Preliminary authentication of apple juices using untargeted UHPLC-HRMS analysis combined to chemometrics

Katy Dinis, Lucie Tsamba, Freddy Thomas, Eric Jamin, Valérie Camel

Preliminary authentication of apple juices using untargeted UHPLC-HRMS analysis combined to chemometrics

Katy Dinis, Lucie Tsamba, Freddy Thomas, Eric Jamin, Valérie Camel

PII: S0956-7135(22)00291-2

DOI: https://doi.org/10.1016/j.foodcont.2022.109098

Reference: JFCO 109098

To appear in: *Food Control*

Received Date: 8 December 2021

Revised Date: 4 April 2022

Accepted Date: 8 May 2022

Please cite this article as: Dinis K., Tsamba L., Thomas F., Jamin E. & Camel Valé., Preliminary authentication of apple juices using untargeted UHPLC-HRMS analysis combined to chemometrics, *Food Control* (2022), doi: https://doi.org/10.1016/j.foodcont.2022.109098.

Credit Author Statement:

Katy Dinis: Conceptualization, Investigation, Formal analysis, Writing – Original Draft, **Lucie Tsamba:** Conceptualization, Validation, Writing – Review and editing, Project administration **Freddy Thomas:** Funding acquisition **Eric Jamin:** Writing – Review and editing, Funding acquisition, Project administration **Valérie Camel:** Writing – Review and editing, Supervision

Preliminary authentication of apple juices using untargeted UHPLC-HRMS analysis combined to chemometrics

Katy Dinis [a,b], Lucie Tsamba [a]*, Freddy Thomas [a], Eric Jamin [a], Valérie Camel [b]

[a] Eurofins Analytics France, 9 rue Pierre Adolphe Bobierre, B.P. 42301, F-44323, Nantes Cedex 3, France

[b] UMR SayFood, Université Paris-Saclay, INRAE, AgroParisTech, 91300 Massy, France

* Corresponding author: Eurofins Analytics France, 9 rue Pierre Adolphe Bobierre, B.P. 42301, F-44323 NANTES Cedex 3, France, tel.: +33 2 51 82 55 39, fax: +33 2 51 83 21 11. E-mail address: LucieTsamba@eurofins.com

Abstract

In this work, apple juice samples from different farming and production processes (direct and concentrated juices; organic and conventional juices) were analyzed by ultra-high performance liquid chromatography coupled to high resolution mass spectrometry (UHPLC-HRMS). A workflow was developed and implemented for data processing using the Workflow4Metabolomics (W4M) platform. First, features were detected using XCMS, and next data filtration steps were applied leading to the removal of nearly 50% of the detected features. Intra- and inter-batch correction was then performed, followed by chemometric tools (PCA, PLS-DA, OPLS-DA, ANOVA). The developed approach successfully discriminated apple juice samples in two distinct scenarios simultaneously (direct *vs.* concentrated juices and organic *vs.* conventional juices). PCA highlighted the reproducibility of the method and confirmed the efficiency of batch corrections. OPLS-DA models showed good quality metrics, particularly after feature selection for organic *vs.* conventional juices discrimination (almost 80% of predictive ability). Based on ANOVA and OPLS-DA results, 24 features were retained as significantly discriminant. Among them, some compounds were identified as amino-acids and derivatives, using additional MS/MS experiments and online databases. An independent data set was used to evaluate their potential as marker compounds, with promising results obtained. Further investigation is needed to validate such an untargeted method and its routine application to detect apple juice adulteration and confirm its authenticity.

Keywords

Food authenticity, High resolution mass spectrometry, Liquid chromatography, Metabolomics, PCA, PLS-DA, OPLS-DA

Highlights

- Development of an UHPLC-HRMS metabolomics approach with great potential in juice authentication
- Discrimination between sample groups in two distinct authentication applications
- Relevant markers selected by OPLS-DA and ANOVA were tentatively identified
- Main discriminant compounds were identified as amino-acids and derivatives

1. Introduction

Food fraud is a worldwide issue, and recent crises (like the horse meat scandal in 2013) have sparked interest on food authentication among consumers and food industries [Brooks et al., 2017]. Concerning food fraud and Economically Motivated Adulteration between 1980 and 2010, fruit juices are one of the top ten products most at risk, particularly apple and orange juices [Moore et al., 2012]. Typical frauds on fruit juices include (1) dilution with water, (2) addition of sugars or organic acids, (3) addition of foreign fruits (mostly cheaper ones) and (4) false labeling of the product (cultivar or geographical origin, as well as production mode such as organic) [Vaclavik et al., 2011].

To facilitate the detection of food fraud, the Codex Alimentarius, the Association of the Industry of Juices and Nectars of the European Union (AIJN) and the European Commission have established guidelines and standards to define permitted practices and evaluate the quality and authenticity of juices [Directive 2012/12/EC; CODEX STAN 247-2005; AIJN Code of practice]. However, fruit juices

authentication may be challenging due to their complex chemical composition influenced by several factors such as variety, geographical origin, stage of maturity, storage conditions and processing techniques [Jandric et al., 2014; Cubero-Leon et al., 2018; Dasenaki et al., 2019].

Juice authentication is routinely performed by conventional analytical methods (called targeted methods) that are usually described and validated by the IFU (International Fruit and Vegetable Juice Association) [IFU website]. For example, sugars, organic acids, minerals, phenolic compounds and several volatile compounds are analyzed to authenticate direct apple juice samples [AIJN Code of practice, Wolter et al., 2008]. These methods are sensitive and usually provide low limits of detection and quantification as they have been developed to detect specific compounds or classes of compounds (e.g., molecular markers of foreign fruits or low fruit content). However, these targeted approaches generally focus on a specific fraud and may fail to reveal more sophisticated frauds such as false organic claims [Knolhoff and Croley, 2016; Dasenaki et al., 2019]. The illegal addition of vegetable water (such as water obtained during the grape juice concentration process) to orange juice concentrate, with the false claim of "orange juice not from concentrate", is another illustrative example since conventional $^{18}O/^{16}O$ isotope ratio analysis fails to detect this fraud; in that case, there is an additional health concern related to the presence of allergenic sulphur dioxide [Rinke and Jamin, 2018].

Therefore, it is important to move toward untargeted methods to detect adulteration and confirm authenticity [Dasenaki et al., 2019; Rinke, 2016]. Untargeted methods allow to have an overview of the sample, also called a fingerprint [Medina et al., 2019]. Thousands of compounds can be detected, making them more holistic than the conventional methods [Dasenaki et al., 2019]. These untargeted methods have emerged with the improvements of analytical techniques (e.g., the development of high resolution mass spectrometers) and the use of advanced statistical methods. Nuclear magnetic resonance (NMR) and mass spectrometry (MS) are widely used in the assessment of food authentication using untargeted methodology, in particular liquid chromatography coupled to high resolution mass spectrometry (LC-HRMS) [Cubero-Leon et al., 2014; Sobolev et al., 2019; Esteki et al., 2018; Danezis et al., 2016].

Metabolomics-based approaches using LC-HRMS have already been used in food safety assessment [Knolhoff et al., 2016; Delaporte et al., 2019] and revealed their potential. In the field of food authentication, untargeted LC-HRMS analysis coupled to chemometrics has been used to attest the geographical origin of saffron (100% of the investigated samples were correctly classified) [Rubert et al., 2016]. Using a similar methodology, Cavanna and co-workers have assessed the authentication of durum wheat based on geographical origin, with approximately 90% of samples correctly classified [Cavanna et al., 2020]. Moreover, metabolomics-based methodology using LC-HRMS has already been successfully implemented for juice authentication regarding geographical origin [Diaz et al., 2014] or for adulteration detection and classification of juices types and varieties [Vaclavik et al., 2012; Jandric et al., 2014; Jandric et al., 2017]. Similarly, Dubin et al. used this methodology to authenticate blackcurrant, specifically to detect adulteration with aronia, with a detection limit of 5% aronia concentrate in blackcurrant concentrate [Dubin et al., 2017]. Furthermore, the untargeted methodology was also used for pomegranate juice authentication allowing detection of 1% adulteration [Dasenaki et al., 2019].

Thereby, the metabolomics-based methodology appears to be a method of choice for juice authentication. However, this trend deserves confirmation and further methodological development. In particular, studies with large data sets and/or with models that offer broad applicability are needed to validate the potential of this methodology for food authentication [Cubero-Leon et al., 2018]. Moreover, the authentication of organic food also requires further work due to limited

number of studies regarding this topic, especially in the juice sector [Cuevas et al., 2017; Mihailova et al., 2021], and the lack of reliable analytical techniques to confirm the organic production of a sample [Cuevas et al., 2019].

In this work, apple juice samples from different farming and production processes (organic and conventional, direct and from concentrate juices) were analyzed using untargeted UHPLC-HRMS analysis. A data processing workflow was developed to select relevant features after peak detection. Based on these features, models were built for sample groups discrimination using chemometric tools in two distinct scenarios (direct juice *vs*. concentrated juice, and organic juice *vs*. conventional juice). Chemical markers allowing the discrimination were then tentatively identified, using online and in-house databases as well as UHPLC-HRMS/MS analyses. The discriminant potential of these marker compounds was evaluated using an independent set of samples.

2. Materials and methods

2.1. Reagents and chemicals

Methanol (MeOH), water and formic acid (FA), all LC-MS grade, were purchased from Fisher Scientific.

Some compounds known to be present in apple juice, and routinely analyzed by targeted methods, were purchased from Sigma-Aldrich: alpha-terpineol (purity: >99%), hexyl acetate (purity: 99%), ethyl 2-methylbutyrate (purity: > 98%), limonene (purity: 97%), phloridzin (purity: > 99%) and 2-methylbutyl acetate (purity: > 99%). Hydroxymethylfurfural (purity: 100%) was purchased from ACROS. These commercial standards were considered to assess the ability of our untargeted method to detect them. In addition, they may be considered as possible candidates for markers responsible of the discrimination between our sample groups. With the aim to develop a real untargeted method, these target compounds were not included in our inclusion list in our MS/MS experiments. Individual standard solutions were prepared in methanol with a concentration of 0.2 mg/L for most compounds, and of 0.5 mg/L for hydroxymethylfurfural and phloridzin. These solutions were analyzed using the UHPLC-HRMS analytical conditions described in section 2.3, in order to determine the *m/z* and retention time (RT) of the compounds which will be used to highlight their potential presence in the analyzed samples.

2.2. Samples description and preparation

One hundred and ten apple juice samples from several geographical origins and farming processes were collected (organic and non-organic juices; direct juices, concentrated juices and juices from concentrate). Samples were stored in the freezer until analysis. After thawing, aliquots (5 ml) of samples were centrifuged for 10 min at 4,500 rpm. The supernatant was collected and diluted with water directly into a vial before analysis. Three replicates per sample were prepared. Sample vials were randomized in the analytical sequence. Quality Control (QC) samples (pool of apple juice samples) and diluted QC samples were also prepared and analyzed every 10 injections. The repeated injections of QC samples were used to evaluate analytical performance. Also, analytical blanks were analyzed regularly to check for carry over (every 20 samples). Moreover, these blanks were useful to detect residual peaks corresponding to the mobile phases used.

Samples were analyzed in different batches. The first one contained 24 samples (12 organic and 12 conventional apple juices). The second batch contained 30 samples (15 direct apple juices and 15 concentrated apple juices). The third batch contained 26 samples including organic and conventional juices as well as direct and concentrated juices. Another set of samples coming from a different harvest year was also analyzed in a fourth batch in which MS/MS acquisition was performed; this

batch contained 30 samples (10 concentrated juice samples, 10 conventional direct juice samples and 10 organic direct juice samples). A detailed list of the samples is presented in Table A.1 (Supplementary material).

2.3. Analytical method

Analyses were performed on a ThermoFisher® Vanquish Flex UHPLC system, composed of a binary pump, refrigerated sampler and column oven, connected to a ThermoFisher® QExactive Plus Orbitrap® high resolution mass spectrometer (version 2.9) with a heated electrospray ion source (HESI). The UHPLC separation was achieved using a C18 Hypersil Gold column (150 x 2.1 mm, 1.9 µm) at a 0.3 mL/min flow-rate. The column temperature was set to 30°C. The mobile phases were water acidified with 0.1% FA (A), and MeOH acidified with 0.1% FA (B), with the following linear gradient elution: 0-2 min, B: 3%; 2-20 min, B: 3-98%; 20-24 min: B: 98%; 24-24.1 min, B: 98-3%; 24.1-32 min, B: 3%. The injection volume was 1 µL.

Raw data were acquired using TraceFinder software (version 3.1, ThermoFisher®). MS data were acquired in positive ion mode (ESI+) with a mass range set at *m/z* 120-1000 in full scan mode and with a resolution of 70,000. The parameters applied on the electrospray ion source are presented in Table A.2 (Supplementary material); MS data was acquired in centroid mode. The MS detector was weekly calibrated using the Pierce™ positive and negative ion calibration solution purchased from Thermo Fisher Scientific.

For MS/MS acquisition, full scan data-dependent analyses were carried out using an inclusion list. This inclusion list was established after the data processing of the first three batches where several features were identified as discriminant (24 features for both studies). The resolution was set at 17,500. An isolation window of ± 1 uma was used to select the *m/z* of interest at the expected retention time of the features (± 1 min). Three normalized collision energies were applied (10; 30 and 60 eV) for the MS/MS spectrum acquisition.

2.4. Data processing

Raw data files were analyzed using the Workflow4Metabolomics (W4M) platform (version 3.0) [Giacomoni et al., 2015] after conversion of the data files to mzXML format using ProteoWizard [Chambers et al., 2012]. The main steps of data processing are: (1) peak detection; (2) retention time alignment; (3) peaks grouping; (4) peak annotation; (5) data filtration and normalization, and (6) chemometric analysis. The first four steps were performed using functions of the XCMS (an acronym for various forms (X) of chromatography mass spectrometry) package [Smith et al., 2006] on the W4M platform as illustrated in Fig. 1.

The features, defined by their *m/z* and retention time, and their intensities in different samples were used for the statistical analysis as commonly reported [Cavanna et al., 2018]. The chemometric methods used were principal component analysis (PCA) for exploratory purpose, as well as partial least squares - discriminant analysis (PLS-DA), and orthogonal partial least squares - discriminant analysis (OPLS-DA) in order to build models for discrimination and classification of samples groups. Also, analysis of variance (ANOVA) and biosigner were used to reduce the number of features selected for models building, which may improve models quality.

2.4.1.Peak detection and alignment (XCMS)

All the data from the first three analytical batches were processed simultaneously as illustrated in Fig. 1. The XCMS phase includes the following steps. First, the peak detection and extraction is achieved using the “findChromPeaks” function with the centWave method [Tautenhahn et al., 2008].

During this step, the chromatograms are described as a 2D-matrix where each peak is described by a combination of its *m/z* value and retention time (RT), called "feature". The selected retention time for each peak is the time corresponding to its apex of the intensity value. Then, the "groupChromPeaks" function is used to group the extracted peaks across all the samples. This step is applied to group ions with close RT between the samples. After this step, *m/z* and RT values are averaged in the data matrix. The peaks are next aligned using the "adjustRtime" function to correct the RT across the samples and then grouped again. Finally, the "fillChromPeaks" function is used to identify features where there is no intensity value for some samples and the signal is integrated in the region of the determined feature to avoid missing values. The XCMS parameters for each step were optimized from the QC samples and are presented in Table A.3 (Supplementary material). A data matrix is then generated, giving the area of each peak for each feature and for each sample. Thus, the features are the variables of the models presented in this study.

2.4.2. Data filtration and batch correction

To perform the subsequent data filtration step, the data matrix was split in three distinct data matrices, as shown in Fig. 1, corresponding to the three initial analytical batches, in order to perform filtration steps within each analytical batch. These filtration steps were needed to remove irrelevant information as the number of features detected by XCMS was very high (about 20,000).

First, all peaks corresponding to the dead volume and the column flush were excluded, which means that all the features with a retention time lower than 1.7 min were removed from the data matrix. Then, features that mainly result from blank analyses were removed by calculating the fold change in blanks and samples analyses. For a feature, when the ratio of samples fold change over blanks fold change is lower than 4, this feature is deleted. In this way, between 15% and 20% of the detected features were removed. Finally, features showing a poor stability (relative standard deviation (RSD) higher than 30%) according to QC analyses were also excluded. Similarly, features for which the ratio of RSD pool over RSD sample is higher than 1.25 were deleted. At the end of this step, about 10,000 features remained. Analytical signal drift within the analytical batch was corrected using a LOESS regression model using the QC sample injections, employing the *Batch Correction* module on the W4M platform.

Then, the three data matrices corresponding to the three sample batches were merged (see Fig. 1) and a second batch correction was applied to correct analytical signal drift between analytical batches by the use of the QC sample injections.

The data matrix was then normalized using the Probabilistic Quotient Normalization method (PQN) [Dieterle et al., 2006] using the QC samples. Its purpose is to limit potential dilution effects that can affect restricted regions of the data. First, the median of each feature in QC samples is calculated, providing a reference vector. Then, the values for each ion in samples are divided by this reference vector. A median of the ratios for each sample is generated. Finally, initial values of each sample are divided by the ratios median.

Prior to chemometrics analysis, the data matrix was Pareto scaled. Then, the data matrix was split to create two distinct authentication studies: the first one contained samples from batches 2 and 3 to evaluate the discrimination between pure and concentrated juice samples (58 samples in the data set); the second study contained samples from batches 1 and 3 to evaluate the discrimination between organic and conventional juice samples (54 samples in the data set).

2.4.3.Chemometrics

Multivariate statistical analyses were performed on the W4M platform using unsupervised and supervised techniques. PCA was first performed to have an initial visualization of the data sets and to detect outliers. In order to evaluate the ability of this methodology to discriminate the apple juice samples, PLS-DA and OPLS-DA were used. These models were built using a 7-fold cross validation; by this way, each data set was divided into 7 different parts. Each model was next built using 6 parts (train set) and tested using the 7^{th} part (test set); this step was then iterated until all the parts were used as test set. The cross validation procedure permitted to determine the optimal number of latent variables (LV) to build the PLS-DA and OPLS-DA models [Ballabio and Consonni, 2013; Wold et al., 2001]. A new LV was added if the Q2Y obtained with this LV was greater than 0.01. Indeed, the Q2Y was calculated from the ratio of PRESS (predictive residual sum of squares) including the new LV over RSS (residual sum of squares) calculated from the model with the previous LV [Wold et al., 2001]. The quality of the built models was assessed by the goodness of fit (R2X), the proportion of the response matrix variance explained by the model (R2Y) and the predictive performance of the model (Q2Y). These three metrics have values between 0 and 1. The higher they are, the better the performance of the model. The Q2Y metric is particularly important here, as it represents the prediction efficiency of the model. An empirical value of 0.4 for Q2Y has been previously established to judge the quality of the model [Worley and Powers, 2012].

An analysis of variance (ANOVA) was also performed to select significant features between the two studied groups (pure *vs* concentrated juices and organic *vs* conventional juices); a maximum accepted p-value of 0.01 was chosen in order to select significant features. The features identified by the ANOVA were used to build new PLS-DA and OPLS-DA models to compare models quality with lower features.

The biosigner tool [Rinaudo et al., 2016] present on the W4M platform was also used for feature selection. Briefly, this algorithm allows to obtain the smallest number of features which have the most significant contribution in models performance (this module performed PLS-DA, Random Forest (RF) and Support Vector Machine (SVM) models) after performing several iterations. The iterations stop when the number of significant features remains equal to that of the previous iteration. Again, the features selected by biosigner were used to build new PLS-DA and OPLS-DA models.

2.4.4. Annotation

Significant features were selected based on their results after the chemometric tools used, particularly OPLS-DA and ANOVA results. After having investigated the MS spectra of those discriminant features, the adduct type of the observed *m/z* was identified which permitted to determine the exact mass of the compound and consequently to suggest molecular formulas. In order to tentatively annotate these features that discriminate the samples, the online databases HMDB [Wishart et al., 2018] and FooDB [FooDB, 2021] (as we are studying apple juice samples) were used. Moreover, the obtained MS/MS spectra were used to confirm the annotation by comparing them to two spectral databases: mzCloud and MassBank. In addition, some commercial compounds known to be present in apple juices were analyzed thanks to available standards as detailed in section 2.1, enabling to build an in-house database.

3. Results and discussion

3.1. Study 1: Authentication of pure apple juices

3.1.1. Principal component analysis

PCA is the most common unsupervised multivariate statistical technique [Medina et al., 2019b; Oliveri and Simonetti, 2016] used for exploratory purposes. It was used here to evaluate the

reproducibility of three replicates of the same sample. For this study, 58 samples were considered with three replicates per sample (resulting in a total of 174 samples). PCA was applied on two distinct data matrices, containing either all values (i.e., including separate triplicate values) or only a single value (being the mean of the three replicates) for each sample. In both cases, no outlier was observed on the PCA score plots, so that we considered the three replicates to be reproducible. Consequently, only the average of sample triplicates was considered for the following statistical analyses.

As shown in Fig. A.1 (Supplementary material), the first three principal components explained about 50% of the variance (PC1: 27%; PC2: 14%; PC3: 8%). The replicates of the QC samples were fairly close on the PCA scores plot, showing a good system stability during the analysis. A slight dispersion was noticed in Fig A.1a, with two subsequent groups for the QC samples, in line with the two distinct analytical batches; this observation highlights an analytical drift not completely corrected. Interestingly, a trend seemed to appear for the discrimination between the two groups of samples (single strength *vs.* both concentrated juices and juices from concentrate) on the PC3 axis, even though no clear separation could be achieved.

Conversely, group separation of fruit juices based on the type of fruit were already reported using PCA on UHPLC-HRMS data, with a distinct cluster for apple juices [Vaclavik et al., 2012]. Guo et al. also reported group separation of fresh squeezed apple juices based on varieties by performing a PCA on their concentrations in 23 polyphenols [Guo et al., 2013]. Therefore, it can be assumed that the apple juice production method has fewer differences in the UHPLC-HRMS fingerprint, which explains why no group separation was observed in our PCA.

3.1.2. Classification and prediction models: PLS-DA and OPLS-DA

PLS-DA and OPLS-DA models were built using the features left after the between batch correction (9,234 features) and from the 58 samples.

PLS-DA models were already reported for classification purpose of orange juices, with a satisfactory classification rate of samples regarding geographical origin (after cross-validation, the model showed a 100% classification capacity) [Diaz et al., 2014]; unfortunately, in our study PLS-DA model remained unsatisfactory (data not shown). The obtained model was built using two latent variables and had a goodness of prediction of 0.37 and a goodness of fit of 0.27. OPLS-DA models were previously found interesting for discrimination of Saffron sample origins [Rubert et al., 2016]; results from our data also showed samples discrimination between single strength and both concentrated juices and juices from concentrate, as presented in Fig. 2.

The model metrics indicated that the OPLS-DA model was quite satisfactory (as shown in Fig. 2) with a goodness of fit of about 50% and a prediction capacity of about 60%. This OPLS-DA model was built using 9,236 features, so that these metrics might be improved with fewer features. In their study on Saffron, Rubert and co-workers reported that the best OPLS-DA model was obtained using 8 features (of about 5,000 features detected) with 85% of prediction capacity and 97% of goodness of fit [Rubert et al., 2016]. Nevertheless, on Fig. 2 it can be observed that one concentrated juice sample was really close to the direct juice samples group. This observation can lead to incorrect classification or prediction of samples. Consequently, further data processing was tested to improve the modeling. Moreover, the high number of features used for building our OPLS-DA model may have induced overfitting. It was thus important to reduce the number of features used for this model.

3.1.3. Feature selection using ANOVA before PLS-DA and OPLS-DA

Selection was needed to reduce as much as possible the number of features to be compare with other analytical batches. This is necessary if we want to implement this methodology as a routine analysis method for apple juice authentication assessment. ANOVA and similar t-test have already been used to identify and select significantly different features between groups of samples [Llano et al., 2018; Bat et al., 2018].

Performing an ANOVA proved to be greatly helpful: from the about 10,000 features obtained at the end of the filtration steps, the ANOVA identified almost 2,000 significantly different features between the two sample groups. Again, the PLS-DA model gave unsatisfactory results (data not shown). The model was built using two latent variables and had a predictive ability of 0.58. In the score plot, the two sample groups were not differentiated. Conversely, the OPLS-DA model obtained with these identified features was again quite satisfactory (based on the metrics values), with more variance being explained by the first latent variable (30% instead of 13% previously). However, the separation of the two groups was not improved, being even quite worse (Fig. A.2a of Supplementary material).

3.1.4. Feature selection using biosigner before PLS-DA and OPLS-DA

The biosigner module present on the W4M platform was also tested for features selection since it allows selecting the fewest number of features to build discrimination models. Accordingly, only 20 features were identified by this tool here. Unfortunately, with these 20 features, the resulting OPLS-DA model showed a worse discrimination of the two groups, even though the direct juice samples stayed close together (Fig. A.2b of Supplementary material). The metrics of the model clearly decreased, confirming the low quality of this model. The PLS-DA model obtained was also unsatisfactory (data not shown). One latent variable was used to build this model and a predictive ability of 0.27 was obtained.

Almost all features selected by the biosigner tool were also selected by the ANOVA (90%). The 20 features seemed thus to be discriminant, but it can be emphasized that as the number of samples was quite low (58 samples), the features selected were not sufficiently discriminant to improve the OPLS-DA model quality. A larger set of reference samples would be required to establish a robust routine model.

3.2. Study 2: authentication of organic apple juices

3.2.1. Principal component analysis

In this study, 54 samples were used to build the PCA. As in the previous study, the reproducibility of the triplicates was evaluated using the PCA scores plots. As the replicates showed to be reproducible, the average of the three replicates per samples was used for the next chemometric analysis. PCA scores plots of the filtered data using the mean of the triplicates are shown in Fig. A.3 of Supplementary material. A good system stability was also observed for this study since the replicates of the QC sample were clustered on the PCA scores plot. As in the previous study, two groups of QC samples could be distinguished, showing that the analytical drift was not completely corrected. However, the correction seemed to be better than in the first study because the QC replicates were less dispersed.

Group separation between organic and conventional juice samples was not achieved by PCA. Using the first three principal components, about 60% of the variance was explained (PC1: 32%; PC2: 15% and PC3: 8%). Cuevas and coworkers also reported previously that PCA did not allow to separate organic from conventional orange juices using UHPLC-HRMS analysis [Cuevas et al., 2017].

3.2.2. Classification and prediction models: PLS-DA and OPLS-DA

In order to build models for sample classification, PLS-DA and OPLS-DA analysis were performed (Fig. A.4. of Supplementary material). PLS-DA and OPLS-DA models were both built using a 7-fold cross validation on the 54 samples of the study. OPLS-DA enabled a clear separation between the two groups. This is in line with another study where organic and conventional juices were discriminated using an OPLS-DA model, with a specificity and sensitivity of nearly 90% for both sample classes after cross-validation [Cuevas et al., 2017]. OPLS-DA models satisfactorily discriminated organic and conventional carrot samples analyzed by UHPLC-HRMS, with a classification rate of about 80% using a validation data set [Cubero-Leon et al., 2018].

The predictive ability of the OPLS-DA model was good (Q2Y: 0.746). As this model was obtained with a high number of features (near 8400 features), it could be hypothesized that it could be improved with a reduced number of features. Further works should focus on the external assessment of the models performance, which was not allowed by the number of samples in this study.

3.2.3. Feature selection using ANOVA before PLS-DA and OPLS-DA

In this study, the ANOVA found 1,422 significantly different features between the organic and conventional juice samples (from almost 10,000 features detected). PLS-DA and OPLS-DA models obtained from these features are presented in Fig. 3.

The new models obtained with a reduced number of features showed similar metrics compared to the previously obtained models (Fig. A.4 of Supplementary material). The discrimination between the two sample groups was still not observed using PLS-DA model. On the contrary, OPLS-DA model showed a clear separation. The predictive ability of this model indicated that it had a good performance (Q2Y: 0.785) and it was slightly better than the OPLS-DA obtained with all the features. The percentage of variance explained by the first LV had increased to 24% with the feature selection.

3.2.4. Feature selection using biosigner before PLS-DA and OPLS-DA

Again, the biosigner tool was used to find the smallest number of most significant features. This module found 48 features. To evaluate whether these selected features were the most significant, PLS-DA and OPLS-DA models were built using a 7-fold cross-validation. In contrast to the results obtained from the feature selection using ANOVA, the obtained PLS-DA and OPLS-DA models were not improved. The metrics showed that models were quite worse than with all the features (Fig. A.5 of Supplementary material).

Only 14 of the 48 features selected by the biosigner tool were also selected by the ANOVA. Most of the features selected by this module were chosen based on their performance using SVM models. SVM models can perform very well but they require lots of data (ideally thousands of samples). In this study, there were only 54 samples, so the features selected using SVM models were not discriminant enough to increase the PLS-DA and OPLS-DA model metrics.

3.3. Tentative identification of discriminant features in both studies

The number of features remained high, even after selection with ANOVA (about 1,500 features). To reduce this number while keeping the most discriminant features, it was decided to filter them according to their VIP (Variable Importance on Projection) value calculated during the construction of the PLS-DA and OPLS-DA models. The VIP value of a feature indicates its importance on the model building: the higher VIP value, the more discriminating the feature is [Wold et al., 2001]. In a previous study, filtration based on the VIP value successfully selected 8 features out of about 5,000 features

detected [Rubert et al., 2016]. Other authors reported the use of VIP values to select discriminant features by retaining 25 features (out of about 5,000 features detected) that were further tentatively identified [Cavanna et al., 2020]. Cubero-Leon and colleagues also used a similar criterion (VIP greater than 1) applied to remove features contributing to other variability (year of harvest); they were able to build successful OPLS-DA models to discriminate between organic and conventional carrot samples [Cubero-Leon et al., 2018]. In the literature, different VIP values between 1 and 2 have been used as a filtration criterion. In this work, a filter was applied to keep features having a VIP value greater than 1, as proposed in different articles [Gorrochategui et al., 2016; Pezzati et al., 2020]. After this filtration, about 150 features remained for both authentication applications considered in our work.

To reduce the number of features to be identified, both results from ANOVA and from OPLS-DA were used. We focused on the features with the highest VIP and the lowest p-value, and attempted to identify them using online databases (HMDB and FooDB). Based on this strategy, less than 15 features were selected in each study for further identification, as indicated in Table 1 and Table A.4 (Supplementary material). Few of these features were also selected by the biosigner tool. Examples of chromatograms for one feature identified as discriminating for each study are shown in Fig. 4 and Fig. A.6 (Supplementary material).

Some features had the same retention time, being either coeluted chromatographic peaks or fragments and/or adducts of a unique compound. The observation of the MS spectra permitted to identify features which correspond to a same molecule (Table 1); in particular, the presence of certain adducts such as $[M+NH_4]^+$ and $[M+K]^+$ allowed to attribute the adduct type of the observed *m/z*. This was mostly the case for the features identified in the second study. As presented in Table 1 and Table A.4, a majority of features were still unknown as no matches were obtained on the online databases used. For some features, several compounds matched the exact mass defined. It is interesting to observe that results from HMDB and FooDB were really close, being a good starting point to annotate features but still insufficient: the compounds of interest may not be present on these databases.

In order to improve the annotation of these discriminant features, MS/MS acquisitions were performed on a new set of samples (30 samples containing concentrated juice samples, conventional direct juice samples and organic direct juice samples). Only few results were obtained from the databases search, either with the monoisotopic mass or with the proposed molecular formula. Interestingly, two amino acids (methionine and isoleucine or norleucine) could be proposed as discriminant markers between organic and conventional apple juice (Table 1); this result seems realistic since those compounds were already reported in apple juice samples [Ma et al., 2018]. In particular a biosynthesis pathway leading to isoleucine formation in ripening apple fruit has been recently reported (Sugimoto et al., 2021). The same methodology was applied for the features identified in the authentication of pure apple juices with N-(1-deoxy-1-fructosyl)phenylalanine proposed as a marker (Table A.4. of Supplementary material). Xu et al. also reported an amino-acid (L-glutamine) to discriminate from concentrate and not from concentrate orange juices [Xu et al., 2020]. Further investigation is needed to improve the annotation of the identified discriminant features, probably by using other online databases or building in-house database.

During the MS/MS experiments, a full scan analysis was also acquired. It was thus possible to use these independent acquisitions to evaluate the potential of the discriminant features to serve as marker compounds. It is noteworthy that the previously selected features were successfully

observed on this fourth analytical batch (only one feature was missing because it has a low intensity); nevertheless, only a few of them still showed a trend in the discrimination of sample groups. In particular, for the authentication of direct apple juices, 5 features still showed a difference in intensity between the two sample groups; these features might be used as markers compounds for the concentrated juice characteristic. On the other hand, for the authentication of organic apple juices, no trend was observed for the discrimination of the two samples groups by observing the intensity of the features between the two sample groups.

It can be emphasized that these independent acquisitions permitted to highlight some marker compounds as they were characteristic of the process type used (juice concentration). For the organic juice characteristic, these new samples came from a different harvest year, which may explain that the discriminant features found previously may fail to discriminate these new acquisitions. Cubero-Leon and co-workers reported that the harvest year was one of the most important variabilities in their studied samples [Cubero-Leon et al., 2018]. On the contrary, Diaz et al. identified a biomarker for orange origin which seems to be independent from the harvest year [Diaz et al., 2014]. Further investigation is thus needed to find reliable features for the authentication of organic apple juice samples and to confirm the use of the 5 features for the authentication of direct apple juice samples.

Based on the analysis of standards, two detected features might be assigned to phloridzin (p-value: 5.83 E-03) and alpha-terpineol (p-value: 1.08 E-04) according to their *m/z* and retention time; unfortunately, these two features were outside the list of Table 1. It is not surprising that phloridzin was not a discriminant compound as it is a naturally present molecule in apples, with varying concentrations depending on different factors such as variety or processing technology used [Spinelli et al., 2016]. The remaining standards were not detected in our samples, possibly because they were not concentrated enough to be observed, while the other one (hydroxymethylfurfural) routinely analyzed by LC-UV may not be present in the juice samples analyzed.

4. Conclusions

This work presents a methodology combining untargeted LC-HRMS analysis and chemometric tools to authenticate apple juice samples. The OPLS-DA models showed good performance in sample classification, especially for the discrimination between organic and conventional sample juices (nearly 80% of predictive ability). To confirm their classification and prediction performance, further validation of these models using an external data set is required.

Coupling the results of ANOVA and OPLS-DA seems to be an interesting methodology to determine the discriminant features as it permitted to reduce the number of detected features (from almost 10,000 features detected to about 150 features) while keeping significant and discriminant features. According to the chemometric tools used (OPLS-DA and ANOVA) about 20 features have been identified as significantly discriminant and tentatively identified for the first time. Some compounds were tentatively annotated as amino-acids and derivatives, and a few markers were confirmed by MS/MS experiments. Interestingly, application of our analytical method to a new set of samples showed that some features retained a tendency to discriminate between the two groups of samples, mainly for authentication of direct apple juices.

The main additional research concerns the annotation workflow, which is the most time-consuming part of this methodology. By building an in-house database, the identification of marker compounds can be faster as the obtained mass spectra will be better compared than by using an online database, the same instrument being used. By identifying the compounds responsible for the discrimination,

492 they could be analyzed in a routine analysis for apple juice authentication. Further investigation is
493 needed to correctly identify the compounds by the analysis of standards.

494 The proposed analytical methodology enabled, for the very first time, the authentication of apple
495 juice samples in two distinct scenarios using a single analysis (organic *vs*. conventional samples and
496 single strength juice *vs*. both concentrated juice and juice from concentrate samples). Other
497 chemometric models could be developed to implement juice discrimination based on variety and/or
498 geographical origin, in addition to the scenarios presented here.

499

500 **Acknowledgement**

501 The authors warmly thank Dr. Peter Rinke from SGF for kindly providing them with several samples of
502 apple juices used in this study. They are thankful for the financial support provided by the
503 Association Nationale Recherche et Technologie (ANRT) through the CIFRE program (CIFRE
504 n°2018/0937).

505

506 **Conflict of interest**

507 The authors declare that they have no commercial or financial relationships that could have influence
508 the research conducted in this paper.

509

510 **Appendix A. Supplementary Data**

511

References

AIJN Code of Practice (2020), AIJN European Fruit Juice Association.

Ballabio, D., Consonni, V., 2013. Classification tools in chemistry. Part 1: linear models. PLS-DA. Analytical Methods 5, 3790. https://doi.org/10.1039/c3ay40582f

Bat, K.B., Vodopivec, B.M., Eler, K., Ogrinc, N., Mulič, I., Masuero, D., Vrhovšek, U., (2018). Primary and secondary metabolites as a tool for differentiation of apple juice according to cultivar and geographical origin. *LWT – Food Science and Technology,* 90, 238–245. https://doi.org/10.1016/j.lwt.2017.12.026

Brooks, S., Elliott, C.T., Spence, M., Walsh, C., Dean, M., (2017). Four years post-horsegate: an update of measures and actions put in place following the horsemeat incident of 2013. *npj Science of Food*, 1. https://doi.org/10.1038/s41538-017-0007-z

Cavanna, D., Righetti, L., Elliott, C., & Suman, M. (2018). The scientific challenges in moving from targeted to non-targeted mass spectrometric methods for food fraud analysis: A proposed validation workflow to bring about a harmonized approach. *Trends in Food Science and Technology*, 80, 223-241. https://doi.org/10.1016/j.tifs.2018.08.007

Cavanna, D., Loffi, C., Dall'Asta, C., Suman, M., (2020). A non-targeted high-resolution mass spectrometry approach for the assessment of the geographical origin of durum wheat. *Food Chemistry,* 317, 126366. https://doi.org/10.1016/j.foodchem.2020.126366

Chaleckis, R., Meister, I., Zhang, P., Wheelock, C.E., (2019). Challenges, progress and promises of metabolite annotation for LC–MS-based metabolomics. *Current Opinion in Biotechnology*, 55, 44-50. https://doi.org/10.1016/j.copbio.2018.07.010

Chambers, M.C., MacLean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., ... Mallick, P., (2012). A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology, 30*(10), 918‑920. https://doi.org/10.1038/nbt.2377

Codex Alimentarius: Codex General Standard for Fruit Juices and Nectars (2005) CODEX STAN 247-2005

Cuberon-Leon, E., Peñalver, R., Maquet, A., (2014). Review on metabolomics for food authentication. *Food Research International*, 60, 95-107. https://doi.org/10.1016/j.foodres.2013.11.041

Cubero-Leon, E., De Rudder, O., Maquet, A., (2018). Metabolomics for organic food authentication: Results from a long-term field study in carrots. *Food Chemistry,* 239, 760–770. https://doi.org/10.1016/j.foodchem.2017.06.161

Cuevas, F.J., Pereira-Caro, G., Moreno-Rojas, J.M., Muñoz-Redondo, J.M., Ruiz-Moreno, M.J., (2017). Assessment of premium organic orange juices authenticity using HPLC-HR-MS and HS-SPME-GC-MS combining data fusion and chemometrics. *Food Control*, 82, 203–211. https://doi.org/10.1016/j.foodcont.2017.06.031

Cuevas, F.J., Pereira-Caro, G., Muñoz-Redondo, J.M., Ruiz-Moreno, M.J., Montenegro, J.C., Moreno-Rojas, J.M., (2019). A holistic approach to authenticate organic sweet oranges (Citrus Sinensis L. cv Osbeck) using different techniques and data fusion. *Food Control,* 104, 63–73. https://doi.org/10.1016/j.foodcont.2019.04.012

Danezis, G.P., Tsagkaris, A. S., Camin, F., Brusic, V., Georgiou, C.A., (2016). Food authentication: Techniques, trends & emerging approaches. *Trends in Analytical Chemistry,* 85, 123–132. https://doi.org/10.1016/j.trac.2016.02.026

Dasenaki, M.E. & Thomaidis, N.S. (2019). Quality and authenticity control of fruit juices - A review. *Molecules,* 24, 1014. https://doi.org/10.3390/molecules24061014

Delaporte, G., Cladiere, M., Jouan-Rimbaud Bouveresse, D., Camel, V., (2019). Untargeted food contaminant detection using UHPLC-HRMS combined with multivariate analysis: Feasibility study on tea. *Food Chemistry*, 277, 54‑62. https://doi.org/10.1016/j.foodchem.2018.10.089

Diaz, R., Pozo, O.J., Sancho, J.V., Hernandez, F., (2014). Metabolomic approaches for orange origin discrimination by ultra-high performance liquid chromatography coupled to quadrupole time-of-flight mass spectrometry. *Food Chemistry*, 157, 84-93. https://doi.org/10.1016/j.foodchem.2014.02.009

Dieterle, F., Ross, A., Schlotterbeck, G., Senn, H., (2006). Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Analytical Chemistry*, 78(13), 4281‑4290. https://doi.org/10.1021/ac051632c

Directive 2012/12/EC, (2012). Council Directive Relating to Fruit Juices and Certain Similar Products Intended for Human Consumption of 19 April 2012

Dubin, E., Dumas, A.-S., Rebours, A., Jamin, E., Ginet, J., Lees, M., Rutledge, D.N., (2017). Detection of Blackcurrant Adulteration by Aronia Berry Using High Resolution Mass Spectrometry, Variable Selection and Combined PLS Regression Models. *Food Analytical Methods,* 10, 683–693. https://doi.org/10.1007/s12161-016-0638-8

Esteki, M., Simal-Gandarab, J., Shahsavaria, Z., Zandbaafa, S., Dashtakia, E., Heydenc, Y.V., (2018). A review on the application of chromatographic methods, coupled to chemometrics, for food authentication. *Food Control*, 93, 195-182. https://doi.org/10.1016/j.foodcont.2018.06.015

FooDB, (2021). Food Database, https://foodb.ca/ accessed on November 24th 2021

Giacomoni, F., Le Corguillé, G., Monsoor, M., Landi, M., Pericard, P., Pétéra, M., ... Caron, C., (2015). Workflow4Metabolomics: A collaborative research infrastructure for computational metabolomics. *Bioinformatics, 31*(9), 1493–1495. https://doi.org/10.1093/bioinformatics/btu813

Gorrochategui, E., Jaumot, J., Lacorte, S., Tauler, R, (2016). Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: overview and workflow. *Trends in Analytical Chemistry*, 82, 425−442. https://doi.org/10.1016/j.trac.2016.07.004

Guo, J., Yue, T., Yuan, Y., Wang, Y., (2013). Chemometric classification of apple juices according to variety and geographical origin based on polyphenolic profiles. *Journal of Agricultural and Food Chemistry*, 61, 6949−6963. https://doi.org/10.1021/jf4011774

IFU, (2021). International Fruit and Vegetable Juice Association, https://ifu-fruitjuice.com/ accessed on May 28th 2021

Jandric, Z., Roberts, D., Rathor, M.N., Abrahim, A., Islam, M., Cannavan, A., (2014). Assessment of fruit juice authenticity using UPLC-QToF MS: A metabolomics approach. *Food Chemistry*, 148, 7–17. https://doi.org/10.1016/j.foodcont.2018.06.015

590 Jandric, Z., Islam, M., Singh, D.K., Cannavan, A., (2017). Authentication of Indian citrus fruit/fruit
591 juices by untargeted and targeted metabolomics. *Food Control*, 71, 181–188.
592 https://doi.org/10.1016/j.foodcont.2015.10.044

593 Knolhoff, A. M., Croley, T. R., (2016). Non-targeted screening approaches for contaminants and
594 adulterants in food using liquid chromatography hyphenated to high resolution mass
595 spectrometry. *Journal of Chromatography A*, 1428, 86‑96.
596 https://doi.org/10.1016/j.chroma.2015.08.059

597 Llano, S.M., Muñoz-Jiménez, A.M., Jiménez-Cartagena, C., Londoño-Londoño, J., Medina, S., (2018).
598 Untargeted metabolomics reveals specific withanolides and fatty acyl glycoside as tentative
599 metabolites to differentiate organic and conventional *Physalis peruviana* fruits. *Food Chemistry*,
600 244, 120–127. https://doi.org/10.1016/j.foodchem.2017.10.026

601 Ma, S., Neilson, A.P., Lahne, J., Peck, G.M., O'Keefe, S.F., Stewart, A.C., (2018). Free amino acid
602 composition of apple juices with potential for cider making as determined by UPLC-PDA. *Journal*
603 *of the Institute of Brewing*, 124, 467–476. https://doi.org/10.1002/jib.519

604 Medina, S., Pereira, J.A., Silva P., Perestrelo, R., Câmara, J.S., (2019a). Food fingerprints – a valuable
605 tool to monitor food authenticity and safety. *Food Chemistry*, 278, 144-162.
606 https://doi.org/10.1016/j.foodchem.2018.11.046

607 Medina, S., Perestrelo, R., Silva, P., Pereira, J., Câmara, J.S., (2019b). Current trends and recent
608 advances on food authenticity technologies and chemometric approaches. *Trends in Food Science*
609 *& Technology*, 85, 163-176. https://doi.org/10.1016/j.tifs.2019.01.017

610 Mihailova, A., Kelly, S.D., Chevallier, O.P., Elliott, C.T. (2021). High-resolution mass spectrometry-
611 based metabolomics for the discrimination between organic and conventional crops: A review.
612 *Trends in Food Science & Technology*, 110, 142-154. https://doi.org/10.1016/j.tifs.2021.01.071

613 Moore, J., Spink, J., Lipp, M., (2012). Development and Application of a Database of Food Ingredient
614 Fraud and Economically Motivated Adulteration from 1980 to 2010. *Journal of Food Science*, 77,
615 R118-R126. https://doi.org/10.1111/j.1750-3841.2012.02657.x.

616 Oliveri, P., & Simonetti, R., (2016). Chemometrics for Food Authenticity Applications. In G. Downey
617 (Eds.), *Advances in Food Authenticity Testing* (pp.701-728). Woodhead Publishing is an imprint of
618 Elsevier

619 Pezzatti, J., Boccard, J., Codesido, S., Gagnebin, Y., Joshi, A., Picard, D., Gonzalez-Ruiz, V., Rudaz, S.,
620 (2020). Implementation of liquid chromatography - high resolution mass spectrometry methods
621 for untargeted metabolomic analyses of biological samples: a tutorial. *Analytical Chimica Acta,*
622 1105, 28e44. https://doi.org/10.1016/j.aca.2019.12.062.

623 Rinke, P., (2016). Tradition Meets High Tech for Authenticity Testing of Fruit Juices. In G. Downey
624 (Ed.), *Advances in Food Authenticity Testing* (pp.625-665). Woodhead Publishing is an imprint of
625 Elsevier

626 Rinke, P., & Jamin, E., (2018). Fruit juices. In Morin, J.-F., Lees, M. (Eds.), *FoodIntegrity Handbook: A*
627 *guide to food authenticity issues and analytical solutions*, 1st ed. Eurofins Analytics France.
628 https://doi.org/10.32741/fihb

Rinaudo, P., Boudah, S., Junot, C., Thévenot, E.A., (2016). biosigner: A New Method for the Discovery of Significant Molecular Signatures from Omics Data. *Frontiers in Molecular Biosciences* 3. https://doi.org/10.3389/fmolb.2016.00026

Rubert, J., Lacina, O., Zachariasova, M., Hajslova, J., (2016). Saffron authentication based on liquid chromatography high resolution tandem mass spectrometry and multivariate data analysis. *Food Chemistry,* 204, 201–209. https://doi.org/10.1016/j.foodchem.2016.01.003

Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R., Siuzdak, G., (2006). XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Analytical Chemistry*, 78, 779-787. https://doi.org/10.1021/ac051437y

Sobolev, A.P., Thomas, F., Donarski, J., Ingallina, C., Circi, S., Marincola, F.C., Capitani, D., Mannina, L., (2019). Use of NMR applications to tackle future food fraud issues. *Trends in Food Science & Technology*, 91, 347-353. https://doi.org/10.1016/j.tifs.2019.07.035

Spinelli, F.R., Dutra, S.V., Carnieli, G., Leonardelli, S., Drehmer, A.P., Vanderlinde, R., (2016). Detection of addition of apple juice in purple grape juice. *Food Control,* 69, 1–4. https://doi.org/10.1016/j.foodcont.2016.04.005

Sugimoto, N., Engelgau, P., Jones, A.D., Song, J., Beaudry, R., (2021). Citramalate synthase yields a biosynthetic pathway for isoleucine and straight- and branched-chain ester formation on ripening apple fruit. *PNAS,* 118(3), e2009988118. https://doi.org/10.1073/pnas.2009988118

Tautenhahn, R., Bottcher, C., Neumann, S., (2008). Highly sensitive feature detection for high resolution LC/MS*. BMC Bioinformatics*, 9. https://doi.org/10.1186/1471-2105-9-504

Vaclavik, L., Schreiber, A., Lacina, O., Cajka, T., (2012). Liquid chromatography-mass spectrometry-based metabolomics for authenticity assessment of fruit juices. *Metabolomics*, 8, 793-803. https://doi.org/10.1007/s11306-011-0371-7

Wishart, D.S., Feunang, Y.D., Marcu, A., Guo, A.C., Liang, K., Vázquez-Fresno, R., … Scalbert, A., (2018). HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Research,* 46, D608–D617. https://doi.org/10.1093/nar/gkx1089

Wold, S., Sjöström, M., Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems,* 58, 109–130.

Wolter, C., Gessler, A., Winterhalter, P., (2008). Aspects when evaluating apple-juice aroma. Fruit processing, 64-80

Worley, B., Powers, R., 2012. Multivariate Analysis in Metabolomics. Current Metabolomics 1, 92–107. https://doi.org/10.2174/2213235X11301010092

Xu, L., Xu, Z., Kelly, S., Liao, X., (2020). Integrating untargeted metabolomics and targeted analysis for not from concentrate and from concentrate orange juices discrimination and authentication. *Food Chemistry*, 329, 127130. https://doi.org/10.1016/j.foodchem.2020.127130

Figure Captions

Fig. 1. Workflow of the data treatment using W4M* (RSD: relative standard deviation) ** text in italic refers to W4M functions.*

Fig. 2. Scores plot for OPLS-DA obtained with cross-validation (blue circles, concentrated juices and juices from concentrate; red crosses, direct juices). The black ellipse represents 95% of the variability, the blue and red ellipses represent 95% of the multivariate distributions of the sample groups.

Fig. 3. (a) Scores plot of PLS-DA and (b) scores plot of OPLS-DA obtained after features selection using ANOVA (blue circles: organic juice samples; red crosses, conventional juice samples). The black ellipse represents 95% of the variability, the blue and red ellipses represent 95% of the multivariate distributions for each sample groups.

Fig. 4. Chromatogram of feature 13 for authentication of organic apple juices (black, organic juice samples; red, conventional juice samples)

Table 1: Discriminant features for authentication of organic apple juices (compounds confirmed based on MS/MS data are indicated in bold characters).

# Compound	# Feature	Detected *m/z*	Adduct type	RT (min)	p-value	VIP	Characteristic	Monoisotopic mass	Proposed molecular formula	Proposed compounds (FooDB)	Proposed compounds (HMDB)
1	1*	132.1019	$[M+H]^+$	3.32	1.0E-05	10.7	Conv > Org	131.0947	$C_6H_{13}NO_2$	Leucine, **Isoleucine**, 6-Deoxyfagomine, Alloleucine, **Norleucine**, Aminocaproic acid, Alanine betaine	Leucine, **Isoleucine**, 6-Deoxyfagomine, Alloleucine, **Norleucine**, Aminocaproic acid, Methylvaline, N-(2-Hydroxyethyl)-morpholine
	3	133.1052	M+1	3.32	1.0E-05	2.8					
2	2*	133.0317	$[M+H]^+$	1.95	4.4E-08	4.0	Conv > Org	132.0245	$C_3H_2F_2N_4$	n.a.	n.a.
									C_8H_3FN	n.a.	n.a.
									$C_8H_4O_2$	2,4,6-Octatriynoic acid	n.a.
									$C_5H_8O_2S$	n.a.	3-Methyl sulfolene
3	4*	150.0583	$[M+H]^+$	1.95	3.8E-08	10.9	Conv > Org	149.0510	$C_3H_5F_2N_5$	n.a.	n.a.
	5*	151.0616	M+1	1.95	3.2E-08	2.5			$C_8H_6FN_2$	n.a.	n.a.
	6*	152.0541	M+2	1.95	3.4E-08	2.2			$C_5H_{11}NO_2S$	**Methionine**	**Methionine**, Penicillamine
4	7	245.0767	$[M+H]^+$	2.47	5.6E-05	3.6	Conv > Org	244.0693	$C_9H_{12}N_2O_6$	*Pseudouridine, Uridine***	*Pseudouridine, Uridine***
	8	267.0585	$[M+Na]^+$	2.47	6.5E-05	2.8			$C_6H_4N_{12}$	n.a.	n.a.
5	9	271.1149	$[M+H]^+$	6.93	4.5E-06	2.2	Conv > Org	270.1077	$C_9H_{14}N_6O_4$	n.a.	n.a.
									$C_{10}H_{10}N_{10}$	n.a.	n.a.
									$C_{11}H_{16}N_3O_5$	n.a.	n.a.
6	10	331.1724	$[M+H]^+$	12.98	1.3E-05	3.8	Conv > Org	330.1652	$C_{12}H_{22}N_6O_5$	n.a.	n.a.
									$C_{13}H_{18}N_{10}O$	n.a.	n.a.
									$C_{14}H_{24}N_3O_6$	n.a.	n.a.
									$C_{11}H_{26}N_2O_9$	n.a.	n.a.
7	11	433.2040	$[M+H]^+$	13.86	1.8E-05	3.4	Conv > Org	432.1968	$C_{16}H_{28}N_6O_8$	n.a.	n.a.
									$C_{17}H_{24}N_{10}O_4$	n.a.	n.a.
									$C_{18}H_{30}N_3O_9$	n.a.	n.a.

									$C_{24}H_{32}O_5S$	S-Furanopetasitin	S-Furanopetasitin
8	12	495.1744	$[M+H]^+$	13.86	2.5E-06	2.2	Conv > Org	494.1671	$C_{24}H_{24}N_5O_7$	n.a.	n.a.
									$C_{23}H_{28}NO_{11}$	n.a.	n.a.
									$C_{22}H_{22}N_8O_6$	n.a.	n.a.
									$C_{21}H_{26}N_4O_{10}$	n.a.	n.a.
9	13	591.1677	$[M+H]^+$	9.58	9.8E-06	3.4	Org > Conv	590.1606	$C_{24}H_{26}N_6O_{12}$	n.a.	n.a.
									$C_{25}H_{22}N_{10}O_8$	n.a.	n.a.
									$C_{28}H_{30}O_{14}$	Maysin 3'-methyl ether	n.a.

n.a.: not applicable

** These features were detected using biosigner*

*** invalid based on MS/MS data*

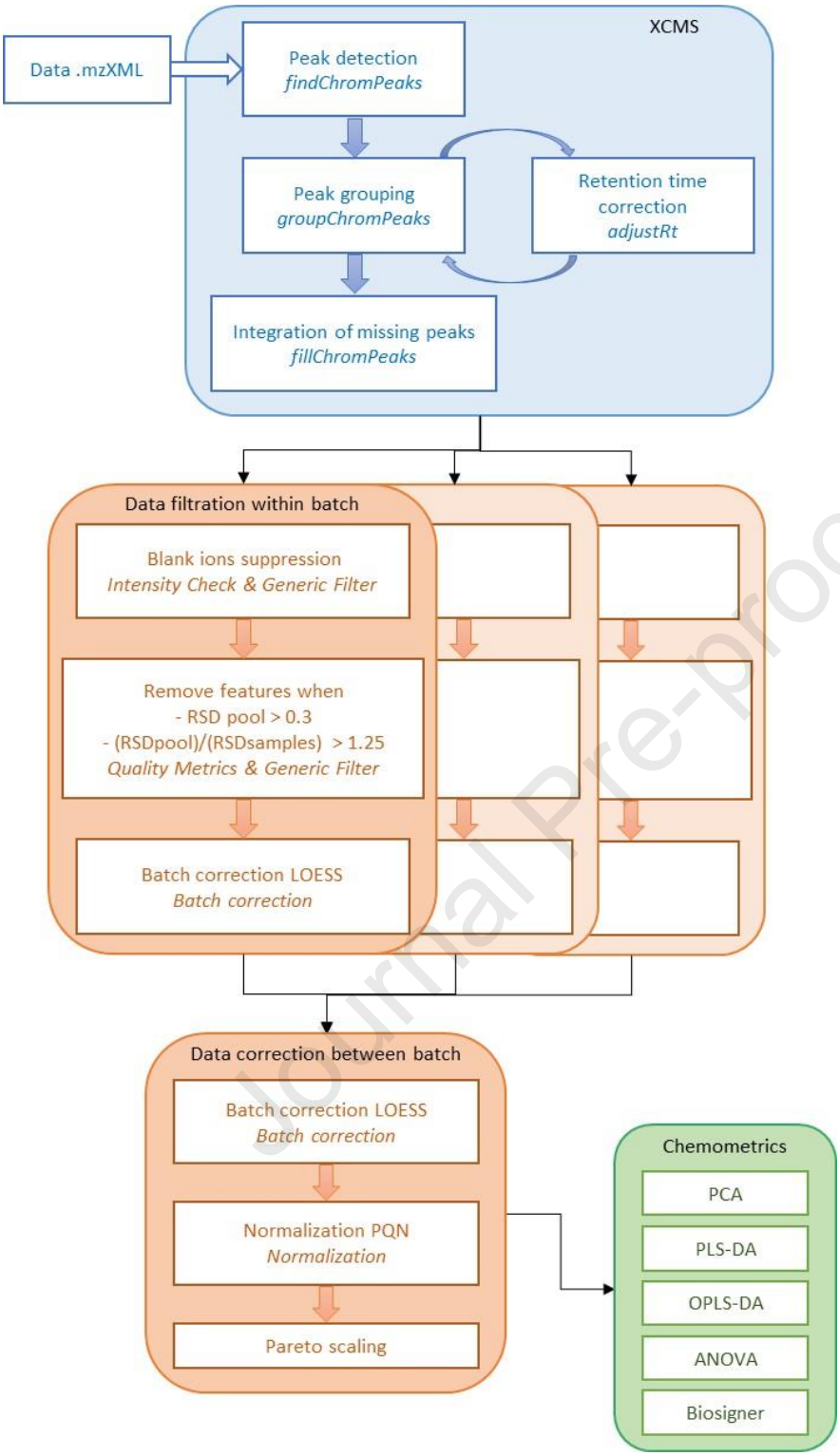

Fig. 1. Workflow of the data treatment using W4M* (RSD: relative standard deviation) ** text in italic refers to W4M functions.*

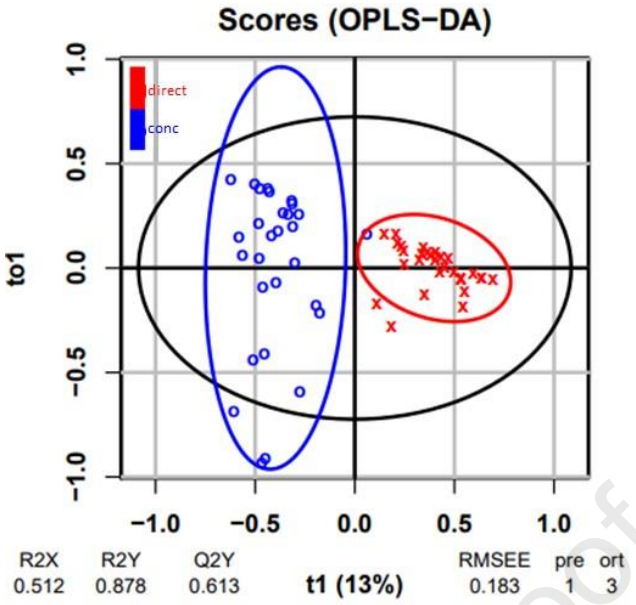

Fig. 2. Scores plot for OPLS-DA obtained with cross-validation (blue circles, both concentrated juices and juices from concentrate; red crosses, direct juices). The black ellipse represents 95% of the variability, the blue and red ellipse are the Mahalanobis ellipse of the sample groups.

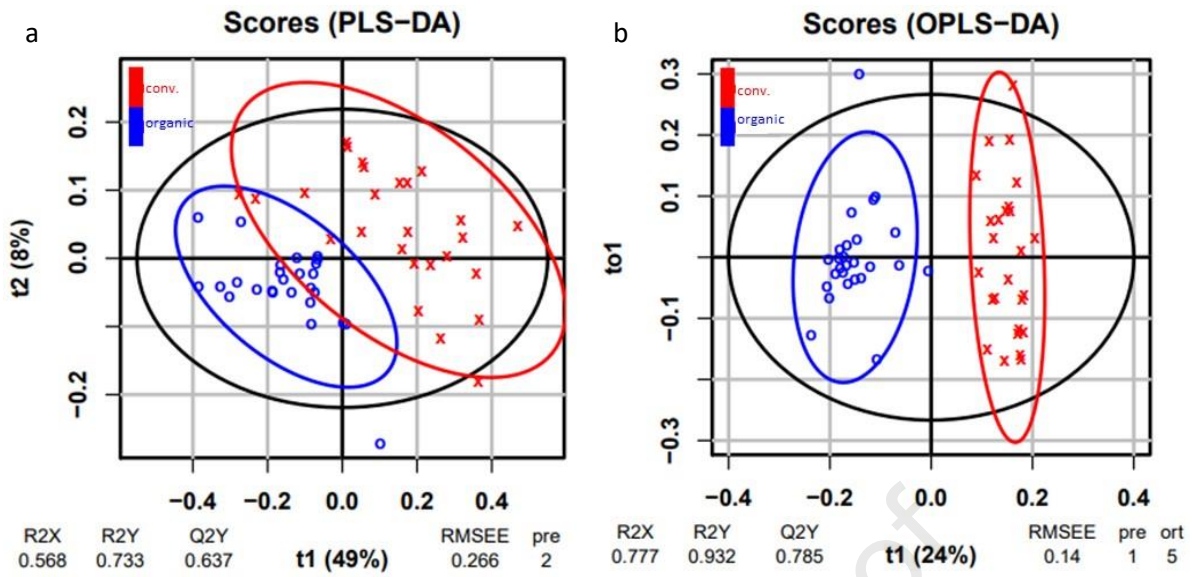

Fig. 3. (a) Scores plot of PLS-DA and (b) scores plot of OPLS-DA obtained after features selection using ANOVA (blue circles: organic juice samples; red crosses, conventional juice samples). The black ellipse represents 95% of the variability, the blue and red ellipses represent 95% of the multivariate distributions for each sample groups.

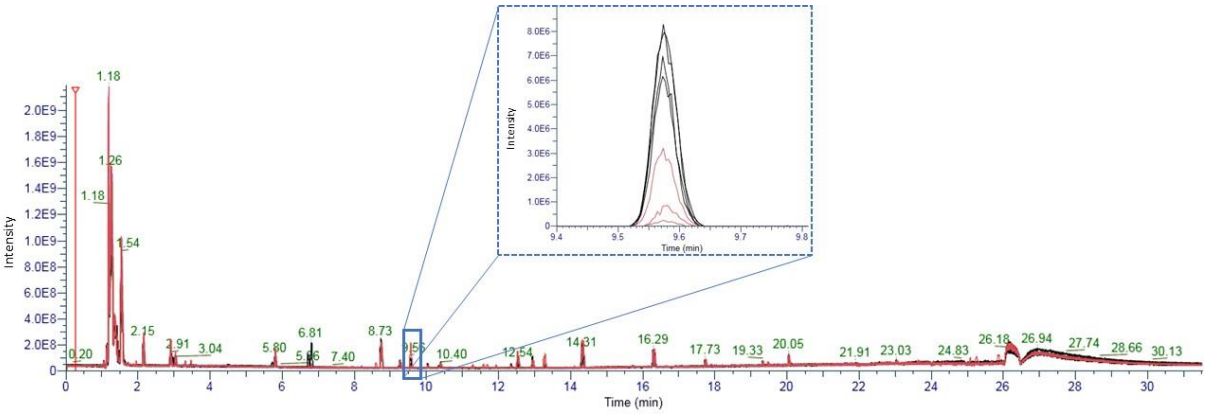

Fig. 4. Chromatogram of feature 13 for authentication of organic apple juices (black, organic juice samples; red, conventional juice samples)

Highlights

- Development of an UHPLC-HRMS metabolomics approach with great potential in juice authentication
- Discrimination between sample groups in two distinct authentication applications
- Relevant markers selected by OPLS-DA and ANOVA were tentatively identified
- Main discriminant compounds were identified as amino-acids and derivatives

Conflict of Interest and Authorship Conformation Form

Please check the following as appropriate:

- o All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.
- o This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.
- o The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript
- o The following authors have affiliations with organizations with direct or indirect financial interest in the subject matter discussed in the manuscript:

Author's name	Affiliation
No affiliation	