Untargeted metabolomic LC-HRMS fingerprinting of apple cultivars for the identification of biomarkers related to resistance to rosy apple aphid

Rosa M. Alonso-Salces, Luis A. Berrueta, Beatriz Abad-García, Andrea Sasía-Arriba, Carlos Asensio-Regalado, Enrique Dapena, and Blanca Gallo

Journal of Agricultural and Food Chemistry 2022 70 (41), 13071-13081

DOI: 10.1021/acs.jafc.2c00738

Untargeted metabolomic LC-HRMS fingerprinting of apple cultivars for the identification of biomarkers related to resistance to rosy apple aphid

Rosa M. Alonso-Salces[1,*], Luis A. Berrueta[2], Beatriz Abad-García[3], Andrea Sasía-Arriba[2], Carlos Asensio-Regalado[2], Enrique Dapena[4], Blanca Gallo[2]

[1] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), CIAS-IIPROSAM, Facultad de Ciencias Exactas y Naturales, Universidad Nacional de Mar del Plata (UNMdP), Funes 3350, B7602AYL Mar del Plata, Argentina

[2] Departamento de Química Analítica, Facultad de Ciencia y Tecnología, Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU), Apdo 644, E-48080 Bilbao, Spain.

[3] Servicio Central de Análisis, Servicios Generales de Investigación (SGIker), Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU), Bº Sarriena s/n, E-48940 Leioa, Spain.

[4] Programa de Fruticultura, Servicio Regional de Investigación y Desarrollo Agroalimentario (SERIDA), Ctra. de Oviedo s/n, Apdo. 13, E-33300 Villaviciosa, Spain.

[*] Corresponding author:
Dr. Rosa María Alonso-Salces
E-mail: rosamaria.alonsosalces@gmail.com

Abstract

Liquid chromatography high resolution mass spectrometry fingerprinting together with pattern recognition techniques were used to determine the metabolites involved in the susceptibility of apple cultivars to rosy apple aphid (RAA). Pre-processing of UHPLC-ESI-QToF/MS raw data of resistant and susceptible apple cultivars was carried out with XCMS and CAMERA packages. Univariate statistical tools and multivariate data analysis highlighted significant different profiles of the apple metabolomes according to their tolerance to RAA. Optimized and cross-validated PLS-DA and OPLS-DA models confirmed *trans*-4-caffeoylquinic acid and 4-*p*-coumaroylquinic acid as biomarkers for the identification of resistant and susceptible apple cultivars to RAA, and disclosed that only hydroxycinnamic acids are involved in the disease susceptibility of cultivars. In this sense, the final steps of the biosynthesis of caffeoylquinic (CQA) and *p*-coumaroylquinic (*p*-CoQA) acids become decisive since the isomerization of 5-CQA to 4-CQA is favored in resistant cultivars whereas 5-*p*-CoQA to 4-*p*-CoQA in susceptible cultivars.

Keywords:

Malus domestica, Dysaphis plantaginea, plant resistance, UHPLC-QToF/MS, chemometrics, hydroxycinnamic acid

1. Introduction

Dysaphis plantaginea Pass. (Hemiptera: Aphididae), causal agent of Rosy apple aphid (RAA), is one of the major damaging insects affecting apple, *Malus domestica* (Borkh.), in Europe, North America, North Africa and Asia.[1] RAA leads to serious damage on shoots, leaves and fruits that remain small and deformed, reducing their commercial value and leading to significant economic losses. The biological control of RAA through naturally occurring predators is still not satisfactorily effective in reducing the aphid population, therefore the main control approach relies on insecticide sprays.[2] Therefore, it is urgent to find new strategies for the sustainable control of RAA.

Sustainable control approaches for the selection of new resistant cultivars are the marker-assisted breeding programs, which reduces simultaneously the collateral effects of pesticides.[1, 3] Regarding the resistance enhancement to different pests and diseases, several breeding programs have been developed.[4] Among them, a cider apple breeding program was started in 1999 by the Regional Service for Agri-Food Research and Development (SERIDA) of Asturias (Spain), which aimed the implementation of new cultivars of cider apple of high interest in terms of fruit quality; resistance to the fungus *Venturia inaequalis* (Cke), causal agent of apple scab, and RAA; low susceptibility to fire blight, caused by the bacterium *Erwinia amylovora* (Burrill); and regular bearing.[4] The apple cultivar 'Florina' is resistant to apple scab and RAA, only slightly susceptible to fire blight, and tolerant to red mite *Panonychus ulmi*.[5] The resistance of 'Florina' to RAA is typified by both tolerance and antibiosis.[6] Susceptible cultivars when infested by RAA show a typical leaf and shoot deformations, which are not noticed in 'Florina'. However, it is observed a decrease in the fertility and an increase in the mortality of RAA fed on 'Florina' leaves, as well as RAA moving from the leaves to the stems. This suggests that leaves release repellent compounds and/or hinder stylet penetration.[3] A single dominant resistance gene (*Dp-fl*) located at the distal end of linkage group 8 of the apple genome was identified studying the inheritance of 'Florina' resistance to RAA in segregating progenies.[7]

Plant resistance mechanisms and plant responses to biotic-stress have been associated with phenolic compounds. The accumulation of benzoic acid in apple fruit after inoculation with *Nectria galligena,* causing latent infections, was observed.[8] A fast and localized accumulation of phenylpropanoids was related to scab resistance.[9-11] Furthermore, flavan-3-o1s concentrations were larger in apple leaf tissues of scab-resistant cultivars.[9] Moreover, the phloridzin/flavanol ratio appeared to be higher in cultivars susceptible to scab, whereas two *p*-coumaric acid derivatives were in higher levels in cultivars with the polygenic resistance character.[12] Besides, a lower susceptibility to scab and fire blight was related to higher amounts of 3-hydroxyphlorizin.[13] Hydroxycinnamic acids, particularly 4-caffeoylquinic acid (4-CQA) and 4-*p*-coumaroylquinic acid (4-*p*-CoQA), were proved to be related to apple cultivar resistance to RAA using a target profiling approach.[14] Forty individual phenolic compounds (flavan-3-ols, hydroxycinnamic acids, dihydrochalcones and flavonols) were determined in the apple juice of an experimental population derived from a controlled cross of 'Meana' and 'Florina' (MxF) created and maintained by SERIDA breeding program. This population was selected because of the particular characteristics of the parents: 'Florina', due to its resistance to scab and RAA and its high tolerance to fire blight; and 'Meana' is a bitter sharp cider apple cultivar included among cultivars of the Protected Denomination of Origin (PDO) 'Sidra de Asturias'/'Sidra d'Asturies' (EU No PDO-ES-0260-AM01 – 31.10.2017) with a high content of phenolic compounds. The tolerance of apple cultivars to RAA was concluded to involve the phenylpropanoid pathway; the isomerization of hydroxycinnamic acids being the metabolic reactions playing a key role in determining that a cultivar is resistant or susceptible to RAA. In such a profiling approach, only the compounds determined were considered as possible biomarkers, and the presence of other compounds related to RAA resistance were not able to be disclosed. Moreover, profiling is time consuming since it requires that all analyzed compounds have to be identified by comparison of their retention time, UV-visible and MS spectra with those of the standards available and data in literature, as well as quantified (including peak integration and external standard calibration). In the present study, an untargeted metabolomic

approach is proposed to identify biomarkers related to diseases, in particular to the resistance of apple cultivars to RAA. This is a fingerprinting approach that presents several advantages: (*i*) it is less time consuming, since the identification and quantitation steps required in target analysis for profiling is not performed; (*ii*) compound identification is only carried out for the biomarkers revealed by multivariate data analysis; and last but not least (*iii*) it allows the discovery of new unknown biomarkers related to the disease under research.

2. Material and methods

2.1. Chemicals, solvents and standards

Water, methanol, acetonitrile and formic acid were of Optima® LC/MS grade (Fisher Scientific, Fair Lawn, NJ, USA). Glacial acetic acid provided by Merck (Darmstadt, Germany) was of Suprapur® quality. Sodium fluoride (Fluka Chemie, Buchs, Switzerland) and ascorbic acid (Panreac, Barcelona, Spain) were of ACS grade. Leucine Enkephalin acetate hydrate, sodium formate solution, caffeic acid, 5′-O-caffeoylquinic acid and *p*-coumaric acid were provided by Sigma-Aldrich Chemie (Steinheim, Germany). Standard stock solutions of phenolic compounds were prepared in methanol and dilutions of stock solutions in methanol-water-acetic acid (30:69:1, v/v/v).

2.2. Samples

Apples were harvested at the optimum stage of maturity at SERIDA in Villaviciosa (Asturias, Spain) in two harvest seasons. The set of samples studied included a population of 130 individuals from a cross of 'Meana' x 'Florina' and the two parents. Three apple batches of each cultivar were processed as previously described.[14] Quality control (QC) samples consisted of a pool of 1 mL-aliquot of each juice sample analyzed in every chromatographic sequence. An aliquot of 0.5 mL of each sample was diluted with 1.5 mL of methanol–water–acetic acid (30:69:1, v/v/v) with 2 g/L of ascorbic acid (w/v), vortexed and filtered through a 0.45 μm PTFE filter (Waters, Milford, MA, USA) prior to injection into the ultrahigh-performance liquid chromatography-diode array detector-

electrospray ionization quadrupole time of flight/mass spectrometer (UHPLC-DAD-ESI-QToF/MS) system.

2.3. UHPLC-DAD-ESI-QToF/MS analysis

UHPLC-DAD-ESI-QToF/MS analysis were performed using a Waters ACQUITY UPLC™ system, equipped with a binary solvent delivery pump, an autosampler, a column compartment, a DAD detector, and a Waters SYNAPT^{TM} G2 HDMS spectrometer. A reverse-phase column (Waters Acquity UPLC BEH C18, 100 mm × 2.1 mm i.d., particle size 1.7 μm) and and a pre-column (Waters Acquity UPLC BEH C18 VanGuard^{TM}, 1.7 μm) were used at 40 °C. Mobile phases were 0.1 % (v/v) acetic acid in water (A) and 0.1 % (v/v) acetic acid in methanol (B). The elution conditions applied were 0–1.60 min, 2 % B isocratic; 1.60–2.11 min, 0–8 % B linear gradient; 2.11–8.80 min, 8 % B isocratic; 8.80–9.80 min, 8–10 % B linear gradient; 9.80–17.00 min, 10 % B isocratic; 17.00–22.00 min, 10–20 % B linear gradient; 22.00–23.40 min, 20–23 % B linear gradient; 23.40–28.40 min, 23–35 % B linear gradient; 28.40–30.40 min, 35–51 % B linear gradient; 30.40–31.40 min, 51–100 % B linear gradient; 31.40–32.40 min, 100 % B isocratic; and finally recondictioning of the column with 2 % B. Flow rate was 0.35 mL/min; injection volume, 5 μL; and autosampler temperature, 4 °C. UV-visible spectra were recorded from 210–500 nm (20 Hz, 1.2 nm resolution).

All MS data acquisitions were performed on a Waters SYNAPT^{TM} G2 HDMS with a quadrupole time of flight (QToF) configuration equipped with an ESI source operating in positive or negative ion modes. The capillary voltage was set to 1.0 kV (ESI+) or 0.5 kV (ESI−). Nitrogen was used as the desolvation and the cone gas at flow rates of 1000 L/h and 10 L/h respectively. The source and desolvation temperatures were 120 ºC and 500 ºC respectively. Leucine-enkephalin solution (2 ng/μL) in formic acid 0.1% (v/v) in acetonitrile-water (50:50, v/v) was used for the lock mass correction (*m/z* 556.2771 and 278.1141 were monitored at scan time 0.3 s, scan frequency 10 s, scans to average 3, mass window ± 0.5 Da, cone voltage 9 V, at a flow rate 10 μL/min). Data

acquisition was recorded in the mass range 50–1200 *u* in resolution mode (FWHM ≈ 18000) with a scan time of 0.1 s and an inter-scan delay of the 0.024 s, and automatically corrected during acquisition based on the lock mass. Before analysis, the mass spectrometer was mass calibrated with the sodium formate solution. For instrument control, data acquisition and processing Waters MassLynx^{TM} software Version 4.1 was used.

For MS^{E} mode analysis, the cone voltage was set to 20 V (ESI+) or 30 V (ESI−) and the quadrupole operated in a wide band RF mode only. Two discrete and independent interleaved acquisition functions were automatically created. The first function, set at 4 eV in both the trap cell of the T-Wave and the transfer cell, collects low energy or unfragmented data, while the second function collects high energy or fragmented data using 4 eV in the trap cell and a collision ramp from 10 to 40 eV in the transfer cell. Argon gas was used for CID and data were recorded in centroid mode. MS/MS experiments were performed at the optimum cone voltages (from 10 to 40 V) which produced the maximum intensity for protonated molecule $[M+H]^+$ or deprotonated molecule $[M–H]^-$ in previous MS full scan experiments. Different collision energies were tested from 10 eV (low energy scans) to 40 eV (high energy scans) in the centroid mode. MS/MS data were collected at a range of *m/z* 50–1200 in the same conditions as described above. The identity of biomarkers was confirmed by means of the UHPLC-DAD-ESI-QToF-MS strategy previously reported.[15]

2.4. Phenotypic analysis of the resistance level to RAA

The response to RAA was evaluated after infestation with aphids at a greenhouse in Villaviciosa (Asturias, Spain). The number of replicates per individual plant genotype varied from three to eight depending on the cultivar/crossing. The same plant number of ‘Florina’ and ‘Golden Delicious’ cultivars were used as resistant and susceptible controls respectively. Plants were cultivated and infested as previously described.[14] Aphids for infestation were field-collected from different apple cultivars to capture some of the natural variability. Individuals from each cultivar were reared

separately on susceptible apple plants. Thus, several distinct populations of RAA were maintained in the laboratory.

To assess damage on plants, observations were made once a week from the day after the infestation to the end of the experiment, 21 days later. Shoot damage was coded from 0 to 5 based on Rat-Morris (1993): 0, no damage; 1, leaf slightly curled at the edge; 2, leaf curled longitudinally; 3, typical RAA leaf rolling; 4, from 2 to 5 typically-rolled leaves; and 5, more than 5 typically-rolled leaves.[6] Plants exhibiting shoot damage classes of 0, 1 or 2 were considered resistant and classes 3 to 5, susceptible.

2.5. Data analysis

2.6.1. Pre-processing and pre-treatment of raw data

UHPLC-ESI-QToF/MS data were converted from Waters files (*.raw*) to machine-independent data format NetCDF files (*.cdf*) using DataBridge 3.5 converter from MassLynx (Waters, Milford, USA) and grouped according to the sample category (resistant, susceptible). The NetCDF files were pre-processed using XCMS 1.42.0 (Metlin, La jolla, CA, USA) for R package (V 3.2.2), in order to convert the three-dimensional LC-MS data (retention time, *m/z*, abundance) into a table of time-aligned detected features (variables), i.e. each feature is a pair of retention time and *m/z* values, and their signal abundances for each sample. The workflow consisted of several steps: peak picking, peak grouping, peak alignment, missing value imputation, filtering and QC correction. The *centWave* feature selection algorithm was used for peak identification, with the following parameters for the function *xcmsSet*: peak width ranging from 3 to 90 s and 10 ppm of mass tolerance. Peaks were grouped to match detected features across samples by the *group* function before peak alignment step using *retcor* function. Then, the *group* function (bandwidth, 10 s) regrouped aligned peaks after retention time correction, and the *fillPeaks* function created a list with all these features (*m/z*–retention time pairs) and the peak areas. The *fillPeaks* function is applied to manage missing values, which can be due to three main reasons: (*i*) metabolites not present in all

the analyzed samples, (*ii*) metabolites in low concentration leading to poor signals close the analytical background, and (*iii*) non-identified metabolite peaks at the feature detection step because the algorithm's criteria are not fulfilled.[16] Peaks were labelled as *MxxxxTyyyy*, with *xxx* referring to its nominal mass and *yyyy* to the corrected retention time in seconds. Metabolite features were defined as ions with unique *m/z* and retention time values. The CAMERA 3.10 package for R was used to annotate isotope and adduct peaks.[17] The data matrix generated, containing 3593 features, was exported as a text file (*.csv*).

2.6.2. Signal drift correction

Along the run of the sample sequence, the MS signal intensity drops as a results of the contamination or dirtying of the ion source components of the mass spectrometer, which has to be considered. The procedure employed to correct the MS signal drift was based on a correction factor for each feature which will vary from sample to sample, namely featured-based signal correction.[18] The correction factor was obtained from the linear regression of the feature signal (peak area) in the QC samples against the injection order. The corrected signal was calculated according to Eq. 1.

$$x'_{i,j} = \frac{x_{i,j}}{f_{i,j}} \cdot x'_{i,QC-1} \quad (1)$$

where $x'_{i,j}$ is the corrected signal of the feature i in the sample j, and $x_{i,j}$ is the pre-processed signal. The correction factor $f_{i,j}$ is calculated as the theoretical signal value of feature i interpolating the order of injection of sample j in the linear regression. The result is multiplied by $x'_{i,QC\text{-}1}$, which is the corrected signal for feature i in the first QC sample injected, *QC-1*, in order to recover the original feature dimensionality.

2.6.3. Statistical data analysis

Datasets containing both the QC samples and the apple cultivar samples or only the latter samples were analyzed by univariate procedures (Shapiro-Wilk test, Levene test, ANOVA, Fisher index and box and whisker plots), and multivariate techniques: unsupervised such as Principal Component

Analysis (PCA) and Hierarchical Cluster Analysis (HCA); and supervised such as Partial Least Square-Discriminant Analysis (PLS-DA) and its modification with Orthogonal Projections to Latent Structures (O-PLS).[19, 20] Data analysis was performed by means of the statistical software packages SPSS Statistics 17.0 (IBM Corporation, Armonk, NY, 1993-2007), Statistica 6.1 (StatSoft Inc., Tulsa, OK, 1984–2004), The Unscrambler 9.1 (Camo Process AS, Oslo, Norway, 1986–2004), SIMCA 14.1 (Umetrics, Umeå/Malmö, Sweden, 2015) and Mass Profiler Professional B14.8 (Agilent Technologies, Barcelona, Spain, 2016). The data matrix of original features was logarithmic transformed: $\log_{10}(x_{ij} + 1)$. Then, data was scaled (autoscaling, mean-centering, Pareto scaling) prior to perform multivariate data analysis using leave-one out or 3-fold cross-validation.[21]

The presence of outliers in the dataset was analyzed by PCA. PCA allows reduction of the number of features retaining the maximum amount of variability present in data in order to provide a partial visualization of data structure in a reduced dimension.

HCA is a pattern recognition technique that is used to reveal the structure residing in a data set and disclose the natural groupings existing between samples characterized by the values of a set of measured features. Sample similarities were calculated on the basis of the Euclidean distance, and the Wards hierarchical agglomerative method was used to establish clusters.[22] Likewise, feature clustering was also studied, and a heatmap was constructed to visualize clusters of samples and features simultaneously.

In PLS-DA, the optimal number of PLS components are estimated by cross-validation by plotting the PRESS (predicted residual error sum of squares) or RMSEP (root mean square error in the prediction) against the number of PLS components. Sometimes there are several almost equivalent local minima on the curve; the first one should be preferred to avoid overfitting (according to the principle of parsimony). The model with the smallest number of features should be accepted from among equivalent models on the training set. In PLS-DA, once the number of PLS components is optimized, the predictions in the training-test set are represented in a box and whisker plot in order

to define the half of the distance between the quartiles as the boundary. The weighted regression coefficients (Bw) of the PLS-components indicate the importance of the features on the model: the larger the regression coefficient (in absolute value), the higher the influence of the feature on the PLS-DA model.[23] Binary classification models can lead to artifacts if they are not used and validated properly.[24] The reliability of the classification models achieved was studied in terms of recognition ability (percentage of the samples in the training set correctly classified during the modeling step) and prediction ability (percentage of the samples in the test set correctly classified by using the model developed in the training step).[19]

O-PLS is a modification of NIPALS PLS algorithm to improve interpretation of PLS models and to reduce model complexity.[20] O-PLS removes systematic variation from the input data X not correlated to the response feature Y by separating the non-correlated variation in X, which is orthogonal to Y, from the correlated variation. This improves the interpretational ability of the resulting models. PLS-DA with O-PLS modification is known as OPLS-DA (Orthogonal Projections to Latent Structures Discriminant Analysis). The number of PLS components in the OPLS-DA model is reduced to a single component, and the number of orthogonal components is estimated by leave-one-out cross-validation and the eigenvalue approach. The loadings of the PLS component and the VIP (variable importance in the projection) value for each feature indicate the most relevant features in the OPLS-DA model, as well as the S-plot (feature covariance vs feature correlation).

3. Results

3.1. Apple cultivars and their resistance to RAA

The apple juices of the cultivars ‘Meana’ (susceptible to RAA) and ‘Florina’ (resistant to RAA) and the progeny derived from the cross of these two cultivars, MxF, were studied by an untargeted metabolomics approach based on UHPLC-DAD-ESI-QToF/MS analysis and chemometrics to

identify the metabolites involved in the resistance of apple cultivars to RAA. Regarding MxF descendants, 55 cultivars were found to be resistant and 75 cultivars were susceptible.

3.2. Data pre-processing

UHPLC-ESI(+)-QToF/MS^E (low energy function) raw data of the apple cultivars studied were pre-processed as described in the experimental section. A total of 3593 mass spectral features composed the raw dataset. Retention time alignment and peak integration were carried out using XCMS package, and the grouping of the feature lists using CAMERA package, which allowed to automatically group all features derived from the same compound and annotate the type of ion species (protonated molecule, sodium adduct, etc.) in order to help in the elucidation workflow.

3.3. QC correction and feature selection

QC samples were included 10 times at the beginning of the sequence to confirm that the analytical system was stabilized before the sample batch was analyzed, and afterwards, every 5 injections to assess its performance and observe gradual changes in instrument sensitivity over time. The relative standard deviation (%) for the peak areas of the standard compounds were less than 10 % (n = 5), and for the retention time shift, less than 5 % (n = 5); and the mass error was less than 3 ppm (n = 5), indicating a satisfactory system performance.[25, 26] This is a common practice to obtain repeatable and interpretable LC-MS metabolomics data,[27, 28] since QC samples were representative of all metabolites present in the experimental set. Apple juice samples were randomly injected so that the sample groups were affected to the same extent. The MS signal intensities for each sample were corrected using the factor calculated using the QC data. The effect of the signal drift correction was evaluated by comparing the slope of the linear regressions of the sum of intensities for all features against the number of injection before and after being corrected. Moreover, PCA was performed to the whole dataset in order to check the behavior of the QC samples as a measure of the technical variability. The corrected data yielded a slope value close to zero when plotted against the injection number, and QC samples were grouped in the PCA score plot.

After the signal drift correction, the quality criterion applied to select the features retained in the dataset for further statistical data analysis was the coefficient of variation (CV). Those features in the QC samples with CV lower or equal to 30%, which is considered an acceptable repeatability in biomarker analysis,[29] were kept in the final dataset (1983 features).

3.4. Data pre-treatment

Once the raw data of each apple juice sample was pre-processed and corrected, the challenge of extracting the relevant information from the resulting large dataset was faced. An appropriate data pre-treatment is required for the statistical data analysis since large differences in the concentrations of metabolites are not proportional to their relevance in the biological process under study.[21] Metabolomic fingerprinting data obtained by liquid chromatography coupled to high resolution mass spectrometry presents a heteroscedastic noise structure: the variance increases as the signal intensity increases. These experimental and instrumental noise sources can affect negatively the statistical and chemometric techniques used for data analysis.[28] Indeed, PLS-DA requirement of homoscedasticity makes data pre-treatment mandatory. The pre-processed and corrected data was transformed using a log-based transformation with the aim of stabilizing the variance respect to peak intensity and reduce heteroscedasticity. The normal distribution of all features was also assessed. The normality and homoscedasticity of the features were compared before and after the logarithmic transformation of the data by Shapiro-Wilk and Levene tests respectively, confirming that both normality and homoscedasticity of the data was improved using the log transformation (Table S1). Actually, using corrected pre-processed data, the clustering of the samples according to their tolerance to RAA in PCA was not so evident, and PLS-DA models attained were less robust and performed worse (data not shown). Further data analysis was carried out on the logarithmic transformed data.

3.5. Pattern recognition

The analysis of variance (ANOVA) performed on the log-transformed data matrix consisting of 1983 mass spectral features disclosed significant differences for certain variables between resistant and susceptible apple cultivars to RAA. The Fisher index was calculated to establish the discriminant capacity of the variables one by one. The most discriminant variables were 274 features that presented the highest Fisher weights ($p < 0.05$), but their box and whisker plots showed an overlap between intensity ranges of the two classes. Thus, none of the variables measured was able to discriminate between the resistant and susceptible categories by itself. Therefore, it was necessary to move on to multivariate data analysis in order to differentiate apple cultivars according to their tolerance to RAA.

A tool provided by XCMS package for the generation of a candidate marker list is the volcano plot that includes the fold change and p-value. In the volcano plot (Fig. S1), up and down regulated molecular features related to the tolerance of apple cultivars to RAA were displayed as red and blue squares in the positive and negative $\log_2$(fold change) values respectively. Twenty six features were preselected by this tool with fold change ≥ 1.5 and p-value ≤ 0.05 using Student t-test and Benjamini-Hochberg multiple testing correction (Table S2).

Unsupervised data analysis by PCA was performed to visualize the data looking for trends and groupings, and to identify possible outliers. PCA was performed on the autoscaled and pareto-scaled data (Fig. S2). The first three principal components accounted for similar percentages of the total variability present in the data, i.e. 37% with autoscaled data and 40% with pareto-scaled data. The bidimensional plot of the sample scores in the space defined by the three first principal components (PCs) showed a natural separation of resistant and susceptible apple cultivars mainly due to PC-3 with autoscaled data and PC-2 with pareto-scaled data. Table S3 gathers the most influent features on these PCs. Pareto scaling reduced the influence of noise variables on the multivariate model compared to autoscaling, as previously observed.[16]

Cluster analysis, being another unsupervised pattern recognition technique, highlights the existence of natural groupings between samples characterized by a set of measured features, as well as between variables. The results achieved by HCA were represented in combined dendrograms of samples and variables as a heatmap (Fig. S3). Two clusters were observed according to apple cultivar tolerance to RAA; one cluster contained only resistant cultivars, and the other one, only susceptible cultivars. Regarding the variable dendogram, two clusters of the most discriminant variables were discerned. The heatmap revealed that the intensity values of the variables in one cluster were higher for one category and lower for the other category, whereas in the other cluster, the opposite occurred. These results agreed with the up and down regulation of molecular features observed in the volcano plot (Fig. S1 and Table S2). PCA and HCA results suggested that there were notable differences among resistant and susceptible cultivars and that the LC-MS data enclosed valuable information to attain cultivar differentiation according to the established categories.

PLS-DA and OPLS-DA models were developed to extract the useful knowledge contained in the LC-MS data related to the tolerance of apple cultivars to RAA (Tables 1 and 2). PLS-DA models were built using autoscaled or pareto-scaled data. In both cases, the optimized models included one PLS component. The recognition and prediction abilities in the cross-validation using autoscaling were 93 % and 70 % for resistant cultivars and 86 % and 70 % for susceptible cultivars, respectively; and using pareto-scaling, 96 % and 80 % for resistant cultivars and 90 % and 80 % for susceptible cultivars, respectively. The fact that the difference between the recognition and prediction abilities was larger with autoscaled data confirmed that pareto scaling of LC-MS data provided more robust methods than autoscaling. This is due to the fact that autoscaling normalizes the data giving the same importance to all variables, thus noise is amplified and modelled, leading to less stable classification models that perform worse in prediction.[21] The model obtained with autoscaled data presented more variables with high Bw (absolute value) among the most influent variables, i.e. 76 variables due to 19 chromatographic peaks (Table S4). In contrast, the model

attained with pareto-scaled data displayed 40 features that belonged to 6 chromatographic peaks. Hence, the latter model was less complex and easier to interpret. Further PLS-DA models were developed regarding only the most influential variables, achieving more than 90 % of correct classifications for both categories (Table 1). Since recognition and prediction abilities in cross-validation were close to each other (being the former higher than the latter), the models are considered feasible and not random, as well as well-represented by the samples in the dataset. The most influent variables on the PLS-DA model obtained with pareto-scaled data were 27 features corresponding to 2 chromatographic peaks, while on the model afforded with autoscaled data, 40 variables due to 6 chromatographic peaks (Table S4). Definitely, pareto-scaling of LC-MS data provided more straightforward interpretable models. PCA and PLS-DA results confirmed that autoscaling is not appropriate for MS spectral data, as also observed using a different MS technique.[30]

A cross-validated OPLS-DA model was built using pareto-scaled data in order to highlight the biomarkers to differentiate between resistant and susceptible apple cultivars to RAA (Fig. S4). The most relevant features responsible for the maximum variation between the two categories were selected using the S-plot and VIP value of the variables (Fig. 1). The analysis of the features selected gave as a result a list of 25 candidates (Table 2), which belonged to 2 different compounds. Eight of these features were adducts and/or fragments of a compound at 11.1 min and the other 17 features of a compound at 16.6 min. This was confirmed by extracting the chromatograms of these ions from the total ion current (TIC) of the low energy function in positive ion mode and checking that chromatographic retention time and peak shape matched (Fig. S5). The signal differences of this compound as a function of the established categories were verified by displaying the extracted chromatogram of the marker candidate ions (Fig. S6).[26]

All features selected by OPLS-DA model were included among the most important variables (highest Bw in absolute value) on the PLS-DA models achieved (Table S4), as well as among those variables with the highest loadings (in absolute value) on PC-2 and PC-3 of the PCA models

afforded with pareto-scaled data and autoscaled data respectively (Table S3), and the features grouped in the two clusters observed in the variable dendogram of HCA (Fig. S3). The features in each cluster presented loadings and Bw with opposite signs. Looking for the chemical interpretation of this observation, it was disclosed that compound at 11.1 min was present in higher concentrations in resistant apple cultivars, whereas compound at 16.6 min was contained in higher amounts in susceptible cultivars (Fig. S6). The fact that different pattern recognition techniques achieved similar results implies that the models were stable and robust, that the data contained information related to apple cultivar tolerance to RAA, and that the established categories were well represented by the sample set. Moreover, from the multivariate data analysis results, it can be figure out that the type of scaling of the data and the removal of features due to noisy MS spectral regions without MS peaks of interest can considerably improve the classification models provided by the chemometric techniques, and simplify their interpretation and the identification of biomarkers. The analysis of UHPLC-ESI(−)-QToF/MS^E (low energy function) data led to the same two biomarkers at 11.1 min and 16.6 min (data not shown)

3.6. Structure elucidation of biomarkers

Several of the selected features were identified by UHPLC-DAD-ESI-QToF/MS using UV-vis and MS spectral data (Tables 3, 4 and S5). The chromatographic peaks at 11.1 min and 16.6 min presented the same UV spectra as the standard *trans*-5-caffeoylquinic acid (5-CQA) and *p*-coumaric acid respectively. The MS^E spectral data in positive ion mode revealed that peak at 11.1 min was due to a caffoylquinic acid (CQA) isomer, and peak at 16.6 min, to a *p*-coumaroylquinic acid (*p*-CoQA) isomer. Concerning peak at 11.1 min, the highest relative abundance (RA) was presented by the feature at *m/z* 163, which is a typical intense fragment of CQA isomers due to the dehydrated caffeic acid moiety in positive ion mode. The features with the second and fourth highest RA at *m/z* 377 and *m/z* 355 corresponded to the sodium adduct $[M+Na]^+$ and the protonated molecular ion $[M+H]^+$ of CQA isomers. The feature with the third highest RA at *m/z* 760, which was tentatively attributed to the adduct $[2M-3H+MeOH+Na]^+$, presented an intensity 19-fold lower than the most

412 intense feature. The other thirteen features exhibited even lower RA; some of them were tentatively
413 assigned to CQA adducts: $[2M–Quinic+MeOH+H]^+$ at *m/z* 549, $[2M–H+MeOH+Na]^+$ at *m/z* 762,
414 $[2M–2H+MeOH+Na]^+$ at *m/z* 761, $[2M–4H+MeOH+Na]^+$ at *m/z* 759, $[3M–2H+MeOH+Na]^+$ at *m/z*
415 1115, $[2M–H+H2O+H]^+$ at *m/z* 726 and $[2M–2H+H2O+H]^+$ at *m/z* 725. Fragments with weak but
416 detectable intensity due to the loss of two protons of the caffeic acid moiety were observed in the
417 positive ion mode MS spectra of 5-CQA and caffeic acid standards. Regarding peak at 16.6 min, the
418 most intense feature at *m/z* 361 corresponded to the sodium adduct $[M+Na]^+$ of *p*-CoQA isomers.
419 The other seven features displayed intensities 30-fold lower than that, and only the ions at *m/z* 731
420 and *m/z* 649 could be tentatively assigned to the adducts $[2M+MeOH+Na]^+$ and [2M-
421 Coumaric+3MeOH+H_2O+Na]$^+$ of *p*-CoQA isomers, repectively.

422 MS/MS experiments in negative ion mode using as the precursor ion the deprotonated molecular
423 ion $[M–H]^–$ at *m/z* 353 for peak at 11.1 min yielded fragment ions at *m/z* 173 (100), 179 (70), 191
424 (40) and 135 (40). For peak at 16.6 min, the precursor ion $[M–H]^–$ at *m/z* 337 produced fragment
425 ions at *m/z* 173 (100), 163 (20) and 191 (5). The fragmentation patterns of both deprotonated
426 molecular ions confirmed that the compounds were 4-acyl isomers, i.e. 4-CQA at 11.1 min and 4-*p*-
427 CoQA at 16.6 min (Fig. S7). The fragment ion at *m/z* 173 ([quinic acid−H−H_2O]$^–$), being usually
428 the base peak, is characteristically formed in the negative ion mode when the cinnamoyl group is
429 bonded to the quinic moiety at position 4, as already noted by other authors using different types of
430 mass spectrometers.[31-33]

431 Although *cis* isomers of chlorogenic acids (CGAs) lead to the same fragments than the more
432 common *trans* isomers, *cis* and *trans* isomers are easily resolved by chromatography.[34] *Cis*-5-acyl
433 and *cis*-1-acyl CGAs, being more hydrophobic, elute later than their *trans* isomers on endcapped
434 C18 and phenylhexyl packings. The opposite occurs with *cis*-3-acyl and *cis*-4-acyl CGAs respect to
435 their *trans* isomers.[31, 34, 35] The presence of other chromatographic peak at 4.3 min with the same

fragmentation pattern as peak at 11.1 min allowed to confirm the *trans* conformation of the latter, being tentatively identified as *trans*-4-CQA (Fig. S7).

4. Discussion

UHPLC-DAD-ESI-QToF/MS data of apple cultivars combined with pattern recognition techniques confirmed that the hydroxycinnamic acids *trans*-4-CQA and 4-*p*-CoQA are related to apple tree tolerance to RAA, as proposed in our previous profiling study.[14] Susceptible apple cultivars contained higher levels of 4-*p*-CoQA than resistant ones; and resistant cultivars presented higher contents of the *trans*-4-CQA than susceptible ones. The enzymes involved in the early stages of the phenylpropanoid pathway, which synthesized hydroxycinnamic acids and shikimate esters, are known;[36] however those involved in the last steps are still not completely disclosed. In higher plants, the primary route for caffeoylquinic acid formation appeared to be via *p*-coumaroyl-CoA by the combined activities of the hydroxycinnamoyl CoA shikimate hydroxycinnamoyl transferase (HCT), hydroxycinnamoyl CoA quinate hydroxycinnamoyl transferase (HQT) and *p*-coumarate 3´-hydroxylase.[37] The 3´-hydroxylation is catalyzed on *p*-coumaric acid conjugates with shikimic or quinic acids, but not on the free *p*-coumaric acid.[38, 39] Therefore, since shikimic derivatives have not been reported in apple tissues to date, the main route for CQA formation involves only *p*-CoQAs. The 5-caffeoylshikimic acid and 5-CQA were the major enzymatic products found *in vitro* studies on HCT and HQT of Robusta coffee, and it was concluded that the subsequent 3- or 4-isomerization of 5-CQA occurred non-enzymatically in solution, whereas the level of isomerization occurring *in vivo* is still unclear.[40] The present results of the untargeted metabolomics approach, revealing 4-CQA and 4-*p*-CoQA as biomarkers for the identification of resistant and susceptible apple cultivars to RAA, confirm the importance of the final steps of hydroxycinnamic acid biosynthesis, in particular of CQA and *p*-CoQA, on the tolerance of apple cultivars to RAA. In resistant cultivars the formation of CQAs assisted through the activity of *p*-coumarate 3´-hydroxylase on *p*-CoQA does take place, while in susceptible cultivars this step is not favored leading to the accumulation of higher contents of *p*-CoQAs. Indeed, the ratio of the total contents of CQAs respect to *p*-CoQAs is

462 higher in the resistant cultivars than in the susceptible ones.[14] Considering that 5-CQA contents is
463 similar in both types of cultivars,[14] the isomerization of 5-CQA to 4-CQA is promoted in resistant
464 cultivars. The facts that the contents of 5-*p*-CoQA are similar in the two types of cultivars and 4-*p*-
465 CoQA is present in higher concentrations in susceptible cultivars,[14] indicates that the isomerization
466 of 5-*p*-CoQA to 4-*p*-CoQA occurs to a greater extent in this type of cultivars. Angeli et al (2006)
467 suggested the release of repellent compounds in apple tree leaves when resistant cultivars were
468 infected by RAA.[3] However, taking into account that the present results were obtained from the
469 analysis of apple juices, the natural and original differential chemical composition of each apple
470 cultivars genetically defined seems to already determine the tolerance of each apple cultivar to
471 RAA. Further studies should be carried out to evaluate whether higher concentrations of these
472 compounds are synthesized in the different apple tissues when apple trees are infected by RAA.
473 Moreover, the results of current study disclosed that only hydroxycinnamic acids are responsible for
474 the resistance of apple cultivars to RAA, and that not any other family of compounds is involved in
475 the disease susceptibility of the cultivars. This data contributes to the knowledge on the multiple
476 specific roles of hydroxycinnamic acids in plants.[37, 41, 42]

477 **Abbreviations Used**

478 Rosy apple aphid (RAA), caffeoylquinic acid (CQA), *p*-coumaroylquinic acid (*p*-CoQA),
479 chlorogenic acids (CGAs), hydroxycinnamoyl CoA shikimate hydroxycinnamoyl transferase
480 (HCT), hydroxycinnamoyl CoA quinate hydroxycinnamoyl transferase (HQT), quality control
481 (QC), ultrahigh-performance liquid chromatography-diode array detector-electrospray ionization
482 quadrupole time of flight/mass spectrometer (UHPLC-DAD-ESI-QToF/MS), Principal Component
483 Analysis (PCA), Hierarchical Cluster Analysis (HCA), Partial Least Square-Discriminant Analysis
484 (PLS-DA), Orthogonal Projections to Latent Structures (O-PLS), weighted regression coefficients
485 (Bw), principal component (PC), PLS-component (PLS-component), relative abundance (RA).

486

Supporting Information

Volcano plot of resistant and susceptible of the UHPLC-ESI(+)-QToF/MS^E data of resistant and susceptible apple cultivars to RAA (Figure S1), PCA score plot of resistant and susceptible apple cultivars to RAA (Figure S2), heatmap of combined dendrograms obtained by HCA of the data of resistant and susceptible apple cultivars to RAA (Figure S3), OPLS-DA score plot of resistant and susceptible apple cultivars to RAA (Figure S4), extracted TIC chromatograms of the selected biomarkers related to the tolerance of apple cultivars to RAA (Figure S5), extracted TIC chromatograms of the main features in resistant and susceptible apple cultivars (Figure S6), structure of the identified biomarkers (Figure S7), normality and homoscedasticity tests before and after the logarithmic transformation of UHPLC-ESI-QToF/MS^E data (Table S1), candidate biomarker list provided by volcano plot (Figure S1) related to the tolerance of apple cultivars to RAA (Table S2), PCA loadings for the most important variables responsible for the grouping of apple cultivars according to their tolerance to RAA (Table S3), weighted regression coefficients (Bw) for the most important variables on the PLS-DA model built to discriminate between apple cultivars according to their tolerance to RAA (Table S4), biomarkers selected by pattern recognition techniques related to the tolerance of apple cultivars to RAA using UHPLC-DAD-ESI(−)-QToF/MS^E data (Table S5).

Acknowledgments

Technical and staff support provided by SGIker (UPV/EHU, MICINN, GV/EJ, ESF) is gratefully acknowledged.

References

(1) Miñarro, M.; Dapena, E. Tolerance of some scab-resistant apple cultivars to the rosy apple aphid, *Dysaphis plantaginea*. *Crop Protec.* **2008**, *27* (3-5), 391-395.

(2) Andreev, R.; Kutinkova, H.; Rasheva, D. Non-chemical control of *Aphis spiraecola* patch. and *Dysaphis plantaginea* pass. on apple. *J. Biopestic.* **2012**, *5*, 239-242.

512 (3) Angeli, G.; Simoni, S. Apple cultivars acceptance by *Dysaphis plantaginea* Passerini
513 (Homoptera: Aphididae). *J. Pest Sci.* **2006**, *79* (3), 175-179.
514 (4) Dapena, E.; Blázquez, M. D. Improvement of the resistance to scab, rosy apple aphid and fire
515 blight in a breeding programme of cider apple cultivars. *Acta Hortic.* **2004**, *663*, 725-728.
516 (5) Lespinasse, Y.; Olivier, J. M.; Lespinasse, J. M.; Le Lezec, M. Florina Quérina: la résistance du
517 pommier à la tavelure. *Arboric. Fruit* **1985**, (378), 43-47.
518 (6) Rat-Morris, E. Development of the rosy apple aphid *Dysaphis plantaginea* Pass. on a tolerant
519 apple cultivar 'Florina'. *IOBC/WPRS Bull.* **1993**, *16* (5), 91-100.
520 (7) Dapena, E.; Miñarro, M.; Blázquez, M. D. Evaluation of the resistance to the rosy apple aphid
521 using a genetic marker. *Acta Hortic.* **2009**, *814*, 787-790.
522 (8) Noble, J. P.; Drysdale, R. B. The role of benzoic acid and phenolic compounds in latency in
523 fruits of two apple cultivars infected with *Pezicula malicorticis* or *Nectria galligena*. *Physiol. Plant*
524 *Pathol.* **1983**, *23*, 207-216.
525 (9) Michalek, S.; Mayr, U.; Treutter, D.; Lux-Endrich, A.; Gutmann, M.; Feucht, W.; Geibel, M.
526 Role of flavan-3-ols in resistance of apple trees to *Venturia inaequalis*. *Acta Hortic.* **1998**, *484*, 535-
527 539.
528 (10) Mikulič Petkovšek, M.; Štampar, F.; Veberič, R. Accumulation of phenolic compounds in
529 apple in response to infection by the scab pathogen, *Venturia inaequalis*. *Physiol. Mol. Plant.*
530 *Pathol.* **2009**, *74* (1), 60-67.
531 (11) Mikulič Petkovšek, M.; Slatnar, A.; Štampar, F.; Veberič, R. Phenolic compounds in apple
532 leaves after infection with apple scab. *Biol. Plant.* **2011**, *55* (4), 725-730.
533 (12) Picinelli, A.; Dapena, E.; Mangas, J. J. Polyphenolic pattern in apple tree leaves in relation to
534 scab resistance. A preliminary study. *J. Agric. Food Chem.* **1995**, *43* (8), 2273-2278.
535 (13) Hutabarat, O. S.; Flachowsky, H.; Regos, I.; Miosic, S.; Kaufmann, C.; Faramarzi, S.; Alam,
536 M. Z.; Gosch, C.; Peil, A.; Richter, K.; et al. Transgenic apple plants overexpressing the chalcone 3-

537 hydroxylase gene of Cosmos sulphureus show increased levels of 3-hydroxyphloridzin and reduced
538 susceptibility to apple scab and fire blight. *Planta* **2016**, *243* (5), 1213-1224.
539 (14) Berrueta, L. A.; Sasía-Arriba, A.; Miñarro, M.; Antón, M. J.; Alonso-Salces, R. M.; Micheletti,
540 D.; Gallo, B.; Dapena, E. Relationship between hydroxycinnamic acids and the resistance of apple
541 cultivars to rosy apple aphid. *Talanta* **2018**, *187*, 330-336.
542 (15) Ramirez-Ambrosi, M.; Abad-Garcia, B.; Viloria-Bernal, M.; Garmon-Lobato, S.; Berrueta, L.
543 A.; Gallo, B. A new ultrahigh performance liquid chromatography with diode array detection
544 coupled to electrospray ionization and quadrupole time-of-flight mass spectrometry analytical
545 strategy for fast analysis and improved characterization of phenolic compounds in apple products. *J.*
546 *Chromatog. A* **2013**, *1316*, 78-91.
547 (16) Di Guida, R.; Engel, J.; Allwood, J. W.; Weber, R. J. M.; Jones, M. R.; Sommer, U.; Viant, M.
548 R.; Dunn, W. B. Non-targeted UHPLC-MS metabolomic data processing methods: a comparative
549 investigation of normalisation, missing value imputation, transformation and scaling. *Metabolomics*
550 **2016**, *12* (5), 93.
551 (17) Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T. R.; Neumann, S. CAMERA: An integrated
552 strategy for compound spectra extraction and annotation of liquid chromatography/mass
553 spectrometry data sets. *Anal. Chem.* **2012**, *84* (1), 283-289.
554 (18) Kamleh, M. A.; Ebbels, T. M. D.; Spagou, K.; Masson, P.; Want, E. J. Optimizing the use of
555 quality control samples for signal drift correction in large-scale urine metabolic profiling studies.
556 *Anal. Chem.* **2012**, *84* (6), 2670-2677.
557 (19) Berrueta, L. A.; Alonso-Salces, R. M.; Héberger, K. Supervised pattern recognition in food
558 analysis. *J. Chromatog. A* **2007**, *1158* (1-2), 196-214.
559 (20) Trygg, J.; Wold, S. Orthogonal projections to latent structures (O-PLS). *J. Chemom.* **2002**, *16*
560 (3), 119-128.

561 (21) van den Berg, R. A.; Hoefsloot, H. C. J.; Westerhuis, J. A.; Smilde, A. K.; van der Werf, M. J.
562 Centering, scaling, and transformations: improving the biological information content of
563 metabolomics data. *BMC Genomics* **2006**, *7*, 142-142.
564 (22) Massart, D. L.; Kaufman, L. *The Interpretation of Analytical Chemical Data by the Use of*
565 *Cluster Analysis*; John Wiley & Sons, 1983.
566 (23) Esbensen, K. H.; Guyot, D.; Westad, F.; Houmøller, L. P. *Multivariate data analysis - In*
567 *practice: An introduction to multivariate data analysis and experimental design*; Camo Process AS,
568 2002.
569 (24) Kjeldahl, K.; Bro, R. Some common misunderstandings in chemometrics. *J. Chemom.* **2010**,
570 *24* (7-8), 558-564.
571 (25) Kalogiouri, N. P.; Aalizadeh, R.; Thomaidis, N. S. Application of an advanced and wide scope
572 non-target screening workflow with LC-ESI-QTOF-MS and chemometrics for the classification of
573 the Greek olive oil varieties. *Food Chem.* **2018**, *256*, 53-61.
574 (26) Díaz, R.; Pozo, O. J.; Sancho, J. V.; Hernández, F. Metabolomic approaches for orange origin
575 discrimination by ultra-high performance liquid chromatography coupled to quadrupole time-of-
576 flight mass spectrometry. *Food Chem.* **2014**, *157*, 84-93.
577 (27) Hoyos Ossa, D. E.; Gil-Solsona, R.; Peñuela, G. A.; Sancho, J. V.; Hernández, F. J.
578 Assessment of protected designation of origin for Colombian coffees based on HRMS-based
579 metabolomics. *Food Chem.* **2018**, *250*, 89-97.
580 (28) Veselkov, K. A.; Vingara, L. K.; Masson, P.; Robinette, S. L.; Want, E.; Li, J. V.; Barton, R.
581 H.; Boursier-Neyret, C.; Walther, B.; Ebbels, T. M.; et al. Optimized preprocessing of ultra-
582 performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved
583 information recovery. *Anal. Chem.* **2011**, *83* (15), 5864-5872.
584 (29) Bonnefoy, C.; Fildier, A.; Buleté, A.; Bordes, C.; Garric, J.; Vulliet, E. Untargeted analysis of
585 nanoLC-HRMS data by ANOVA-PCA to highlight metabolites in *Gammarus fossarum* after in
586 vivo exposure to pharmaceuticals. *Talanta* **2019**, *202*, 221-229.

587 (30) Martínez-Lozano Sinues, P.; Alonso-Salces, R. M.; Zingaro, L.; Finiguerra, A.; Holland, M.
588 V.; Guillou, C.; Cristoni, S. Mass spectrometry fingerprinting coupled to National Institute of
589 Standards and Technology Mass Spectral search algorithm for pattern recognition. *Anal. Chim. Acta*
590 **2012**, *755*, 28-36.
591 (31) Viacava, G. E.; Roura, S. I.; Berrueta, L. A.; Iriondo, C.; Gallo, B.; Alonso-Salces, R. M.
592 Characterization of phenolic compounds in green and red oak-leaf lettuce cultivars by UHPLC-
593 DAD-ESI-QToF/MS using MS^E scan mode. *J. Mass Spectrom.* **2017**, *52* (12), 873-902.
594 (32) Alonso-Salces, R. M.; Guillou, C.; Berrueta, L. A. Liquid chromatography coupled with
595 ultraviolet absorbance detection, electrospray ionization, collision-induced dissociation and tandem
596 mass spectrometry on a triple quadrupole for the on-line characterization of polyphenols and
597 methylxanthines in green coffee beans. *Rapid Comm. Mass Spec.* **2009**, *23* (3), 363-383.
598 (33) Clifford, M. N.; Knight, S.; Surucu, B.; Kuhnert, N. Characterization by $LC\text{-}MS^n$ of four new
599 classes of chlorogenic acids in green coffee beans: Dimethoxycinnamoylquinic acids,
600 diferuloylquinic acids, caffeoyl-dimethoxycinnamoylquinic acids, and feruloyl-
601 dimethoxycinnamoylquinic acids. *J. Agric. Food Chem.* **2006**, *54* (6), 1957-1969.
602 (34) Clifford, M. N.; Kirkpatrick, J.; Kuhnert, N.; Roozendaal, H.; Salgado, P. R. $LC\text{-}MS^n$ analysis
603 of the *cis* isomers of chlorogenic acids. *Food Chem.* **2008**, *106* (1), 379-385.
604 (35) Clifford, M. N.; Wu, W.; Kuhnert, N. The chlorogenic acids of *Hemerocallis*. *Food Chem.*
605 **2006**, *95* (4), 574-578.
606 (36) Vogt, T. Phenylpropanoid biosynthesis. *Mol. Plant* **2010**, *3* (1), 2-20.
607 (37) Niggeweg, R.; Michael, A. J.; Martin, C. Engineering plants with increased levels of the
608 antioxidant chlorogenic acid. *Nat. Biotechnol.* **2004**, *22* (6), 746-754.
609 (38) Abdulrazzak, N.; Pollet, B.; Ehlting, J.; Larsen, K.; Asnaghi, C.; Ronseau, S.; Proux, C.;
610 Erhardt, M.; Seltzer, V.; Renou, J. P.; et al. A coumaroyl-ester-3-hydroxylase insertion mutant
611 reveals the existence of nonredundant meta-hydroxylation pathways and essential roles for phenolic
612 precursors in cell expansion and plant growth. *Plant Physiol.* **2006**, *140* (1), 30-48.

613 (39) Hoffmann, L.; Besseau, S.; Geoffroy, P.; Ritzenthaler, C.; Meyer, D.; Lapierre, C.; Pollet, B.;
614 Legrand, M. Silencing of hydroxycinnamoyl-coenzyme a shikimate/quinate
615 hydroxycinnamoyltransferase affects phenylpropanoid biosynthesis. *Plant Cell* **2004**, *16* (6), 1446-
616 1465.
617 (40) Lallemand, L. A.; Zubieta, C.; Lee, S. G.; Wang, Y.; Acajjaoui, S.; Timmins, J.; McSweeney,
618 S.; Jez, J. M.; McCarthy, J. G.; McCarthy, A. A. A structural basis for the biosynthesis of the major
619 chlorogenic acids found in coffee. *Plant Physiol.* **2012**, *160* (1), 249-260.
620 (41) Clé, C.; Hill, L. M.; Niggeweg, R.; Martin, C. R.; Guisez, Y.; Prinsen, E.; Jansen, M. A. K.
621 Modulation of chlorogenic acid biosynthesis in *Solanum lycopersicum*; consequences for phenolic
622 accumulation and UV-tolerance. *Phytochemistry* **2008**, *69* (11), 2149-2156.
623 (42) Leiss, K. A.; Maltese, F.; Choi, Y. H.; Verpoorte, R.; Klinkhamer, P. G. Identification of
624 chlorogenic acid as a resistance factor for thrips in chrysanthemum. *Plant Physiol.* **2009**, *150* (3),
625 1567-1575.
626 (43) Alseekh, S.; Aharoni, A.; Brotman, Y.; Contrepois, K.; D'Auria, J.; Ewald, J.; C. Ewald, J.;
627 Fraser, P. D.; Giavalisco, P.; Hall, R. D.; et al. Mass spectrometry-based metabolomics: a guide for
628 annotation, quantification and best reporting practices. *Nature Methods* **2021**, *18* (7), 747-756.

629

630 **Funding**

631 This work was supported by the Instituto Nacional de Investigación y Tecnología Agraria y
632 Alimentaria INIA with ERDF funds [Projects RTA2012-00118-C03-01 and -03, RTA2014-00090-
633 C03-01 and -03, and RTA2017-00102-C03-01 and 03].

634

635 **Figure captions**

636 **Figure 1.** S-plot (a) and VIP values plot (b) obtained by OPLS-DA of the UHPLC-ESI(+)-
637 QToF/MS^E data of apple cultivars in order to identify biomarkers related to their tolerance to RAA.
638 The most discriminant features with p[1] ≥ 0.09 or p[1] ≤ –0.09, p(corr)[1] ≥ 0.08 or p(corr)[1] ≤
639 –0.08, and VIP > 3.4 are highlighted in blue (features with higher relative abundances in resistant
640 cultivars) and red (features with higher relative abundances in susceptible cultivars).

641

642 **Tables**

643 **Table 1.** PLS-DA models for the discrimination of apple cultivars according to their tolerance
644 to RAA.[a]

PLS-DA model	Features	Scaling	Category	n	prior prob	Classification abilities	
						% R	% P
1) 1 PLS-comp	1983	Autoscaling	R	56	0.42	92.9	69.6
Boundary: 0.49625			S	76	0.58	85.5	69.7
2) 1 PLS-comp	1983	Pareto	R	56	0.42	96.4	80.4
Boundary: -0.11347			S	76	0.58	89.5	80.3
*3)*1 PLS-comp	76	Autoscaling	R	56	0.42	98.2	96.4
Boundary: 0.47871			S	76	0.58	92.1	90.8
4) 1 PLS-comp	40	Pareto	R	56	0.42	100	98.2
Boundary: -0.00037			S	76	0.58	93.4	92.1

645 [a] Abbreviations: PLS-DA, partial least-squares discriminant analysis; n, number of samples; prior
646 prob, prior probability; PLS-comp, number of selected PLS components; % R, percentage of
647 recognition ability; % P, percentage of prediction ability in cross-validation; Category codes: 1,
648 resistant (R); 0, susceptible (S).

649

650 **Table 2.** Candidate biomarker list provided by the S-plot and VIP values of the OPLS-DA
651 model built to discriminate apple cultivars according to their tolerance to RAA.[a,b,c]

OPLS-DA model				
Components	**Explained variance (%)**	**$R^2(X)$**	**$R^2(Y)$**	**Q^2**
1 PLS-comp + 2 O-comp	37.1	0.371	0.999	0.800
PLS-comp 1	11.6	0.116	0.999	0.800
O-comp 1	20.3	0.203		
O-comp 2	5.6	0.056		

#	**VIP[1+2+0]**	**VIP[1] cv SE**	**Feature**	***m/z***	**RT (min)**	**Fold change**	***p*-value**
1	4.36	0.99	M361T996	361.0898	16.59	1.55	2.9E-06
2	4.19	1.26	M528T991	527.6296	16.52	2.08	4.9E-06
3	4.18	1.54	M696T992	696.1810	16.54	2.50	3.5E-05
4	4.09	1.15	M730T992	730.1172	16.54	1.92	5.6E-06
5	3.96	0.90	M569T992	569.1185	16.53	1.80	4.9E-06
6	3.82	1.04	M649T1000	649.1545	16.66	1.83	9.3E-06
7	3.61	0.59	M731T994	731.1270	16.57	1.79	2.4E-06
8	3.53	0.55	M502T999	502.1115	16.65	1.66	1.4E-07
9	4.58	1.09	M728T667	727.6641	11.12	-2.46	1.2E-06
10	4.30	1.03	M549T667	549.1068	11.12	-1.93	1.3E-07
11	4.28	1.35	M1115T667	1115.1936	11.12	-2.22	9.3E-06
12	4.20	1.46	M727T666	726.6576	11.10	-2.28	4.9E-06
13	4.17	1.56	M760T668	760.0969	11.13	-1.71	4.9E-06
14	4.14	1.44	M726T668	726.1545	11.13	-1.91	5.0E-06
15	4.07	1.41	M729T667	728.6685	11.12	-2.17	2.9E-06
16	3.85	1.22	M730T665	730.1666	11.09	-1.96	1.1E-06
17	3.85	1.14	M163T668	163.0398	11.14	-1.36	4.9E-06
18	3.85	1.04	M377T667	377.0846	11.11	-1.42	1.1E-05
19	3.75	1.10	M726T667	725.6477	11.11	-1.82	7.2E-08
20	3.73	0.99	M759T668	759.0911	11.13	-2.05	4.9E-06
21	3.66	0.69	M746T665	746.1332	11.08	-1.78	1.4E-06
22	3.64	1.34	M761T669	761.1000	11.14	-1.60	2.9E-05
23	3.59	1.40	M725T666	725.1497	11.10	-1.78	2.8E-05
24	3.57	1.12	M355T667	355.0996	11.12	-1.51	2.9E-05
25	3.48	1.14	M762T669	762.1146	11.15	-1.66	3.1E-04

652 [a] Abbreviations: PLS comp, PLS component; O-comp, orthogonal component; VIP, variable
653 importance in the projection; VIP cv SE, VIP cross validation standard error; RT, retention time;
654 $R^2(X)$, fraction in the training set of total variation of X explained in the components; $R^2(Y)$,
655 fraction in the training set of total variation of Y modeled by X in the component, i.e. multiple

656 correlation coefficient (goodness of fit in the training set); Q^2, fraction of total variation of Y
657 predicted by the component in crossvalidation (goodness of fit in the test set).
658 [b] Number of features, 1983; data scaling, pareto-scaling.

659 [c] Criteria to identify potential candidate biomarkers: $p[1] \geq 0.09$ or $p[1] \leq -0.09$, $p(corr)[1] \geq 0.08$ or
660 $p(corr)[1] \leq -0.08$, and VIP > 3.4.

661 [d] Features presenting higher relative abundances in resistant cultivars are highlighted in blue; and
662 those in susceptible cultivars are in red.

663

664 **Table 3.** Biomarkers related to the tolerance of apple cultivars to RAA selected by pattern
665 recognition techniques using UHPLC-DAD-ESI(+)-QToF/MS^E data.[a]

		ESI(+)-QToF/MS			
#	Feature[b]	Exp. Acc. Mass *m/z*	Error (mDa)	Formula	Tentative ion assignment
1	M731T994	731.1270	-89	$C_{33}H_{40}O_{17}Na$	$[2M+MeOH+Na]^+$
2	M730T992	730.1172			
3	M696T992	696.1810			
4	M649T1000	649.1545	-77	$C_{26}H_{42}O_{17}Na$	$[2M-Coumaric+3MeOH+H_2O+Na]^+$
5	M569T992	569.1185			
6	M528T991	527.6296			
7	M502T999	502.1115			
8	M361T996	361.0898	-0.1	$C_{16}H_{18}O_8Na$	$[M+Na]^+$
9	M1115T667	1115.1936	-92	$C_{49}H_{56}O_{28}Na$	$[3M-2H+MeOH+Na]^+$
10	M762T669	762.1146	-84	$C_{33}H_{39}O_{19}Na$	$[2M-H+MeOH+Na]^+$
11	M761T669	761.1000	-90	$C_{33}H_{38}O_{19}Na$	$[2M-2H+MeOH+Na]^+$
12	M760T668	760.0969	-86	$C_{33}H_{37}O_{19}Na$	$[2M-3H+MeOH+Na]^+$
13	M759T668	759.0911	-84	$C_{33}H_{36}O_{19}Na$	$[2M-4H+MeOH+Na]^+$
14	M746T665	746.1332			
15	M730T665	730.1666			
16	M729T667	728.6685			
17	M728T667	727.6641			
18	M727T666	726.6576			
19	M726T668	726.1545	-46	$C_{32}H_{38}O_{19}$	$[2M-H+H_2O+H]^+$
20	M726T667	725.6477			
21	M725T666	725.1497	-43	$C_{32}H_{37}O_{19}$	$[2M-2H+H_2O+H]^+$
22	M549T667	549.1068	-54	$C_{26}H_{29}O_{13}$	$[2M-Quinic+MeOH+H]^+$
23	M377T667	377.0846	-0.3	$C_{16}H_{18}O_9Na$	$[M+Na]^+$
24	M355T667	355.0996	-3.3	$C_{16}H_{19}O_9$	$[M+H]^+$
25	M163T668	163.0398	0.3	$C_9H_7O_3$	$[Caffeic-H2O+H]^+$

666 [a] Abbreviations: Exp. Acc. Mass, experimental accurate mass; Caffeic, caffeic acid; Coumaric,
667 coumaric acid; Quinic, quinic acid.

668 [b] The UV-visible spectrum of chromatographic peak at 11.1 min presents a wavelength maximum at
669 323 nm and a shoulder at 300 nm (characteristic of caffeoylquinic acid isomers); and peak at
670 16.6 min, wavelength maximum at 311 nm (characteristic of *p*-coumaroylquinic acid isomers).

671

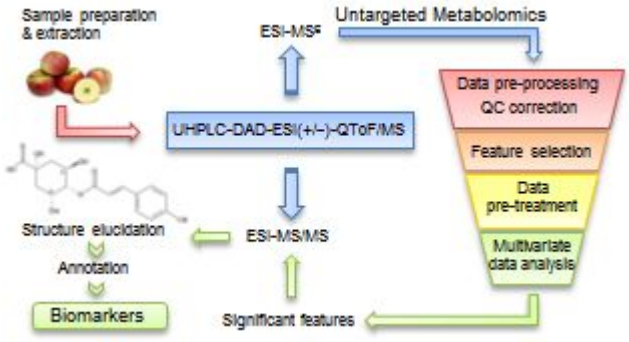
672 **Table 4.** Annotation and documentation of metabolites related to the tolerance of apple cultivars to RAA.[a]

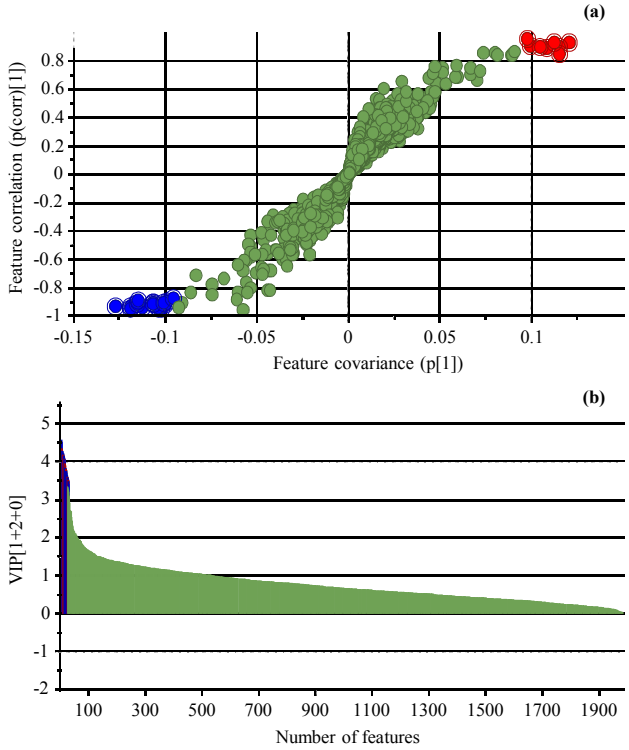
#	RT (min)	UV bands (nm)	Molecular formula	ESI(+) Theor. *m/z*	ESI(+) Exp. *m/z*	ESI(+) Error (mDa)	ESI(+) MS/MS frag./add.	ESI(–) Theor. *m/z*	ESI(–) Exp. *m/z*	ESI(–) Error (mDa)	ESI(–) MS/MS frag./add.	Putative metabolite	Metabolite class	Iden. level[b]	Reference ID
1	11.12	300 sh, 324	C16H18O9	355.1029	355.1030	0.1	377.0850 $[M+Na]^+$	353.0873	353.0874	0.1	707.1811 $[2M–H]^–$	*trans*-4-O-caffeoylquinic acid	Phenolic acid	B(i)	CAS no: 905-99-7
							163.0398 [Caffeic–H_2O+H]$^+$				191.0551 [Quinic–H]$^–$				PubChem: 9798666
							145.0291 [Caffeic–$2H_2O$+H]$^+$				179.0314 [Caffeic–H]$^–$				ChEBI: 75491
							117.0347 [Caffeic–$2H_2O$–CO+H]$^+$				173.0444 [Quinic–H_2O–H]$^–$				ChemSpider 22912773
											135.0439 [Quinic–2CO–H]$^–$				
2	16.57	309	C16H18O8	339.1080	339.1096	1.6	361.0896 $[M+Na]^+$	337.0923	337.0911	-1.2	191.0641 [Quinic–H]$^–$	4-*p*-coumaroylquinic acid	Phenolic acid	B(i)	CAS no: 1108200-72-1
							147.0435 [Coumaric –H_2O+H]$^+$				173.0446 [Quinic–H_2O-H]$^–$				PubChem: 5281766
							119.0504 [Coumaric –H_2O –CO+H]$^+$				163.0387 [Quinic–CO–H]$^–$				ChEBI: 1945
															ChemSpider 30785511

673 [a] Abbreviations: See Tables 2 and 3; sh, shoulder; Theor., theoretical accurate mass; Exp, experimental accurate mass; MS/MS frag./add., MS/MS
674 fragments/adducts; Iden. level, identification level.

675 [b] Identification level: (A) standard or NMR, B(i) confident match based on MS/MS, B(ii) confident match using in-silico MS/MS approaches, B(iii)
676 partial match based on MS/MS, C(i) confident match based on MS^n, C(ii) confident match using in-silico MS^n approaches, C(iii) partial match based
677 on MS^n, and (D) MS only.[43]

678 **Table of Contents graphic**

679 


(a)
Feature correlation (p(corr)[1])
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-0.15
-0.1
-0.05
0
0.05
0.1
Feature covariance (p[1])
(b)
VIP[1+2+0]
5
4
3
2
1
0
-1
-2
100
300
500
700
900
1100
1300
1500
1700
1900
Number of features