

# Metabolite profiling for plant functional genomics

Oliver Fiehn<sup>1\*</sup>, Joachim Kopka<sup>2</sup>, Peter Dörmann<sup>1</sup>, Thomas Altmann<sup>1</sup>, Richard N. Trehewey<sup>2</sup>, and Lothar Willmitzer<sup>1</sup>

<sup>1</sup>Max Planck Institute of Molecular Plant Physiology, 14424 Potsdam, Germany. <sup>2</sup>Metanomics GmbH & Co KGaA, Tegeler Weg 33, 10589 Berlin, Germany.

\*Corresponding author ([fiehn@mpimp-golm.mpg.de](mailto:fiehn@mpimp-golm.mpg.de)).

Received 17 April 2000; accepted 21 August 2000

**Multiparallel analyses of mRNA and proteins are central to today's functional genomics initiatives. We describe here the use of metabolite profiling as a new tool for a comparative display of gene function. It has the potential not only to provide deeper insight into complex regulatory processes but also to determine phenotype directly. Using gas chromatography/mass spectrometry (GC/MS), we automatically quantified 326 distinct compounds from *Arabidopsis thaliana* leaf extracts. It was possible to assign a chemical structure to approximately half of these compounds. Comparison of four *Arabidopsis* genotypes (two homozygous ecotypes and a mutant of each ecotype) showed that each genotype possesses a distinct metabolic profile. Data mining tools such as principal component analysis enabled the assignment of "metabolic phenotypes" using these large data sets. The metabolic phenotypes of the two ecotypes were more divergent than were the metabolic phenotypes of the single-loci mutant and their parental ecotypes. These results demonstrate the use of metabolite profiling as a tool to significantly extend and enhance the power of existing functional genomics approaches.**

Keywords: functional genomics, *Arabidopsis thaliana*, metabolite profiling, cluster analysis, metabolomics, bioinformatics

In the post-genomic era, elucidation of gene function is a main focus. Plant functional genomics<sup>1</sup> couples the generation of transgenic and mutant plants to the multiparallel analysis of gene products such as mRNA<sup>2</sup> and proteins<sup>3</sup>. However, these methods do not provide direct information about how a change in mRNA or protein is coupled to a change in biological function. As a result of a multiplicity of regulatory interactions at all levels in plant cells, a change at one level in the complex network does not necessarily lead to a particular change in function or phenotype. Instead, single point mutations might often lead to complex responses at the level of the whole organism. In applying the profiling concept, it is crucial to perform unbiased (metabolite) analyses in order to define precisely the biochemical function of plant metabolism<sup>4</sup>. Such analyses complement existing functional genomics methodologies while offering a direct link between a gene sequence and the function of the metabolic network in plants. Furthermore, metabolite profiling can elucidate links and relationships that occur primarily through regulation at the metabolic level. Finally, a broad metabolic analysis may address public concerns about the safety and value of plant genetically modified organisms.

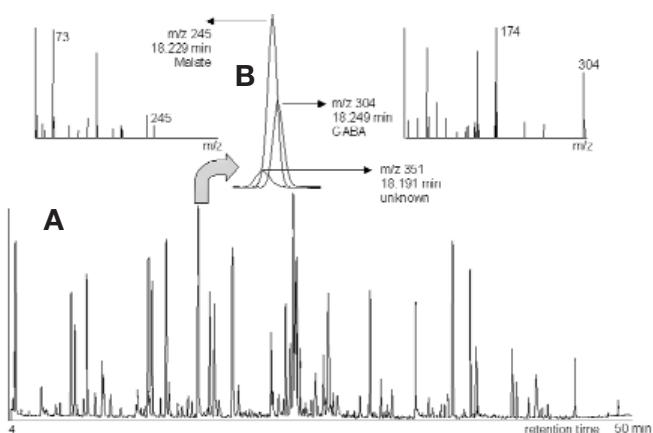
To become established as a robust tool, metabolite profiling must be fast, reliable, sensitive, and suitable for automation, as well as covering a significant number of metabolites. A range of analytical technologies enhances the sensitivity and universality of mass spectrometry by chromatographic separations. To date, however, metabolic screening approaches using mass spectrometry are rarely used in plant research<sup>5,6</sup>. For the most part, the use of multi-target profiling has been limited to rapid clinical detection of human diseases<sup>7</sup>. We judged gas chromatography coupled to electron-impact quadrupole mass spectrometry (GC/MS) to be the most mature technology capable of fulfilling the required criteria.

The methodology described here allows the detection and quantification of more than 300 compounds from a single plant leaf extract.

## Results and discussion

Plant leaf extracts yield 326 quantifiable compounds. Metabolite extraction from *Arabidopsis* leaf tissue was done using methanol and heat, thereby rapidly inhibiting enzymatic activity. We added internal standards in order to correct for minor variations occurring during sample preparation and analysis. A single fractionation step into a lipophilic and a polar phase was followed by solvent evaporation and derivatization for increasing metabolite stability and volatility as reported<sup>8</sup>. Briefly, the lipid phase was trimethylsilylated for the analysis of total fatty acids, fatty alcohols, sterols, and aliphatics, whereas the polar phase was methoximated and trimethylsilylated for the analysis of hydroxy- and amino acids, sugars, sugar alcohols, organic monophosphates, (poly)amines, and aromatic acids. Metabolite sizes were in the range of ethylene glycol (62 AMU) to trisaccharides (504 AMU). Optimal reaction conditions were established as a compromise between reaction completeness and the maintenance of labile compound integrity (data not shown). We chose analytical parameters as a compromise between separation efficiency, column capacity, and column long-term stability. This GC/MS approach is extremely powerful for plant metabolite profiling (Fig. 1). Hundreds of different compounds were detected in parallel, some of which had severely overlapping peaks that are deconvoluted by selective ion traces (Fig. 1B). Compound identification was performed by comparison of mass spectra and retention times with those obtained with commercially available reference compounds. A major advantage of mass spec-

## RESEARCH ARTICLES

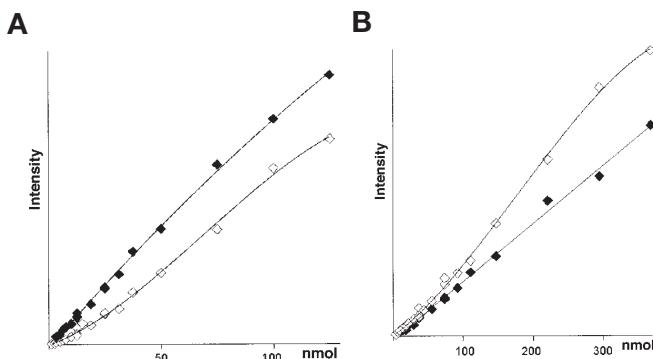


**Figure 1.** Metabolite profiling by GC/MS. Base peak intensity GC/MS chromatogram of the polar fraction of a leaf extract from the *Arabidopsis dgd1* mutant (A). Target metabolites are identified by exact retention times and their corresponding mass spectra (B) as shown for the co-eluting peaks of malate,  $\gamma$ -aminobutyric acid (GABA), and an unidentified compound. m/z, Ratio of mass to charge.

trometry is that unknown peaks can be determined as reliably as known target analytes without prior knowledge of their exact chemical structure. In *Arabidopsis* extracts, after rigorous comparison of mass spectra with commercial libraries<sup>9</sup>, about half of the detected peaks currently remain unidentified. High-throughput peak finding was done by matching mass spectra within a 0.25 min wide time window around the predicted retention times for each target compound. Fluctuations in the relative retention times of sugars and hydroxy acids were found to lie within 0.02 min of the predicted time and, thus, occasional false positive identifications were corrected by setting a postacquisition threshold for the deviation of the retention time of 0.04 min (0.07 min threshold for amino compounds). False positive identifications were automatically qualified as not determined and excluded from further calculations or manually corrected, where necessary. In total, 326 compounds were found in the *Arabidopsis thaliana* leaf extracts (101 polar and 63 lipophilic identified compounds, plus 113 polar and 49 lipophilic compounds of unknown chemical structure). A complete list of the mass spectra of our current target compounds and sample preparation protocols can be downloaded from our website<sup>10</sup>.

With respect to quantification, we followed two approaches. Relative amounts of the various compounds were obtained by normalizing the intensity of individual ion traces (that are indicative for the respective compound even in the presence of co-eluting compounds) to the response of internal reference compounds, and further, to 1 mg of plant leaf fresh weight. For quantification a linear relationship between metabolite amount and the analytical signal is crucial. Internal calibration curves confirmed that this assumption holds true over two to three orders of magnitude when 11 stable isotopic labeled compounds were added to 32 different *Arabidopsis thaliana* leaf extracts (Fig. 2). Because of matrix effects, up to twofold differences were found between external and internal calibration, but no differences were found between mutant and wild-type C24 plants. Calibration linearity was also confirmed for 50 metabolites of unidentified chemical structure both by diluting derivatized plant samples and by derivatizing different volumes of a single plant extract (data not shown).

The stable isotope internal calibration curves were also used to determine the absolute amounts of certain metabolites. Table 1 summarizes the absolute mean values for these compounds as determined for leaf extracts of 18 individual *Arabidopsis thaliana* C24 wild-type plants. Graphs in Figure 2 and data contained in Table 1 confirm that the profiling method established here allows the deter-



**Figure 2.** Metabolite calibrations. Calibration curves for determination of dynamic ranges and absolute concentrations using stable isotope-labeled metabolites. Open symbols, external calibration; filled symbols, internal calibration. (A)  $^{13}\text{C}_6$ -Glucose. (B)  $d_4$ -ethanolamine.

mination of both relative and absolute quantities. However, it is important to stress that for the vast majority of applications of the profiling technology, the absolute value is unimportant, rather the relative value is sufficient.

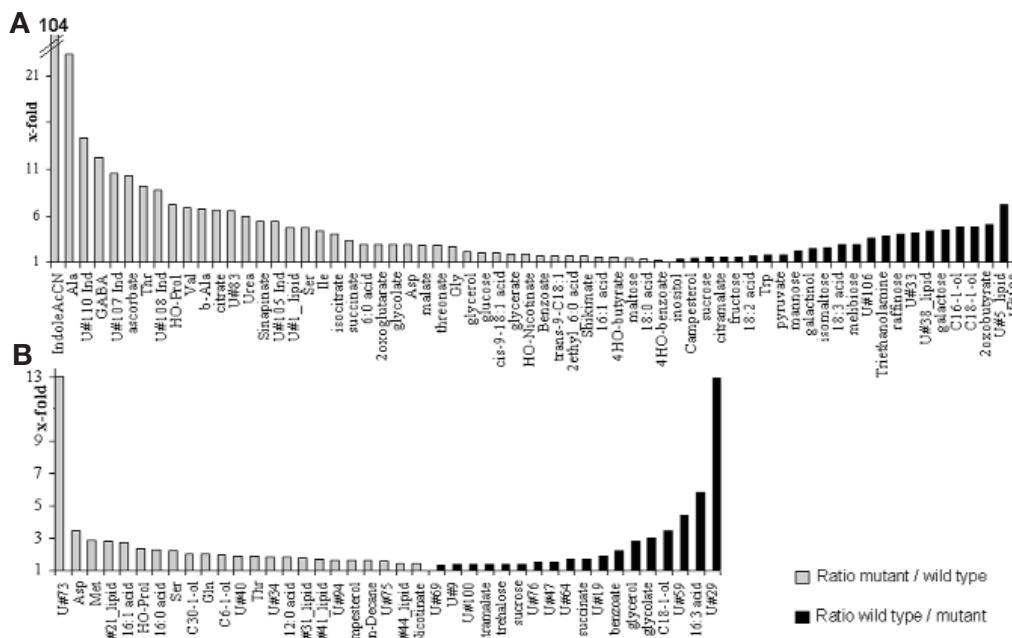
Another key factor for any analytical technique is reproducibility. The reproducibility of the whole process was tested in order to determine the potential contribution of variability in the analytical method to the observed variation between different biological samples. In order to estimate the influence of the sample preparation and the analytical device on variability, two samples of *Arabidopsis* C24 wild type were combined directly following extraction and divided into seven aliquots. Each aliquot was taken separately through the sample preparation procedure and after GC/MS analysis of the polar metabolites. The mean of these deviations was  $8\% \pm 6\%$ , and 110 of these compounds showed even lower deviations ( $5\% \pm 2\%$ ). This is at least as accurate as comparable functional genomic methods at the protein level<sup>11</sup> and clearly more accurate than differential analysis of expression using cDNA microarrays<sup>12</sup>. Therefore, we conclude that the variability introduced into the analytics by the sample preparation and the actual measurement is small and can be tolerated.

In order to get an insight into the biological variability, 18 plants of *Arabidopsis thaliana* genotype C24 (wild type) were grown in the phytotron side by side under identical conditions and harvested at the same time. Absolute values were determined for 11 metabolites based on isotope-labeled internal calibration curves. As evident from Table 1, the variability due to the biological variability is in clear excess of the variability due to the overall analytical precision. This finding indicates the metabolic flexibility of plants. For the

**Table 1.** Biological variation and analytical precision in *Arabidopsis thaliana* C24 WT plants

Chemical	Average contents (nmol/mg FW) <sup>a</sup> , n = 18	Biological variation (% s.d.), n = 18	Analytical reproducibility (% s.d.), n = 7
Ethylene glycol	1.2 ± 0.3	26	6
Alanine	144 ± 56	38	12
Valine	38 ± 6	17	5
Ethanolamine	63 ± 14	23	6
Glycerol	24 ± 12	49	2
Leucine	25 ± 7	29	8
Benzoic acid	0.8 ± 0.3	40	10
Aspartic acid	79 ± 34	43	5
Glutamic acid	199 ± 75	38	3
Glucose	119 ± 67	56	5
Sucrose	598 ± 180	30	2

<sup>a</sup>FW, fresh weight.



**Figure 3. Significant metabolite differences in plant genotypes.** Alterations in mean metabolite levels (*t*-test,  $p < 0.01$ ) of (A) the *dgd1* mutant and (B) the *sdd1-1* mutant compared to their respective parental wild-type backgrounds (Col-2 and C24). For the *dgd1* mutant, 153 significant alterations in metabolites were found. For visual clarity, only 67 of the metabolites are presented that were selected either by their physiological importance or by their metabolite alteration exceeding a factor of 3. For *sdd1-1*, all 41 of the significant alterations are shown.

remaining 140 compounds only relative quantifications were performed. Again the biological variability was found to be on average  $\sim 40\%$  s.d. These findings therefore show that the biological variability seen between genetically identical plants grown under identical conditions is the largest source for the variability observed.

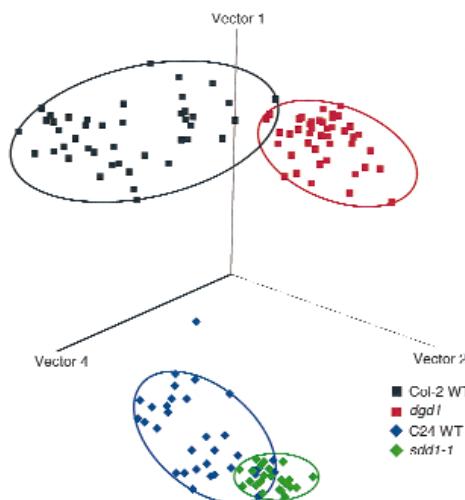
For application in the framework of genomic approaches, one single individual can process 60 samples per working day. Using our protocols, three GC/MS machines are needed for the processing of the 60 samples. It is obvious that this figure can be easily amplified by increasing the number of machines and persons involved.

**Mutants and parental ecotypes display large metabolic differences.** The power of metabolite profiling was tested for its ability to distinguish between ecotypes using various *Arabidopsis thaliana* genotypes, which are supposedly genetically characterized by the presence of several hundred allelic differences, and two mutants with these ecotype backgrounds. One of the mutants should display a severe visible phenotype, whereas the other mutant should grow and develop essentially indistinguishable from the parental wild-type background. The two ecotypes chosen by us were Col-2 and C24. On the genetic background of Col-2, the *dgd1* mutant was chosen, which is characterized by a 90% reduction in the galactolipid digalactosyldiacylglycerol (DGD)<sup>13</sup>. As a consequence of the reduced levels of DGD, the mutant is impaired in photosynthesis and is hypersensitive to light stress<sup>14</sup>, and thus served as an example of a rather severe phenotype. The gene affected was recently cloned and shown to encode a galactosyl transferase (DGD synthase). Because the mutant was backcrossed four times with the parental ecotype, Col-2, most of the original mutant DNA was replaced by Col-2 DNA. By transformation of this line with wild-type genomic DNA fragments carrying the *DGD1* gene or with the *DGD1* cDNA, we could demonstrate<sup>15</sup> that not only the DGD lipid phenotype but also the growth defect were complemented. Therefore, all effects other than deficiency in DGD biosynthesis are believed to be secondary effects. The second mutant used in this study, *sdd1-1*, carries a point mutation in a regulatory gene involved in the control of stomatal development<sup>16</sup>. Like *dgd1*, *sdd1-1* was also backcrossed four times with

its parental ecotype, C24. The lack of *SDD1* gene function causes a two- to fourfold increase in stomatal density; however, the mutant displays no other visible phenotype, and therefore was chosen to represent a mild mutant phenotype. Thus, *sdd1-1* was selected as a morphological mutant for analysis to gain information about the potential metabolic changes caused by the increased stomatal density that result in enhanced gas exchange properties (increased CO<sub>2</sub> uptake and H<sub>2</sub>O release) of the leaves.

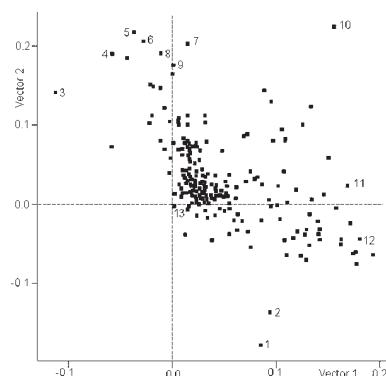
Mutant plants were grown in parallel with their corresponding wild-type plants until the flowering stage (defined by the presence of an inflorescence stem about 7 cm in height) in a controlled environment under standard conditions. All plants were randomly distributed within the growth chamber to eliminate a potential contribution of position effects. For analysis of

each genotype samples from fully expanded rosette leaves were taken from 28–45 individual plants. Individual processing of these samples resulted in 28–45 individual profiles per genotype. After GC/MS analysis, data normalization, and data validation, Student's *t*-tests were carried out for statistical analysis<sup>17</sup>. To achieve high result reliability, we used *t*-test probability limits of  $p < 0.01$  in our evaluations.



**Figure 4. Metabolic phenotype clustering.** Clusters found after principal component analysis (PCA) of log-scaled polar metabolite data of 151 samples originating from four plant genotypes. Single-loci mutants show metabolic phenotypes distinct from wild-type plants (WT). Basic vectors in PCA span an *n*-dimensional space to give best sample separation. Each point represents a linear combination of all the metabolites from an individual sample. Vectors 1, 2, and 4 were chosen for best visualization of genotype separation and include 62% of the total information content derived from metabolite variances.

## RESEARCH ARTICLES



**Figure 5. Metabolite impacts on clustering results. The contribution of individual polar metabolites to the PCA vector calculation is computed by linear combination. The closer to zero, the less influence of a metabolite on linear combination is found. Vector 1 predominantly separates plants from C24 and Col-2 genetic backgrounds, whereas vector 2 separates Col-2 ecotype plants from the corresponding *dgd1* mutants. Vector 4 contributed most to the separation of C24 wild-type plants from the *sdd1-1* mutants, although for reasons of clarity this vector is not shown in this figure. Examples of metabolite identity are numbered: 1 = isomaltose; 2 = U#106; 3 = proline; 4 = serine; 5 = threonine; 6 = pyroglutamic acid; 7 = glutamate; 8 = β-alanine; 9 = phenylalanine; 10 = U#107 (indole derivative); 11 = ascorbate; 12 = γ-hydroxybutyric acid lactone; 13 = U#73.**

The loss of activity of a single enzyme in the *dgd1* mutant resulted in a dramatic alteration in the metabolite composition: in comparison to the corresponding Col-2 wild-type plants, the levels of 153 out of 326 quantified metabolites were significantly different (Fig. 3A) in the *dgd1* mutant plants. The metabolic differences between Col-2 wild type and the *dgd1* mutant are quite complex and at present can only be partially explained. For example, some amino acids and citrate cycle intermediates are increased in the *dgd1* mutant, possibly indicating an increase in citrate cycle activity. Furthermore, indole-3-acetonitrile and several unidentified indole derivatives were increased in the mutant. Indole-3-acetonitrile is the precursor of the plant hormone indole-3-acetic acid (IAA), which itself did not reach detectable amounts by our profiling approach. The differences in IAA metabolism may reflect a hormone-controlled mechanism induced by the growth retardation of the *dgd1* mutant. Concomitant with the reduction in DGD lipid, the amount of the fatty acid 16:3 is decreased in the mutant, which can be explained by a change in the relative amounts of different forms of the substrate of the DGD synthase reaction, monogalactosyl diacylglycerol. The apparent reduction of galactose content in the *dgd1* mutant may reflect the downregulation of overall galactose biosynthesis as a response to the block in galactolipid biosynthesis. Furthermore, the concomitant reductions of inositol, galactinol, raffinose, and melibiose point toward a reduced flux through the biosynthesis of the carbohydrates of the galactinol family. It is obvious from its metabolic profile that a wide range of enzymes and pathways have been affected by the *dgd1* mutation. This analysis demonstrates the power of the metabolite profiling method to identify and quantify previously overlooked alterations, allowing a more comprehensive interpretation of the consequences of genetic modifications.

The second mutant that we tested, *sdd1-1*, is deficient in a subtilisin-like serine protease likely to be involved in the processing of a proteinaceous component of a signal transduction pathway controlling stomatal development. Apart from exhibiting increased stomatal density and stomatal cluster development, the *sdd1-1* mutant does not display any other obvious visible morphological alterations. However, it does have a slight retardation in seedling

establishment after sowing, which results in a three- to four-day delay in flowering under the conditions used. In contrast to the drastic alterations observed in the *dgd1* mutant, there were fewer variations in metabolite levels in the *sdd1-1* mutant when compared to the corresponding wild type. Significant differences were found in 41 metabolites (Fig. 3B), but only a few compounds were altered more than twofold. None of these changes are obviously linked to the elevated stomatal density in *sdd1-1*. Metabolite levels were expected to be increased for osmotically active components (as compensatory reactions to elevated transpiration) or carbohydrates (because of a raised net CO<sub>2</sub> uptake mediated by the enhanced stomatal density). Metabolite profiling, however, revealed neither a net increase in osmolytes nor in primary products of photosynthesis. The most dramatic difference between the *sdd1-1* mutant and the wild type occurred for two hydrophilic substances of unknown identity. The 13-fold reduction in substance U#29 and the concomitant 13-fold increase in U#73 may be indicative of a close metabolic relationship. The effects of the *sdd1-1* mutation on lipophilic metabolites are as difficult to understand as the alteration in polar metabolism. There is a significant change in leaf fatty acid composition: One of the most abundant fatty acids in *Arabidopsis*, 16:3, is decreased by more than fivefold in mutant plants, whereas 16:1 and 16:0 are increased and most of the other fatty acids remained unaffected. It has been shown recently by means of gene silencing that decreases in 16:3 levels lead to an improved thermo tolerance in transgenic plants<sup>18</sup> by modification of membrane function. Therefore, it is an interesting finding that the single-loci mutations tested here, *sdd1-1* and *dgd1*, also lead to a decrease of 16:3.

In both mutants alterations in many metabolites were observed. As stated above, half of the scored metabolites are of unknown structure. Because metabolite profiling reveals in cases such as *sdd1-1* that the most dramatic changes occur in unknown metabolites, further analyses, including structure elucidation, can be focused on a small number of compounds. In addition, new plant metabolites from unknown pathways can be detected by a non-target profiling approach. Triethanolamine is not commonly known as a plant endogenous metabolite in standard biochemical pathways<sup>19,20</sup> but is of widespread use as an organic solvent. The fact that significantly decreased levels of triethanolamine were observed in *dgd1* plants compared to Col-2 wild-type plants strongly argues against it being a contaminant, and rather suggests that it is produced by the plant biosynthetic machinery.

**Principal component analysis reveals four clusters.** Data interpretation of mean metabolite levels is difficult not only because biochemical pathways are linked and highly regulated but also because information gets lost in the process of averaging. Each individual plant represents a unique biological system; thus, it is to be expected that metabolite correlations to gene functions will be more clearly distinguished by multivariate data mining techniques<sup>21</sup>. Data mining tools reduce data complexity by focusing on the information content of a given data set. Two methods were applied: hierarchical component analysis (HCA) and principal component analysis (PCA)<sup>22</sup>. Both methods use all metabolite data from a plant sample to compute an individual metabolic profile and simultaneously compare this profile with all other plant metabolic profiles. As a first example, calculation of pattern recognition was based on the metabolic profiles of the polar compounds. In HCA, this pattern recognition is performed by calculation of Euclidean distances resulting in groups of samples (clusters) that show multivariate similarity. By examining the corresponding HCA dendrogram we found two main clusters for each of the *Arabidopsis* ecotypes. Each of these clusters was further divided into two subclusters corresponding to wild-type and mutant plants (data not shown). This genotype clustering was confirmed by PCA pattern recognition, which in some ways is an even more useful approach for the identification of gene function from metabolic profiles. By an *n*-dimensional

vector approach, PCA finds those basic vectors (eigenvectors) that give best overall sample separation. On the basis of total variances, vectors are determined by linear combination of all metabolite data. The resulting vectors are ordered by decreasing amount of total variance resulting in a minimum of loss of information content when data are visualized. Each sample can then be represented in a two- or three-dimensional space spanned by these vectors. When all samples of a genotype accumulate in the same cluster, this cluster can be regarded as a specific "metabolic phenotype." After application of PCA algorithms to the Col-2 / *dgd1* / C24 / *sdd1-1* experimental data set of polar compounds, four different clusters were found that are identical with the four plant genotypes (Fig. 4). For visualization, vector 4 was chosen instead of vector 3, which had nearly the same information content but was less powerful in separating C24 WT from *sdd1-1* samples. Plants with the Col-2 genetic background were quite dissimilar from C24 plants, whereas the difference between the two wild types and their corresponding mutant metabolic phenotypes were smaller or even partially overlapping (C24 WT / *sdd1-1*). This finding corresponds well to the results obtained from Student's *t*-tests of individual metabolites, where metabolite differences were both more abundant and more extreme for the Col2 WT / *dgd1* samples when compared with the C24 WT / *sdd1-1* samples.

Furthermore, PCA data can be used to analyze which metabolites exert the largest influence on the basic vector calculation (Fig. 5). For example, for computing the most powerful PCA vectors 1 and 2, many metabolites had values near zero, indicating that only minor variances were observed. However, some metabolites such as isomaltose, unknown #106, serine, threonine,  $\beta$ -alanine, and the unknown indole derivative #107 had a comparatively strong impact on the calculation of PCA vector 2, which separated predominantly Col-2 WT from *dgd1* plants. These compounds also demonstrated  $p < 0.01$  in the *t*-test comparison. Additionally, PCA vector 2 was strongly influenced by metabolites that were not significantly different in *t*-tests, either because these metabolites did not match the *t*-test threshold (pyroglutamic acid,  $p = 0.015$ ; phenylalanine,  $p = 0.048$ ; glutamate,  $p = 0.022$ ) or because they were not detectable in one of the two genotypes being compared, which causes *t*-tests to fail (proline, ascorbate). Analysis of PCA vector loading supplies information for the interpretation of metabolic profiles that extends the results obtainable by classical *t*-tests. For ease of visualization, vector 4 was left out in this presentation.

The ability to assign plant samples to groups using PCA of metabolic profiles offers an exciting perspective for plant functional genomics. On one hand, such groups are likely to be defined predominantly by different genotypes, and on the other hand, the use of PCA enables the defining elements of metabolic profiles to be distinguished. Furthermore, with metabolic analysis, response of metabolic networks to changes in single-gene loci is demonstrably complex, indicating how important it will be to have good methodologies in functional genomics that are capable of distinguishing cause from effect. Metabolite profiling is a valuable additional tool in the plant functional genomics repertoire and is worthy of wide application within and beyond the plant kingdom.

## Experimental protocol

*Arabidopsis* plants were grown on GS 90 standard soil in growth chambers in a 16 h light / 8 h dark photoperiod, changing from 60% humidity and 20°C during the day to 75% humidity and 18°C at night. Light intensity was fixed to 120  $\mu\text{mol}/\text{m}^2/\text{s}$ . After approximately 8 h of the photoperiod, 300 mg fresh weight rosette leaves were harvested randomly from trays that had alternate lines of pots containing wild-type and transgenic plants ( $n = 43$  (Col-2 WT), 45 (*dgd1*), 35 (C24 WT), and 28 (*sdd1-1*)). Extraction and fractionation was performed as reported recently<sup>8</sup>. Lipids were transmethylated by adding

900  $\mu\text{l}$  chloroform and 1 ml methanol including 3% (vol/vol) sulfuric acid at 100°C for 4 h. Sulfuric acid was removed using three 4 ml portions of water. The lipophilic phase was dried over anhydrous sodium sulfate and carefully concentrated to about 80  $\mu\text{l}$ . Before analysis, 20  $\mu\text{l}$  of pyridine plus 20  $\mu\text{l}$  of N-methyl-N-trimethylsilyl-trifluoroacetamide were added.  $^{13}\text{C}_{12}$ -Sucrose,  $^{13}\text{C}_6$ -glucose, d<sub>8</sub>-glycerol, d<sub>4</sub>-ethanolamine, d<sub>6</sub>-ethylene glycol, d<sub>3</sub>-aspartate,  $^{13}\text{C}_5$ -glutamate, d<sub>4</sub>-alanine, d<sub>8</sub>-valine, d<sub>3</sub>-leucine, and d<sub>5</sub>-benzoic acid were obtained from Campro Scientific (Emmerich, Germany) and used for exact quantification. GC/MS was performed using a GC 8000/Voyager mass spectrometer system (ThermoQuest, Manchester, UK). Peak finding and quantification of selective ion traces was accomplished using the instrument's MassLab FindTarget software. PCA and HCA pattern recognition was performed using the Pirouette software (Infometrix, Woodinville, WA) with  $\log_{10}$  data transformation and mean-center preprocessing. Principal component analysis was performed with cross-validation. Hierarchical component analysis was performed using Euclidean distances with complete linkages.

## Acknowledgments

This project was funded by the Max-Planck-Society. We thank Frank Kose, Una Griebel, and Antje Feller for their support in carrying out laboratory and computer work, Urte Schlüter for providing C24 WT and *sdd1-1* mutant plants, and Megan McKenzie for revising the manuscript.

1. Somerville, C. & Somerville, S. Plant functional genomics. *Science* **285**, 380–383 (1999).
2. Baldwin, D., Crane, V. & Rice, D. A comparison of gel-based, nylon filter and microarray techniques to detect differential RNA expression in plants. *Curr. Opin. Plant Biol.* **2**, 96–103 (1999).
3. Santoni, V. et al. Use of a proteome strategy for tagging proteins present at the plasma membrane. *Plant J.* **16**, 633–641 (1998).
4. Trethewey, R.N., Krotzky, A.J. & Willmitzer, L. Metabolic profiling: a Rosetta stone for genomics? *Curr. Opin. Plant Biol.* **2**, 83–85 (1999).
5. Katona, Z.F., Sass, P. & Molnár-Perl, I. Simultaneous determination of sugars, sugar alcohols, acids and amino acids in apricots by gas chromatography-mass spectrometry. *J. Chromatogr. A* **847**, 91–102 (1999).
6. Adams, M.A., Chen, Z.L., Landman, P. & Colmer, T.D. Simultaneous determination by capillary gas chromatography of organic acids, sugars, and sugar alcohols in plant tissue extracts as their trimethylsilyl derivatives. *Anal. Biochem.* **266**, 77–84 (1999).
7. Duez, P., Kumps, A. & Mardens, Y. GC-MS profiling of urinary organic acids evaluated as a quantitative method. *Clin. Chem.* **42**, 1609–1615 (1996).
8. Fiehn, O., Kopka, J., Trethewey, R.N. & Willmitzer, L. Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Anal. Chem.* **72**, 3573–3580 (2000).
9. McLafferty, F.W., Stauffer, D.A., Loh, S.Y. & Wesdemiotis, C. Unknown identification using reference mass spectra. Quality evaluation of databases *J. Am. Soc. Mass Spectrom.* **10**, 1229–1240 (1999).
10. Metabolite profiling (Max Planck Institute of Molecular Plant Physiology). <http://www.mpimp-golm.mpg.de/fiehn/index-e.html>.
11. Gygi, S.P. et al. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags *Nat. Biotechnol.* **17**, 994–999 (1999).
12. Vingron, M. & Hoheisel, J. Computation aspects of expression data *J. Mol. Med.* **77**, 3–7 (1999).
13. Dörmann, P., Hoffmann-Benning, S., Balbo, I. & Benning, C. Isolation and characterization of an *Arabidopsis* mutant deficient in the thylakoid lipid digalactosyl diacylglycerol. *Plant Cell* **7**, 1801–1810 (1995).
14. Härtel, H., Lokstein, H., Dörmann, P., Trethewey, R.N. & Benning, C. Photosynthetic light utilization and xanthophyll cycle activity in the galactolipid deficient *dgd1* mutant of *Arabidopsis thaliana*. *Plant Physiol. Biochem.* **36**, 407–417 (1998).
15. Dörmann, P., Balbo, I. & Benning, C. *Arabidopsis* galactolipid biosynthesis and lipid trafficking mediated by DGD1 *Science* **284**, 2181–2184 (1999).
16. Berger, D. & Altmann, T. A subtilisin-like serine protease involved in the regulation of stomatal density and distribution in *Arabidopsis thaliana*. *Gene Dev.* **14**, 1119–1131 (2000).
17. Mead, R., Curnow, R.N. & Hasted, A.M. (eds). *Statistical methods in agriculture and experimental biology*, Edn. 2. (Chapman & Hall, London, 1993).
18. Murakami, Y., Tsuyama, M., Kobayashi, Y., Kodama, H. & Iba, K. Trienoic fatty acid and plant tolerance of high temperature. *Science* **287**, 476–479 (2000).
19. GenomeNet database. (Institute for Chemical Research, Kyoto University, Japan). <http://www.genome.ad.jp/>
20. What Is There? Interactive metabolic reconstruction on the WEB. (Argonne Computational Biology Group, Chicago, IL). <http://wit.mcs.anl.gov/WIT2/>
21. Zweigert, G. Knowledge discovery in gene-expression-microarray data: mining the information output of the genome. *Trends Biotechnol.* **17**, 429–436 (1999).
22. Jurs, P.C. Pattern recognition used to investigate multivariate data in analytical chemistry. *Science* **232**, 1219 (1986).