

# Characterization of the natural variation in *Arabidopsis thaliana* metabolome by the analysis of metabolic distance

Benyamin Houshyani · Patrick Kabouw ·  
Dorota Muth · Ric C. H. de Vos · Raoul J. Bino ·  
Harro J. Bouwmeester

Received: 8 August 2011 / Accepted: 7 October 2011  
© The Author(s) 2011. This article is published with open access at Springerlink.com

**Abstract** Metabolite fingerprinting is widely used to unravel the chemical characteristics of biological samples. Multivariate data analysis and other statistical tools are subsequently used to analyze and visualize the plasticity of the metabolome and/or the relationship between those samples. However, there are limitations to these approaches for example because of the multi-dimensionality of the data that makes interpretation of the data obtained from untargeted analysis almost impossible for an average human being. These limitations make the biological information that is of prime importance in untargeted studies be partially exploited. Even in the case of full exploitation, current methods for relationship elucidation

focus mainly on between groups variation and differences. Therefore, a measure that is capable of exploiting both between- and within-group biological variation would be of great value. Here, we examined the natural variation in the metabolome of nine *Arabidopsis thaliana* accessions grown under various environmental conditions and established a measure for the metabolic distance between accessions and across environments. This data analysis approach shows that there is just a minor correlation between genetic and metabolic diversity of the nine accessions. On the other hand, it delivers so far in *Arabidopsis* unexplored chemical information and is shown to be biologically relevant for resistance studies.

**Electronic supplementary material** The online version of this article (doi:10.1007/s11306-011-0375-3) contains supplementary material, which is available to authorized users.

B. Houshyani · R. J. Bino · H. J. Bouwmeester (✉)  
Plant Sciences Group, Laboratory of Plant Physiology,  
Wageningen University, P.O. Box 658, 6700 AR Wageningen,  
The Netherlands  
e-mail: harro.bouwmeester@wur.nl

P. Kabouw  
Terrestrial Ecology, NIOO-KNAW, Droevendaalsesteeg 10,  
6708 PB Wageningen, The Netherlands

D. Muth · R. C. H. de Vos  
Plant Research International, Droevendaalsesteeg 1,  
6708 PB, Wageningen, The Netherlands

D. Muth · R. C. H. de Vos  
Centre for Biosystems Genomics, P.O. Box 98,  
6700 AB Wageningen, The Netherlands

D. Muth · R. C. H. de Vos  
Netherlands Metabolomics Centre, Einsteinweg 55,  
2333 CC Leiden, The Netherlands

**Keywords** *Arabidopsis* · Metabolomics · Metabolite fingerprinting · Metabolic distance · Natural variation · Metabolome diversity and metabolome plasticity

## 1 Introduction

Metabolites in living organisms together constitute the metabolome, which gives specific eco-physiological properties to the organism enabling it to interact with its kin and other species in the ecosystem. The signaling between plants (Bouwmeester et al. 2003, 2007; Belz 2007), and other living organisms such as predators and pollinators (Raguso and Pichersky 1999; Hilker and Meiners 2006; Cheng et al. 2007; Arimura et al. 2005) and defense against biotic agents (Treutter 2005; Rullmann et al. 2002; Kliebenstein et al. 2005; Ehrlich and Raven 1964; Cona et al. 2006; Chong et al. 2009) are among the interactions in which metabolites play a pivotal role. The metabolome of an organism, however, is not a stable entity. Many different sources of variation including genetic, and the biotic

and abiotic environment shape the metabolome resulting in a phenomenon that is referred to as metabolome plasticity. A number of approaches have been introduced to study this plasticity (Schripsema 2010). Metabolite fingerprinting or profiling, which is the unbiased global scanning of the metabolome, is being widely used in metabolomics research to unravel the metabolite composition (or metabolite “fingerprint”) of biological samples. Multivariate analysis of these metabolite fingerprints is subsequently used to reduce the dimensionality of the fingerprint data to a number of components that explain the maximum variation, i.e. principal components (PCs). This is followed by visualization of the data and drawing conclusions based on the clustering of samples (Kashif et al. 2009; Garcia-Perez et al. 2008; Aliferis and Jabaji 2009).

Due to limitations in human imagination and visualization power, the conclusions about clustering are usually based on just two or rarely three PCs. Hence, although some of the additional PCs may contain information about relevant biological variation and are thus important for understanding the metabolic relationships between samples, they are usually excluded from the analysis. Even in the case of clear visual separation between clusters of samples along all PCs, there are no methods for the quantification of the distance between these clusters making quantitative comparisons between groups of samples impossible.

The lack of a measure to describe the distance between clusters can be partially circumvented by application of principal component analysis (PCA) and hierarchical cluster analysis (HCA) simultaneously. However, the experimental design (such as a complete block design), the dimensionality of the data (in case of multivariate data), the relationship between and the ratio of inter- and intra-group variation, different levels of resolution to define the clusters and the method used for the clustering (Almeida et al. 2007) make HCA dendrograms not always compatible with the PCA clustering or the treatment structure. This is because HCA not only takes into account the biological sources of variation (such as genotype and a treatment effect) but also non-biological or undesired sources of variation (such as technical variation and block effects). The non-biological sources of variation may hinder the dendrogram calculation if they have a comparable or greater influence on the variation than the biological sources. Some researchers filter their data statistically and select those data points with significant difference among predefined groups of samples (Boccard et al. 2007). Others use an arithmetic mean analysis and make dendrograms with a representative of each predefined group (Kim et al. 2009). As a consequence of these filtering techniques, biological information and/or the insight in the possible causes of variation may be lost. Therefore, a measure for the distance between clusters of samples that incorporates both between- and within-cluster biological variation would be of great value.

For example, such a measure can be applied to determine the metabolic distance between genotypes, including genetically engineered organisms and their wild type relatives. The metabolic distance can also be used for correlation analysis between genetic and metabolic diversity.

In the present study, we examined the natural variation and plasticity in the metabolome of nine *A. thaliana* accessions in response to four different growing conditions. The objectives were: (1) to show the potential of metabolite fingerprinting and multivariate data analysis to characterize the effect of more than one source of variation on the diversity and plasticity of the metabolome, (2) to show the potential of metabolite fingerprinting and multivariate data analysis to establish the metabolic distance between accessions and different environmental conditions, (3) to estimate the correlation between the genetic and metabolic diversity of the nine accessions. Untargeted metabolite fingerprinting using three types of analytical platforms was employed to produce fingerprints of a wide range of metabolites in the nine accessions. A number of statistical methods were applied subsequently to the fingerprint data. Metabolites that contribute to the differences between the most diverged accessions were tentatively identified and the biological relevance of the observed differences in metabolic profiles of accessions assessed using a number of bioassays with biotic agents.

## 2 Materials and methods

### 2.1 Plant material

Nine accessions of *A. thaliana* (supplementary information Table 1) were selected, based on habitat geographical distribution and variation in volatile headspace profile (Snoeren et al. 2010). Accessions were sown in four environments: on soil in a climate chamber (CC), a controlled-conditions greenhouse (GH), an uncontrolled-conditions greenhouse (UC) and on hydroponics in the climate chamber (HY). Supplementary information Table 2 lists the environmental conditions.

Seeds were sown in pot soil (heated to 60 °C overnight before use; Lentse potgrond BV, Lent, The Netherlands) and placed in a climate chamber. Seedlings at stage 1.02 (Boyes et al. 2001), with two rosette leaves >1 mm, were transplanted to plastic containers (12 cm diameter, four seedlings per container) filled with the same soil. Containers were distributed randomly on shelves in CC, GH and UC and watered twice a week. For HY, seeds were sown on rock wool units fixed on a floating structure on the hydroponic solution (Tocquin et al. 2003). The rock wool absorbed the required water and nutrition for germination and growth. The hydroponics solution was refreshed weekly and aerated continuously.

Stage 3.70–3.90 plants (Boyes et al. 2001) with 70–100% rosette formation were cut from the surface of the soil or rock wool. Roots were cut from the hypocotyl at the rock wool subsurface. Six biological replicates of shoot material were used for each accession in each environment. Each shoot biological replicate consisted of a pool of four plants that had been growing in the same pot in CC, GH and UC or were selected randomly in HY. For hydroponically grown roots, four biological replicates were used that consisted of the roots of six randomly pooled plants.

All biological replicates were flash frozen in liquid nitrogen, lyophilized for 72 h, homogenized in a steel jar containing two steel balls shaken at  $20\text{ s}^{-1}$  by a MM300 mixer mill (Retsch) for 45 s at  $21^{\circ}\text{C}$  and ambient humidity. Samples were stored dry at  $4^{\circ}\text{C}$  until extraction for chemical analysis.

## 2.2 Extract preparation and LC-TOF-MS analysis

The protocol of Keurentjes et al. (2006) was followed for extraction of semi-polar metabolites with some modifications. Fifty milligrams of shoot or 12.5 mg of root material, both lyophilized and homogenized, were mixed with 2 ml (for shoots) or 0.5 ml (for roots) of ice-cold 75% methanol acidified with 0.1% (v/v) formic acid. After vortexing for 5 s, sonication for 15 min and centrifugation (2,500 rpm) for 10 min, the extracts were filtered through syringe filters (Minisart SRP 4,  $0.45\text{ }\mu\text{m}$ , Sartorius Stedim Biotech) and collected in glass vials. The filtered extract ( $150\text{ }\mu\text{l}$ ) was transferred to a glass insert ( $300\text{ }\mu\text{l}$ ) in a screw neck glass vial (1.5 ml) and then analyzed. In both positive and negative mode analyses, shoot samples were grouped in three sample blocks each containing two randomly selected biological replicates of each accession in each environment. A mixed sample of the nine accessions was passed through the same extraction procedures and used as technical replicates. They were analyzed at the beginning, the end and as every 15th sample in the injection sequence.

Liquid chromatography was performed on a Waters Acquity Ultra Performance Liquid Chromatography system (Waters, Milford, MA, USA). Five microlitre of extract was injected automatically on an Acquity UPLC BEH  $\text{C}_{18}$  column ( $150 \times 2.1\text{ mm}$  i.d.,  $1.7\text{ }\mu\text{m}$  particle size) (Waters), held at  $50^{\circ}\text{C}$  with a mobile phase flow of  $0.4\text{ ml min}^{-1}$ . The mobile phase consisted of water and acetonitrile containing 20 mM formic acid. The gradient applied started at 100% water for 0.5 min and subsequently changed to 10% acidified acetonitrile in 1 min, then rose linearly to 25% in 4 min, 65% in 3.5 min and 95% in 5 min, which was held for 6 min. Before the next run the column was equilibrated with starting conditions for 3 min.

Compounds eluting from the column were detected by a Waters LCT Premier TOF MS (Waters, Milford, MA, USA) equipped with a Z-spray interface and an electrospray ionization (ESI) source. The analysis was performed in both negative and positive ion modes in the range of  $m/z$  80–1,000 in separate runs, using a scan time of 200 ms. The parameters of the source were: desolvation gas temperature of  $400^{\circ}\text{C}$ , nitrogen gas flow of  $500\text{ l h}^{-1}$ , capillary spray voltage of 2.5 keV, source temperature of  $120^{\circ}\text{C}$ , cone voltage of 50 eV, nitrogen gas flow of  $50\text{ l h}^{-1}$ , and aperture 1 voltage of 8 eV. The mass spectrometer was calibrated with 5 mM sodium formate in iso-propanol/water (9:1). A  $1\text{ }\mu\text{g ml}^{-1}$  leucine enkephalin solution in acetonitrile/water (1:1) containing 0.1% formic acid, infused at a flow rate of  $0.02\text{ ml min}^{-1}$ , was used as a lock mass to continuously recalibrate the mass accuracy in both electrospray modes. The sampling rate of the lock mass solution was 0.4 s every 2 s. MassLynx software version 4.1 (Waters) was used to control the instruments and for data analysis.

## 2.3 Extract preparation and GC-TOF-MS analysis

Ten mg of shoot or 5 mg of root material, both lyophilized and homogenized, were weighed for extraction of polar metabolites. The instrument and protocol described in Fu et al. (2009) were used for extraction, derivatization and data acquisition by GC-TOF-MS with minor changes. Shoot samples were grouped in three sample blocks, each containing two randomly selected biological replicates of each accession in each environment. Derivatized extracts ( $25\text{ }\mu\text{l}$ ) were injected ( $2\text{ }\mu\text{l}$ ) with an Optic3 injector (ATAS) at  $70^{\circ}\text{C}$  with a gradient of  $6^{\circ}\text{C s}^{-1}$  to  $240^{\circ}\text{C}$ . A split flow of 10 (1 ml:11 ml) was used for shoot or 5 (1 ml:6 ml) for root material with a column flow of  $2\text{ ml min}^{-1}$  in a GC6890 N gas chromatograph (Agilent Technologies) on a ZB-50 capillary column ( $30\text{ m} \times 0.32\text{ mm}$  i.d.,  $0.25\text{ }\mu\text{m}$  DF; Phenomenex). The column temperature was  $70^{\circ}\text{C}$  for 2 min with a gradient of  $10^{\circ}\text{C min}^{-1}$  to  $310^{\circ}\text{C}$  and a final time of 3 min. The GC was coupled to a Pegasus III time-of-flight mass spectrometer (LECO) and compounds were detected at a scanning rate of 20 spectra per second ( $m/z$  50–600).

## 2.4 Data processing and analysis

The data of all analytical platforms were processed using MetAlign software (Lommen 2009) for peak detection and alignment of the data points. An in-house script called MetAlign Output Transformer (METOT; Plant Research International, Wageningen) was used for data filtration,

missing value replacement, and data quality and analytical technique reproducibility verification. The post-METOT data matrix was subjected to multivariate mass spectra reconstruction (MMSR) for data size reduction and putative compound mass spectrum reconstitution (Tikunov et al. 2005). MMSR relates thousands of ion fragments in a chromatogram to their parental metabolites by clustering them based on retention time and peak intensity pattern across samples into reconstructed metabolites. These mass clusters were used for further analyses and putative identification of metabolites.

For multivariate data analysis, the intensity values of reconstructed metabolites were normalized by the dry weight of the sample. Subsequently, metabolite intensities of each sample were  $^{10}\log$ -transformed and scaled by dividing by the standard deviation of the metabolite intensities of the corresponding sample. An integrated dataset was constructed by combining the shoot data of three analytical platforms. Values in the integrated dataset were derived in the same manner with an additional scaling by the standard deviation of samples for a reconstructed metabolite subsequent to dry weight normalization. The ordination diagrams in CANOCO (ter Braak 1988) (Biometris, Wageningen, NL) were used to visualize the variation in the sample profiles. Detrended correspondence analysis (DCA) was used to check the gradient length (L) of the explanatory variables (accession and environment) and accordingly choose between the linear ( $L < 4$ ) or unimodal ( $L > 4$ ) ordination techniques (Smilauer 2003). In addition to the first two ordines (or PCs), the third and fourth ordines were incorporated into the analysis if they individually explained more than 10% of the total variation and the technical replicates were grouped along those ordines on the scores plot. Partitioning of explanatory variable effects such as accession, environment and their interaction on the observed variation was performed and tested statistically ( $P$ -value  $< 0.05$ ) by Monte-Carlo permutation (MCP) test using the partial redundancy analysis (RDA) function of CANOCO (Smilauer 2003).

CANOCO scores and loading plots of shoot datasets were superimposed separately for three analytical methods which resulted in three biplots. Subsequently, 10 reconstructed metabolites that fitted more than 55% into the ordination space and showed to be accession-specific on the biplot were selected for further analysis.

### 2.5 Correlation between genetic and metabolic distance

A matrix of 149 genome-wide distributed SNP (single nucleotide polymorphism) markers from the Borevits lab (<http://borevitzlab.uchicago.edu>) was used to calculate the genetic distance between the nine accessions. The “Jukes

& Kantor” distance and complete linkage clustering were determined using TREECON v.1.3b (Van De Peer and De Wachter 1997).

To compute the inter-accession metabolic distances, the inter-sample Euclidean distances in an ordination diagram were examined by taking the sample scores on the selected ordines of the PCA scores plots (Kabouw et al. 2009). The inter-sample Euclidean distance matrices were computed for all platforms. The resulting matrices were used in an ANOSIM (analysis of similarity) by the program PAST (Hammer et al. 2001) to calculate the R-values as a measure for the metabolic distance between accessions. The Pearson correlation coefficient between two matrices ( $r$ ) was determined by a Mantel test (10,000 permutations).

### 2.6 In Silico Identification of Reconstructed Metabolites

Putative identification of selected metabolites from the shoot LC–MS datasets was done through the following steps: (1) Elimination of adduct ions; From all  $m/z$  ratios reconstructed in the centrotypes only mono-isotopic signals were selected and the respective neutral mass of the molecule was calculated using the mass spectrometry adduct calculator (<http://fiehnlab.ucdavis.edu/staff/kind/Metabolomics/MS-Adduct-Calculator/>). (2) Molecular formula assignment; Putative molecular formulas for an accurate mass were predicted by the elemental composition tool of MassLynx (Waters) with a 5 ppm tolerance. (3) Molecular formula screening; The Seven Golden Rules software ([http://fiehnlab.ucdavis.edu/projects/Seven\\_Golden\\_Rules/Software/](http://fiehnlab.ucdavis.edu/projects/Seven_Golden_Rules/Software/)) was used for heuristic filtering of the obtained molecular formulas (Kind and Fiehn 2007). Remaining possible molecular formulas were scored by the software according to their isotopic abundance error. (4) Molecular formula ranking; The five highest ranking molecular formula were prioritized based on prior identification in *A. thaliana*, Brassicaceae species or other plant species, presence in the Dictionary of Natural Products (<http://dnp.chemnetbase.com>) and their score.

The Golm Metabolome Database (GMD@CSB.DB MSRI, [http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/msri/gmd\\_msri.html](http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/msri/gmd_msri.html)), NIST library and an in-house mass spectral database for GC-TOF–MS were used to putatively identify reconstructed metabolites from GC-TOF–MS analysis, using NIST MS Search v.2.0. Both mass spectra and retention indices of the reconstructed metabolites were used to search for putative candidate metabolites already reported in *A. thaliana*. Matching factor (MF) and reverse matching factor (RMF) (Davies 1998) were exploited to select the best matching metabolite identity.



## 2.7 Metabolite identification with HPLC–PDA–QTOF–MS/MS system

Metabolites detected by LC–TOF–MS were further annotated using an HPLC–PDA–QTOF–MS/MS system from Waters (Moco et al. 2006). Ten microlitre of the extracts were automatically injected on an HPLC Luna C18 analytical column (150 × 2.0, 3 mm) (Phenomenex). The chromatographic phases composition were water/formic acid (0.1% v/v) (A) and acetonitrile/formic acid (0.1% v/v) (B). The separation was performed at 40°C with a flow of 0.19 ml min<sup>-1</sup> in a gradient starting with 5% of B which linearly increased to 75% in 45 min, then up to 90% in 2 min and continued isocratic for 5 min. The column was then equilibrated for 16 min under starting conditions.

The HPLC was linked to a PDA detector (Waters 2996) and a QTOF Ultima mass spectrometer (Waters Corporation). The ionization source parameters were: capillary voltage 2.75 keV, cone voltage 35 eV, source temperature 120°C and desolvation temperature 250°C. Cone gas and desolvation gas flows were 50 and 600 l h<sup>-1</sup>, respectively. MS/MS measurements were made with 0.40 s of scan duration and 0.10 s of interscan delay with increasing collision energies according to the following program: 5, 10, 15, 30, 50 eV (ESI positive), or 10, 15, 25, 35, and 50 eV (ESI negative). Leucine enkephalin was used as a lock mass and was continuously sprayed into a second ESI source using an LKB Bromma 2150 HPLC pump, and sampled every 10 s.

## 2.8 Bioassays

Accessions An-1, Cvi, Eri and Col-0 were selected to conduct bioassays. Inoculation by the powdery mildew pathogen *Oidium neolycopersici* and assessment of infection was performed according to Bai et al. (2008). Inoculation by the downy mildew pathogen *Hyaloperonospora arabidopsidis* isolates Emoy2, Waco9 and Cala2 was done and infection assessed according to Van Damme et al. (2009). For *Botrytis cinerea* inoculation the protocol of Ferrari et al. (2003) was used with minor modifications. Plants were placed in darkness for 24 h after inoculation and subsequently kept at 9 h photoperiod. Scoring was done 3 d after inoculation by visual determination of the area of the lesions on the inoculated leaves.

Western flower thrips, *Frankliniella occidentalis* (Pergande), was reared according to De Vos et al. (2005). Sixteen thrips were transferred to each of twelve Petri dishes (9 cm) used as replicates. Each Petri dish contained two detached leaves of each accession on a 1.5% agar medium (Technical No.3). Every two leaves of the same accession were randomly distributed on four sides of the Petri dish keeping the same distance from each other. The

number of thrips present on each accession was counted at 0.5, 1, 2, 3, 4, 5, 6, 21, 22, 23 and 24 h. The average number of observed thrips on each accession in the last four hours was used to compare the thrips attraction of accessions by generalized linear model analysis using PASW statistics 17.

## 3 Results and discussion

### 3.1 The phylogenetic relationship between accessions

Complete linkage clustering using 149 genome-wide SNPs resolved the phylogenetic relationship between the 9 *A. thaliana* accessions of this study (supplementary information Fig. 1). The level of confidence was estimated using a bootstrap of 500 replications. A separation between An-1, C-24, Cvi and Kyo-1 versus WS, Ler, Kond, Col-0 and Eri occurred at a low stringency level and formed two major clades with relatively close genotypes (accessions). Accessions An-1, Eri, Col-0 and C-24 diverged earliest from the rest of the genotypes in the clades indicating a larger genetic distance between them and the rest of the clade members.

### 3.2 Metabolome analyses

Shoot samples from all environments and root samples from hydroponics were subjected to three different profiling methods: LC–TOF–MS in positive and negative mode, and GC–TOF–MS. The number of entities (masses or reconstructed metabolites) after each step of the data processing workflow with MetAlign, METOT and MMSR is given in supplementary information Table 3. Reproducibility of the analyses and data processing were verified by two approaches: The approach of Vorst et al. (2005) and a PCA approach. Supplementary information Fig. 2 illustrates the methods and graphs of the first approach using root data of LC–MS negative mode—after MetAlign and METOT preprocessing—as an example. The amplitude scatter plots of all mass peaks of two technical replicates were made and compared with the same plot of two biological replicates (supplementary information Fig. 2a, b as an example). The scatter plots showed that there is a close linear relationship between all signal intensities of any two technical replicates of an analytical technique (supplementary information Fig. 2a). For biological replicates there is also a linear relationship albeit with lower correlation (supplementary information Fig. 2b). The scatter plots also showed absence of hypo-alignment, as is concluded from the absence of satellite clouds around the diagonal axes of the aligned masses in the scatter plot. These clouds can be due to peaks of one technical replicate

being aligned with noise peaks in the other technical replicate, instead of with its corresponding mass peak (Vorst et al. 2005). Their absence indicates that masses were aligned efficiently and/or that misaligned masses were eliminated by METOT. supplementary information Fig. 2c shows that more than 80% of the ion masses are present in at least five of the six technical replicates. Around 80% of the ions that are detected in all six technical replicates show a variability in their measured mass of less than 5 ppm (supplementary information Fig. 2d) and the amplitude variation for more than 90% of the masses present in all six technical replicates was less than 20% (supplementary information Fig. 2e).

For the second approach, we visualized the variation in metabolite fingerprints of the technical and biological replicates by PCA (supplemental information Fig. 2f, as an example). The technical replicates clustered more closely together than the biological replicates in the examined PCs showing that there was relatively minor variation due to extraction and instrument artifacts compared with the biological variation.

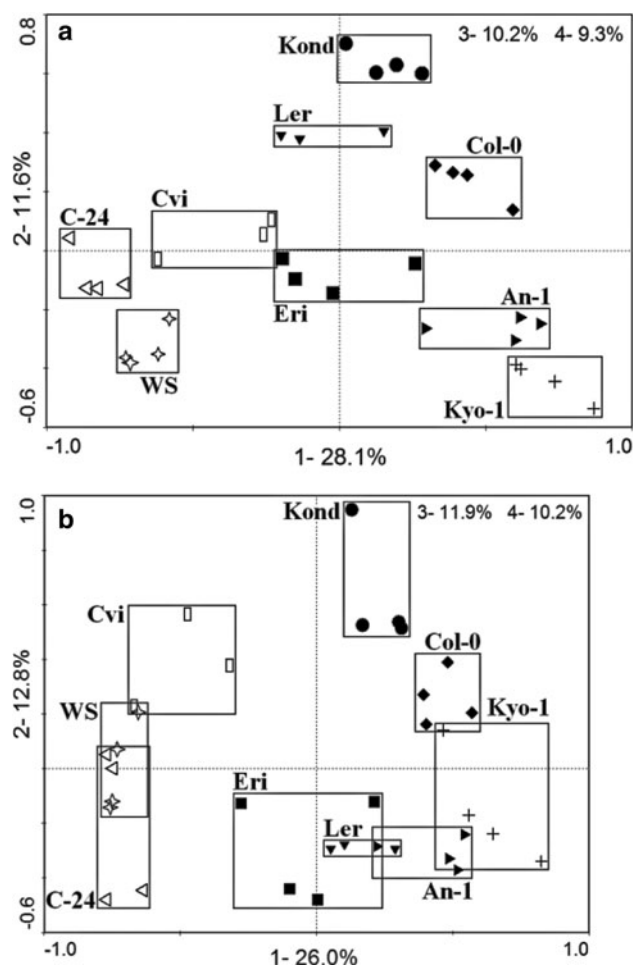
### 3.3 Visualization and quantification of the variation due to source effects

The gradient of the explanatory variables (accession and environment) was computed by DCA for all datasets. All of the gradients were short ( $L < 4$ ) suggesting a linear metabolite fingerprint response to the gradient of the underlying variation source(s). Accordingly, PCA was used for all data visualizations and analyses in which one or more independent gradients (PCs) represent predictors for fitting the regression model.

With the exception of the GC-TOF-MS data, all PCAs of shoot datasets showed separate clustering of HY samples from samples grown in the other environments along PC1 and PC2 (see supplementary information Fig. 2f as an example). Clustering of all hydroponically grown genotypes on one side of PC1 implies that there is a pronounced effect of this growth condition compared to the other sources of variation within the shoot LC-MS data. In other words, this environment has a stronger effect on metabolite composition than the other environments and its effect is stronger than the genotypic effect. Within the HY environment, however PCA showed clear clustering of the accessions along PC1 and PC2 in both LC-MS (Fig. 1 as an example) and GC-TOF-MS analyses (data not shown). In both root LC-MS modes, about 40% of the variation between accessions was explained by the first two PCs. In positive mode, all accessions were separated along PC1 and PC2 with C-24, Kyo-1 and Kond positioned at the extremes of the two dimensional PCA plot (Fig. 1a). In negative mode some accessions clustered together

suggesting a close metabolic relationship (An-1 and Kyo-1, Eri and Ler, and WS and C-24) (Fig. 1b). PC3 (10.2 and 11.9% for positive and negative mode, respectively) and PC4 (9.3 and 10.2%) explained a similar amount of variation as PC2 (11.6 and 12.8%). Therefore, PC3 and PC4 were also used for ANOSIM and metabolic distance calculation, but only if the technical replicates did not separate along these PCs as much as the biological replicates.

Multivariate statistical representation of metabolite fingerprints by using PCA as shown in Fig. 1 allows the visualization of the natural variation and plasticity in the metabolome. However, information about the magnitude of the variation caused by a known factor is difficult, if not impossible, to obtain in this way. Therefore, RDA was used to estimate the effect size of a single or multiple source(s) of variation such as accession or accession and



**Fig. 1** PCA scores plots of the root metabolite profile of nine accessions grown in hydroponics, analyzed by LC-MS in positive mode (a) and negative mode (b). Numbers along the axes indicate the PC number and percentage of explained variation. Boxes approximate the boundaries of within accession variation and illustrate clustering of samples belonging to the accession. open triangle An-1, filled diamond Col-0, open triangle C-24, open rectangle Cvi, filled square Eri, filled circle Kond, + Kyo-1, filled triangle

environment on the metabolome in a multi-factorial experiment. Furthermore, in partial RDA, one or more of the explanatory variables can be defined as cofactor to remove the associated effect from the solution of the ordination model. For the root dataset only a single source of variation (accession) was defined and the explained variation in each dataset was calculated accordingly (supplementary information Table 4). Accession explained a bigger portion of the variation in both LC–MS datasets (79.0 and 80.6% in positive and negative mode, respectively) compared with the GC–TOF–MS dataset (53.9%). For shoot datasets, first the effect of sample block was removed by partial RDA. This resulted in environment and accession explaining together 38.8, 35.2 and 63.0% of the total variation in the GC–TOF–MS, LC–MS positive mode and LC–MS negative mode datasets, respectively. Using sample block and environment or sample block and accession as cofactor group in RDA, the explained variation by accession and environment was calculated, respectively (supplementary information Table 4). In conclusion, environment had a larger influence than the accession on the LC–MS determined metabolic variation in the shoot. For the GC–TOF–MS platform, which mainly detects polar primary metabolites, the environment effect was smaller than the accession effect.

Likewise the environment-accession interaction was defined as an explanatory variable and sample block, environment and accession main effects were added to the cofactor group. Environment-accession interaction explained a minor part of the total variation in GC–TOF–MS and LC–MS positive mode datasets, respectively, while it had no significant effect on the variation in the LC–MS negative mode dataset (supplementary information Table 4). In all datasets, the environment caused a change in the position of samples (Fig. 2c as an example with LC–MS negative mode dataset). This change is an indication of the capacity of each accession to respond to perturbations in the environment (metabolome plasticity). However, plant species might differ in the degree of phenotypic plasticity when exposed to the same environmental change (De Jong 2005). The amount of variation explained by the accession-environment interaction can be a measure for the difference between accessions in their ability to respond to a perturbation. The fact that the accession-environment interaction explained only little metabolite variation suggests that the metabolic responses of the different *A. thaliana* accessions towards changes in the environment is quite similar.

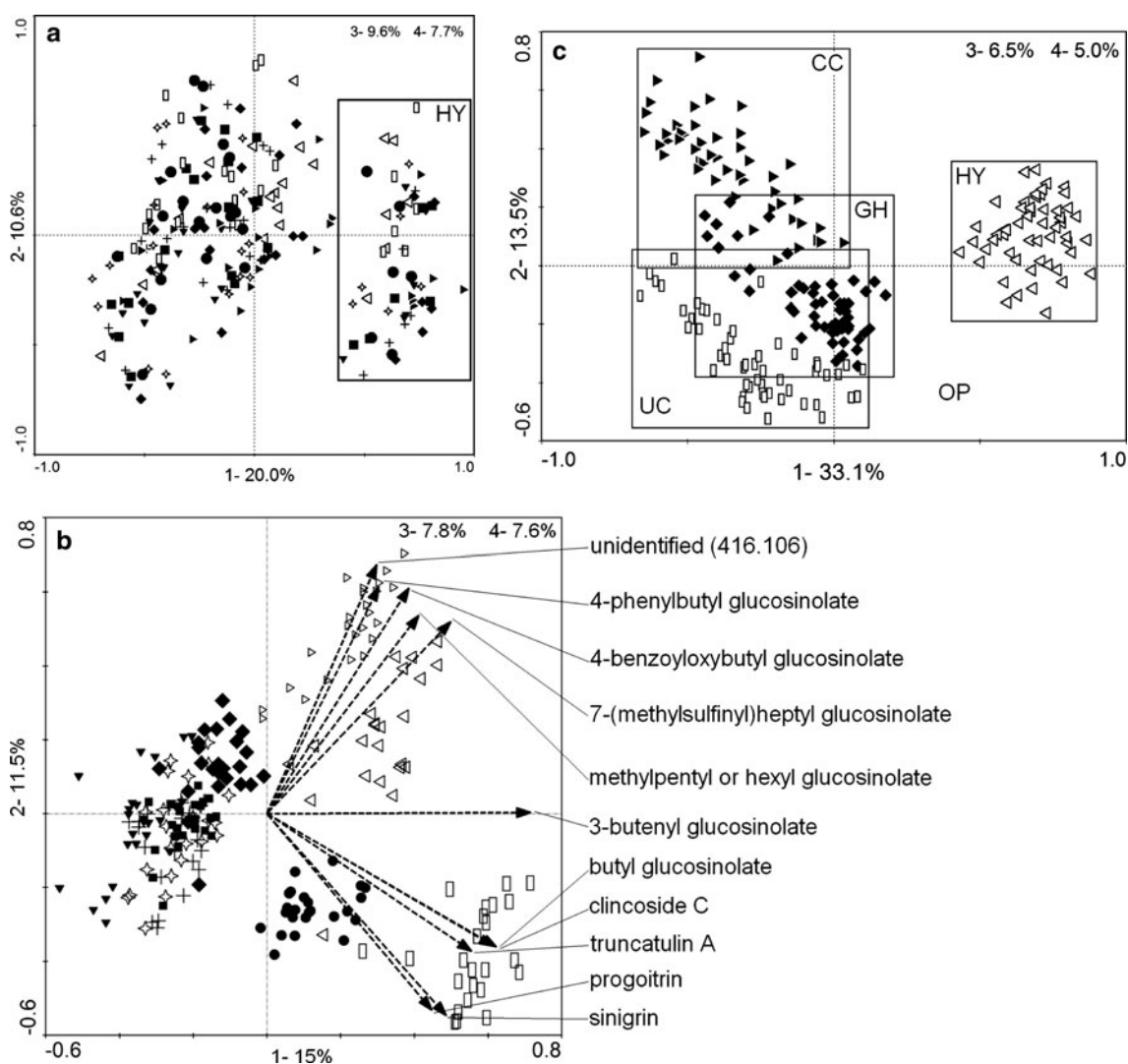
### 3.4 Metabolic distance between accessions

The PCA scores plots show that biological replicates of the same accession cluster together (Fig. 1). In many cases, however, the large within-accession variation (variation

between biological replicates; indicated by boxes in Fig. 1) or small between-accession metabolic variation caused merging of samples belonging to different accessions along one or more PCs. The merging of clusters happened along different PCs (also along PC3 and PC4, which are not shown in Fig. 1). To take the relative position of samples to each other along all PCs, the scores of the samples on these PCs were exploited as sample properties. A matrix containing all these scores was used to calculate the Euclidian distance between samples (inter-sample distance). Using the inter-sample distance matrix, an R-value was assigned by the ANOSIM permutation test. This R-value is a geometric function that describes the distance between pairs of a priori groups of samples belonging to two accessions in the hyperplane. R-values can range from +1 to –1, with +1 indicating the maximal divergence of two accessions and –1 indicating that one group is the child of the other. R = 0 occurs if the position of both accessions in the hyperplane completely overlaps. Table 1 shows the pair wise R-values for the root metabolite fingerprints obtained by LC–MS negative and positive mode analyses. Based on these R-values, for example accessions Kyo and An-1 had a very similar metabolome in both analytical methods ( $R_{\text{Positive mode}} = 0.07$  &  $R_{\text{Negative mode}} = 0.03$ ) (Table 1). Ler and Kond, on the other hand, were maximally different for the positive mode metabolic profile, whereas for the negative mode metabolic profile they were less different ( $R_{\text{Positive mode}} = 1.00$ ;  $R_{\text{Negative mode}} = 0.33$ ) (Table 1). On the contrary, Ler and An-1 were maximally different for the negative mode metabolic profile and less different for the positive mode metabolic profile ( $R_{\text{Positive mode}} = 0.48$ ;  $R_{\text{Negative mode}} = 1.00$ ) (Table 1).

The use of more than two PCs in the R-value calculation resulted in quite a different interpretation of the metabolic distance between some accessions compared with the two dimensional PCA plot. As an extreme example, the root samples of accessions C-24 and Col-0 clearly separated along the first PC in both LC–MS positive and negative mode (Fig. 1). However, the R-value of the negative mode analysis showed that these accessions are metabolically more similar in the hyperplane than what can be observed on a plain PCA plot ( $R_{\text{Negative mode}} = 0.08$ ) (Table 1). The distance is larger in positive mode (0.39) but still much smaller than for example the distance between C-24 and Cvi which in Fig. 1 are closer together than C-24 and Col-0.

Ranking of the average distance of one accession to all other accessions was used as a measure for its metabolic divergence from the rest of the accessions. Accessions Cvi and C-24, for example, were the most and the least diverging accessions, respectively, with regard to their root metabolite profiles obtained by both LC–MS modes ( $R_{\text{Positive mode average}} = 0.98$  &  $R_{\text{Negative mode average}} = 0.85$  for Cvi and  $R_{\text{Positive mode average}} = 0.45$  &



**Fig. 2** PCA plots of the nine accessions grown in four environments, analyzed by LC–MS of shoot in negative-mode. **a** scores plot. **b** partial PCA biplot (superimposed scores and loadings plots) with environment and sample block as cofactor. Dashed arrows represent the 11 metabolites that more than 55% of their influence was represented by the first two PCs. The accurate mass is given in parentheses for unidentified masses. **c** partial PCA scores plot with accession and sample block as cofactor. Boxes approximate the boundaries of within environment variation and illustrate clustering of

samples belonging to the environment. Numbers along the axes indicate the PC number and percentage of variation explained. Accessions in **a**, **b**: open triangle An-1, filled diamond Col-0, open triangle C-24, open rectangle Cvi, filled square Eri, filled circle Kond, + Kyo-1, filled triangle Ler, open diamond WS; Environments in **c**: CC climate chamber, GH controlled-conditions greenhouse, UC uncontrolled-conditions greenhouse, HY hydroponics in the climate chamber

$R_{\text{Negative mode average}} = 0.20$  for C-24) (Table 1). Average metabolic distance depended on the analytical method used as was exemplified by accessions An-1 and Kond that changed from second to eighth and eighth to second as a result of the analytical method. This indicates the importance of the use of several different analytical platforms for elucidation of metabolic relationships between accessions.

### 3.5 Metabolic distance in complex datasets

So far, we analyzed the metabolic distance for genetic differences only. It would be of interest to look also at the

metabolic distance between genotypes under different treatments. However, this implies the involvement of more than one source of variation and that makes such analyses complicated. Partial analysis, however, allows for reduction of the complexity by removing the effect of an undesired source of variation from the dataset. In Fig. 2, three PCA plots of the shoot LC–MS negative mode data are shown (216 samples). If none of the sources of variation (sample block, accession, environment) was excluded from the model, only separate clustering of hydroponically grown samples could be observed (Fig. 2a). Hence, the first PC explains only the considerable change in metabolite



**Table 1** R-values (metabolic distance) obtained by the ANOSIM permutation test using root LC–MS positive mode (values above the *diagonal*) and negative mode (values below the *diagonal*) data, based on the first four principal components of PCA

		LC–MS positive mode									
		An-1	Col-0	C-24	Cvi	Eri	Kond	Kyo-1	Ler	WS	Average+
LC–MS negative mode	An-1		0.34	0.20	1.00	0.44	0.69	0.07	0.48	0.55	0.47 (8)
	Col-0	0.73		0.39	1.00	0.79	0.99	0.51	0.98	0.76	0.72 (4)
	C-24	0.35	0.08		1.00	0.31	0.60	0.25	0.46	0.42	0.45 (9)
	Cvi	1.00	1.00	0.61		0.83	1.00	1.00	1.00	1.00	0.98 (1)
	Eri	0.94	0.35	0.13	1.00		1.00	0.70	0.37	1.00	0.68 (6)
	Kond	0.50	0.26	0.08	0.33	0.42		0.57	1.00	0.65	0.81 (2)
	Kyo-1	0.03	0.34	0.04	0.87	0.32	0.27		0.48	0.30	0.49 (7)
	Ler	1.00	0.57	-0.02	1.00	0.87	0.33	0.30		1.00	0.72 (3)
	WS	0.68	0.36	0.30	1.00	0.72	0.30	0.36	1.00		0.71 (5)
	Average-	0.65 (2)	0.46 (6)	0.20 (9)	0.85 (1)	0.59 (4)	0.31 (8)	0.32 (7)	0.63 (3)	0.59 (4)	

The numbers in parentheses indicate the ranking of the corresponding accession with regard to the average of its distances with the other accessions

fingerprint in the HY samples compared with the other environments.

Partial PCA of the LC–MS negative mode dataset with sample block and environment or sample block and accession as cofactor removed the cofactors' effect from the dataset and resulted in cluster formation based on the remaining sources of variation (Fig 2b, c). Hence, partial analysis reduced the complexity of the datasets in which more than one known variation source (accession and environment in this case) was responsible for the metabolic differences. Moreover, partial analysis enabled having more biological replicates for each accession or environment by combining the data from different environments or accessions, respectively, hence giving better insight in the within-accession or within-environment biological variation.

Subsequently, the scores of the samples on PC 1 to 4 in these partial PCA analyses were used to calculate the distance between a priori groups (accessions or environments) of samples in the hyperplane and successively the average distance as a measure for metabolic divergence (Table 2 and supplementary information Table 5). The total of the average R-values provides a measure for the overall metabolic distance of an accession or environment from the others across all analytical methods. Among all accessions and across all analytical methods, Cvi ( $R_{\text{total}} = 4.93$ ; ranking 1st) was the metabolically most diverged accession (Table 2). Accessions Kyo-1 ( $R_{\text{total}} = 2.41$ ; ranking 7th) and Eri ( $R_{\text{total}} = 2.63$ ; ranking 8th) had the lowest total R-value which makes them the least metabolically diverged accessions or in other words metabolically average accessions that in most of the analyses located close to the center of the hyperplane or in between most of the other accessions (Table 2; Fig. 2b).

Analogously among environments across all analytical methods, Hydroponics ( $R_{\text{total}} = 2.43$ ) was the most distant environment from the others with the largest total R-value (Supplementary Information Table 5) as was also concluded from Fig. 2c. The total R-value of greenhouse ( $R_{\text{total}} = 1.63$ ) samples was the smallest, indicating that this environment was the most “average” growth condition. (Supplementary Information Table 5 and Fig. 2c).

### 3.6 Correlation between metabolic distance matrices

The metabolic distance matrices (matrix of R-values) can be used also to analyze pattern similarity between analytical methods. A Mantel test showed a significant correlation between the metabolic distance matrices of shoot LC–MS positive and negative mode ( $r = 0.70$ ,  $P$ -value  $< 0.01$ ) and between root LC–MS positive and negative mode ( $r = 0.47$ ,  $P$ -value  $< 0.05$ ) (Table 3). Interestingly, there was also a significant correlation between the metabolic distance matrices of shoot LC–MS negative mode and GC-TOF–MS ( $r = 0.61$ ,  $P$ -value  $< 0.05$ ) (Table 3).

A significant correlation between inter-accession metabolic distance matrices for the analytical methods indicates that the metabolite profiles obtained with these methods give a similar picture of the relationship between the accessions. This correlation does not imply that there is similarity in the measured metabolites by those methods. The lack of a significant correlation between some analytical methods, on the other hand, indicates that metabolic distances can be also method dependent and that the analytical methods are complementary, rather than redundant, in covering the metabolome. This is confirmed by the fact that the ranking of accessions across uncorrelated

**Table 2** Average R-values (metabolic distance) of accessions in different analytical methods obtained by the ANOSIM permutation test on the inter sample distances

	An-1	Col-0	C-24	Cvi	Eri	Kond	Kyo-1	Ler	WS	PC <sup>a</sup>	Variance explained <sup>b</sup>
Root LC-MS <sup>+</sup>	0.47 (8)	0.72 (3)	0.45 (9)	0.98 (1)	0.68 (6)	0.81 (2)	0.49 (7)	0.72 (3)	0.71 (5)	4	0.593
Root LC-MS <sup>-</sup>	0.65 (2)	0.46 (6)	0.20 (9)	0.85 (1)	0.59 (4)	0.31 (8)	0.32 (7)	0.63 (3)	0.59 (4)	4	0.609
Root GC-MS	0.57 (5)	0.60 (4)	0.65 (3)	1.00 (1)	0.34 (8)	0.37 (7)	0.38 (6)	0.38 (6)	0.73 (2)	3	0.535
Shoot LC-MS <sup>+</sup>	0.38 (3)	0.29 (5)	0.53 (2)	0.58 (1)	0.25 (9)	0.27 (6)	0.32 (4)	0.25 (9)	0.25 (9)	3	0.239
Shoot LC-MS <sup>-</sup>	0.89 (2)	0.65 (5)	0.78 (4)	0.97 (1)	0.54 (8)	0.85 (3)	0.57 (7)	0.59 (6)	0.51 (9)	3	0.343
Shoot GC-MS	0.29 (6)	0.23 (9)	0.31 (4)	0.55 (1)	0.23 (9)	0.46 (2)	0.33 (3)	0.30 (5)	0.28 (7)	4	0.441
Total	3.25 (2)	2.95 (4)	2.92 (5)	4.93 (1)	2.63 (7)	3.07 (3)	2.41 (8)	2.87 (6)	3.07 (3)		

Numbers in parentheses show the ranking of the accession within the row

<sup>a</sup> The number of principle components (PC) used to calculate the Euclidean distances for ANOSIM

<sup>b</sup> % variation explained by the PCs used

**Table 3** Mantel statistics “r”, for the correlation between different datasets

	r	P-value
Root LC-MS <sup>+</sup> Vs Root LC-MS <sup>-</sup>	0.47	0.04*
Root LC-MS <sup>+</sup> Vs Root GC-MS	0.24	0.19
Root LC-MS-Vs. Root GC-MS	0.39	0.08
Shoots LC-MS <sup>+</sup> Vs. Shoot LC-MS <sup>-</sup>	0.70	0.004*
Shoots LC-MS <sup>+</sup> Vs. Shoot GC-MS	0.38	0.08
Shoots LC-MS-Vs. Shoot GC-MS	0.61	0.01*
Hyd. shoots LC-MS-Vs Hyd roots LC-MS <sup>-</sup>	0.18	0.25
Hyd. shoots LC-MS <sup>+</sup> Vs Hyd roots LC-MS <sup>+</sup>	0.29	0.08
Hyd. shoots GC-MS Vs Hyd roots GC-MS	0.39	0.10

P-value calculated by 10,000 permutations  
Hyd Hydroponics

metabolic distance matrices varies more than across correlated matrices (Table 2). This also implies the importance of using different analytical methods for metabolome characterization. Table 3 provides some guidelines for the choice of analytical methods that avoid generation of redundant analytical data for metabolic relationship dissection. A combination of LC-MS positive mode and GC-TOF-MS analysis of derivatized polar extracts would, at least for *A. thaliana* accessions of this study, give reasonable coverage of the metabolome while avoiding too much redundancy.

No significant correlation was observed between metabolic distance matrices of shoot and root, in all three analytical methods (Table 3). The lack of a similarity between metabolite profiles of shoot and root has been shown in a number of studies (Kabouw et al. 2009; Van Dam et al. 2009). However, here we show that there is also a lack of correlation between the matrices of metabolic distances of accessions based on root and shoot. This complements prior studies (Kabouw et al. 2009; Van Dam et al. 2009) and indicates that the metabolic relationship between accessions is dependent on the tissue under study.

### 3.7 Chemical relevance of metabolic distance

Key in estimating the metabolic distance between a priori groups of samples (accessions or environments) is the quantity of the analytical information and relevance of the calculated distance. The metabolite profiling using three analytical platforms was aimed at increasing the quantity of the information from different classes of metabolites. A metabolite identification strategy was followed to evaluate the chemical relevance of the calculated metabolic distance. Shoot metabolites responsible for separation of accessions were identified in silico as the first strategy. Hereto, we used partial PCA plots eliminating the effect of the environment from the dataset to find the metabolites responsible for grouping of the accessions. Then, the scores and loadings plots with the first two PCs were superimposed for shoot GC-MS and LC-MS positive and negative mode datasets separately. The biplots of shoot GC-MS data (supplementary information Fig. 3) and LC-MS negative mode data (Fig. 2b) are shown as examples. As a rule, only those metabolites were included into the ordination diagram that more than 55% of their influence was

represented by the first two PCs. Consequently, ten metabolites were pinpointed in the GC–MS data that contributed most to the clustering of accessions Cvi and Kond on one sides of the scores plot (supplementary information Fig. 3 and supplementary information Table 6). They were identified and annotated as monosaccharides (fructose, 1-methyl- $\alpha$ -D-glucopyranoside and glucopyranose), a disaccharide (sucrose), an amino acid (L-glutamic acid) and its derivative (pyroglutamic acid), and four other, yet unknown, compounds that we could not unambiguously identify. All of them have previously been detected in *A. thaliana* (Fiehn et al. 2000). ANOVA showed that there is a statistical significant difference between the accessions ( $\alpha = 0.05$ ) for these ten metabolites. The fact that these primary metabolites were among the most discriminating compounds suggests a fundamental difference in central carbon metabolism of Cvi and Kond compared with the other accessions, including Col-0. This difference is reflected in the divergence ranking of Cvi and Kond in the GC–MS shoot analysis (Table 2).

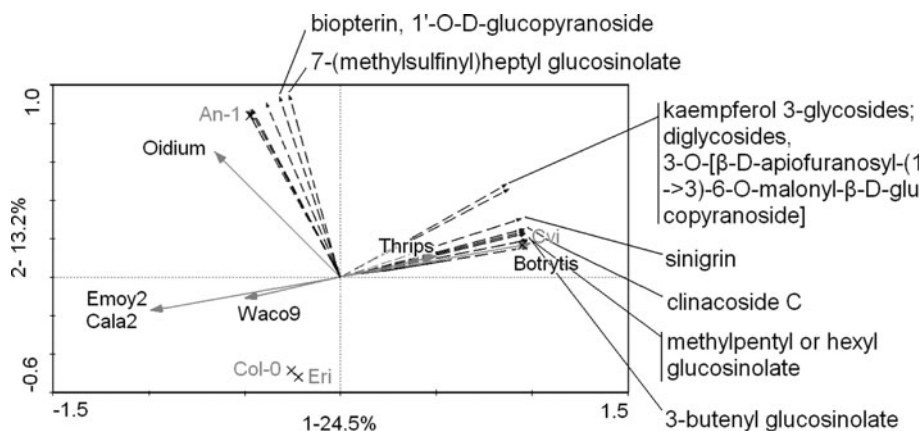
In the shoot LC–MS negative mode dataset, 11 metabolites were most responsible for the separate grouping of the accessions in the PCA biplot (Fig. 2b). The most likely elemental compositions of the corresponding parent ion mass of all these could be calculated by Seven Golden Rules (Kind and Fiehn 2007). Library search in the KnapSack database ([http://kanaya.naist.jp/knapsack\\_jsp/top.html](http://kanaya.naist.jp/knapsack_jsp/top.html)) and the Dictionary of Natural Products (<http://dnp.chemnetbase.com>) resulted in the annotation of 10 compounds of which eight were denoted as glucosinolates (supplementary information Table 7). Glucosinolate profiles of *A. thaliana* have indeed been reported to vary

across accessions (Kliebenstein et al. 2001). All putatively identified glucosinolates have been reported before in Col-0 (Matsuda et al. 2010) or other Brassicaceae genera (KnapSack). The remaining three compounds with parent ion masses of 789.222 (tentatively identified as truncatulin A from *medicago truncatula*) 338.093 (tentatively identified as clinacoside C) and 416.106 have so far not been reported in Brassicaceae or any plant species.

Analogously, in the shoot LC–MS positive mode dataset, ten metabolites responsible for separate grouping of An-1, Cvi and C-24 on the PCA biplot (not shown) were pin-pointed, of which nine could be tentatively identified. Only two of them, diptocarpilidine and 3-carboxytyrosine, have been reported before in at least one Brassicaceae species (supplementary information Table 7). The other *m/z* signals corresponded mainly to flavonoid derivatives and nitrogen containing compounds that have not been reported before in *A. thaliana*. All of the accession-specific metabolites in the LC–MS analyses were confirmed to be statistically different across accessions by ANOVA ( $\alpha = 0.05$ ).

### 3.8 MS/MS fragmentation of the most relevant metabolites

MS/MS experiments were performed to further identify the differential metabolites. Mass and retention time-directed spectra were registered at different collision energies (from 5 up to 50 eV) for each compound and combined into one MS/MS spectrum. Not all of the selected ion masses could be fragmented perhaps due to structure stability although their presence was confirmed



**Fig. 3** Ordination plot of accessions tested with a range of biotic agents: *Peronospora parasitica* isolates (Emoy2, Cala2, Waco9); *Oidium neolycopestici*; *Botrytis cinerea* and *Frankliniella occidentalis* (thrips). The resistance or repellence level of accessions to biotic agents (grey vectors) were set as explanatory variables and abundance of metabolites (black vectors) as response variables in the RDA plot. Indicated metabolites are accession-specific identified in the present

study and the rest of the metabolites were not identified in the present study but all of them correlated with resistance. Numbers along the axes indicate the ordinate number and percentage of variation explained. X indicates the position of accessions on the scores plot with respect to their resistance level. Vectors pointing in the same direction are positively correlated and those pointing in opposite directions are negatively correlated

in the samples of independent experiments. The fragmentation pattern of 14 of the 21 ion masses was obtained successfully. The fragmentation pattern in combination with the corresponding retention time, isotopic pattern and UV–Vis spectrum allowed for further confirmation of the structure.

In negative mode, MS/MS confirmed the identity of six glucosinolates with their characteristic fragment signal at  $[M-H]^-$  96.9595 Da ( $\sim 97$ , supplementary information Table 7). The isotopic pattern of the other putative glucosinolates, which failed to fragment in MS/MS analysis, confirmed the presence of sulfur making it likely they are indeed glucosinolates.

A phenolic compound with the elemental formula of  $C_{37}H_{42}O_{19}$  and accurate mass of  $[M-H]^-$  789.224 Da was also detected in negative mode which was putatively annotated as an isomer of truncatulin A and not reported before in Brassicaceae. MS/MS analysis of the deprotonated molecule confirmed the presence of ferulic acid, guaiacylglycerol and gentiic acid linked via two pentose sugars. (supplementary information Fig. 4a). This is the first report of this mass and elemental composition in *A. thaliana*. Hydroxybenzoic acids (such as gentiic acid) combined with a guaiacylglycerol group have been found to occur as breakdown products of lignin (Katayama et al. 1981). Ferulic acid is a constituent of lignocellulose that crosslinks the lignin and polysaccharides conferring rigidity to the cell walls. These products may have some relation with defense (Fayos et al. 2006) but its presence in some *A. thaliana* accessions remains to be further confirmed and explained.

In positive mode, three of the nitrogen containing compounds were annotated as alkaloids that were not reported before in *A. thaliana* (supplementary Information Table 7). After analysis of MS/MS spectra it was possible to correlate the fragments and characteristic neutral losses with the structures proposed for one alkaloid, diptocarpilidine and also for one amino acid derivative, 3-carboxytyrosine, which have been reported before in Brassicaceae. To date, only the indolic alkaloid, camalexin (a phytoalexin), has been identified in *A. thaliana* (Glazebrook and Ausubel 1994; Hansen and Halkier 2005). However, the presence of a multitude of alkaloid biosynthetic gene homologues in the *A. thaliana* genome may suggest that more alkaloids can potentially be synthesized (Facchini et al. 2004). Annotated alkaloids were particularly present in accessions Cvi, C-24 and An-1 and the fact that they have not been reported before may be simply due to their low abundance in Col-0.

Two putative flavonoid derivatives were identified in LC-TOF-MS positive mode (supplementary Information Table 7). One of them was registered with accurate mass  $[M + H]^+$  947.2816 Da and was annotated *in silico* as an

anthocyanin derivative not reported before in the Brassicaceae. Analysis of the MS/MS spectra (supplementary information Fig. 5a) shows this is a different flavonoid derivative, namely kaempferol 7-*O*-rhamnoside 3-*O*-rhamnosyl-(synapoyl)glucoside (supplementary information Fig. 5b). Flavonoids have been studied to a large extent in *A. thaliana* and variants of this class of secondary metabolites have been identified including derivatives of kaempferol 3-glycosides (Sever et al. 2010; Matsuda et al. 2010). Here we report two new putative flavonoids in *A. thaliana* detected in LC-MS positive mode, with one of them also annotated as kaempferol glycoside by MS/MS fragmentation. The fact that these flavonoids have so far escaped identification in *A. thaliana* may be again due to their absence in the mostly studied accession Col-0.

### 3.9 Biological relevance of metabolic distance

Metabolites mediate the interaction of plants with their environment including biotic agents (Macel and Klinkhamer 2009; Kashif et al. 2009; Mahatma et al. 2009; Forlani 2010). Therefore metabolites are highly adaptive and hence strongly vary between genotypes and habitats/environments of the accessions (Menezes and Jared 2002). Hence, our strategy to evaluate the biological context of the observed metabolic divergence between accessions was to relate their differential metabolic profiles to the interaction with biotic agents. We characterized accessions in a non-induced metabolic state. The obtained metabolite fingerprints thus represent the innate immune or general defense system without a requirement for induction. Based on their metabolic distances, four accessions were selected that best represent the observed metabolome variation across the nine accessions. Cvi and An-1 were selected as the most diverged, Col-0 as the moderately diverged and most studied and Eri as an average accession (Table 2). The four accessions showed segregation with regard to resistance against or attractiveness to different biotic agents, including host and non-host and biotrophic and necrotrophic pathogens as well as an insect (supplementary information Table 8). None of the selected accessions was resistant or susceptible to all of the biotic agents. Subsequently, the resistance or attractiveness levels to the biotic agents were used in RDA as explanatory variables for the abundance of the reconstructed metabolites in the CC-environment samples as bioassays were done under the same conditions. Several metabolites were shown to significantly (MCP test, 10,000 permutations,  $P$ -value  $< 0.01$ ) correlate with the resistance level of the accessions to different biotic agents (Fig. 3), some of which were not identified as accession-specific in our first strategy. Among the correlating metabolites based on the first two PCs two were detected in LC-MS positive mode whereas the rest were detected in



negative mode with glucosinolates dominating the list (Fig. 3). Abundance of positively correlated compounds with *Botrytis* and thrips resistance was shown to be negatively correlated with the resistance against downy mildew isolates, while abundance of positively correlated compounds with *Oidium* resistance was shown to have no correlation with the resistance against other biotic agents (Fig. 3).

Using the metabolic distance based on untargeted metabolite fingerprints, we selected accessions of *A. thaliana* that chemically diverged in the PCA plane. As a consequence of this chemical divergence, these accessions also differed in their interaction with biotic agents. This variation in chemical composition and biotic interaction could be subsequently used to identify compounds correlating with the interaction, some of them had been found to be accession specific in the two dimensional PCA plane (Fig. 3 and supplementary Information Table 7), thus linking the metabolic distance of accessions with their differential response towards biotic stresses.

### 3.10 Correlation between genetic and metabolic distance

An integrated dataset that consisted of shoot metabolite fingerprints from all analytical methods was used in a partial PCA with environment as cofactor, in order to estimate the pair-wise metabolic distance (R-value) among the nine accessions across all analytical techniques (supplementary information Table 9). The 149 genome-wide distributed SNP markers were used to determine the genetic distance among the nine accessions (supplementary Information Table 9). A Mantel test showed that there is a small (correlation coefficient  $r = 0.04$ ) but significant ( $P$ -value  $< 0.01$ ) pattern similarity between the shoot metabolic and genetic distance matrices. There was no significant correlation ( $r = 0.26$  and  $P$ -value = 0.11) between the root metabolic distance (determined using all analytical methods) and the genetic distance matrices.

The weak correlation between the genetic and metabolic distance matrices shows that genetic diversity is not one to one translated into metabolic diversity. The phylogenetic tree of all used accessions (supplementary information Fig. 1) demonstrates that at the resolution level of 0.8, An-1 is the most genetically diverged accession and Eri is the second at a lower resolution (0.7). Moreover, there is a large genetic distance between these two accessions as shown in supplementary information Table 9 and as they belong to the two different major clades of the tree (supplementary information Fig. 1). However, based on the combined metabolite datasets of the shoot, Cvi has the largest metabolic divergence and Eri is considered as an average accession with the least metabolic divergence (shown by the

ranking of the average distance, supplementary Information Table 9). An-1 is the second metabolically diverged accession, which would fit with the fact that it is genetically the most diverged. However, An-1 also has a large metabolic distance from Cvi ( $R = 0.95$ ) while genetically they belong to the same major clade (supplementary information Fig. 1). Such discrepancies imply that the genetic distance of two genotypes does not completely define their relationship and distance at the metabolic level. Our result is in accordance with a number of previous studies on closely related genotypes of two plant species (*Sesamum indicum* and *Oryza sativa*) and on several *Rhizobium* species (Wolde-Meskel et al. 2004; Mochida et al. 2009; Laurentin et al. 2008). We also observed convergence of “metabotypes”, i.e. accessions with no metabolic distance although diverged genetically. As an example in the root dataset, negligible metabolic distance is observed between An-1 and Kyo-1 in both LC-MS modes (Table 1), while the two accessions of each of these pairs belong to the two diverged subclades of the phylogenetic tree (supplementary information Fig. 1). A close metabolic relationship for genetically diverged accessions is in accordance with the hypothesis of phenotypic buffering (Fu et al. 2009). Although a genetic basis underlies the metabolome variation between *A. thaliana* accessions (our data) and mapping populations (Keurentjes et al. 2006), the hypothesis of phenotypic buffering suggests the existence of breakpoints in a system that buffers them against too large effects of genetic variation on the phenotype.

## 4 Conclusion

We characterized the metabolic variation within 9 *A. thaliana* accessions grown under various growing conditions and established a statistical method for estimating a metabolic distance between genotypes or treatments. This method may help to evaluate and compare the effects of genetic (natural variation, breeding and genetic modification) or environmental perturbations on the metabolome. Metabolic distance can be used to quantify the metabolic diversity and plasticity among plant genotypes and environments and could be a useful tool in breeding programs and genetical genomics studies.

**Acknowledgment** This work was funded by Earth and Life Sciences Council of the Netherlands Organization for Scientific Research (NWO-ALW) under the ERGO program (Number 838.06.010). RCHdV acknowledges funding by the Netherlands Metabolomics Centre (NMC) and the Centre for BioSystems Genomics (CBSG), which are under auspices of the Netherlands Genomics Initiative (NGI) and Dr. A.C. Tas from TNO-Zeist, NMC for his help in MS/MS fragments annotation.

**Conflict of interest** None

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

## References

- Aliferis, K. A., & Jabaji, S. (2009). 1H NMR and GC-MS metabolic fingerprinting of developmental stages of *Rhizoctonia solani* sclerotia. *Metabolomics*, 1–13.
- Almeida, J. A. S., Barbosa, L. M. S., Pais, A. A. C. C., & Formosinho, S. J. (2007). Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering. *Chemometrics and Intelligent Laboratory Systems*, 87(2), 208–217.
- Arimura, G. I., Kost, C., & Boland, W. (2005). Herbivore-induced, indirect plant defences. *Biochimica et Biophysica Acta - Molecular and Cell Biology of Lipids*, 1734(2), 91–111.
- Bai, Y., Pavan, S., Zheng, Z., Zappel, N. F., Reinstädler, A., Lotti, C., et al. (2008). Naturally occurring broad-spectrum powdery mildew resistance in a Central American tomato accession is caused by loss of Mlo function. *Molecular Plant-Microbe Interactions*, 21(1), 30–39.
- Belz, R. G. (2007). Allelopathy in crop/weed interactions — an update. *Pest Management Science*, 63(4), 308–326.
- Boccard, J., Grata, E., Thiocone, A., Gauvrit, J. Y., Lantieri, P., Carrupt, P. A., et al. (2007). Multivariate data analysis of rapid LC-TOF/MS experiments from *Arabidopsis thaliana* stressed by wounding. *Chemometrics and Intelligent Laboratory Systems*, 86(2 SPEC. ISS), 189–197.
- Bouwmeester, H. J., Matusova, R., Zhongkui, S., & Beale, M. H. (2003). Secondary metabolite signalling in host–parasitic plant interactions. *Current Opinion in Plant Biology*, 6(4), 358–364.
- Bouwmeester, H. J., Roux, C., Lopez-Raez, J. A., & Becard, G. (2007). Rhizosphere communication of plants, parasitic plants and AM fungi. *Trends in Plant Science*, 12(5), 224–230.
- Boyes, D. C., Zayed, A. M., Ascenzi, R., McCaskill, A. J., Hoffman, N. E., Davis, K. R., et al. (2001). Growth stage-based phenotypic analysis of *Arabidopsis*: A model for high throughput functional genomics in plants. *Plant Cell*, 13(7), 1499–1510.
- 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(2), 179–186.
- Chong, J., Poutaraud, A., & Huguene, P. (2009). Metabolism and roles of stilbenes in plants. *Plant Science*, 177(3), 143–155.
- Cona, A., Rea, G., Angelini, R., Federico, R., & Tavladoraki, P. (2006). Functions of amine oxidases in plant development and defence. *Trends in Plant Science*, 11(2), 80–88.
- Davies, A. N. (1998). The new automated mass spectrometry deconvolution and identification system (AMDIS). *Spectroscopy Europe*, 10(3), 22–26.
- De Jong, G. (2005). Evolution of phenotypic plasticity: Patterns of plasticity and the emergence of ecotypes. *New Phytologist*, 166(1), 101–118.
- De Vos, M., Van Oosten, V. R., Van Poecke, R. M. P., Van Pelt, J. A., Pozo, M. J., Mueller, M. J., et al. (2005). Signal signature and transcriptome changes of *Arabidopsis* during pathogen and insect attack. *Molecular Plant-Microbe Interactions*, 18(9), 923–937.
- Ehrlich, P. R., & Raven, P. H. (1964). Butterflies and plants: A study in coevolution. *Evolution*, 18, 586–608.
- Facchini, P. J., Bird, D. A., & St-Pierre, B. (2004). Can *Arabidopsis* make complex alkaloids? *Trends in Plant Science*, 9(3), 116–122.
- Fayos, J., Bellés, J. M., López-Gresa, M. P., Primo, J., & Conejero, V. (2006). Induction of gentisic acid 5-O-β-D-xylopyranoside in tomato and cucumber plants infected by different pathogens. *Phytochemistry*, 67(2), 142–148.
- Ferrari, S., Plotnikova, J. M., De Lorenzo, G., & Ausubel, F. M. (2003). *Arabidopsis* local resistance to *Botrytis cinerea* involves salicylic acid and camalexin and requires EDS4 and PAD2, but not SID2, EDS5 or PAD4. *Plant Journal*, 35(2), 193–205.
- Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, R. N., & Willmitzer, L. (2000). Metabolite profiling for plant functional genomics. *Nature Biotechnology*, 18(11), 1157–1161.
- Forlani, G. (2010). Differential in vitro responses of rice cultivars to Italian lineages of the blast pathogen *Pyricularia grisea*. 2. Aromatic biosynthesis. *Journal of Plant Physiology*, 167(11), 928–932.
- Fu, J., Keurentjes, J. J. B., Bouwmeester, H., America, T., Verstappen, F. W. A., Ward, J. L., et al. (2009). System-wide molecular evidence for phenotypic buffering in *Arabidopsis*. *Nature Genetics*, 41(2), 166–167.
- García-Pérez, I., Vallejo, M., García, A., Legido-Quigley, C., & Barbas, C. (2008). Metabolic fingerprinting with capillary electrophoresis. *Journal of Chromatography. A*, 1204(2), 130–139.
- Glazebrook, J., & Ausubel, F. M. (1994). Isolation of phytoalexin-deficient mutants of *Arabidopsis thaliana* and characterization of their interactions with bacterial pathogens. *Proceedings of the National Academy of Sciences of the United States of America*, 91(19), 8955–8959.
- Hammer, O., Harper, D. A. T., & Ryan, P. D. (2001). Past: Paleontological statistics software package for education and data analysis. *Palaeontologia Electronica*, 4(1), XIX–XX.
- Hansen, B. G., & Halkier, B. A. (2005). New insight into the biosynthesis and regulation of indole compounds in *Arabidopsis thaliana*. *Planta*, 221(5), 603–606.
- Hilker, M., & Meiners, T. (2006). Early herbivore alert: Insect eggs induce plant defense. *Journal of Chemical Ecology*, 32(7), 1379–1397.
- Kabouw, P., Biere, A., Van Der Putten, W. H., & Van Dam, N. M. (2009). Intra-specific differences in root and shoot glucosinolate profiles among white cabbage (*Brassica oleracea* var. capitata) Cultivars. *Journal of Agricultural and Food Chemistry*, 58(1), 411–417.
- Kashif, A., Federica, M., Eva, Z., Martina, R., Young, H. C., & Robert, V. (2009). NMR metabolic fingerprinting based identification of grapevine metabolites associated with downy mildew resistance. *Journal of Agricultural and Food Chemistry*, 57(20), 9599–9606.
- Katayama, T., Nakatsubo, F., & Higuchi, T. (1981). Degradation of arylglycerol-β-aryl ethers, lignin substructure models, by *Fusarium solani*. *Archives of Microbiology*, 130(3), 198–203.
- Keurentjes, J. J. B., Fu, J., De Vos, C. H. R., Lommen, A., Hall, R. D., Bino, R. J., et al. (2006). The genetics of plant metabolism. *Nature Genetics*, 38(7), 842–849.
- Kim, S. W., Min, S. R., Kim, J., Park, S. K., Kim, T. I., & Liu, J. R. (2009). Rapid discrimination of commercial strawberry cultivars using Fourier transform infrared spectroscopy data combined by multivariate analysis. *Plant Biotechnology Reports*, 3(1), 87–93.
- Kind, T., & Fiehn, O. (2007). Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, 8, 105–125.
- Kliebenstein, D. J., Kroymann, J., Brown, P., Figuth, A., Pedersen, D., Gershenzon, J., et al. (2001). Genetic control of natural variation in *Arabidopsis* glucosinolate accumulation. *Plant Physiology*, 126(2), 811–825.
- Kliebenstein, D. J., Kroymann, J., & Mitchell-Olds, T. (2005). The glucosinolate-myrosinase system in an ecological and evolutionary context. *Current Opinion in Plant Biology*, 8(3 SPEC. ISS), 264–271.

- Laurentin, H., Ratzinger, A., & Karlovsky, P. (2008). Relationship between metabolic and genomic diversity in sesame (*Sesamum indicum* L.). *BMC Genomics*, *9*, 250.
- Lommen, A. (2009). Metalign: Interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. *Analytical Chemistry*, *81*(8), 3079–3086.
- Macel, M., & Klinkhamer, P. G. L. (2009). Chemotype of *Senecio jacobaea* affects damage by pathogens and insect herbivores in the field. *Evolutionary Ecology*, *24*(1), 237–250.
- Mahatma, M. K., Bhatnagar, R., Dhandhukia, P., & Thakkar, V. R. (2009). Variation in metabolites constituent in leaves of downy mildew resistant and susceptible genotypes of pearl millet. *Physiology and Molecular Biology of Plants*, *15*(3), 249–255.
- Matsuda, F., Hirai, M. Y., Sasaki, E., Akiyama, K., Yonekura-Sakakibara, K., Provart, N. J., et al. (2010). AtMetExpress development: A phytochemical atlas of Arabidopsis development. *Plant Physiology*, *152*(2), 566–578.
- Menezes, H., & Jared, C. (2002). Immunity in plants and animals: Common ends through different means using similar tools. *Comparative Biochemistry and Physiology —C Toxicology and Pharmacology*, *132*(1), 1–7.
- Mochida, K., Furuta, T., Ebana, K., Shinozaki, K., & Kikuchi, J. (2009). Correlation exploration of metabolic and genomic diversity in rice. *BMC Genomics*, *10*, 568–668.
- Moco, S., Bino, R. J., Vorst, O., Verhoeven, H. A., De Groot, J., Van beek, T. A., et al. (2006). A liquid chromatography-mass spectrometry-based metabolome database for tomato. *Plant Physiology*, *141*(4), 1205–1218.
- Raguso, R. A., & Pichersky, E. (1999). A day in the life of a linalool molecule: Chemical communication in a plant—pollinator system. Part I: Linalool biosynthesis in flowering plants. *Plant Species Biology*, *14*(2), 95–120.
- Rulhmann, S., Leser, C., Bannert, M., & Treutter, D. (2002). Relationship between growth, secondary metabolism, and resistance of apple. *Plant Biology*, *4*(2), 137–143.
- Schripsema, J. (2010). Application of NMR in plant metabolomics: Techniques, problems and prospects. *Phytochemical Analysis*, *21*(1), 14–21.
- Sever, A., Einhorn, J., Brunelle, A., & Laprave, O. (2010). Localization of flavonoids in seeds by cluster time-of-flight secondary ion mass spectrometry imaging. *Analytical Chemistry*, *82*(6), 2326–2333.
- Smilauer, J. L. A. P. (2003). *Multivariate analysis of ecological data using CANOCO*. Cambridge: Cambridge University Press.
- Snoeren, T. A. L., Kappers, I. F., Broekgaarden, C., Mumm, R., Dicke, M., & Bouwmeester, H. J. (2010). Natural variation in herbivore-induced volatiles in *Arabidopsis thaliana*. *Journal of Experimental Botany*, *61*(11), 3041–3056.
- ter Braak, C. J. F. (1988). CANOCO—an extension of DECORANA to analyze species—environment relationships. *Vegetatio*, *75*(3), 159–160.
- Tikunov, Y., Lommen, A., De Vos, C. H. R., Verhoeven, H. A., Bino, R. J., Hall, R. D., et al. (2005). A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. *Plant Physiology*, *139*(3), 1125–1137.
- Tocquin, P., Corbesier, L., Havelange, A., Pieltain, A., Kurtem, E., Bernier, G., et al. (2003). A novel high efficiency, low maintenance, hydroponic system for synchronous growth and flowering of *Arabidopsis thaliana*. *BMC Plant Biology*, *3*, 2.
- Treutter, D. (2005). Significance of flavonoids in plant resistance and enhancement of their biosynthesis. *Plant Biology*, *7*(6), 581–591.
- Van Dam, N. M., Tytgat, T. O. G., & Kirkegaard, J. A. (2009). Root and shoot glucosinolates: A comparison of their diversity, function and interactions in natural and managed ecosystems. *Phytochemistry Reviews*, *8*(1), 171–186.
- Van Damme, M., Zeilmaker, T., Elberse, J., Andel, A., De Sain-van Der Velden, M., & Van Den Ackerveken, G. (2009). Downy mildew resistance in arabidopsis by mutation of homoserine kinase. *Plant Cell*, *21*(7), 2179–2189.
- Van De Peer, Y., & De Wachter, R. (1997). Construction of evolutionary distance trees with TREECON for Windows: Accounting for variation in nucleotide substitution rate among sites. *Computer Applications in the Biosciences*, *13*(3), 227–230.
- Vorst, O., de Vos, C. H. R., Lommen, A., Staps, R. V., Visser, R. G. F., Bino, R. J., et al. (2005). A non-directed approach to the differential analysis of multiple LC-MS-derived metabolic profiles. *Metabolomics*, *1*(2), 169–180.
- Wolde-Meskel, E., Terefework, Z., Lindstrom, K., & Frostegard, A. (2004). Metabolic and genomic diversity of rhizobia isolated from field standing native and exotic woody legumes in southern Ethiopia. *Systematic and Applied Microbiology*, *27*(5), 603–611.