG listed compounds, indicating a more general problem in plant metabolomics: a lack of structural information. Of the 17 molecular formulas found for the NE genotypes, two were related to down-stream products of the methylerythritol phosphate (MEP) pathway where the PcISPS knock-down was induced ($C_5H_{12}O_7P_2$ for dimethylallyl pyrophosphate / isopentenyl pyrophosphate (DMAPP / IPP, pyrophosphate = PP) and $C_{10}H_{20}O_7P_2$ for geranyl-PP (GPP)). Five formulas were related to saccharide metabolism, glycolysis, or pentose-phosphate metabolism ($C_6H_{12}O_6$ for hexose, $C_{12}H_{22}O_{11}$ for hexose di-saccharide, $C_6H_{10}O_7$ for hexuronic acid, $C_6H_{13}O_9P$ for hexose-P, $C_7H_{15}O_{10}P$ for heptose-P). Furthermore, the formulas for malate, C_{20}- and C_{22}-fatty acids, two flavonoids, two hydroxybenzoate derivatives, and one phytosterol were found. Five flavonoid molecular formulas, three unsaturated fatty acids/α-linoleic acid derivatives including linoleic acid, three phenylpropanoid derivatives, two glucosyl-flavonoids, three quinic acid derivatives, as well as hexose-bisphosphate and dihydroxyacetone phosphate (DHAP) were found on the side of IE genotypes. PEP and glycerol 3-phosphate, the isomer of DHAP, are the first substrates for the non-mevalonate or 2-C-methyl-*D*-erythritol 4-phosphate (MEP) pathway. Higher amounts of DMAPP, the substrate of PcISPS, were observed in NE genotypes. This is consistent with the corresponding RNAi knock-down of this enzyme. Furthermore, flavonoid metabolism, which is derived from phenylpropanoid metabolism, is pronounced in IE plants (Table S1, Fig. S2, and Fig. S3). While hexose and hexose-P are enriched in NE, hexose-bisphosphate is depleted; this implies feedback regulation in this genotype on the level of phosphofructokinase (PFK, EC 2.7.1.11), which is inhibited either by aberrant ATP/AMP ratios or by phospho*enol*pyruvate (PEP) (Kelly & Latzko 1977, Stitt 1990). Furthermore, differential behaviors of these three hexose derivatives implies that the classification of features is not primarily driven by matrix effects as they have similarly strong hydrophilicity and polarizability, i.e. they are features that are typically suppressed by more surfactant molecules in ESI ionization. KEGG metabolic pathway hits that were markedly enriched in the NE genotype are ‘Ubiquinone and other terpenoid-quinone biosynthesis’ as well as ‘Starch and sucrose metabolism’. Major changes in general terpenoid metabolism were not indicated. The IE specific pathway hits ‘Biosynthesis of amino acids’, ‘Flavonoid biosynthesis’, ‘Phenylalanine, tyrosine and tryptophan biosynthesis’, ‘Anthocyanin biosynthesis’, and ‘Phenylpropanoid biosynthesis’ in turn agrees with the compositional results presented above (Table S1, Fig. S2, and Fig. S3).

Mass-difference-based building blocks (MDBs) and mass difference networks (MDiNs)

4,335 molecular formulas could be assigned globally. The MDiNs were reconstructed over the set of theoretical monoisotopic exact masses that pertain to the 4,335 neutral molecular formulas of the poplar dataset (POP-formulas). Literature proposes two ways to generate lists of MDBs. The first one (Breitling *et al.* 2006a) formulates MDBs by means of manual inspection of biochemical reactions. Subtracting the masses of substrate A from product C in a biochemical reaction A+B→C, will result in an MDB that describes substrate B, notwithstanding the existence of various reaction mechanisms (Fig. 1A and Fig. 1B). Such a list of MDBs will be termed as ‘manual MDBs’. Using this approach, enzyme-catalyzed and spontaneous reactions were curated manually and converted into a first set of 248 mass differences (Table S2). The second approach (Weber & Viant 2010) mines MDBs from biochemical reactions listed in databases.

This approach is the basis for a second MDB set which is created from KEGG reaction-pairs (rpairs) that were downloaded from KEGGAPI, (similar to Weber & Viant (2010); http://www.KEGG.jp/KEGG/rest/KEGGapi.html). This set will be termed as ‘rpair-MDBs’ for convenience. Ultimately, 301 unique rpair-MDBs shared 99 entries with the list of manual MDBs. The MDiNs that were generated by cross-linking the POP-formulas either with manual or rpair-MDBs encompassed 63,608 (NE|IE: 18,143|15,537) and 65,180 (NE|IE: 19,102|16,072) edges respectively. As MDBs are derived from biochemical reactions, the following sections will interpret and discuss them as both, building blocks and reactions.

It was then possible to investigate whether the edges of specific MDBs were significantly associated to all IE– and NE-formulas as compared to their frequencies across the entire MDiNs. These specific up- or down-regulated MDBs form the basis of MDEA. In MDEA, the MDiN itself is the reference population, and as MDB-frequency counts are discrete, MDB-NE/IE association can be tested against a hypergeometric distribution (Fisher Exact Test; see method section for detailed description).

Prenylation-MDBs directly link the metabolism to the modification of isoprenoid biosynthesis

Mass difference enrichment analysis (MDEA) was used to attribute each MDB with a Z-score expressing the over- ($Z \geq 2$) or underrepresentation ($Z \leq -2$) of MDBs with either IE or NE nodes (molecular formulas) A scatter plot of Z(IE) over Z(NE)-scores (Fig. 2B) enables a swift visual summary of the types of MDBs associated to either genotype. Forty MDBs out of 248 (16%) of the manually curated MDB list were observed to be over-represented in the NE genotype (Fig. 2B, Fig. S2).

Three of the MDBs describe mass-differences that pertain to mono-, di-, and tri-prenylation reactions, with mono-prenylation yielding the highest Z-score (Z=4.54) of the entire dataset. The molecular formulas of two hemiterpene glucosides (HTG) were annotated. Namely 4-hydroxy-2-methyl-2-buten-1-yl-O-glucopyranoside ($C_{11}H_{20}O_7$) (Ward *et al.* 2011) and 2-C-methyl-D-erythritol-O-4-b-D-glucopyranoside (Gonzalez-Cabanelas *et al.* 2015), which were recently characterized in Arabidopsis. Both compounds were upregulated in the NE genotype (Table S1) and possessed a high incidence with the three aforementioned mentioned prenylation MDBs (Fig. 2C, Fig. 3A).

Our analysis revealed that hexose and other discriminant sugars in the isoprene-free cellular background frequently connect to nodes of unknown identity and strong phenotypic discrimination by means of prenylation MDBs; reasonably, the molecular formulas fit to prenylated glycosides (Fig. 2C, Fig. 3A). The same three prenylation reactions, plus two KEGG-specific carotenoid-epoxide rpair-MDBs were observed to be associated to the NE genotype, using both MDB sets.

MDBs mirror oxidative stress responses in the NE genotype

Dicarboxylic acids (DCA) form an MDB-class that is present only in the manually curated list. Eight out of ten DCA-MDBs were calculated to be enriched in the metabolome of the NE genotype (Fig. 2B), a finding that would have been missed by conservative database-driven analysis. The DCAs azelaic acid (Z=2.13) and pimelic acid (Z=3.37) are produced either by lipoxygenases or via spontaneous fragmentation of oxidized lipids (Zoeller *et al.* 2012). The manually curated MDBs included 15 FA reactions, 11 of which (73%) were over-represented in NE, while four were simultaneously depleted in the IE genotype (Fig. 2B). The majority of these FA MDBs were acids with less than ten carbon atoms, which is a chain length that corresponds well with the expected residuals of oxilipin break-down (Table S2). Although biochemical functions regarding the remaining DCAs are not yet described, their corresponding MDBs can be hypothesized to be markers for variants of the oxilipin pathway, which is the primary known source for plant DCAs. Corresponding DCA precursors can be unsaturated FAs other than linoleic acid, likely with alternative double bond positions (Zoeller *et al.* 2012). High oxidative stress in plants leads to increased concentrations of malondialdehyde (MDA), a marker for lipid hydroperoxidation (Moore & Roberts 1998). MDA is formed through a spontaneous radical reaction (Pryor *et al.* 1976), which was translated into an MDB (Table S2). Its over-representation (Z=2.77) in isoprene-free lines

(Fig. 2B) matches previously published work, which shows the accumulation of MDA and H_2O_2 in the NE plants (Behnke *et al.* 2010a). MDA itself is too small and too unstable to be detected using FT-ICR-MS. Oxidative stress-related MDBs also dominate the KEGG-based MDEA results and therewith confirm both the “curated MDB”-based interpretations and literature-based knowledge/hypotheses. MDBs that pertained to α-linoleic acid residuals (3 out of 4 MDBs) and oxilipin metabolism (3 out of 3 MDBs), one of three MDA reactions and one of two peroxidation reactions, were significantly enriched in association to NE nodes.

Network triangle-motifs aid lifting curation induced biases

The curation of MDB lists induces biases towards the curator’s assumptions, however, as MDBs are mass differences there is no guarantee for them to reflect the exact biochemical reaction they were intended to describe. One MDB might likewise be the sum of smaller MDBs, i.e. more elementary building blocks or reactions.

Specific MDB patterns that were strongly enriched in the NE genotype were detected: (i) the decarboxylative condensation of adipic acid is equivalent to a prenylation followed by an hydroxylation; (ii) the condensation of decanoic acid (Z=4.34) can alternatively be described by a di-prenylation followed by H_2O addition; and (iii) the decarboxylative condensation of ketohexanoic acid (Z=4.42) is equivalent to a two-step reaction involving prenylation and H_2O addition. A subgraph containing these MDBs was extracted while allowing only nodes connected to at least one upregulated NE node (Fig. 2C). The three MDB groups formed triangle-motifs, which are similarity-indicating network motifs that are commonly used for the calculation of the clustering coefficient (Fig. 2D, Fig. 3) (Barabasi & Oltvai 2004). These triangles offer alternative biochemical interpretations: the decarboxylative condensation of adipate is isomeric to the condensation of 2-methylbut-2-en-1,4-diol (detected in Arabidopsis) (Ward *et al.* 2011), which might also occur in a two-step reaction of prenylation

(Z_{NE} = 4.54) followed by hydroxylation. Furthermore, the connection of the above-mentioned HTG to glucose (an NE node) and $C_{11}H_{20}O_6$ was found using that exact triangle motif (Fig. 2D, Fig. 3A). Additionally, nodes that participate in triangles (Fig. 2D, Fig. 3A) all share at least one characteristic composition, in this case hexose. The highest FA Z-score (Z=4.34) was obtained by the condensation of decanoic acid (depleted in IE Z=-2.67), which forms a triangle-motif with diprenylation and hydrolysis. Again, this triangle connected to molecular formula annotations that fit prenyl glycosides, monoterpenes, and carbohydrates (Fig. 2E, Fig. 3B). Consequently, the high Z-score of decanoic acid is partly due to its close relationship to the geranylation MDB. Here, triangle motifs helped overcome self-induced biases upon MDB-interpretation.

MDEA reveals global compositional reprogramming in the NE genotype

Hemiterpene glucosides were identified in Arabidopsis when plants were grown under nitrogen limitation, showing an interdependent regulation of the N and C metabolism (Ward *et al.* 2011). MDEA yielded a striking number of 96 under-represented MDBs in the NE genotype (Fig. 2B, Table S2), each of them containing nitrogen, with 56 of those pertaining to amino acids (56/59 amino acid MDBs). The compositional evaluations (Fig. 1, Fig. S2), that indicate a loss of CHNOP compounds in NE, agree with these findings. Closely linked to amino acids are the one- and two-step transamination MDBs, which were also depleted in the NE samples. The major source for N-metabolism is ammonia, which enters primary metabolism in the form of carbamoyl-phosphate (Masclaux-Daubresse *et al.* 2010). The corresponding MDB is strongly underrepresented in the metabolome of NE (Z=-4.71) while phosphorylation is enriched (Z=2.51). The other metabolic entry site for ammonia is glutamine/glutamate (Bernard & Habash 2009, Chellamuthu *et al.* 2014). With a Z-score of -

5.04, glutamine is the second most underrepresented NE-MDB, directly following formimine transfer (Z=-5.15) which is tetrahydrofolate-dependent.

MDBs from the NE genotype highlight alterations in the phospho*enol*pyruvate-dependent metabolism.

The TCA cycle makes use of TCAs, DCAs and 2-oxo-acids. Eleven aliphatic and/or acidic 2-oxo-acid MDBs were enriched in the NE genotype (Fig. 2B). The highest (second overall) Z-score (Z=4.42) of this MDB group was obtained by the decarboxylative addition of ketohexanoic acid. As previously stated, this mass difference forms a triangle with the prenylation MDB (Fig. 2F, Fig. 3C).

Four 2-oxo-acid MDBs, pertaining to pyruvic acid, hydroxypyruvic acid, and 2-ketosuccinic acid, as well as erythrose and transphosphorylation, are known to be related to PEP and pyruvate and were found to be enriched in NE poplar leaves. Notably, transphosphorylation describes the biosynthesis of PEP from oxaloacetate, which can be synthesized from malate. The latter tricarboxylic acid cycle metabolite accumulated in NE leaves. The 2-keto-succinic acid MDB (Z=3.4) is equivalent to the decarboxylative addition of oxaloacetate, which is strongly associated to the NE genotype while its free ion was not detected. Yet another NE-MDB pertained to 2-oxo-glutarate, which next to oxaloacetate, malate and succinate, represents a tricarboxylic acid cycle (TCA) intermediate. Furthermore, the enrichment of the condensation and decarboxylative addition of ketoisovaleric acid in the NE genotype establishes an additional link to pyruvate and the TCA cycle.

Phenolic MDBs are characteristic for the isoprene emitting genotype

Seven out of 13 IE-genotype-associated MDBs pertained to the metabolism of aromatics and shikimate (Fig. 2B, Table S2). Earlier metabolomic and transcriptomic experiments on NE poplars showed the diminished production of phenolics when isoprene is absent, compared to their isoprene-emitting homologs (Behnke *et al.* 2010a, Kaling *et al.* 2015). This result coincides with the up-regulation of the IE-nodes of dehydroquinate and quinate, two intermediates of the shikimate pathway (Fig. 2G).

MDEA-driven cropping of MDiNs improves the visual localization of metabolic pathways

Full MDiNs, reconstructed with hundreds of MDBs, often allow for neither visual nor graph-theoretical network analyses as they tend to resemble a ball of wool and do thus not possess any appreciable network structure/topology. MDB-based biochemical interpretation aside, MDEA is helpful as a means of data-driven dimensionality reduction for network visualization (Fig. 4A). Here, hydroxylation and hydration were used in addition to the top six MDBs that were over-represented in the NE genotype for the extraction of sub-graphs enriched with upregulated NE nodes (Fig. 4A). This approach resulted in the formation of five sub-graphs (separated into the elemental compositions CHO, CHOP, CHNO, CHOS and CHNOS), which were affected by the genetic modification (Fig. 1, Fig. S2), whereby annotations of CHO, CHOS and CHOP exhibited many discriminant nodes that were upregulated in the metabolome of isoprene-free lines. Breitling *et al.* (2006a) previously stated that MDiNs might represent a means for the characterization of pathways within the known metabolic realm. The MDiA approach enabled the visual detection of the shikimate pathway which underpinned its association with the IE genotype within the extracted CHO sub-graph (Fig. 4B). Per definition, MDiNs must contain metabolic pathways, and analyzing how building blocks are processed aside, MDEA and the introduced triangle-motifs (Fig. 2) are useful tools to narrow down pathways that are both known and unknown.

The MDiN incidence matrix describes the origin of the unknown metabolic space

MDBs were shown to be highly valuable for the interpretation of metabolomic MS-data, especially if the majority of molecular formulas are unknown. However, all results obtained so far describe relationships within this poplar dataset itself. The next question of interest was whether the same context held true if the poplar metabolome (POP) was related to a widely accepted database (KEGG). To this end, an incidence matrix was constructed, which is a network representation where nodes are mapped against their respective edges (Table S3). Figure 5 describes how KEGG-metabolites are transformed into discriminant unknown compounds. Notable differences between the poplar genotypes were found for terpene nodes (Fig. 5). The NE plants displayed a more pronounced terpene metabolism mainly conjugating different carbohydrates to terpenes and *vice versa* (farnesylation and geranylation on carbohydrate containing compounds).

IE plants preferentially conjugated aromatic moieties with carbohydrates that likely resulted in (poly)phenolic-glycosides. Additionally, the incidence matrix shows that IE plants use N-aromatics and amino acids more frequently, which corresponds to the MDEA results that were mined in POP.

The initial molecular formula annotations already indicated differences in phosphorous-containing compounds between the genotypes (Fig. 1, Fig. S2). The incidence matrix confirmed this observation and additionally revealed different usage patterns. Plants with missing isoprene bio-catalysis performed FA condensations and geranylation, whereas isoprene producers preferably conjugated aromatic metabolites. The introduced methodology facilitates the definition of targeted strategies to investigate very specific aspects of a largely unknown metabolism. Ultimately, it allows the mining of candidate enzymes by querying databases for the combined information of source compound class and acting MDB for future studies on gray poplar.

Tandem mass spectrometric measurements validate MDEA findings in the NE genotypes

As described in the introduction, the main focus of this work is to demonstrate how MDiA can integrate the entirety of acquired m/z data, be it of high or low ion abundance, into the analysis of biochemical building block usage patterns. Fragmentation experiments on single m/z species are only possible if a given feature can be isolated using a quadrupole filter. For this exact reason, Morreel *et al.* (2014) fragmented only the most abundant m/z signal per full FT-ICR MS scan. One of our initial hypotheses was that MDEA can reveal information about the rate at which a building block/metabolite is invested by distinct genotypes to synthesize their specific metabotype. As single feature identification is costly in many different ways, and the aim of this work is to introduce the concept of MDiA based metabolome contextualization, multiple adjacent selected ion monitoring (SIM) MS/MS (Southam *et al.* 2007) scans were performed on leafs of both the NE and IE genotypes. Markers of interest suffer from less penalization using the SIM-MS/MS approach as compared to smaller SIM windows conventionally used for MS/MS. The larger SIM windows allow for marker features of low abundance to contribute to the daughter-ion space. The analysis was focused on the mass range 245 to 455 m/z, which equals the mass range of the network shown in Figure 2. MDEA was used to mine neutral losses that were significantly associated to the NE and IE markers as compared to all parent-daughter ion pairs of a given SIM window.

The MDEA results of the SIM-MS/MS experiments yielded an overrepresentation of 49 MDBs in the NE genotype using the curated MDB-list (Fig. 6A, Table S4). The correlation coefficient between all full-scan and SIM-MS/MS MDB Z-scores from the NE samples was 0.76 (Fig. 6E, Table S4).

Twenty-five MDBs were significantly over-represented in both the full scan and the SIM-MS/MS results (Fig. 6A). The three MDBs, which were part of the triangle motifs in Figure 3, namely the prenylation, decarboxylative condensation of adipic acid, and of ketohexanoic acid, were also over-represented in the SIM-MS/MS data (Fig. 6B). This observation directly validates the full-scan MDEA data in which MDEA established a direct linkage between the metabotype and the genotype. Additionally, it substantiates the presence of unknown prenylated compounds in NE poplars (Fig. 3).

Close similarities between conventional full-scan and SIM-MS/MS results were observed in the DCA MDBs, where seven out of the eight over-represented full-scan MDBs were also associated to the NE poplars in the SIM-MS/MS data (Fig. 6B).

Additionally, 17 2-oxo acid MDBs, of which 11 described the cleavage of aliphatic 2-oxo acids, were over-represented in the SIM-MS/MS spectra of NE plant extracts. This is in agreement with the observation that ten out of those 17 2-oxo acid MDBs were also over-represented in the full-scan MDB results of this genotype (Fig. 6B).

The MDBs that only were over-represented in the SIM-MS/MS results pertained to four two carbon 2-oxo acid building blocks (pyruvate-related), two hydroxyphenylpyruvate MDBs, and one ketoglutarate MDB (Table S4). Six FA related MDBs were enriched in the NE SIM-MS/MS experiments (Table S4). Four of them overlapped with MDBs that were enriched in the full-scan MDEA results of the NE genotype (Fig. 6B). Two of those MDBs describe butanoic acid reactions (type A and B), one hexanoic type A reaction and one dodecanoic acid type A reaction (Fig. 6B). Another observation that confirms the results obtained by the incidence matrix is the overrepresentation of six carbohydrate MDBs in the SIM-MS/SM NE poplar results. The results of the incidence matrix show that those MDBs were used for terpene modifications to form unknown terpene glycosides in NE poplar (Fig. 5B). A high overall correlation between the Z-score profiles of the NE incidence matrix and of the NE SIM-MS/MS (Fig. 6E) was observed.

MS/MS experiments validate and extend the conjugation of phenolics in the IE genotype

Forty-two MDBs were over-represented in the SIM-MS/MS data set of IE poplars. The correlation coefficient between the MDB Z-scores of full-scan and SIM-MS/MS MDEA analysis in the IE poplars was 0.68 (Fig. 6E, Table S4). Eight MDBs yielded significant Z-scores in both measurement types (Fig. 6C). Five of those MDBs described phenolic reactions (including two aromatic 2-oxo-acids), one characterizes the condensation of quinate and the remaining two described ketoglutarate reactions (Fig. 6D). These results do not only validate the full-scan MDB results (Fig. 2), they also complement the known literature showing higher contents of phenolics in leaves of IE compared to leaves of the NE genotypes (Behnke *et al.* 2010a, Way *et al.* 2013, Kaling *et al.* 2015). Additionally, four phenolic MDBs and three aromatic 2-oxo acid MDBs were over-represented in the SIM-MS/MS results, thus further complementing the full-scan MDEA results in IE poplars.

Discussion

Interpretation of m/z feature statistics

The present analysis shows that 97% of the assigned sum formulas in the leaf extracts of IE and NE poplars do not match to KEGG listed compounds, highlighting a more general problem in plant metabolomics; a lack of structural information. Although the amount of KEGG-hits is limiting, some interpretations are still possible: while hexose and hexose-P are enriched in leaves of NE poplars, hexose-bisphosphate is depleted; this implies feedback regulation in this genotype on the level of phosphofructokinase (PFK, EC 2.7.1.11), which is inhibited either by aberrant ATP/AMP ratios or by phospho*enol*pyruvate (PEP) (Kelly & Latzko 1977, Stitt 1990). Furthermore, a differential behavior of these three hexose derivatives implies that the classification of features is not primarily driven by matrix effects. DHAP is depleted in the NE genotype as well, which supports that interpretation. The hexose-scenario implies a consequent increase in PEP flux, while PEP itself could not be observed as an ion; increase of PEP flux would however confirm previous observations of 1-deoxy-*D*-xylulose 5-phosphate synthase (DXS; EC 2.2.1.7) feedback inhibition by DMAPP (Ghirardo *et al.* 2014). The observed decrease in DHAP goes hand in hand with a decrease in pentose-P, which generally indicates a limited energy metabolism in NE plants. As highlighted above, leaves of NE poplars are enriched in hexose and hexose-P and depleted in hexose-PP levels, a scenario which implies feedback regulation of PFK by PEP. PEP itself – a highly reactive/unstable metabolite - could not be detected. Increased levels of PEP can be well explained by a backlog of metabolites resulting from strong reduction of metabolic flux through the MEP pathway as consequence of the negative feedback regulation of DXS by increased DMAPP levels (Ghirardo *et al.* 2015). As phenylpropanoids were found to be downregulated in NE poplar leaves, the PEP-traversing C-flux might have been redirected to the anaplerotic TCA cycle whose metabolites involve TCAs, DCAs and 2-oxo-acids. Malate is apparently enriched in this genotype, but this observation does not allow any further interpretation pertaining to the TCA cycle mass flux. Yet, the MDBs of 11 2-oxo-acids, among them 2-keto-succinic acid, as well as transphosphorylation and pyruvate-related MDBs were found to substantially contribute as building blocks for NE upregulated metabolites. The entire context of these findings suggests 13C-fluxomic analyses to be a hotspot of future investigations and thus exemplifies how hypotheses that can be of use for future investigations can be generated using non-targeted metabolomics and MDiA.

Prenylation-MDBs directly link the metabolism to the modification of isoprenoid biosynthesis

It is known that the chloroplastic DMAPP pool is much larger in NE as compared to IE leaves (Ghirardo *et al.* 2014), yet only 'ubiquinone- and terpenoid-quinone biosynthesis' is listed among the addressed terpenoid KEGG pathway maps (Fig. 1, Fig. S3). As these pathways are located in the cytosol and not in the plastids, the KEGG pathway hits do not appropriately reflect alterations in the MEP pathway in NE plants in the near proximity to PcISPS (Cheng *et al.* 2007). This result represents a major phenological/contextual connection between the metabotype and absence of isoprene biosynthesis in NE plants. DMAPP and GPP were found to be upregulated in the NE genotype (Table S1) among the few KEGG annotations. DMAPP is the substrate of PcISPS, and together with its isomer IPP forms the C5 building blocks of GPP and farnesyl pyrophosphate (FPP), which are the major di- and tri-prenylating agents for phenolic compounds (Shen *et al.* 2012), zeatin biosynthesis (Mok *et al.* 2000), and for the posttranslational modification of proteins (Zhang & Casey 1996). The knock-down of PcISPS in poplar results in a strong accumulation of DMAPP (Ghirardo *et al.* 2014), confirming the present up-regulation of isoprenoid intermediates in the NE genotypes. This is in accordance with Weise *et al.* (2013) who showed that acid hydrolysis, a commonly used technique for the quantification of DMAPP available for isoprene synthesis, results in higher amounts of DMAPP in gray poplar than is quantified by LC/MS, which indicates a substantial pool of unknown prenylated compounds (Weise *et al.* 2013). The three types of MDiA and MDEA performed agree that this pool of prenylated compounds is likely constituted of various kinds of hemiterpene glucosides. The POP-MDEA finds prenylation, geranylation, and farnesylation among the top ranked MDBs. The POP-KEGG incidence matrix agrees with this finding as theoretical terpenoids preferentially connected to NE-upregulated compounds using various carbohydrate MDBs, while various carbohydrates used farnesylation and geranylation to connect to regulated features. Carbohydrate containing compounds were further shown to connect to a wide range of fatty acids in NE and phenylpropanoids in IE. Earlier metabolomic and transcriptomic experiments on NE poplars showed the diminished production of phenolics when isoprene is absent compared to their isoprene-emitting homologs (Behnke *et al.* 2010a, Kaling *et al.* 2015). This result coincides with the up-regulation of the IE-nodes of dehydroquinate and quinate, two intermediates of the shikimate pathway (Fig. 2G). These results were confirmed by the results of the SIM-stitch MS/MS approach, where prenylation, but not geranylation and farnesylation, was found to be of importance. The NE-leaves' lack in phenylpropanoids was confirmed as well. The high incidence of carbohydrate MDBs in the SIM-stitch MS/MS approach is not shown in Fig. 6 but can be viewed in Table S4, where condensations of glucose, rhamnose, and erythrose were top ranked for NE markers. Poplar trees are known for their high production of phenolics, such as flavonoids, phenolic glycosides, and phenylpropanoids (Babst *et al.*, 2010, Boeckler *et al.*, 2011). These compounds differ drastically in their glycosylation patterns, and because of that neutral losses of carbohydrates are often observed in MS/MS experiments of plant extracts (Kachlicki *et al.*, 2008; Abreu *et al.*, 2011). These results complement the known literature showing higher contents of phenolics in leaves of IE compared to leaves of the NE genotypes (Behnke *et al.* 2010a, Way *et al.* 2013, Kaling *et al.* 2015).

MDBs mirror oxidative stress responses in the NE genotype

Interestingly, the linoleic acid derivatives, which are commonly interpreted to have an association to oxidative stress (op den Camp *et al.* 2003, Moller *et al.* 2007) are higher in IE. This observation contradicts existing knowledge, which implies higher oxidative stress in plants when cell internal isoprene is absent (Behnke *et al.* 2010a). Indeed, MDEA results correct preemptive interpretation based on a small set of KEGG metabolite hits as follows:

although biochemical functions regarding most DCAs are not yet described, their corresponding MDBs can be hypothesized to be markers for variants of the oxilipin pathway, which is the primary known source for plant DCAs. This is supported by the incidence-matrix (Fig. 5B) that visualizes the compositional relationship of DCAs with fatty acids (FA) via a clustering of FA KEGG-nodes, and the NE MDBs succinate transfer to aldehydes and pimelate condensation. As Table S4 confirms, DCA-neutral losses are jointly characteristic for NE-features in SIM-stitch MS/MS. Furthermore, it may be speculated whether the DCA-MDBs represent unstable metabolic intermediates, similar to pimeloyl-CoA (Streit & Entcheva 2003), which prevents the detection of the free metabolites with MS techniques. This finding clearly supports the assumption that mass difference analysis can drastically improve the interpretation of metabolomic data because it has the capability to describe spontaneous reactions of metabolites *in vivo*, as is the case for MDA.

Conclusion

Three types of MDiNs and MDEAs were used: firstly, the full scan MDiN, where markers for both genotypes were assigned using OPLS classification. Here, MDEA was performed on the full scan MDiN and significant Z-scores pertained to the general differences in metabolome setup between both investigated genotypes. The reference edge population of all marker-associated MDBs was the entire set of MDB-edges found in the MDiN. This approach was largely data driven, but knowledge driven in that two different types of MDB sets were used, the curated list and the KEGG list. Both approaches yielded a consistent metabolomic context of the investigated genotypes. The approach of Breitling *et al.* (2006a, 2006b) is flexible in a way that manual curation can lead to mass differences addressing building blocks that are not listed as free reactants in e.g. KEGG. The approach by Weber and Viant (2010) has the advantage that each MDB (rpair) can be associated to a set of enzymes, which – knowing some compositional and structural properties of the detected features – can be narrowed down to more specific enzyme sets. These can be targeted in future proteomics and transcriptomics experiments as well as on UHPLC-ToF-MS data (Forcisi *et al.* 2015).

Secondly, the directed KEGG-POP incidence matrix: this approach is an extension of Walker *et al.* (2014b) as it uses MDBs to connect knowledge base data to experimental non-targeted metabolomics data. Ultimately, this approach enabled Walker *et al.* (2014b) to discover sulfonated lipids that were confirmed via MS/MS. Herein, the full list of neutral KEGG metabolites were considered as substrates (building blocks) for the marker sets of the NE and IE genotypes. The reconstructed MDiN was thus directed and bipartite because all source nodes came from KEGG and all target nodes came from the full scan data (POP). The reference edge population was the entirety of all MDB-edges connecting KEGG to POP.

Thirdly, the directed SIM-MS/MS network: here, edges were directed from parent ions (P) to daughter ions (D) per SIM window and collision energy. All P→D MDBs per spectrum were counted as reference population and P→D MDBs that connected marker P's defined in above to the daughter space were tested for their overrepresentation against the reference population. The consistency of all MDEAs was tested given the Weber-Viant-KEGG list. The correlation matrix for all six Z-score sets (Fig. 6E) clearly shows that MDEAs from the NE genotypes were more similar among themselves in comparison to MDEAs from IE and vice versa. Only the SIM-MS/MS Z-score profiles of IE and NE poplars were more similar to each other than the Z-scores of the full-scan IE MDEA and the MDEA of the IE incidence matrix. The reason for that is that SIM-MS/MS experiments lead to a fragmentation of the entire metabotype, which – as both genotypes are poplar trees – still have a given large basal set of building blocks. Nevertheless, all three investigations delivered consistent results for each respective genotype.

The present study demonstrated how the application of MDiNs can be extended beyond feature annotation and compound identification by probing them for network regions, where nodal genotype differentiation significantly coincides with compositional context. In fact, the complex chemical biology of the two gray poplar genotypes was completely grasped by MDEA, which demonstrates this technique's tremendous potential for 'omics'-based applications, and opens the door for the development of tailor-made targeted techniques beyond the limitations of database knowledge.

Acknowledgments

The authors would like to thank M. Clancy for language editing. This project was financially supported by European Union 7th Framework Programme FP7-Health-2013-Innovation-1 under the grant agreement No.603288 and the German Resarch Foundation (DFG Project No. SCHN653/5-1).

References

Abreu I.N., Ahnlund M., Moritz T. & Albrectsen B.R. (2011) UHPLC-ESI/TOFMS determination of salicylate-like phenolic gycosides in *Populus tremula* leaves. *Journal of Chemical Ecology* **37**, 857-870.

Babst B.A., Harding S.A. & Tsai C.J. (2010) Biosynthesis of phenolic glycosides from phenylpropanoid and benzenoid precursors in populus. *Journal of Chemical Ecology* **36**, 286-297.

Barabasi A.L. & Oltvai Z.N. (2004) Network biology: understanding the cell's functional organization. *Nature Reviews Genetics* **5**, 101-113.

Behnke K., Ehlting B., Teuber M., Bauerfeind M., Louis S., Hasch R. (...) Schnitzler J.P. (2007) Transgenic, non-isoprene emitting poplars don't like it hot. *The Plant Journal* **51**, 485-499.

Behnke K., Grote R., Brüggemann N., Zimmer I., Zhou G., Elobeid M. (...) Schnitzler J.P. (2012) Isoprene emission-free poplars--a chance to reduce the impact from poplar plantations on the atmosphere. *New Phytologist* **194**, 70-82.

Behnke K., Kaiser A., Zimmer I., Brüggemann N., Janz D., Polle A. (…) Schnitzler J.P. (2010a) RNAi-mediated suppression of isoprene emission in poplar transiently impacts phenolic metabolism under high temperature and high light intensities: a transcriptomic and metabolomic analysis. *Plant Molecular Biology* **74**,

Behnke K., Loivamaki M., Zimmer I., Rennenberg H., Schnitzler J.P. & Louis S. (2010b) Isoprene emission protects photosynthesis in sunfleck exposed Grey poplar. *Photosynthesis Research* **104**, 5-17.

Bernard S.M. & Habash D.Z. (2009) The importance of cytosolic glutamine synthetase in nitrogen assimilation and recycling. *New Phytologist* **182**, 608-620.

Boeckler G.A., Gershenzon J. & Unsicker S.B. (2011) Phenolic glycosides of the *Salicaceae* and their role as anti-herbivore defenses. *Phytochemistry* **72**, 1497-1509.

Boyle N.R. & Morgan J.A. (2009) Flux balance analysis of primary metabolism in *Chlamydomonas reinhardtii*. *BMC Systems Biology* **3**, 4.

Breitling R., Ritchie S., Goodenowe D., Stewart M.L. & Barrett M.P. (2006a) *Ab initio* prediction of metabolic networks using Fourier transform mass spectrometry data. *Metabolomics* **2**, 155-164.

Breitling R., Pitt A.R. & Barrett M.P. (2006b) Precision mapping of the metabolome. *Trends in Biotechnology* **24**, 543-548.

Chellamuthu V.-R., Ermilova E., Lapina T., Lüddecke J., Minaeva E., Herrmann C. (...) Forchhammer K. (2014) A Widespread glutamine-sensing mechanism in the plant kingdom. *Cell* **159**, 1188-1199.

Cheng A.X., Lou Y.G., Mao Y.B., Lu S., Wang L.J. & Chen X.Y. (2007) Plant terpenoids: biosynthesis and ecological functions. *Journal of Integrative Plant Biology* **49**, 179-186.

Choi Y.H., van Spronsen J., Dai Y., Verberne M., Hollmann F., Arends I.W. (…) Verpoorte R. (2011) Are natural deep eutectic solvents the missing link in understanding cellular metabolism and physiology? *Plant Physiology* 156 1701-1705.

Forcisi S., Moritz F., Lucio M., Lehmann R., Stefan N. & Schmitt-Kopplin P. (2015) Solutions for low and high accuracy mass spectrometric data matching: a data-driven annotation strategy in nontargeted metabolomics. *Analytical Chemistry* **87**, 8917-8924.

Fulkerson D.R. & Gross O. (1965) Incidence matrices and interval graphs. *Pacific Journal of Mathematics* **15**, 835-855.

Ghirardo A., Wright L.P., Bi Z., Rosenkranz M., Pulido P., Rodriguez-Concepcion M. (…) Schnitzler J.P. (2014) Metabolic flux analysis of plastidic isoprenoid biosynthesis in poplar leaves emitting and nonemitting isoprene. *Plant Physiology* **165**, 37-51.

Gipson G.T., Tatsuoka K.S., Sokhansanj B.A., Ball R.J. & Connor S.C. (2008) Assignment of MS-based metabolomic datasets via compound interaction pair mapping. *Metabolomics* **4**, 94-103.

Gonzalez-Cabanelas D., Wright L.P., Paetz C,. Onkokesung N., Gershenzon J., Rodriguez-Concepcion M. & Phillips M.A. (2015) The diversion of 2-C-methyl-d-erythritol-2,4-cyclodiphosphate from the 2-C-methyl-d-erythritol 4-phosphate pathway to hemiterpene glycosides mediates stress responses in *Arabidopsis thaliana*. *The Plant Journal* **82**, 122-137.

Harary F. (1962) The determinant of the adjacency matrix of a graph. *SIAM Review* **4**, 202-210.

Ideker T., Galitski T. & Hood L. (2001) A new approach to decoding life: systems biology. *Annual Reviews of Genomics and Human Genetics* **2**, 343-372.

Joseph B., Corwin J.A., Baohua L., Atwell S. & Kliebenstein D.J. (2013) cytoplasmic genetic variation and extensive cytonuclear interactions influence natural variation in the metabolome. *eLife* **2**, e00776.

Kachlicki P., Einhorn J., Muth D., Kerhoas L. & Stobiecki M. (2008) Evaluation of glycosylation and malonylation patterns in flavonoid glycosides during LC/MS/MS metabolite profiling. *Journal of Mass Spectrometry* **43**, 572-586.

Kaling M., Kanawati B., Ghirardo A., Albert A., Winkler J.B., Heller W. (…) Schnitzler J.P. (2015) UV-B mediated metabolic rearrangements in poplar revealed by non-targeted metabolomics. *Plant Cell and Environment* **38**, 892-904.

Kampranis S.C., Ioannidis D., Purvis A., Mahrez W., Ninga E., Katerelos N.A. (…)Johnson C.B. (2007) Rational conversion of substrate and product specificity in a Salvia monoterpene synthase: structural insights into the evolution of terpene synthase function. *The Plant Cell* **19**, 1994-2005.

Kankainen M., Gopalacharyulu P., Holm L. & Oresic M. (2011) MPEA - metabolite pathway enrichment analysis. *Bioinformatics* **27**, 1878-1879.

Kelly G.J. & Latzko E. (1977) Chloroplast phosphofructokinase II. Partial purification, kinetic and regulatory properties. *Plant Physiology* **60**, 295-299.

Krumsiek J., Suhre K., Illig T., Adamski J. & Theis F.J. (2011) Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Systems Biology* **5**, 21.

Lucio M., Fekete A., Frommberger M. & Schmitt-Kopplin P. (2011) Metabolomics: High-resolution tools offer to follow bacterial growth on a molecular level. In *Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches* (ed F.J. de Bruijn), pp. 683-69. John Wiley & Sons, Inc, New York.

Masakapalli S.K., Le Lay P., Huddleston J.E., Pollock N.L., Kruger N.J. & Ratcliffe R.G. (2010) Subcellular flux analysis of central metabolism in a heterotrophic Arabidopsis cell suspension using steady-state stable isotope labeling. *Plant Physiology* **152**, 602-619.

Masclaux-Daubresse C., Daniel-Vedele F., Dechorgnat J., Chardon F., Gaufichon L. & Suzuki A. (2010) Nitrogen uptake, assimilation and remobilization in plants: challenges for sustainable and productive agriculture. *Annals of Botany* **105**, 1141-1157. doi:10.1093/aob/mcq028.

Merris R. (1994) Laplacian matrices of graphs: a survey. *Linear Algebra Applications* **197–198**, 143-176.

Mok M.C., Martin R.C. & Mok D.W.S. (2000) Cytokinins: Biosynthesis, metabolism and perception. *In Vitro Cellular & Developmental Biology - Plant* **36**, 102-107.

Moller I.M., Jensen P.E. & Hansson A. (2007) Oxidative modifications to cellular components in plants. *Annual Reviews of Plant Biology* **58**, 459-481.

Moore K. & Roberts L.J. (1998) Measurement of lipid peroxidation. *Free Radical Research* **28**, 659-671.

Moritz F., Janicka M., Zygler A., Forcisi S., Kot-Wasik A., Kot J. (…) Schmitt-Kopplin P (2015) The compositional space of exhaled breath condensate and its link to the human breath volatilome. *Journal of Breath Research* **9**, 027105.

Morreel K., Saeys Y., Dima O., Lu F., Van de Peer Y., Vanholme R. (…) Boerjan W (2014) Systematic structural characterization of metabolites in *Arabidopsis* via candidate substrate-product pair networks. *The Plant Cell* **26**, 929-945.

Müller C., Dietz I., Tziotis D., Moritz F., Rupp J. & Schmitt-Kopplin P. (2013) Molecular cartography in acute *Chlamydia pneumoniae* infections--a non-targeted metabolomics approach. *Analytical and Bioanalytical Chemistry* **405**, 5119-5131.

Nägele T., Mair A., Sun X., Fragner L., Teige M. & Weckwerth W. (2014) Solving the differential biochemical Jacobian from metabolomics covariance data. *PLoS ONE* **9**, e92299.

op den Camp R.G., Przybyla D., Ochsenbein C., Laloi C., Kim C., Danon A. (…) Apel K. (2003) Rapid induction of distinct stress responses after the release of singlet oxygen in *Arabidopsis*. *The Plant Cell* **15**, 2320-2332.

Pichersky E. & Lewinsohn E. (2011) Convergent evolution in plant specialized metabolism. *Annual Reviews of Plant Biology* 62, 549-566.

Pryor W., Stanley J.P. & Blair E. (1976) Autoxidation of polyunsaturated fatty acids: II. A suggested mechanism for the formation of TBA-reactive materials from prostaglandin-like endoperoxides. *Lipids* **11**, 370-379.

Rogers S., Scheltema R.A., Girolami M. & Breitling R. (2009) Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics* **25**, 512-518.

Schwab W. (2003) Metabolome diversity: too few genes, too many metabolites? *Phytochemistry* **62**, 837-849.

Senior J.K. (1951) Partitions and their representative graphs. *American Journal of Mathematics* **73**, 663-689.

Shen G., Huhman D., Lei Z., Snyder J., Sumner L.W. & Dixon R.A. (2012) Characterization of an isoflavonoid-specific prenyltransferase from *Lupinus albus*. *Plant Physiology* **159**, 70-80.

Southam A.D., Payne T.G., Cooper H.J., Arvanitis T.N. & Viant M.R. (2007) Dynamic range and mass accuracy of wide-scan direct infusion nanoelectrospray fourier transform ion cyclotron resonance mass spectrometry-based metabolomics increased by the spectral stitching method. *Analytical Chemistry* **79**, 4595-4602.

Steuer R., Kurths J., Fiehn O. & Weckwerth W. (2003) Observing and interpreting correlations in metabolomic networks. *Bioinformatics* **19**, 1019-26.

Stitt M (1990) Fructose-2, 6-bisphosphate as a regulatory molecule in plants. *Annual Reviews of Plant Biology* **41**, 153-185.

Streit W.R. & Entcheva P. (2003) Biotin in microbes, the genes involved in its biosynthesis, its biochemical role and perspectives for biotechnological production. *Applied Microbiology and Biotechnology* **61**, 21-31.

Subramanian A., Tamayo P., Mootha V.K., Mukherjee S., Ebert B.L., Gillette M.A., (…) Mesirov J.P. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Acadademy of Sciences USA* **102**, 15545-15550.

Tziotis D., Hertkorn N. & Schmitt-Kopplin P. (2011) Kendrick-analogous network visualisation of ion cyclotron resonance Fourier transform mass spectra: improved options for the assignment of elemental compositions and the classification of organic molecular complexity. *European Journal of Mass Spectrometry* **17**, 415-421.

Velikova V., Ghirardo A., Vanzo E., Merl J., Hauck S.M. & Schnitzler J.P. (2014) Genetic manipulation of isoprene emissions in poplar plants remodels the chloroplast proteome. *Journal of Proteome Research* **13**, 2005-2018.

Velikova V., Müller C., Ghirardo A., Rock T.M., Aichler M., Walch A. (…) Schnitzler J.P. (2015) Knocking down of isoprene emission modifies the lipid matrix of thylakoid membranes and influences the chloroplast ultrastructure in poplar. *Plant Physiology* **168**, 859-870.

Walker A., Lucio M., Pfitzner B., Scheerer M.F., Neschen S., de Angelis M.H. (...) Schmitt-Kopplin P. (2014a) Importance of sulfur-containing metabolites in discriminating fecal extracts between normal and type-2 diabetic mice. *Journal of Proteome Research* **13**, 4220-4231.

Walker A., Pfitzner B., Neschen S., Kahle M., Harir M., Lucio M. (…) Schmitt-Kopplin P. (2014b) Distinct signatures of host-microbial meta-metabolome and gut microbiome in two C57BL/6 strains under high-fat diet. *ISME Journal* **8**, 2380-2396.

Ward J.L., Baker J.M., Llewellyn A.M., Hawkins N.D. & Beale M.H. (2011) Metabolomic analysis of *Arabidopsis* reveals hemiterpenoid glycosides as products of a nitrate ion-regulated, carbon flux overflow. *Proceedings of the National Academy of Sciences USA* **108**, 10762-10767.

Watrous J., Roach P., Alexandrov T., Heath B.S., Yang J.Y., Kersten R.D. (…) Dorrestein PC (2012) Mass spectral molecular networking of living microbial colonies. *Proceedings of the National Academy of Sciences* USA **109**, E1743-E1752.

Way D.A., Ghirardo A., Kanawati B., Esperschütz J., Monson R.K., Jackson R.B. (...) Schnitzler JP (2013) Increasing atmospheric CO_2 reduces metabolic and physiological differences between isoprene- and non-isoprene-emitting poplars. *New Phytologist* **200**, 534-546.

Weber R.J.M. & Viant M.R. (2010) MI-Pack: Increased confidence of metabolite identification in mass spectra by integrating accurate masses and metabolic pathways. *Chemometrics and Intelligent Laboratory Systems* **104**, 75-82.

Weise S.E., Li Z., Sutter A.E., Corrion A., Banerjee A. & Sharkey T.D. (2013) Measuring dimethylallyl diphosphate available for isoprene synthesis. *Analytical Biochemistry* **435**, 27-34.

Witting M., Lucio M., Tziotis D., Wagele B., Suhre K., Voulhoux R., Garvis S. & Schmitt-Kopplin P. (2015) DI-ICR-FT-MS-based high-throughput deep metabotyping: a case study of the *Caenorhabditis elegans-Pseudomonas aeruginosa* infection model. *Analytical and Bioanalytical Chemistry* **407**, 1059-1073.

Wishart D.S., Tzur D. & Knox C. (2007) HMDB: the human metabolome database. *Nucleic Acids Research* **35**, Database issue.

Zhang F., Harir M., Moritz F., Zhang J., Witting M., Wu Y. (…) Hertkorn N. (2014) Molecular and structural characterization of dissolved organic matter during and post cyanobacterial bloom in Taihu by combination of NMR spectroscopy and FTICR mass spectrometry. *Water Research* **57**, 280-294.

Zhang F.L. & Casey P.J. (1996) Protein prenylation: molecular mechanisms and functional consequences. *Annual Review of Biochemistry* **65**, 241-269.

Zöller M., Stingl N., Krischke M., Fekete A., Waller F., Berger S. & Müller M.J. (2012) Lipid profiling of the *Arabidopsis* hypersensitive response reveals specific lipid peroxidation and fragmentation processes: biogenesis of pimelic and azelaic acid. *Plant Physiology* 160, 365-378.

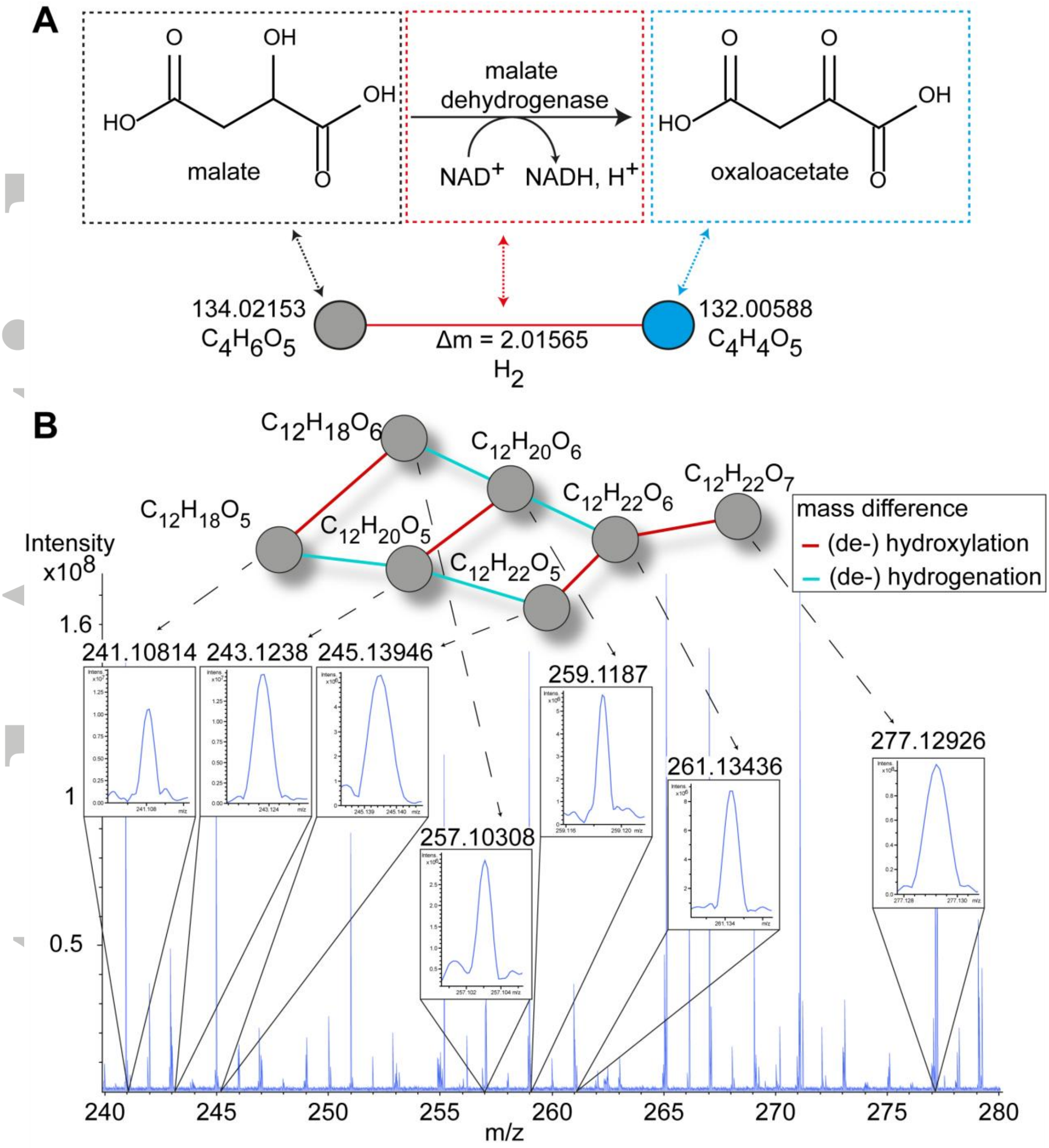


Figure 1. Fresh weight of leaves (black) and roots (white) of A. thaliana plants. *Arabidopsis seedlings were either continuously grown under control conditions or Cu-deficiency or nineteen-days-old control plants were exposed to 5 µM $CdSO_4$ alone or in combination with additional $CuSO_4$ (0.5, 1 or 2 µM Cu extra in comparison to control Hoagland solution) for 72 h. Data are mean ± S.E. of 6 biological replicates. Significant differences ($P<0.05$) after one-way ANOVA test and Tukey correction are indicated with different capital (roots) or small (leaves) letters.*

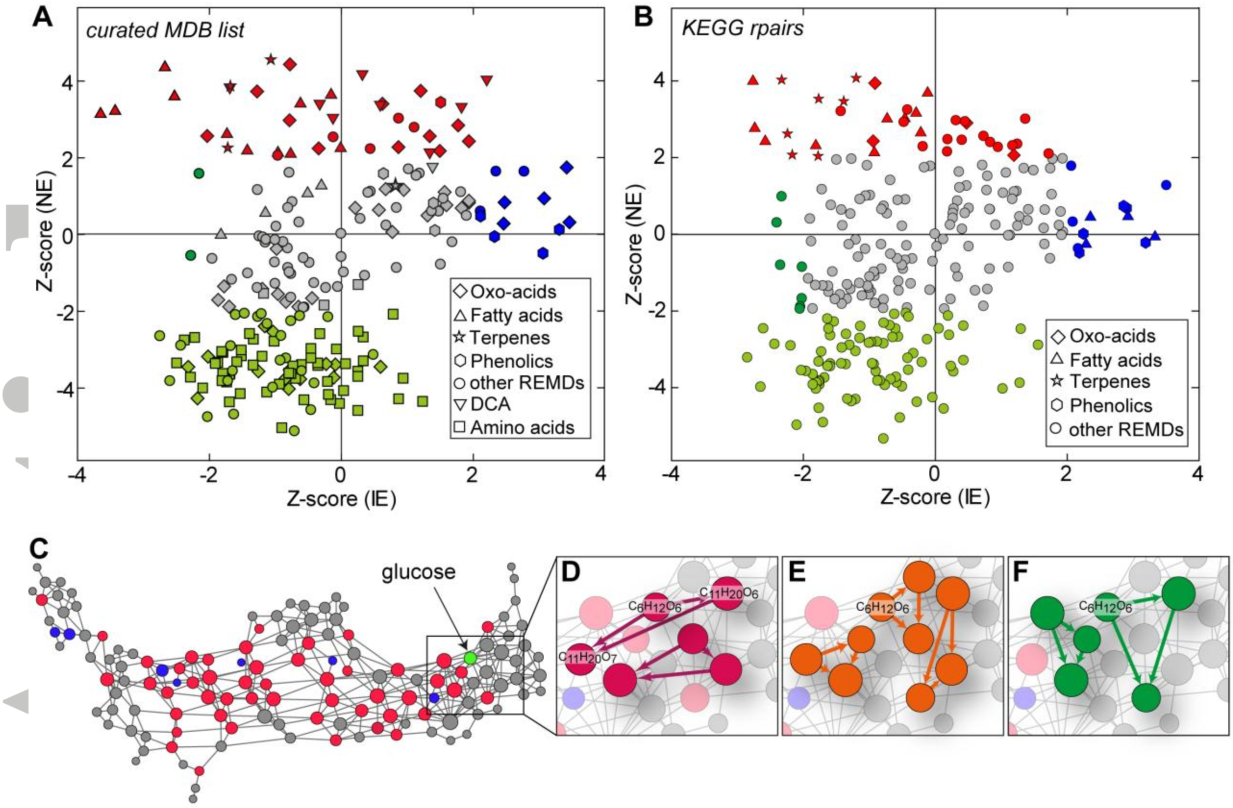


Figure 2. Cd translocation in A. thaliana. *Arabidopsis seedlings were either continuously grown under control conditions or Cu-deficiency or nineteen-days-old control plants were exposed to 5 µM $CdSO_4$ alone or in combination with additional $CuSO_4$ (0.5, 1 or 2 µM Cu extra in comparison to control Hoagland solution) for 72 h. The Cd translocation from roots to leaves was calculated as the concentration in the leaves relative to the concentration in the roots. Data are mean ± S.E. of 6 biological replicates. Significant differences ($P<0.05$) after one-way ANOVA test and Tukey correction are indicated with different letters.*

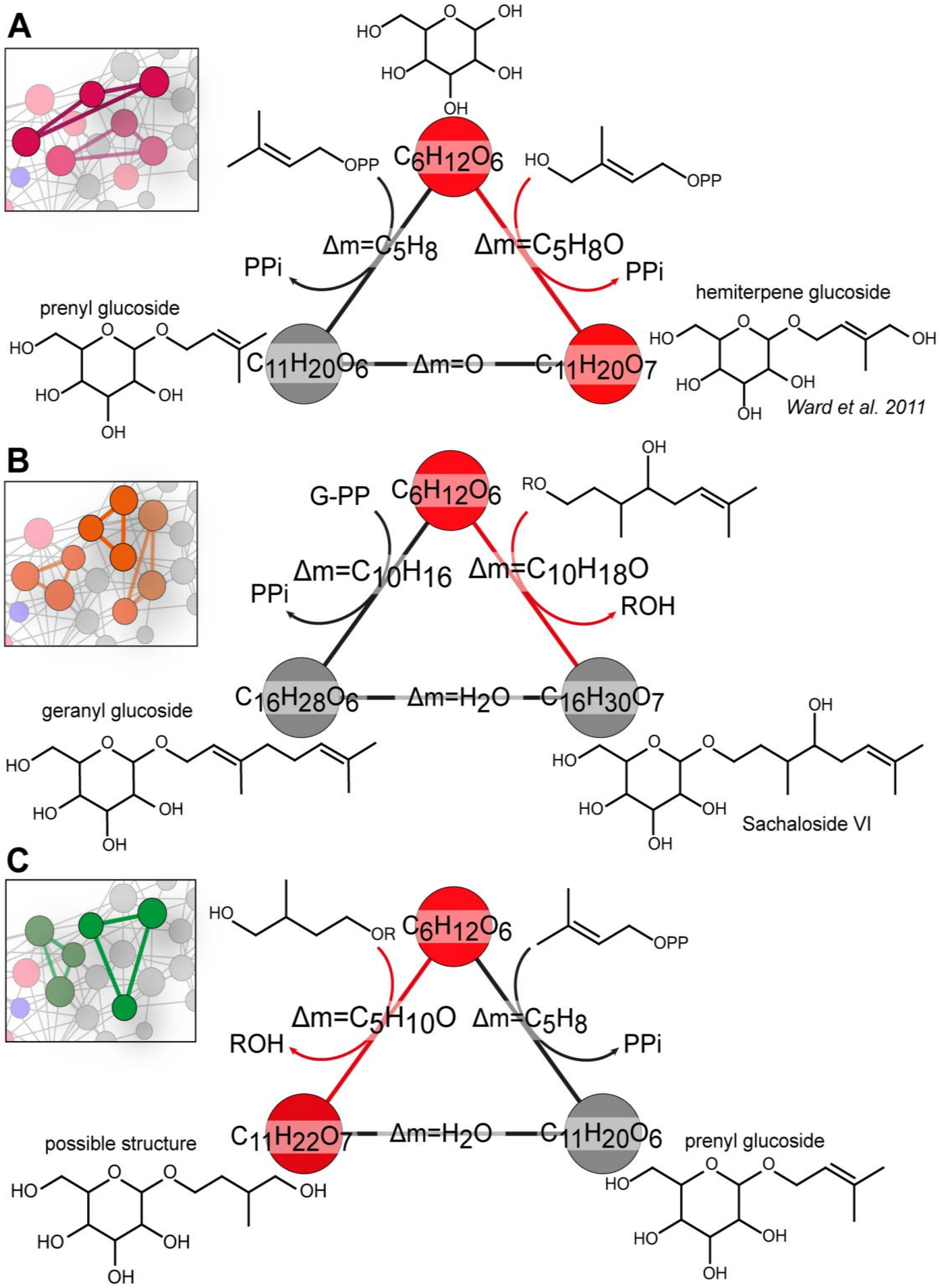


Figure 3. Gene expression levels of SPL7-regulated genes in the leaves and roots of A. thaliana. *Arabidopsis seedlings were either continuously grown under control conditions or Cu-deficiency or nineteen-days-old control plants were exposed to 5 µM CdSO$_4$ alone or in combination with additional CuSO$_4$ (0.5, 1 or 2 µM Cu extra in comparison to control Hoagland solution) for 72 h. Significant differences (P<0.05) after one-way ANOVA test and Tukey correction are indicated with*

colour shading: red for reduction in comparison to control, green for induction in comparison to control and yellow for induction in comparison to control but reduction in comparison to a single 5 µM Cd exposure. The relative fold changes are shown in Table S3. Abbreviations: FSD1, Fe superoxide dismutase1; CCH, Cu chaperone; COPT2, Cu transporter2; ZIP2, zinc-regulated transporter2.

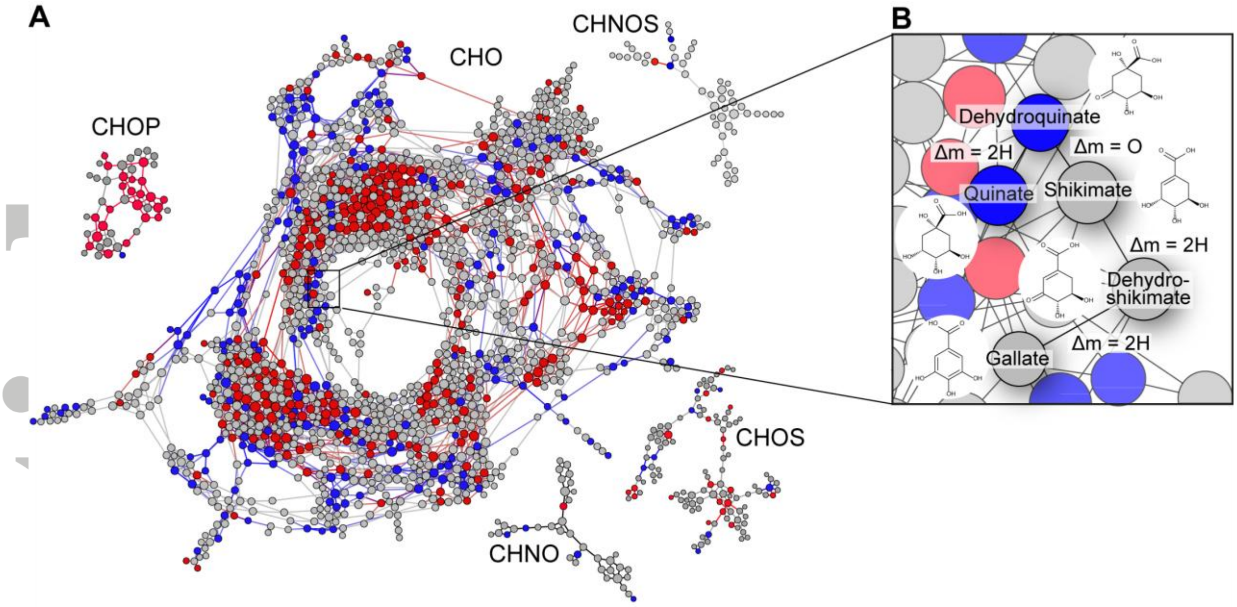


Figure 4. Fresh weight of leaves and roots of A. thaliana wildtype (black) and cad1-3 mutant (white) plants. *Arabidopsis seedlings were either continuously grown under control conditions or nineteen-days-old plants were exposed to 5 µM CdSO$_4$ for 24 h and 72 h. Data are mean ± S.E. of 4 biological replicates. Significant differences (P<0.05) after three-way ANOVA test and Tukey correction are indicated with different small (leaves) or capital (roots) letters.*

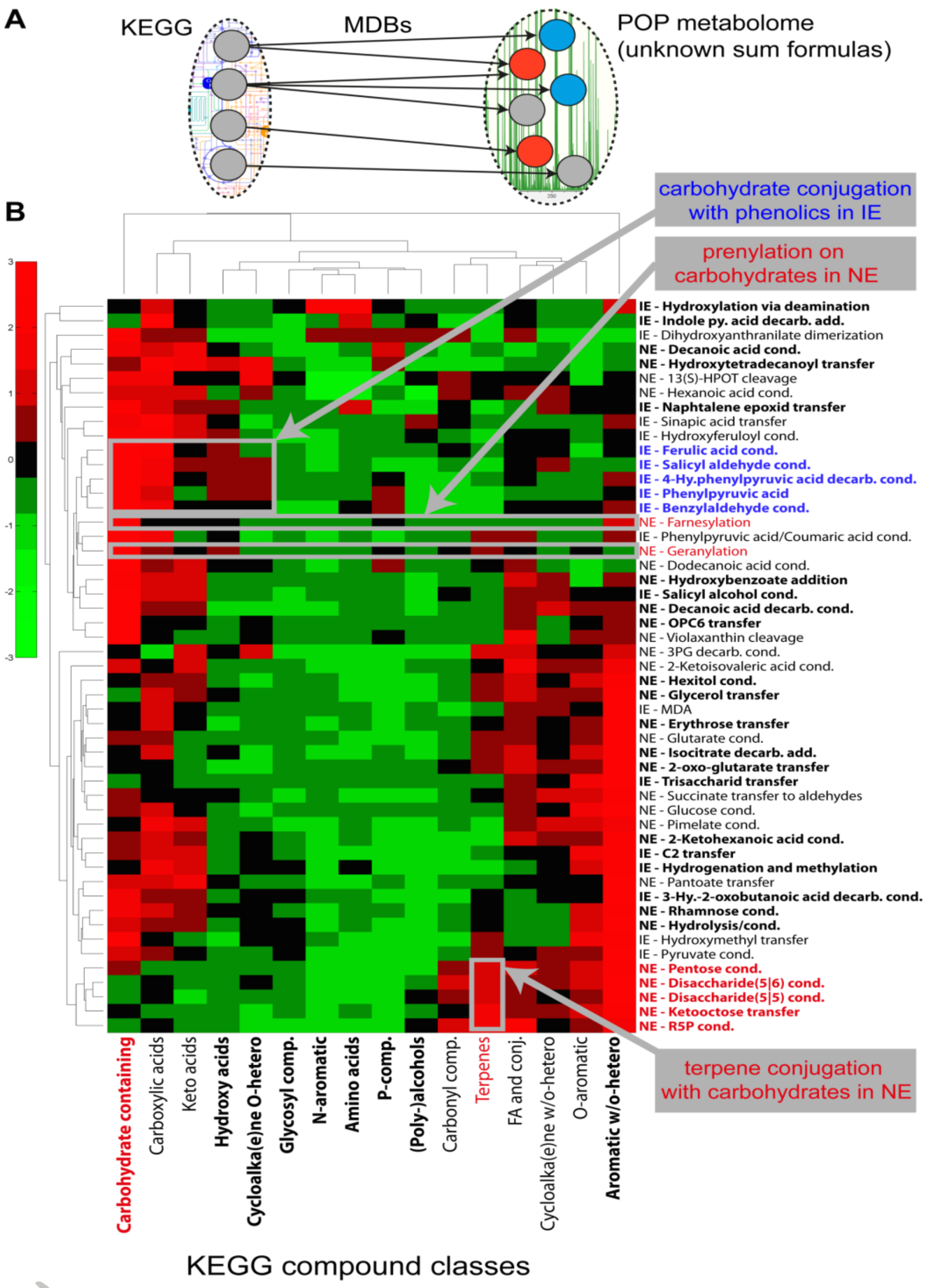
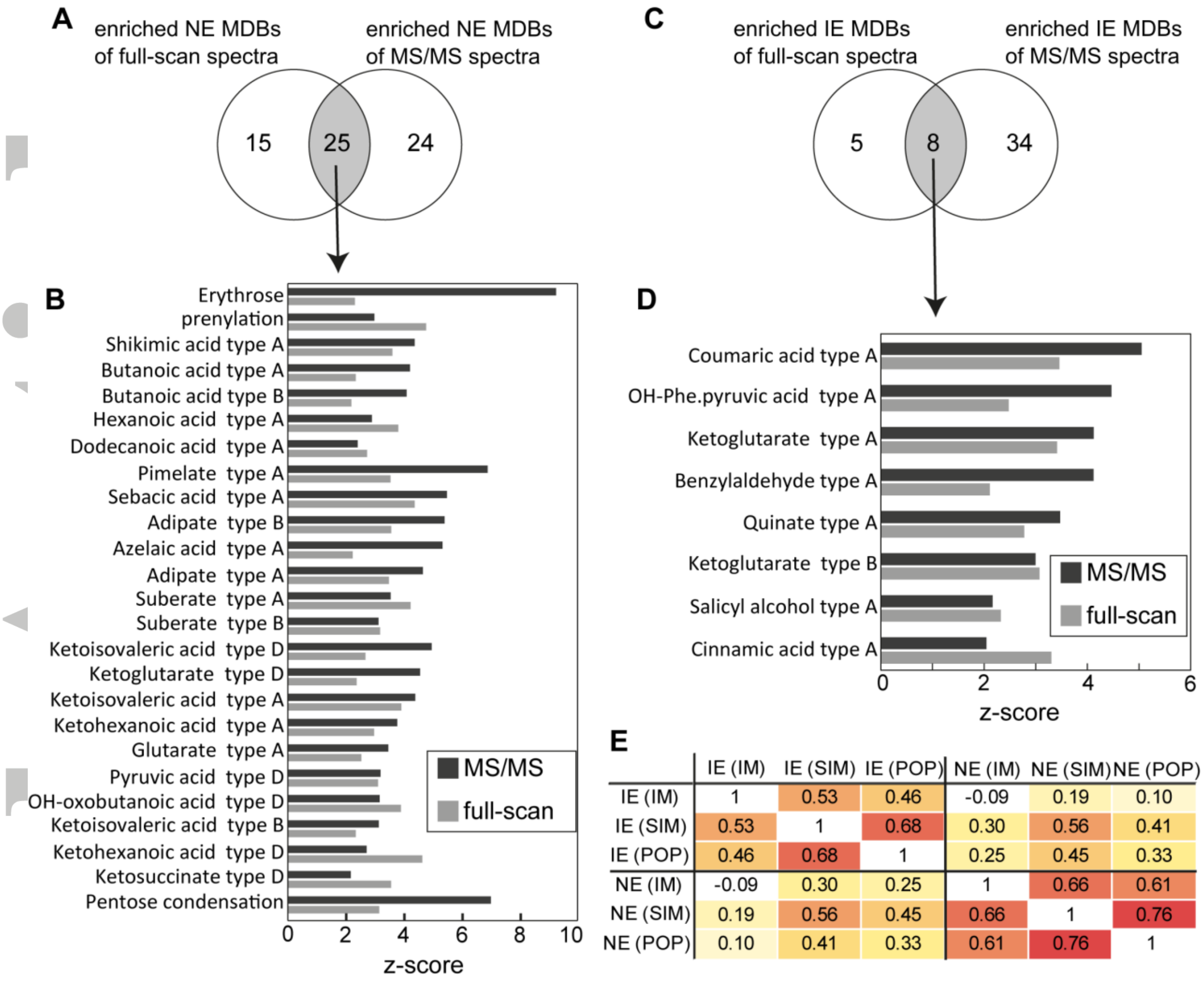


Figure 5. Cd translocation in A. thaliana wildtype (black) and cad1-3 mutant (white) plants. Arabidopsis seedlings were either continuously grown under control conditions or nineteen-days-old plants were exposed to 5 µM CdSO$_4$ for 24 h and 72 h. The Cd translocation was calculated as the concentration in the shoots relative to the concentration in the roots. Data are mean ± S.E. of 7 biological replicates. Significant differences (P<0.05) after two-way ANOVA test and Tukey correction are indicated with different letters.



	IE (IM)	IE (SIM)	IE (POP)	NE (IM)	NE (SIM)	NE (POP)
IE (IM)	1	0.53	0.46	-0.09	0.19	0.10
IE (SIM)	0.53	1	0.68	0.30	0.56	0.41
IE (POP)	0.46	0.68	1	0.25	0.45	0.33
NE (IM)	-0.09	0.30	0.25	1	0.66	0.61
NE (SIM)	0.19	0.56	0.45	0.66	1	0.76
NE (POP)	0.10	0.41	0.33	0.61	0.76	1

Figure 6. Gene expression levels of HMA2 (A) and HMA4 (B) and HMA5 (C) in roots of A. thaliana wildtype and cad1-3 mutant plants. *Arabidopsis seedlings were either continuously grown under control conditions or nineteen-days-old plants were exposed to 5 µM CdSO$_4$ for 24 h. Data are mean ± S.E. of at least 3 biological replicates relative to its own unexposed control (set at 1.00). Significant Cd-induced differences (P<0.05) in expression within each genotype relative to the control after two-way ANOVA test and Tukey correction are indicated with asterisks.*