DOI:10.1021/jf9009137

JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY
ARTICLE

# Plant Phenotype Demarcation Using Nontargeted LC-MS and GC-MS Metabolite Profiling

Vicent Arbona,[†] Domingo J. Iglesias,[‡] Manuel Talón,[‡] and Aurelio Gómez-Cadenas*,[†]

[†]Ecofisiologia i Biotecnologia. Departament de Ciències Agràries i del Medi Natural, Universitat Jaume I, E-12071 Castelló de la Plana, Spain, and [‡]Centro de Genómica, Instituto Valenciano de Investigaciones Agrarias, E-46113 Moncada, Spain

The characterization of the metabolome is a critical aspect in basic research and plant breeding. In this work, the putative application of metabolomics for phenotyping closely related genotypes has been tested. Crude extracts were profiled by LC-MS and GC-MS, and mass data extraction was performed with XCMS software. Result validation was achieved with principal component analysis (PCA). The ability of the profiling methodologies to discriminate plant genotypes was assessed after hierarchical clustering analysis (HCA). Cluster robustness was assessed by a multiscale bootstrap resampling method. A better performance of LC-MS profiling over GC-MS was evidenced in terms of phenotype demarcation after PCA and HCA. Citrus demarcation was similarly achieved independently of the environmental conditions used to grow plants. In addition, when all different locations were pooled in a single experimental design, it was still possible to differentiate the three closely related genotypes. The presented methodology provides a fast and nontargeted workflow as a powerful tool to discriminate related plant phenotypes. The novelty of the technique relies on the use of mass signals as markers for phenotype demarcation independent of putative metabolite identities and the relatively simple analytical strategy that can be applicable to a wide range of plant matrices with no previous optimization.

**KEYWORDS: Citrus; *Arabidopsis*; metabolomics; principal component analysis; hierarchical clustering analysis; mass spectrometry**

## INTRODUCTION

Plant metabolomes have been described as bridges between genotypes and phenotypes, reflecting different biological end points as the downstream result of gene expression. Extensive knowledge on metabolic flows could allow assessment of genotypic or phenotypic differences between plant species. In addition, target metabolites have been analyzed as nutritional and/or agronomical biomarkers to classify different crop cultivars or to optimize growth conditions. Recently, metabolomics has arisen as a key functional genomics tool and has been successfully applied to assess differences in metabolite composition in distinct tomato cultivars (*1*) and introgression lines (*2*) and, more recently, between different *Arabidopsis* species (*3*). Besides its use as a breeding or selection tool, metabolomics techniques have also been used to evaluate stress responses in barley (*4*), *Citrus* (*5*), *Medicago truncatula* (*6*), and *Arabidopsis thaliana* (*7*). In addition, it has been discussed that metabolomes mirror genetic and/or environmental changes to a great extent and, therefore, describe more accurately the phenotype of a given organism (*8*). The use of these techniques could help in the development of rational breeding programs (*9*), especially when desirable agronomical or nutritional traits do not correlate with specific DNA markers. Overall, metabolomics represent a useful tool to evaluate the contribution of environmental and genetic factors to the differences in metabolite composition or content (*10*, *11*).

The final goal of a metabolomics analysis is the identification and quantification of all metabolites in a given organism in addition to the assessment of the metabolic relationships among them. However, this has not been possible to date because most metabolites are yet unknown and the available analytical techniques do not allow such exhaustive metabolite detection (*12*). Therefore, the most widely used techniques, known as metabolite profiling (*13*), consist in the analysis of the maximum number of metabolites in a given sample. Mass spectrometry (MS) coupled to separative techniques such as HPLC, GC, or capillary electrophoresis is used for this purpose (*14*). Indeed, HPLC coupled to hybrid quadrupole time-of-flight mass spectrometers (QTOF) are among the most versatile metabolite profiling techniques because LC is the most compatible technique with biomolecules, and the accurate mass measurement, true isotopic pattern recognition, and high sensitivity provided by QTOF instruments are suitable for calculations on elemental composition of mass signals (*15*).

An important factor in MS-based metabolomics is the number of detected signals. Sample collection and extraction steps are crucial and may definitely influence the results obtained. In this sense, efforts to standardize procedures for fresh tissue handling and extraction have been done to minimize manipulation of plant

*Corresponding author (e-mail aurelio.gomez@uji.es; telephone +34 964 72 94 02; fax +34 964 72 82 16).

 Published on Web 07/29/2009 

material (*13*, *14*). However, it is not clear whether these techniques are suitable for all classes of metabolites or tissues. In addition, the parallel use of different hyphenated techniques such as GC-MS and LC-MS could be a good choice to better profile different classes of compounds. Another important aspect in the optimization of a metabolite profiling method is the extraction procedure because different plant matrices can exhibit a notable complexity when injected without purification. In this sense, a separation technique preceding mass spectrometry is often used, and it is also a critical step in determining the exhaustiveness of the method. However, flow injection electrospray ionization coupled to mass spectrometry has been used as a fingerprinting tool due to the relatively short acquisition time (*12*), although ion suppression upon ionization of crude extracts may reduce the number of detected mass signals (*16*). When liquid chromatography is used, the polarity of the stationary phase as well as the separation gradient used will condition the performance of the mass spectrometer and the results of the metabolite profiling. For example, when reversed phase liquid chromatography coupled to electrospray ionization is used, highly polar compounds and volatiles cannot be detected. Other chromatographic techniques, such as GC, do not allow the injection of aqueous or nonvolatile samples, therefore requiring a derivatization step. Generally, GC-MS techniques offer a relatively more robust chromatography and greater separation efficiency, resulting in reproducible retention times of hundreds of mass peaks, together with the availability of reference compound libraries (*10*, *13*).

The extraction of mass signals from raw data files after acquisition is another limiting step because no absolutely accurate methodology currently exists (see ref *17* for a recent review). When dealing with LC-MS data, different software such as the commercially available Markerlynx (Micromass Ltd., Manchester, U.K.) or freeware such as metAlign (Plant Research International, Wageningen, The Netherlands), MzMine (*18*), or XCMS (*19*) perform the automatic extraction, alignment, and retention time correction of chromatographic peaks within individual mass-to-charge values using different algorithms. Another software package for deconvolution of chromatographic peaks is AMDIS (NIST). This software automatically identifies and annotates mass signals in GC-MS experiments; however, it does not perform alignment and retention time correction of chromatographic features.

The putative application of metabolomics for phenotyping of closely related species has been tested in this work. To make results widely applicable, *Citrus* genotypes have been used as a model due to their complex matrices and also their economical importance. In addition, two external groups have been added to test the performance of the clustering procedure: *Arabidopsis thaliana* L. Heyhn. ecotype Columbia and *Prunus persica* L. To obtain a more comprehensive view of metabolite profiling, two techniques were used: GC-MS and LC-MS. The effectiveness and accuracy to distinguish among species and to highlight metabolite differences are discussed.

## MATERIALS AND METHODS

**Plant Material and Experimental Designs.** Eight-month-old seedlings of *P. persica* L., citrumelo CPB 4475 (*Citrus paradisi* L. Macf. × *Poncirus trifoliata* L. Raf.), Carrizo citrange (*P. trifoliata* L. Raf. × *Citrus sinensis* L. Osb.), and Cleopatra mandarin (*Citrus reshni* Hort. ex Tan.) were purchased from a commercial nursery and immediately transplanted to 2 L pots filled with a mixture of peat moss/perlite/vermiculite (8:1:1) as substrate. Seeds of *A. thaliana* L. Heyhn Columbia-0 ecotype were surface-sterilized with 70% EtOH followed by 20% bleach and, finally, rinsed with sterile water and sown in moistened sterile peat moss. After 2 days at 4 °C to break dormancy, seeds were germinated in a growth chamber at 22 °C under a 10 h photoperiod. After 1 week, seedlings were individually transplanted to 0.2 L pots filled with the same substrate as for woody genotypes and transferred to the greenhouse described under Experiment 1.

**Experiment 1.** After transplanting, *Citrus* and *Prunus* plants were grown for 2 months before harvesting, whereas 1-week-old *Arabidopsis* plants were grown in the same conditions for 5 weeks. Day/night temperatures within the greenhouse were 23 ± 2/18 ± 2 °C with 70−80% relative humidity. Natural light was used (with a photoperiod varying approximately from 10 to 12 h). *Citrus* and *Prunus* seedlings were watered twice a week using 0.5 L of a half-strength Hoagland solution (*26*). *Arabidopsis* pots were arranged in 30 × 60 cm trays and regularly watered from below using 1 L of a $^{1}/_{10}$ dilution of Hoagland nutrient solution as reported previously (*15*).

Throughout the experimental period, all groups of plants were apparently healthy without any visible damage. Fifteen plants of each woody genotype (approximately 1 m tall) were individually harvested, and only adult leaves collected. For *Arabidopsis*, rosettes of 15 groups of plants (three plants per group) were independently harvested. All samples were rinsed with distilled water, blotted dry, and immediately frozen in liquid nitrogen. Tissue was then ground to a fine powder and stored at −80 °C until analyses.

**Experiment 2.** Plants of the three *Citrus* genotypes were cultivated for 3 weeks outdoors, in the experimental field of the Jaume I University (Castellón, Spain; 39° 59′ N, 0° 02′ W) and, therefore, exposed to the variable environmental conditions (during the experimental period, average, maximum, and minimum air temperatures were 20.9, 24.3, and 15.7 °C, respectively). Plants were of the same age as in experiment 1 and were cultivated in the same pots and substrate and with the same watering program described above. Adult leaves of 15 plants per genotype were independently harvested.

**Experiment 3.** Plants identical to those described under Experiment 2 were cultivated for 3 weeks outdoors, in the I.VIA experimental field (Moncada, Valencia, Spain; 39° 32′ N, 0° 23′ W) located 70 km south of Castellón. During the experimental period, average, maximum, and minimum air temperatures were 24.9, 27.9, and 20.4 °C, respectively). In this case, plants were cultivated in the same pots but with perlite as substrate. Plants were watered 5 days a week with 0.5 L of the watering solution described above. Adult leaves of 15 plants per genotype were independently harvested.

**Extraction and Fractionation.** Frozen plant tissue powder (0.5 g) was extracted in 5 mL of 80% MeOH (HPLC grade, Panreac, Barcelona, Spain) using a tissue homogenizer (Ultra-Turrax, Ika-Werke, Staufen, Germany) in an ice bath to prevent sample heating. After centrifugation to pellet debris, supernatant was recovered and evaporated at room temperature using a centrifuge vacuum evaporator (RT 2.2., Jouan, Saint Herblain, France). The dry residue was resuspended in 2 mL of a 40% MeOH solution and subsequently fractionated with C18 cartridges (100 mg, BondElut, Varian Inc., Palo Alto, CA). Nonretained eluates (defined as the polar fraction) from C18 cartridges were collected and evaporated as above. The retained fraction (defined as nonpolar) was eluted with 2 mL of 100% MeOH, collected, and evaporated to dryness.

When LC-MS was used, polar and nonpolar fractions were resuspended in 1 mL of 10% MeOH or 50% MeOH, respectively, and filtered through cellulose acetate filters (0.22 $\mu$m pore size) before injection in the HPLC. For GC-MS analyses, both fractions were reconstituted in 30 $\mu$L of methoxamine hydrochloride (Sigma-Aldrich, Madrid, Spain) in pyridine (20 mg mL$^{-1}$, Sigma-Aldrich) and kept for 90 min at room temperature. After evaporation under vacuum, dry residues were trimethylsylilated by adding 40 $\mu$L of MSTFA (Sigma-Aldrich) followed by incubation at 37 °C for 30 min. For all experiments, 15 independent biological samples were analyzed by LC-MS and 9 by GC-MS to prevent instrument overloading.

**LC and GC Conditions.** As a preliminary program for LC optimization, a 20 $\mu$L aliquot of each fraction was injected onto a HPLC system (Waters Alliance 2690, Milford, MA). UV profiles (at 290 and 350 nm) and mass spectra (between 50 and 900 arbitrary mass units) were collected over a 60 min period using MeOH (solvent A) and 0.01% acetic acid in $H_2O$ (solvent B) and following a linear gradient, increasing solvent A from 5 to 95% (all solvents were of HPLC grade). The gradients were optimized to meet an agreement between short run time and good chromatographic resolution. Therefore, two final programs were chosen: Gradient 1 for

polar fractions was 95:5 (B/A) to 20:80 in 20 min, 10:90 (20−23 min), and 95:5 (23−26 min). A 4 min re-equilibration period was established between injections. Gradient 2 for nonpolar fractions lasted for 35 min and started with a 50:50 (B/A) proportion, in 20 min, 5:95, and then this proportion maintained for 2 min before initial conditions were restored in 5 min. Finally, a period of 8 min was established to re-equilibrate the column. Separations were carried out at room temperature using a 5 $\mu$m Kromasil 100 C18 column (100 × 2.1 mm, Scharlab, Barcelona, Spain) at a flow rate of 0.3 mL $min^{-1}$. Effluents were injected in the QTOF (QTOF I, Micromass Ltd.) through an orthogonal Z-spray electrospray interface using $N_2$ as both nebulization and desolvation gas. Nebulizer and dry gas flows were adjusted to 90 and 800 arbitrary units, respectively. Data were acquired in continuous mode within a mass scan range between 50 and 900 atomic mass units (amu) in both positive and negative electrospray mode at 4000 V spray and 25 V cone voltages.

For GC-MS analyses, polar and nonpolar fractions were independently injected into a Star 3400 CX gas chromatograph coupled to a Saturn ion trap mass spectrometer (Varian). Helium inlet pressure was set at 85 kPa, and the injector, interface, and ion source temperatures were 250, 250, and 200 °C, respectively. Samples (2 $\mu$L) were injected in splitless mode and separated in a 30 m length polymeric column (VF-5 ms, 0.25 mm × 0.10 $\mu$m, Varian) using helium as a carrier gas. After optimization of GC temperature ramps to obtain good peak resolution, a linear temperature gradient from 40 to 280 at 8 °C/min was set in the oven. Mass spectra were collected over a 22−25 min period; the solvent cut was 4 min. Data were recorded within the 100−650 amu.

**Data Acquisition and Analysis.** LC-MS data were acquired and centroided using Masslynx 4.1 software (Micromass Ltd.). Centroidization of raw files was accomplished using an internal lock-mass reference, L-tri-iodotyrosine, that was postcolumn injected into the ion source ($m/z$ 433.8150 [$ESI^+$], 431.8594 [$ESI^-$]) during sample analyses.

The stability of the LC-MS system was assessed by extracting the lock-mass signal intensity and plotting it over time. Signal intensity plots from different samples overlapped without significant variation among them (only samples from the same genotype were considered to discard side effects derived from ion suppression due to different matrix effects). In addition, to assess variations in LC and GC performance, a set of standard analytes were injected in three replicate batches at the beginning, in the middle, and at the end of each sample list. Whereas kinetin, biochanin A, rutin, *o*-anisic acid, ferulic acid, and *N*-(3-indolylacetyl)-L-phenylalanine (Sigma-Aldrich) were used in LC-MS, only the last three compounds were reliably detected in GC-MS. In LC-MS, relative standard deviations (RSD) for peak areas and for retention times varied between 3.80 and 17.34% and between 0.21 and 0.49%, respectively. For GC-MS, the same parameters showed RSD values varying between 0.69 and 9.28% and between 0.05 and 0.21%, respectively.

**Raw Mass Data Preprocessing and Extraction.** Centroided files were subsequently transformed into netCDF format using the Databridge utility in the Masslynx package. Native MS files from Varian Saturn GC-MS instrument were converted into raw Xcalibur 1.4 (Thermo Fisher Scientific, Inc., Waltham, MA) files and subsequently to netCDF using the File Converter tool. After conversion, files were checked for similarity to original files using Insilicos viewer 1.4.5 (Insilicos LLC, Seattle, WA). This conversion is needed prior to preprocessing, peak extraction, retention time correction, and alignment with XCMS (*19*). Files were arranged in one folder that was set as the file source. Peaks were subsequently extracted using the default "matchedFilter" method. The optimized parameters were as follows: signal-to-noise ratio = 4; full width at half-maximum (fwhm) = 30 (for LC data and 20 for GC data); width of the $m/z$ range = 0.1 (step parameter); "bin" was used as the profiling method, which performed adequately for both Micromass and Varian Saturn netCDF centroided data. The $m/z$ difference was set at 0.1. XCMS analyses were carried out under R 2.5.1 environment (www.bioconductor.org) running in an Intel Core2 Duo T7200 1.8 GHz and 2Gb RAM. After peak extraction, grouping and nonlinear retention time correction of peaks was accomplished in three iterative cycles with descending bandwidth (bw). This was accomplished by manually decreasing the bw parameter (from 30 to 5 s). The performance of the alignment and retention time correction procedure was monitored after each round by checking the number of aligned peak groups and by plotting the corrected retention time versus the retention time deviation (in seconds). The resulting peak list was further processed using Microsoft Excel (Microsoft Corp., Redmond, WA). In this final report, the average area, the maximum, and minimum $m/z$ values and corrected retention time are shown (Supporting Information Files 5−10). Absolute peak area values were autoscaled (the mean area value of each feature throughout all samples was subtracted from each individual feature area and the result divided by the standard deviation) as in ref *20* prior to principal component analysis.

**Principal Component Analysis (PCA) and Box−Whisker Plots.** In the Microsoft Excel spreadsheets, all missing peaks (NA) were substituted by a minimum value (*10*) and subsequently subjected to PCA using Ginkgo Analysis System software (this multiplatform application is available at http://biodiver.bio.ub.es/ginkgo/index.html). In these analyses, "mass value retention time" was used as label for each feature (variables) in injected samples (individuals). To assess normality of data sets, area values were normalized as $\log_2$ and directly used to plot box−whisker graphs (see Supporting Inforamtion Files 11 and 12) using SigmaPlot v. 9.0 for Windows (Systat Software, Inc., Chicago, IL).

**Hierarchical Clustering and Bootstrap Test.** The spreadsheets generated above were subjected to hierarchical clustering using the publicly available microarray analysis software DChip (*27*). This software provides a fast and user-friendly way to calculate and visualize sample clusters. Distances between samples and signals in data sets were calculated as 1-correlation; subsequently, a centroid clustering method was performed. Clusters were ordered by tightness. $P$ value thresholds of 0.001 for signal enrichment function and 0.01 for sample were set. The distance cutoff value used to establish the different sample clusters was 0.5. Cluster robustness was assessed with a multiscale bootstrap resampling test using the pvclust R package according to ref *28* on autoscaled data setting a nboot value of 1000 (number of iterations).

## RESULTS AND DISCUSSION

In modern agriculture, the characterization of the metabolome as a phenotyping trait is becoming an essential aspect (*9*). The full development of these techniques would support breeding and selection programs and also facilitate the differentiation between traditional and biotechnology-derived crops (*9*). In the citrus industry, important agronomical traits do not usually correlate with specific DNA markers, and the development of accurate identification methods has been either unsuccessful (*21*) or needed important investment in transcriptomics platforms (*22*). In the present work, metabolite profiles from a set of closely related *Citrus* genotypes were analyzed along with two external groups by LC-MS and GC-MS. The aim of this research was to establish a reliable, reproducible, and nontargeted metabolite profiling methodology to differentiate among different species. In this approach, we analyzed five genotypes: two hybrids, citrumelo CPB4475 and Carrizo citrange, that share one parental; Cleopatra mandarin; and two nonrelated genotypes, *Arabidopsis thaliana* and *Prunus persica*.

**Comparison of Metabolite Profiles.** As a first step in the development of the metabolomics procedure used in this work, LC and GC gradients were optimized to reduce time of chromatographic runs. This prevented overloading of the mass spectrometer with nonvolatile sample residues and a subsequent reduction in sensitivity (see Materials and Methods). In addition, a 4 min solvent delay step was set to avoid accumulation of salts in the ion source. As a result of the optimization process, two short LC gradients were obtained for polar (26 min) and nonpolar (35 min) fractions, respectively, which allowed the analysis of several samples per day without significant reduction in sensitivity and/or accuracy (data not shown). After the inspection of TIC chromatograms from *Citrus* and *Prunus* extracts, it could be stated that these showed an extraordinary complexity when compared to those from *Arabidopsis* (**Figure 1**). It could be also observed that LC-MS profiles of *Citrus* genotypes were very similar among them. However, the analysis of TIC chromatograms did not provide much information regarding differences among plant genotypes, and a deeper examination was needed.

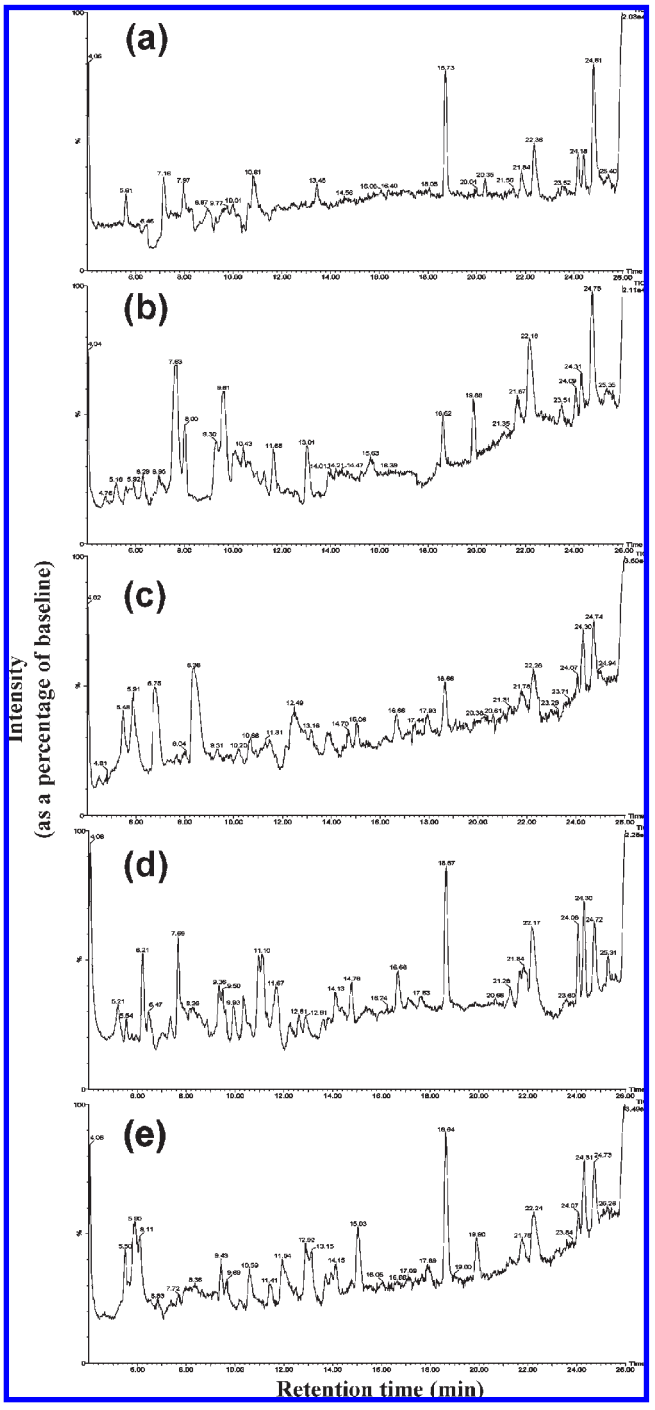


**Figure 1.** Typical TIC chromatograms of the polar fractions acquired with LC-MS in positive ionization mode: (**a**) *Arabidopsis*, (**b**) *Prunus*, (**c**) citrumelo, (**d**) Cleopatra, and (**e**) Carrizo leaf extracts. Mass data were acquired within a mass range of 50−900 amu.

**Intra- and Interspecies Variability.** As a first approach to account for overall variability among plant genotypes, normalized area values from experiment 1 were subjected to PCA (**Figure 2**). As a general aspect, high variability could be explained with only three components. However, not always a high cumulative variability was associated with a better resolution of plant genotypes. As an example, PCA of LC-MS positive mode profiles from nonpolar fractions yielded the highest cumulative variability (91.9% for 1223 mass features) but, however, could not resolve well *Arabidopsis* and *Prunus* samples. This was associated with a more important involvement of polar compounds in the discrimination of plant genotypes. In LC-MS metabolite profiles from polar fractions, component 1 resolved well *Citrus* from non-*Citrus* plant genotypes, whereas, in general, component 2 resolved *Citrus* genotypes. Similar results were observed after PCA of LC-MS metabolite profiles from nonpolar fractions with component 1 showing higher variability values. Analysis of GC-MS metabolite profiles rendered different results. PCA of GC-MS profiles from polar fractions rendered results comparable to those of LC-MS, where *Citrus* and non-*Citrus* genotypes resolved well along component 1 (55.8%) and *Citrus* genotypes resolved along component 2. On the contrary, PCA from nonpolar fractions did not show any clear trend, and although cumulative variability was high (70.9%), no resolution of plant genotypes could be achieved. Similar results, regarding resolution of *Citrus* genotypes, were obtained in LC-MS and GC-MS profiles from polar and nonpolar fractions of *Citrus* genotypes cultivated in two different locations (Supporting Information Files 1 and 3). These results validated those obtained in experiment 1. In addition, a deeper analysis of 135 files obtained from the three experiments (3 locations × 3 *Citrus* genotypes × 15 sample replicates per genotype) associated component 1 with genotype discrimination, whereas component 2 was related to environmentally driven variation (**Figure 3**). These results showed that, for LC-MS profiles, genotype variability ruled over environment, although it was still possible to differentiate different environments within each genotype group.

**Cluster Analysis of Metabolite Profiles.** Hierarchical clustering analysis (HCA) was performed on autoscaled data (*20*). As an example, **Figure 4** shows a green−black−red diagram illustrating the distribution of signal intensities. Clustering of samples from experiment 1 situated citrumelo and Carrizo genotypes very close in all LC-MS-based profiles (**Figure 5**), whereas Cleopatra always clustered close to *Arabidopsis* and *Prunus*, although separately from these two species. In addition, as indicated above, HCA on nonpolar profiles showed a closer relationship between *Arabidopsis* and *Prunus* than that shown by polar profiles. This was correlated with previous results of PCAs. HCA of polar GC-MS-derived data sets grouped well plant genotypes (**Figure 6**), although not as efficiently as those derived from LC-MS. On the contrary, HCA from nonpolar fractions showed two major clusters, *Citrus* and non-*Citrus* genotypes, and only in the latter was it possible to differentiate *Arabidopsis* and *Prunus* genotypes. HCA performed on data from experiments 2 and 3, in which *Citrus* genotypes were cultivated in different environmental conditions, rendered very similar results to those obtained in experiment 1. In general, HCA performed on LC-MS-derived data sets rendered a better resolution of plant genotypes than GC-MS-derived ones (Supporting Information Files 2 and 4). High bootstrap score values (between 80 and 100%) confirmed clustering robustness in the three experiments reported.

A similar approach was followed in ref *7* to identify new compounds with low molecular mass involved in responses to wounding in *Arabidopsis*. In this previous work, a rapid LC gradient (10 min) was used to analyze metabolites and to characterize profiles by means of PCA and HCA. However, only major constituents were detected because ion suppression was high, presumably due to matrix effects. The methodology presented in this paper efficiently differentiated the two genetically related genotypes (citrumelo and Carrizo) despite that, after HCA, they clustered together. As a result of the greater differences with *Citrus*, *Prunus* and *Arabidopsis* occurred together in the HCA but clearly differentiated from each other. It is interesting to note that the other *Citrus* genotype, Cleopatra, was apart from the rest of the *Citrus* species but also from the two

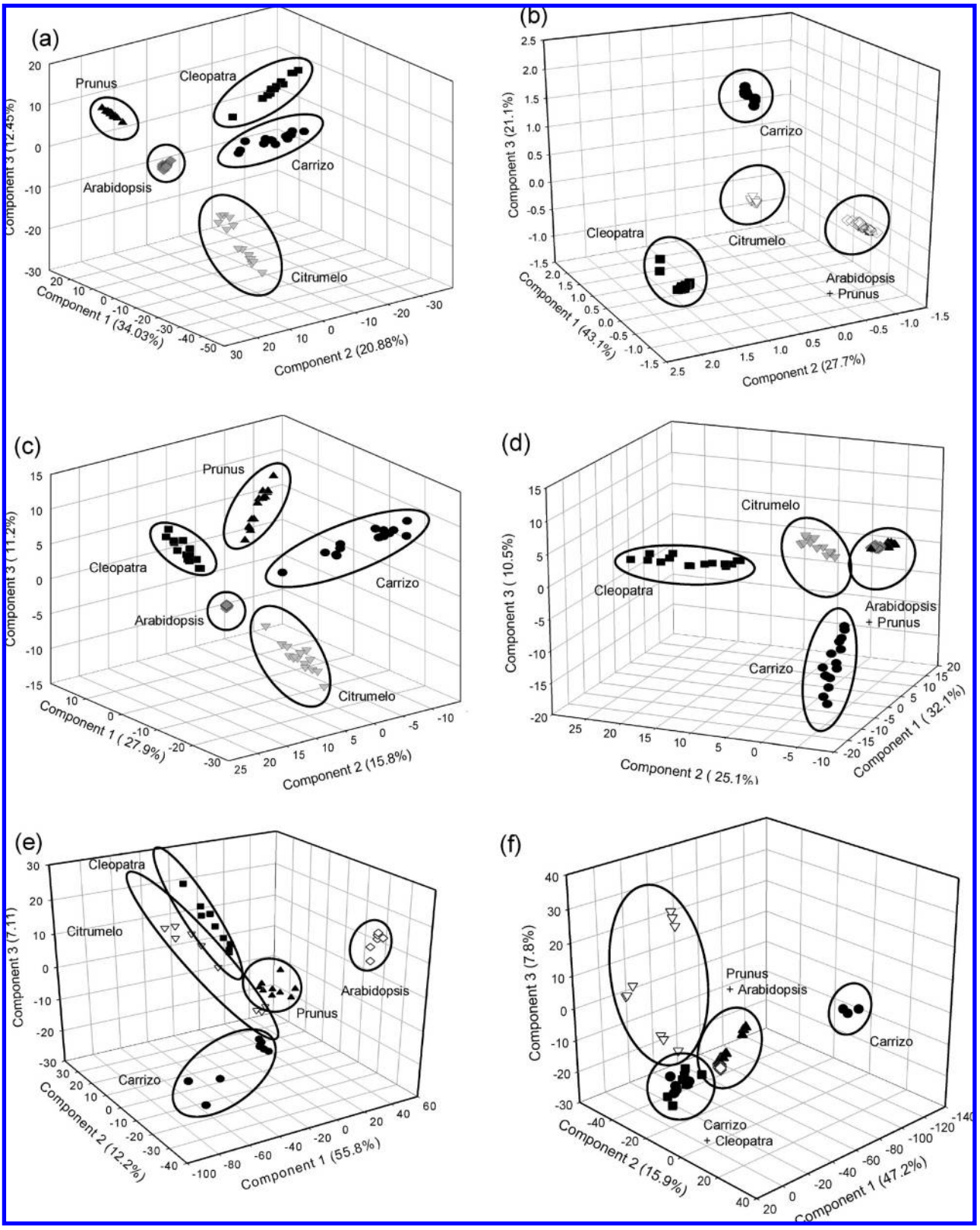


**Figure 2.** PCA plot showing three major sources of variability among citrumelo, Carrizo, Cleopatra, *Prunus*, and *Arabidopsis* genotypes after LC-MS in positive mode of (**a**) polar and (**b**) nonpolar fractions; LC-MS in negative mode of (**c**) polar and (**d**) nonpolar fractions; GC-MS of (**e**) polar and (**f**) nonpolar fractions after derivatization. Percentage of variability explained is given on each axis.

external groups. *Citrus* demarcation was similarly achieved independently of the environmental conditions used to grow plants. Therefore, when *Citrus* genotypes were cultivated in two additional locations (with different environmental conditions) and studied as separate experiments, the ability to discriminate plant genotypes was maintained between the two techniques assayed. Similar results were obtained when all three locations were pooled in a single experimental design.

**Exclusive Features.** A number of chromatographic features, which were only consistently found in one genotype and not in the

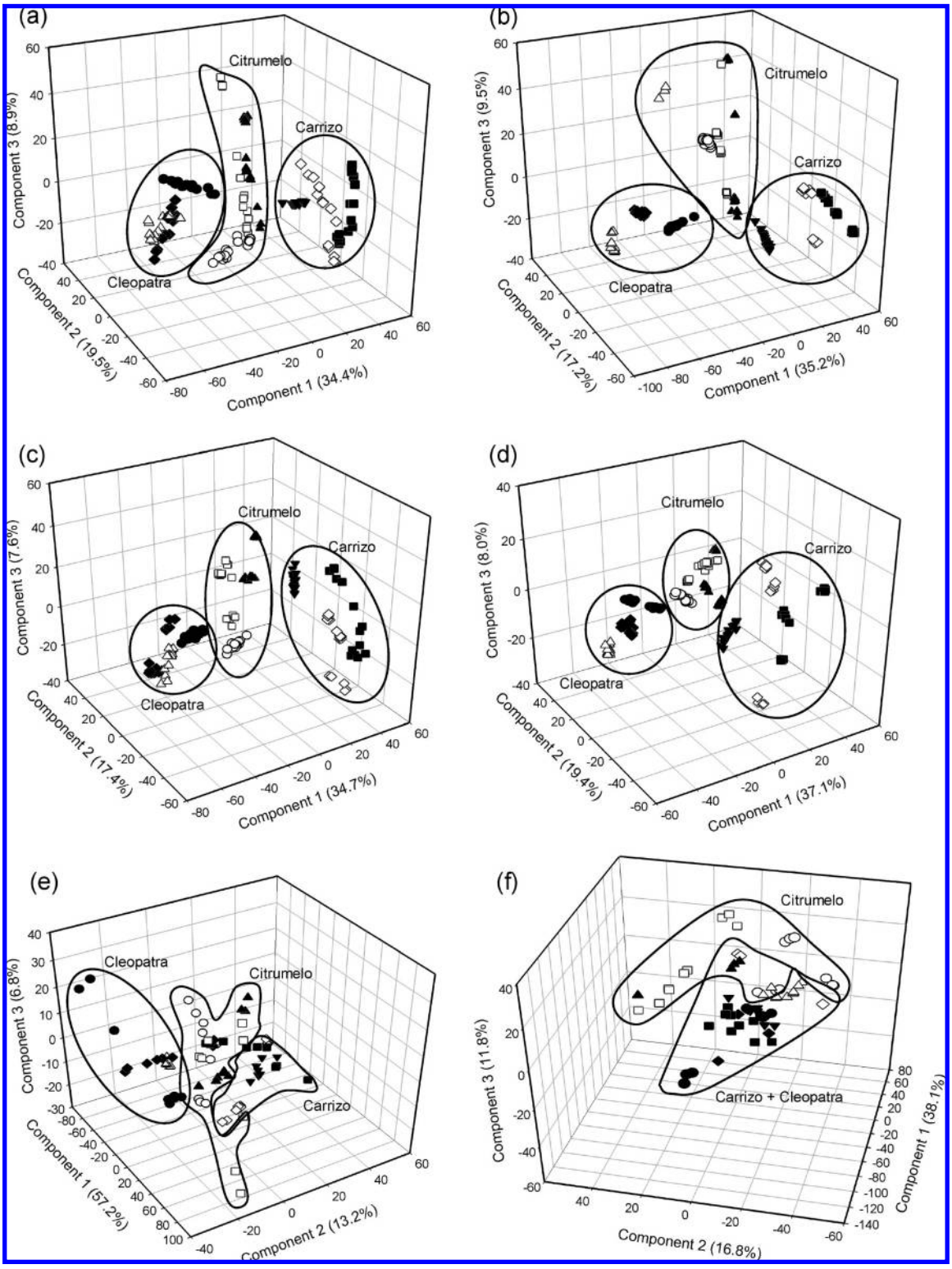


**Figure 3.** PCA plot showing three major sources of variability among citrumelo, Carrizo, and Cleopatra when data from experiments 1, 2, and 3 were pooled together after LC-MS in positive mode of (**a**) polar and (**b**) nonpolar fractions; LC-MS in negative mode of (**c**) polar and (**d**) nonpolar fractions; GC-MS of (**e**) polar and (**f**) nonpolar fractions after derivatization. Within Cleopatra, citrumelo, and Carrizo (△, ○, ●) represent experiment 1, (◆, □, ■) experiment 2, and (◆, ▲, ▼) experiment 3. Percentage of variability explained is given on each axis.

rest, were observed. Features were selected for their representation throughout plant samples. Using MS Excel, cells containing a value different from "NA" were labeled "1", whereas the rest were automatically labeled "0". Selected cells were grouped by genotypes and features chosen by their presence in at least 12 of 15 samples in LC-MS and 7 of 9 samples in GC-MS. Similar numbers of exclusive features were found in polar and nonpolar fractions (**Tables 1**−**3**). However, when each genotype was checked, there was no correlation between the amounts of exclusive features found in both fractions. This might indicate an uneven contribution of polar and nonpolar metabolites to the demarcation of genotypes, as observed previously after PCA and

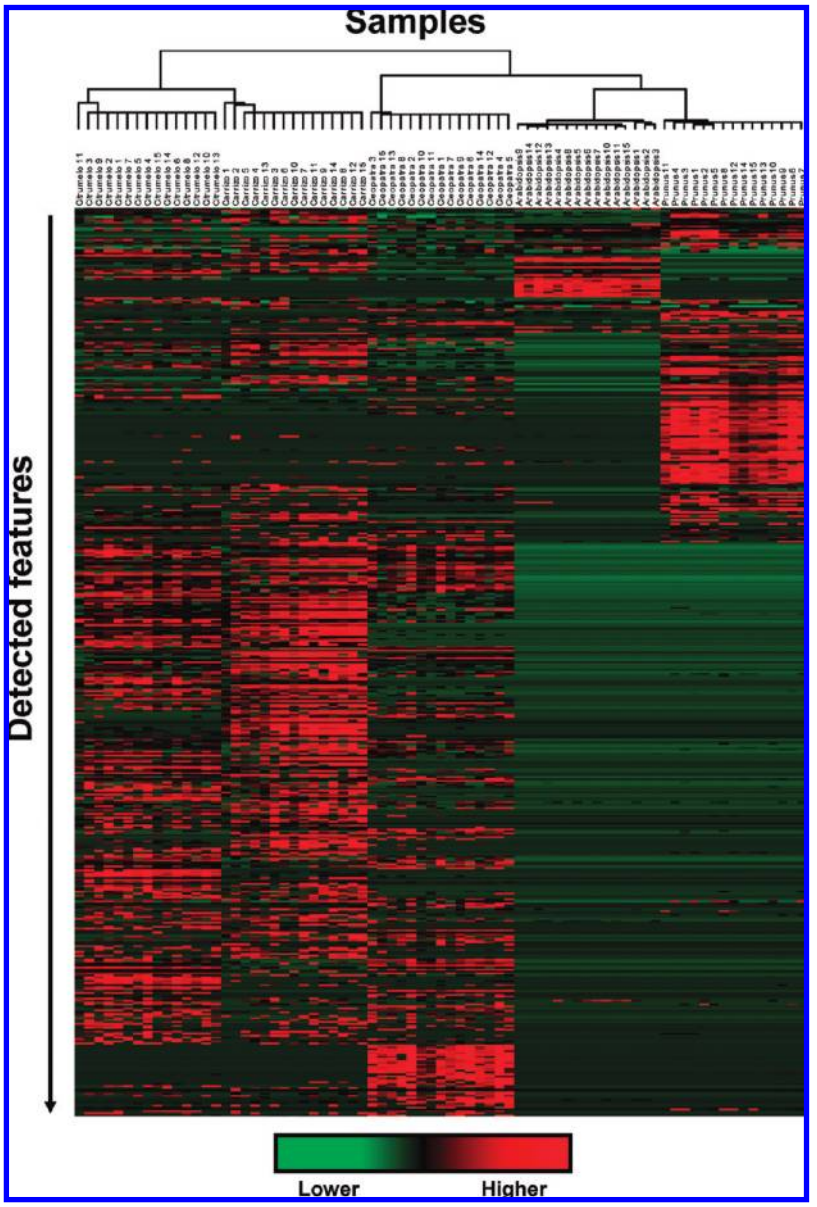


**Figure 4.** Green−black−red diagram from the standardized polar fractions analyzed by LC-MS in negative mode. Intensities are expressed as a color code, where red is the maximum intensity and green is the lowest.

HCA. In the positive polar profiles, Cleopatra and *Prunus* showed the maximum number of exclusive features (133 and 120, representing 11.8 and 10.7% of the total number of detected chromatographic features, respectively), whereas in the nonpolar fraction, Carrizo (150, 12.26%) and Cleopatra (105, 8.56%) showed the maximum number of exclusive features. In the negative ionization profiles of both polar and nonpolar fractions, the amount of exclusive features (**Table 2**) was much lower than in the positive profiles (**Table 1**). This could account for the positive ionization mode having a better performance than negative ionization in differentiating plant genotypes. In this sense, the percentage of exclusive features could be more important because, in most cases, it was related to a better discrimination of genotypes (i.e., in samples from Cleopatra and *Prunus*, a higher percentage of exclusive features was found in the LC-MS positive polar fraction, and, accordingly, both yielded tighter clusters than the rest of genotypes). This is of special interest when compositional differences demarcate cultivars or genetically engineered crops (*9*). In the GC-MS-based profiles from polar fractions, the number of exclusive features per genotype was in the same range as LC-MS profiles, whereas GC-MS profiles from nonpolar fractions yielded, in general, a much lower number (**Table 3**). Nevertheless, differences in the percentage of exclusive features per genotype were found in GC-MS profiles. In the polar fraction, 208 exclusive features were found in *Arabidopsis*, a number much higher than in the rest of genotypes. On the contrary, in the nonpolar fraction, 133 exclusive features were found in citrumelo and only 2 or 3 in the rest of the genotypes. This confirmed the uneven contribution of the polar and nonpolar fractions to the overall metabolite load. Similar numbers of exclusive signals were found in experiments 2 and 3 (data not shown). Despite the importance of exclusive features in genotype demarcation, it should be noted that the amount of these signals will be probably higher than the actual number of exclusive metabolites. Selected features could likely be a combination of metabolites, adducts, and fragments. In addition, it is well-known that GC-MS data produce many mass signals per peak and, therefore, it is particularly difficult to filter out a unique mass for a single metabolite.

As extracted from results, sample fractionation using C18 cartridges appears to be a suitable methodology to evaluate the

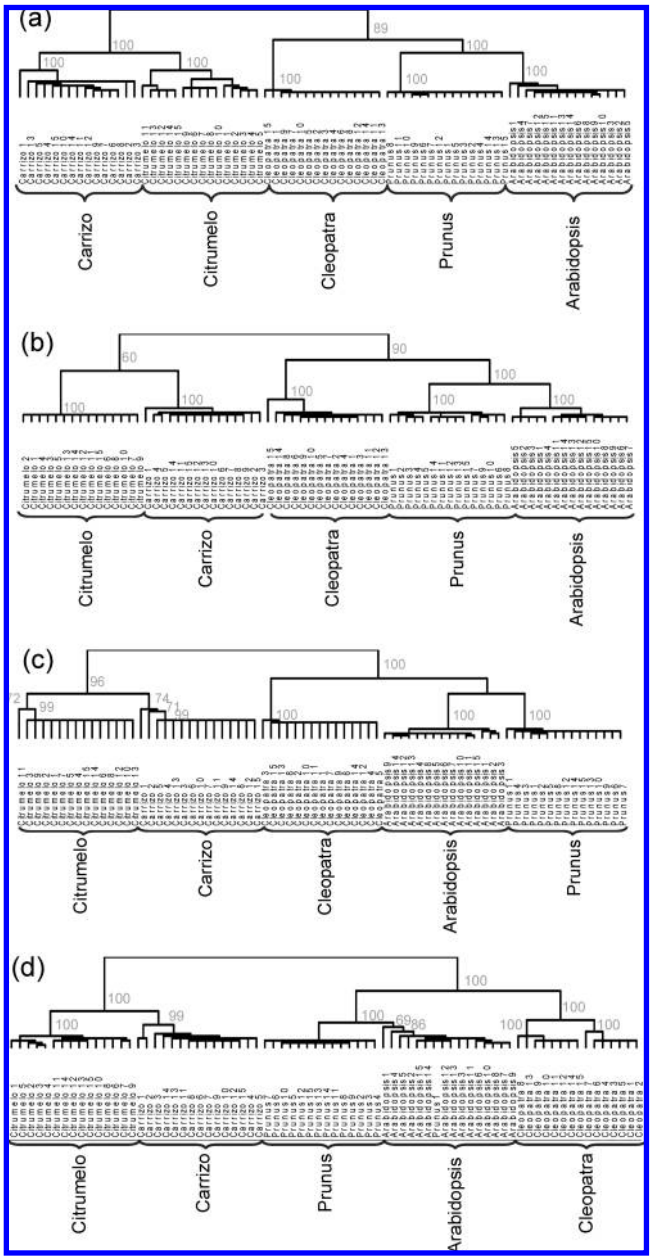


**Figure 5.** Hierarchical trees of the five genotypes based on metabolite profile patterns after extraction of signals corresponding to LC-MS in positive mode of (**a**) polar and (**b**) nonpolar fractions; LC-MS in negative mode of (**c**) polar and (**d**) nonpolar fractions. On selected nodes, values indicate bootstrap score values.

relative contribution of polar and nonpolar metabolites to genotypic differences. It is widely accepted that GC-MS is better suited for metabolites derived from the primary metabolism such as sugars or amino acids, whereas reversed phase LC-MS covers a higher percentage of the secondary metabolism, excluding highly polar compounds. Secondary metabolism exhibits a myriad of chemically diverse metabolites that are specific to the different plant genotypes (such as polyphenols, alkaloids, glucosinolates). However, previous reports highlighted the importance of primary polar metabolites (such as sugars, TCA intermediates, amino acids, and/or polyalcohols) as taxonomical and physiological traits (*23*). To undeniably prove the importance of secondary versus primary metabolism in our experimental system, further work should be directed to identify a subset of metabolites from both types of metabolism and study their relative coverage by LC-MS or GC-MS.

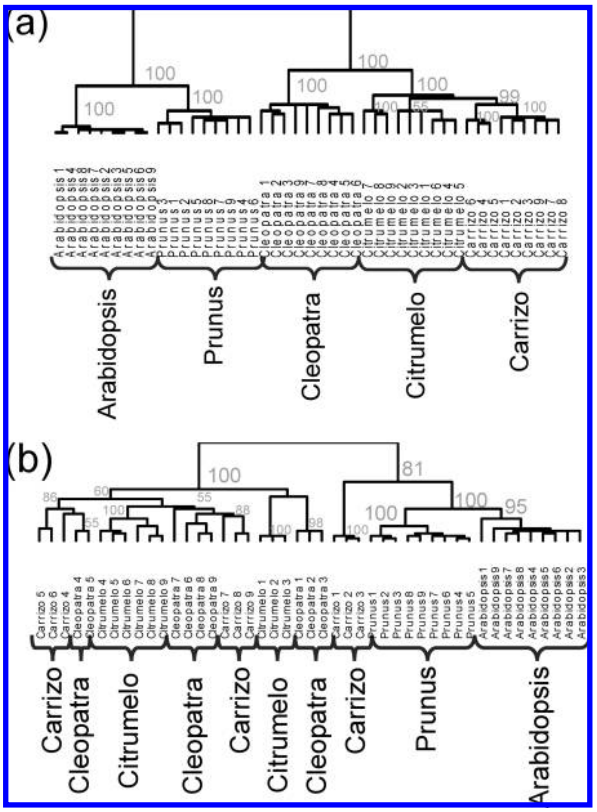


**Figure 6.** Hierarchical trees of the five genotypes based on metabolite profile patterns after extraction of signals corresponding to GC-MS of (**a**) polar and (**b**) nonpolar fractions after derivatization. On selected nodes, values indicate bootstrap score values.

**Table 1.** Exclusive Features from Positive Electrospray LC-MS Data

| | LC-MS positive polar fraction | | LC-MS positive nonpolar fraction | |
|---|---|---|---|---|
| | amount[a] | percentage[b] | amount[a] | percentage[b] |
| Carrizo | 87 | 7.75 | 150 | 12.26 |
| Ccitrumelo | 54 | 4.81 | 100 | 8.18 |
| Cleopatra | 133 | 11.85 | 105 | 8.59 |
| *Arabidopsis* | 37 | 3.30 | 74 | 6.05 |
| *Prunus* | 120 | 10.70 | 44 | 3.60 |
| total | 1122 | | 1223 | |

[a] Total amount of exclusive features in positive LC-MS data from experiment 1. [b] Percentage of exclusive features in each genotype from the total of detected features.

**Table 2.** Exclusive Features from Negative Electrospray LC-MS Data

| | LC-MS negative polar fraction | | LC-MS negative nonpolar fraction | |
|---|---|---|---|---|
| | amount[a] | percentage[b] | amount[a] | percentage[b] |
| Carrizo | 9 | 1.93 | 7 | 1.83 |
| citrumelo | 7 | 1.50 | 13 | 3.39 |
| Cleopatra | 17 | 3.64 | 31 | 8.09 |
| *Arabidopsis* | 9 | 1.93 | 2 | 0.52 |
| *Prunus* | 35 | 7.49 | 8 | 2.09 |
| total | 467 | | 383 | |

[a] Total amount of exclusive features in negative LC-MS data from experiment 1. [b] Percentage of exclusive features in each genotype from the total of detected features.

**Table 3.** Exclusive Features from GC-MS Data

| | GC-MS polar fraction | | GC-MS nonpolar fraction | |
|---|---|---|---|---|
| | amount[a] | percentage[b] | amount[a] | percentage[b] |
| Carrizo | 18 | 0.67 | 2 | 0.10 |
| citrumelo | 2 | 0.07 | 133 | 6.74 |
| Cleopatra | 20 | 0.75 | 3 | 0.15 |
| *Arabidopsis* | 208 | 7.77 | 2 | 0.10 |
| *Prunus* | 28 | 1.05 | 0 | 0.00 |
| total | 2677 | | 1973 | |

[a] Total amount of exclusive features in GC-MS data from experiment 1. [b] Percentage of exclusive features in each genotype from the total of detected features.

It was beyond the scope of this work to annotate mass signals as individual metabolites as they were used only as markers for genotype classification. The metabolic differences highlighted in this work are consistent with previous data where significant physiological differences were found among the same *Citrus* genotypes in response to flooding and salt stress (*24*, *25*). In those studies, citrumelo and Carrizo *Citrus* genotypes exhibited similar physiological responses to stress, whereas Cleopatra showed a more specific behavior. Under salinity stress, Cleopatra exhibited a more tolerant behavior than that of Carrizo and citrumelo, which was related to a rapid reduction of net photosynthetic rate, stomatal conductance, and performance of PSII and photosynthetic efficiency. The other *Citrus* genotypes showed a lower capability to adjust stomata and PSII to the new conditions (*25*). Under soil flooding, Carrizo and citrumelo showed similar responses, whereas Cleopatra was unable to activate the cellular antioxidant machinery (*24*).

## ACKNOWLEDGMENT

We thank Dr. Dierk Scheel, Dr. Edda Von Roepenack-Lahaye, and Dr. Christoph Böttcher from the Leibniz Institute of Plant Biochemistry (Halle-Saale) for their help and critical review of the manuscript. HPLC-QTOF analyses were carried out at the SCIC facility of the Universitat Jaume I.

**Supporting Information Available:** **File 1:** PCA plot showing sources of variability among citrumelo, Carrizo, and Cleopatra in experiment 2 after LC-MS in positive mode of (a) polar and (b) nonpolar fractions; LC-MS in negative mode of (c) polar and (d) nonpolar fractions; GC-MS of (e) polar and (f) nonpolar fractions after derivatization. Percentage of variability explained is given on each axis. **File 2:** Hierarchical trees of the three *Citrus* genotypes based on metabolite profile patterns from experiment 2 after extraction of signals corresponding to LC-MS in positive mode of (a) polar and (b) nonpolar fractions; LC-MS in negative mode of (c) polar and (d) nonpolar fractions; GC-MS of (e) polar and (f) nonpolar fractions after derivatization. On selected nodes, values indicate bootstrap score values. **File 3:** PCA plot showing three major sources of variability among citrumelo, Carrizo, and Cleopatra in experiment 3 after LC-MS in positive mode of (a) polar and (b) nonpolar fractions; LC-MS in negative mode of (c) polar and (d) nonpolar fractions; GC-MS of (e) polar and (f) nonpolar fractions after derivatization. Percentage of variability explained is given in each axis. **File 4:** Hierarchical trees of the three *Citrus* genotypes based on metabolite profile patterns from experiment 3 after extraction of signals corresponding to LC-MS in positive mode of (a) polar and (b) nonpolar fractions; LC-MS in negative mode of (c) polar and (d) nonpolar fractions; GC-MS of (e) polar and (f) nonpolar fractions after derivatization. On selected nodes, values indicate bootstrap score values. **File 5:** XCMS outputs in Excel format from experiment 1 containing average retention time and accurate mass for each feature and area values from LC-MS polar profiles in positive electrospray mode. **File 6:** XCMS outputs in Excel format from experiment 1 containing average retention time and accurate mass for each feature and area values from LC-MS nonpolar profiles in positive electrospray mode. **File 7:** XCMS outputs in Excel format from experiment 1 containing average retention time and accurate mass for each feature and area values from LC-MS polar profiles in negative electrospray mode. **File 8:** XCMS outputs in Excel format from experiment 1 containing average retention time and accurate mass for each feature and area values from LC-MS nonpolar profiles in negative electrospray mode. **File 9:** XCMS outputs in Excel format from experiment 1 containing average retention time and accurate mass for each feature and area values from GC-MS polar profiles. **File 10:** XCMS outputs in Excel format from experiment 1 containing average retention time and accurate mass for each feature and area values from GC-MS nonpolar profiles. **File 11:** Box−whisker plots from LC-MS profiles in experiment 1. **File 12:** Box−whisker plots from GC-MS profiles in experiment 1. This material is available free of charge via the Internet at http://pubs.acs.org.

## LITERATURE CITED

(1) Tikunov, U.; Lommen, A.; de Vos, C. H. R.; Verhoeven, H. A.; Bino, R. J.; Hall, R. D.; Bovy, A. G. A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. *Plant Physiol.* **2005**, *139*, 1125–1137.

(2) Schauer, N.; Semel, Y.; Roessner, U.; Gur, A.; Balbo, I.; Carrari, F.; Pleban, T.; Perez-Melis, A.; Bruedigam, C.; Kopka, J.; Willmitzer, L.; Zamir, D.; Fernie, A. R. Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. *Nat. Biotechnol.* **2006**, *24*, 447–454.

(3) Davey, M. P.; Burrell, M. M.; Woodward, F. I.; Quick, W. P. Population-specific metabolic phenotypes of *Arabidopsis lyrata* ssp. *petraea*. *New Phytol.* **2008**, *177*, 380–388.

(4) Roessner, U.; Patterson, J. H.; Forbes, M. G.; Fincher, G. B.; Langridge, P.; Bacic, A. An investigation of boron toxocity in barley using metabolomics. *Plant Physiol.* **2006**, *142*, 1087–1101.

(5) Djoukeng, J. D.; Arbona, V.; Argamasilla, R.; Gomez-Cadenas, A. Flavonoid profiling in leaves of *Citrus* genotypes under different environmental situations. *J. Agric. Food Chem.* **2008**, *56*, 11087–11097.

(6) Broeckling, C. D.; Huhman, D. V.; Farag, M. A.; Smith, J. T.; May, G. D.; Mendes, P.; Dixon, R. A.; Sumner, L. W. Metabolic profiling of *Medicago truncatula* cell cultures reveals the effects of biotic and abiotic elicitors on metabolism. *J. Exp. Bot.* **2005**, *56*, 323–336.

(7) Boccard, J.; Grata, E.; Thiocone, A.; Gauvrit, J. Y.; Lanteri, P.; Carrupt, P. A.; Wolfender, J. L.; Rudaz, S. Multivariate data analysis of rapid LC-TOF/MS experiments from *Arabidopsis thaliana* stressed by wounding. *Chem. Intell. Lab. Syst.* **2007**, *86*, 189–197.

(8) Allen, J.; Davey, H. M.; Broadhurst, D.; Heald, J. K.; Rowland, J. J.; Oliver, S. G.; Kell, D. B. High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nat. Biotechnol.* **2003**, *21*, 692–696.

(9) Harrigan, G. G.; Martino-Catt, S.; Glenn, K. C. Metabolomics, metabolic diversity and genertic variation in crops. *Metabolomics* **2007**, *3*, 259–272.

(10) Roessner, U.; Luedemann, A.; Brust, D.; Fiehn, O.; Linke, T.; Willmitzer, L.; Fernie, A. R. Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* **2001**, *13*, 11–29.

(11) Robinson, A. R.; Ukrainetz, N. K.; Kang, K. Y.; Mansfield, S. D. Metabolite profiling of Douglas-fir (*Pseudotsuga menziesii*) field trials reveals strong environmental and weak genetic variation. *New Phytol.* **2007**, *174*, 762–773.

(12) Enot, D. P.; Beckman, M.; Draper, J. Detecting a difference—assessing generalisability when modelling metabolome fingerprint data in longer term studies of genetically modified plants. *Metabolomics* **2007**, *3*, 335–347.

(13) Fiehn, O.; Kopka, J.; Dormann, P.; Altmann, T.; Trethewey, R. N.; Willmitzer, L. Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* **2000**, *18*, 1157–1161.
(14) De Vos, R. C.; Moco, S.; Lommen, A.; Keurentjes, J. J.; Bino, R. J.; Hall, R. D. Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. *Nat. Protoc.* **2007**, *2*, 778–791.
(15) von Roepenack-Lahaye, E.; Degenkolb, T.; Zerjeski, M.; Franz, M.; Roth, U.; Wessjohann, L.; Schmidt, J.; Scheel, D.; Clemens, S. Profiling of *Arabidopsis* secondary metabolites by capillary liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry. *Plant Physiol.* **2004**, *134*, 548–559.
(16) Böttcher, C.; Von Roepenack-Lahaye, E.; Willscher, E.; Scheel, D.; Clemens, S. Evaluation of matrix effects in metabolite profiling based on capillary liquid chromatography electrospray ionization quadrupole time-of-flight mass spectrometry. *Anal. Chem.* **2007**, *79*, 1507–1513.
(17) Katajamaa, M.; Orêsiç, M. Data processing for mass spectrometry-based metabolomics. *J. Chromatogr., A* **2007**, *1158*, 318–328.
(18) Katajamaa, M.; Orêsiç, M. Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics* **2005**, *6*, 179.
(19) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* **2006**, *78*, 779–787.
(20) van den Berg, R. A.; Hoefsloot, H. C.; Westerhuis, J. A.; Smilde, A. K.; van der Werf, M. J. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* **2006**, *7*, 142.
(21) Bretó, M. P.; Ruiz, C.; Pina, J. A.; Asins, M. J. The diversification of *Citrus clementina* Hort. ex Tan., a vegetatively propagated crop species. *Mol. Phylogenet. Evol.* **2001**, *21*, 285–293.
(22) Ancillo, G.; Gadea, J.; Forment, J.; Guerri, J.; Navarro, L. Class prediction of closely related plant varieties using gene expression profiling. *J. Exp. Bot.* **2007**, *58*, 1927–1933.
(23) Merchant, A.; Richter, A.; Popp, M.; Adams, M. Targeted metabolite profiling provides a functional link among eucalypt taxonomy, physiology and evolution. *Phytochemistry* **2006**, *67*, 402–408.
(24) Arbona, V.; Hossain, Z.; Lopez-Climent, M. F.; Perez-Clemente, R. M.; Gomez-Cadenas, A. Antioxidant enzymatic activity is linked to waterlogging stress tolerance in Citrus. *Physiol. Plant.* **2008**, *132*, 452–466.
(25) López-Climent, M. F.; Arbona, V.; Pérez-Clemente, R. M.; Gómez-Cadenas, A. Relationship between salt tolerance and photosynthetic machinery performance in *Citrus*. *Environ. Exp. Bot.* **2008**, *62*, 176–184.
(26) Arbona, V.; Lopez-Climent, M. F.; Mahouachi, J.; Perez-Clemente, R. M.; Abrams, S. R.; Gomez-Cadenas, A. Use of persistent analogs of abscisic acid as palliatives against salt-stress induced damage in *Citrus* plants. *J. Plant Growth Regul.* **2006**, *25*, 1–9.
(27) Schadt, E. E.; Li, C.; Ellis, B.; Wong, W. H. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J. Cell Biochem.* **2001**, *Suppl. 37*, 120–125.
(28) Hannah, M. A.; Wiese, D.; Freund, S.; Fiehn, O.; Heyer, A. G.; Hincha, D. K. Natural genetic variation of freezing tolerance in *Arabidopsis*. *Plant Physiol.* **2006**, *142*, 98–112.

**Received March 18, 2009. Revised manuscript received May 12, 2009. Accepted July 15, 2009. This work was funded by Fundació Bancaixa/ Universitat Jaume I and Ministerio de Educación y Ciencia through grants No. P1-1A2007-04 and AGL2007-65437-C04-03/AGR to V.A. and A.G-C., respectively. M.T. and D.I. were supported by grants INCO 015453, AGL2007-65437-C04-01/AGR and INIA RTA04-013/05-247.**