

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/43132876>

Discrimination of Wine Attributes by Metabolome Analysis

ARTICLE *in* ANALYTICAL CHEMISTRY · APRIL 2010

Impact Factor: 5.64 · DOI: 10.1021/ac902678t · Source: PubMed

CITATIONS

58

READS

104

6 AUTHORS, INCLUDING:



Hugo Pena-Cortes

Universidad de Valparaíso (Chile)

66 PUBLICATIONS 2,419 CITATIONS

[SEE PROFILE](#)

Discrimination of Wine Attributes by Metabolome Analysis

Alvaro Cuadros-Inostroza,^{†,‡} Patrick Giavalisco,[‡] Jan Hummel,[‡] Aenne Eckardt,[‡] Lothar Willmitzer,[‡] and Hugo Peña-Cortés^{*,†}

Centro de Biotecnología, Universidad Técnica Federico Santa María, General Bari 699, Valparaíso, Chile, and Max-Planck Institut für molekulare Pflanzenphysiologie, Am Mühlenberg 1, 14476-Potsdam-Golm, Germany

The chemical composition of any wine sample contains numerous small molecules largely derived from three different sources: the grape berry, the yeast strain used for fermentation, and the containers used for wine making and storage. The combined sum of these small molecules present in the wine, therefore, might account for all wine specific features such as cultivar, vintage, origin, and quality. Still, most wine authentication procedures rely either on subjective human measures or if they are based on measurable features, they include a limited number of compounds. In this study, which is based on an untargeted UPLC-FT-ICR-MS-based approach, we provide data, demonstrating that unbiased and objective analytical chemistry in combination with multivariate statistical methods allows to reproducible classify/distinguish wine attributes like variety, origin, vintage, and quality.

The molecules present in a wine cover a large number of metabolites including primary (e.g., sugars, organic acids, amino acids) and secondary metabolites (e.g., flavonoids, anthocyanins, and other pigments). All of these compounds have a strong influence on the quality and character of the wine, and are therefore not only important for the characterization and differentiation of wines, but also for the detection of frauds.^{1–4} The whole chemical composition of a wine reflects the history of the wine producing process, including the grape variety, the yeast strain, the containers used for fermentation and storage, and the enological practice.^{5–15}

* To whom correspondence should be addressed. Phone: +56 32 2654732. Fax: +56 32 2654836. E-mail: hugo.pena@usm.cl.

† Universidad Técnica Federico Santa María.

‡ Max Planck Institut für molekulare Pflanzenphysiologie.

- (1) Bisson, L.; Waterhouse, A.; Ebeler, S.; Walker, M.; Lapsley, J. *Nature*. 2002, 418, 696–699.
- (2) Wittwer, G. 2007, <http://ideas.repec.org/p/cop/wpaper/g-66.html>. Accessed July 2007.
- (3) Castillo-Sánchez, J. J.; Mejuto, J. C.; Garrido, J.; García-Falcón, S. *Food Chem.* 2006, 97, 130–136.
- (4) Cayot, N. *Food Chem.* 2007, 102, 445–453.
- (5) Kliewer, W. M.; Dokoozlian, N. K. *Am. J. Enol. Vitic.* 2005, 56, 170–181.
- (6) Holt, H. E.; Francis, L.; Field, J.; Herderich, J.; Iland, P. G. *Aust. J. Grape Wine Res.* 2008, 14, 162–176.
- (7) Holt, H. E.; Francis, L.; Field, J.; Herderich, J.; Iland, P. G. *Aust. J. Grape Wine Res.* 2008, 14, 191–202.
- (8) Conde, C.; Silva, P.; Fontes, N.; Dias, A.; Tavares, R.; Sousa, M.; Agasse, A.; Delrot, S.; Gerós, H. *Food.* 2007, 1, 1–22.
- (9) Borneman, A.; Chambers, P.; Pretorius, S. *Trends Biotechnol.* 2007, 25, 349–355.

Variety, provenance, production year (vintage), and quality ratings, which are the most important attributes used for the characterization and description of wines should be reflected in the total composition of small molecules in the wine. Interestingly the largest part of previously performed studies focused either on a limited number of molecules to discriminate different wines derived from different cultivars or to predict the origin of the wine samples.^{10–28} Only recently the first two studies emerged applying un-targeted approaches to evaluate the influence of the geographic origin and the species of oak wood on the quality of wines matured in oak barrels¹¹ or to discriminate surface active compounds in champagne aerosols an champagne bulk.²⁹ Both studies demonstrated that the restricted, targeted analysis of specific metabolites misses a large part of the molecular information pertained in the metabolome of a wine (or champagne wines), thus limiting its analytical significance. This is particularly true for wines of highest

- (10) Bautista-Ortin, A. B.; Fernandez, J.; Lopez, J.; Gomez, J. *J. Food Comp. Anal.* 2007, 20, 546–552.
- (11) Gougeon, R. D.; Lucio, M.; Frommberger, M.; Peyron, D.; Chassagne, D.; Alexandre, H.; Feuillat, F.; Voilley, A.; Cayot, P.; Gebefügi, I.; Hertkorn, N.; Schmitt-Kopplin, P. *Proc. Natl. Acad. Sci. U.S.A.* 2009, 106, 9174–9179.
- (12) Burlingame, B. *J. Food Comp. Anal.* 2008, 21, 587–588.
- (13) Lund, S.; Bohlmann, J. *Science*. 2006, 311, 804–805.
- (14) Gawel, R.; Godden, P. W. *Aust. J. Grape Wine Res.* 2008, 14, 1–8.
- (15) Cozzolino, D.; Cowey, G.; Lattey, K. A.; Godden, P.; Cynkar, W. U.; Damberg, R. G.; Jamik, L.; Gishen, M. *Anal. Bioanal. Chem.* 2008, 391, 975–981.
- (16) Diaz, C.; Conde, J. E.; Méndez, J. J.; Pérez, J. P. *Eur. Food Res. Technol.* 2002, 215, 83–90.
- (17) Diaz, C.; Conde, J. E.; Estevez, D.; Perez, S.; Perez, J. P. *J. Agric. Food Chem.* 2003, 51, 4303–4307.
- (18) Cullere, L.; Escudero, A.; Cacho, J.; Ferreira, V. *J. Agric. Food Chem.* 2004, 52, 1653–1660.
- (19) Pérez-Magariño, S.; Ortega-Heras, M.; González-San José, M. L.; Boger, Z. *Talanta*. 2004, 62, 983–990.
- (20) Pazourek, J.; šová, D.; Spanilá, M.; Farková, M.; Novotna, K.; Havel, J. *J. Chromatogr., A* 2005, 1081, 48–54.
- (21) Gürbüüz, O.; Rouseff, J. M.; Rouseff, R. L. *J. Agric. Food Chem.* 2006, 54, 3990–3996.
- (22) Makris, D.; Kallithraka, S.; Mamalosc, A. *Talanta*. 2006, 70, 1143–1152.
- (23) Sobieski, D.; Mulvihill, G.; Broz, J. S.; Augustine, M. P. *Solid State Nucl. Magn. Reson.* 2006, 29, 191–198.
- (24) Campo, E.; Cacho, J.; Ferreira, V. *J. Chromatogr., A* 2007, 1140, 180–188.
- (25) Escudero, A.; Campo, E.; Farin, L.; Cacho, J.; Ferreira, V. *J. Agric. Food Chem.* 2007, 55, 4501–4510.
- (26) Capron, X.; Smeyers-Verbeke, J.; Massart, D. L. *Food Chem.* 2007, 101, 1585–1597.
- (27) Christopher, J.; Bevina, C. J.; Damberg, R. G.; Fergusson, A.; Cozzolino, D. *Anal. Chim. Acta* 2008, 621, 19–23.
- (28) Tao, Y.; Li, H.; Zhang, L. *J. Food Comp. Anal.* 2008, 21, 689–694.
- (29) Liger-Belair, G.; Cilindre, C.; Gougeon, R. D.; Lucio, M.; Jeandet, P.; Gebefügi, I.; Schmitt-Kopplin, P. *Proc. Natl. Acad. Sci. U.S.A.* 2009, 106, 16545–16549.

quality and complexity, where efficient analysis and discrimination has yet not been achieved by means of targeted analysis.

To overcome these limitations we developed a nontargeted metabolome profiling strategy based on ultra high performance liquid chromatography (UPLC) coupled to ultra high resolution mass spectrometry (FT-ICR)^{30,31} to analyze the metabolic composition of a wide array of unfractionated Chilean red wines. To broaden the coverage of detectable metabolites we examined all samples in both negative and positive ionization modes. For the data analysis different multivariate statistical methods were tested and applied, allowing not only to classify but also to discriminate the wines according to cultivar, vintage, provenance, and quality. In addition, we investigated whether our methodology allows identifying measurable, objective biomarkers or molecular patterns related to wine sensory properties (quality), to assist or confirm the rating score provided by wine experts.

EXPERIMENTAL SECTION

Reagents and Standards. All chemicals used in this study were of the highest purity grade available. They were all purchased from Sigma-Aldrich (St. Louis, MO). HPLC water was obtained from a Milli-Q purification system (Millipore, Bedford, MA).

Wine Samples and Preparation. Around 400 monovarietal commercial Chilean wines, elaborated by different vineyards (Viña Concha y Toro, Viña San Pedro, Viña los Vascos, Viña Undurraga, Viña Aresti and Viña Tamaya), from different pure cultivars [(100% Carmenère (CM), 100% Cabernet Sauvignon (CS), 100% Merlot (ME), and 100% Syrah (SY)], from different vintages (2004, 2005, and 2006) and different quality scores, were directly analyzed from each bottle (Table 1 and 2). The wines were provided directly by each vineyard and the wines were chosen by experts of the staff, attending their own sensory parameters/quality criteria and scored in three quality levels: premium-quality (high), reserve-quality (medium), and varietal-quality (low). Next to the directly analyzed wine samples, 3 mL of each wine samples was stored under nitrogen in a hermetically closed glass vials. These samples were kept as back-ups at -20 °C.

UPLC-NanoMate-FT-ICR MS Measurement. The metabolomic profiling analysis of all samples was carried out as described in Figure 1a. The UPLC FT-ICR MS measurements were essentially performed as described previously.³¹ In brief: UPLC separation was performed using a Waters Acquity UPLC system (Waters, Milford, MA, USA), using a HSS T3 C₁₈ reversed phase column (100 × 2.1 mm i.d. 1.8 μm particle size, Waters) at a temperature of 40 °C. The mobile phases consisted of 0.1% formic acid in water (Solvent A) and 0.1% formic acid in acetonitrile (Solvent B). The flow rate of the mobile phase was 400 μL/min, and 2 μL sample were loaded per injection. The gradient was: 0–1 min isocratic flow at 99% A, 1–12 min linear gradient from 95% to 65% A, 12–13.5 min linear gradient from 65% to 20% A, 13.5–14.5 min linear gradient from 20% to 1% A, 14.5–16 min isocratic flow at 1% A, 16–17 min linear gradient

from 1% to 95% A, 17 to 20 min isocratic flow at 95% A, to re-equilibrate the column before to the next sample could be injected.

The UPLC was connected to the FT-ICR via a TriVersa NanoMate (Advion, Ithaca, NY, USA). The UPLC flow rate, which was 400 μL/min, was split 1:1000 with a T-Valve (Advion). One tenth of a percent was directly loaded to the FT-ICR MS, while 99.9% were delivered to waste. 400 nL/min of sample were infused into the MS via a nanospray Chip (Type A, Advion), by applying a voltage of 1.8 kV in the positive and 1.9 kV in the negative ionization mode. Spray sensing was used between min 1 and 17 of the UPLC gradient.

The mass spectra were acquired using the LTQ FT-ICR-Ultra mass spectrometer (Thermo-Fisher, Bremen, Germany). The spectra were recorded using full scan mode, covering a mass range from m/z 100–1300. The resolution was set to 50,000 and the maximum loading time for the ICR cell was set to 250 ms. The transfer capillary temperature was set to 200 °C and the MS spectra were recorded from min 1 to 17 of the UPLC gradient.

Metabolomic Profiling. The metabolomic profiling analysis of all samples was carried out as described in Figure 1a. This process involved: sample taking from the bottle, detection, spectral analysis, generation of a “sample specific” fingerprint, data processing and searching of known masses in GoBioSpace (in house developed databank). The variation in sample taking and mass spectrometry were controlled using spiked standards and assessed using quality control parameters.

All samples were randomized prior to mass spectrometric analyses to avoid any experimental drifts. A number of internal standards (chloramphenicol, costicosterone and ampicilin), added to each sample just prior to injection, were used to control experimental variability. To address overall process variability, metabolomic studies were augmented to include a set of six experimental sample technical replicates, which were randomly measured. Reproducibility analysis for compounds detected in each of these six replicate samples, gave a measure of the combined variation for all process components including sample taking injection and instrument steps. Pearson product-moment correlation coefficient between pairs of technical replicates ranged from 0.87 to 0.97.

The above developed process was used to quantify the metabolomic alterations in red wine samples. Specifically, this study included wine bottles from four different cultivars (CM, CS, ME, and SY), each with six technical replicas (total of samples 216), to test whether wine metabolome provides the information to discriminate between the wine varieties. To evaluate whether wine metabolome could be employed for identification of the wine level quality (score rating previously determined by the corresponding winemakers), 162 Chilean bottles of CS wine (three different wine scores (high, medium, and low), three different vintages (2004, 2005, and 2006), six bottles per year, three different vineyards), were measured.

Data Analysis. Molecular masses, retention time and associated peak intensities were extracted from the raw files, using the SIEVE software (Version 1.2, Thermo-Fisher). The mass and retention time lists were used for searches against the ChemSpider database (<http://www.chemspider.com>), employing the in-house developed database search tool GoBioSpace, (Hummel et al, unpublished).^{30,31} The result files, including the database annota-

(30) Giavalisco, P.; Hummel, J.; Liseć, J.; Inostroza, A.; Catchpole, G.; Willmitzer, L. *Anal. Chem.* **2008**, *80*, 9417–9425.

(31) Giavalisco, P.; Köhl, K.; Hummel, J.; Seiwert, B.; Willmitzer, L. *Anal. Chem.* **2009**, *81*, 6546–6551.

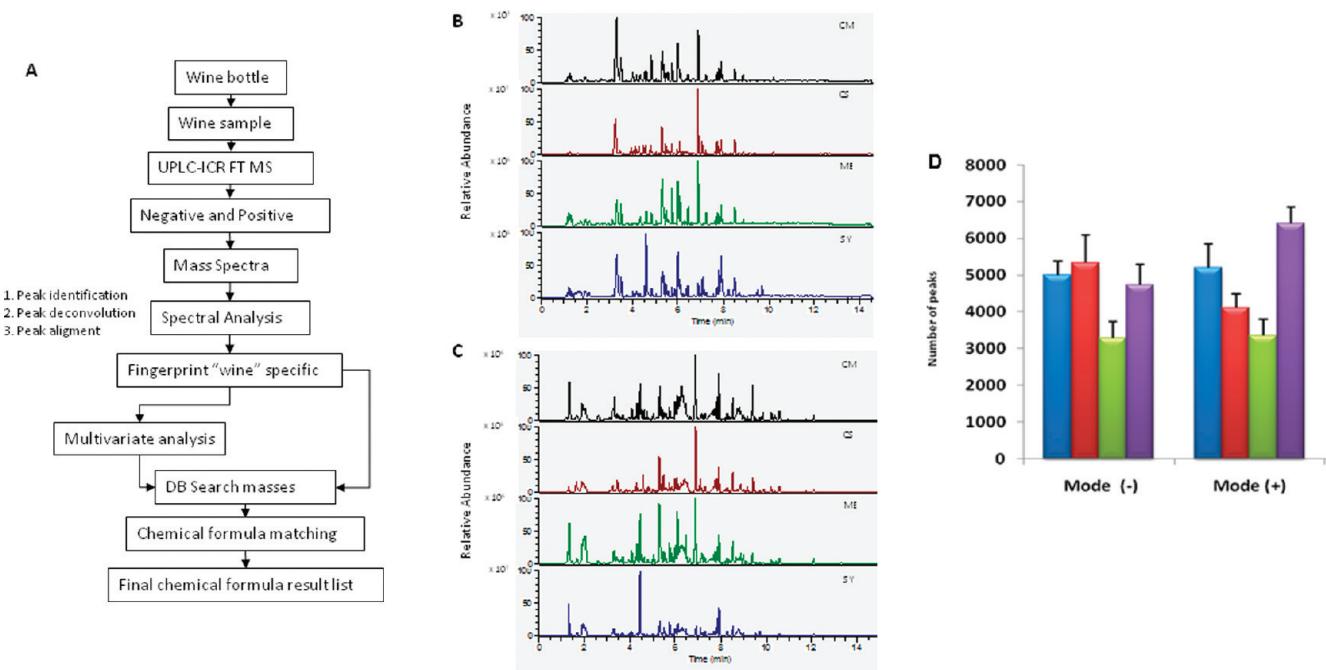


Figure 1. Metabolomic profiling of wine samples. (A), Flowchart of the steps involved in metabolomic profiling of wine samples. This includes wine samples provided in the bottle, sampling and injection in LC, mass spectrometry-based detection, spectral analysis, generation of a “sample specific” fingerprint, data processing and searching of known masses in GoBioSpace (in house developed databank). The variation in sample taking and mass spectrometry were controlled using spiked standards as described in materials and methods. (B) (negative) and (C) (positive) base peak chromatograms representing the wine cultivars Carmenère (CM), Cabernet Sauvignon (CS), Merlot (ME) and Syrah (SY). (D) Number of ion traces between 100 and 1000 m/z in both negative [mode (-)] and positive [(mode +)] mode. CM (blue bar), CS (red bar), ME (green bar) and SY (purple bar). Each result represents the peak average of the total number of bottles analyzed, considering three technical replicates ($n = 636$).

tions of each mass, associated chemical formula, retention time, m/z value, compound ID, and possible substance names were exported as text files and subjected to downstream statistical data analysis or data visualization. Spectra manipulations and peak extractions were performed using Xcalibur (Version 2.06, Thermo Fischer).

Statistical Analysis of Data. Data Normalization. To reduce variation between samples, we divided the individual intensities by the total sample intensity and multiplied the result by 1×10^7 . After that, we log-transformed, scaled and centered the features, that is, the retention time-mass pairs, by using the unit variance transformation. Zero intensity was set to 1×10^{-5} to allow log transformation.

Hierarchical Clustering was performed by using MultiExperiment Viewer (<http://www.tm4.org>) on the normalized data. Euclidian distance and average linkage were used to build the heat maps. For visualization, the color scale was centered about the median intensity of the metabolite matrix.

Principal Component Analysis (PCA) was carried out with SIMCA-P version 11.5 (<http://www.umetrics.com>) on the normalized data. A cutoff of 95% of the explained variance was used to select the optimum number of principal components.

Linear Discriminant Analysis (LDA) was performed using R statistical software (<http://www.r-project.org>). LDA is a supervised statistical method, mostly used in machine learning, that aims to find a linear combination of features that best classify two or more groups of objects. In our case, we followed two approaches to process the wine metabolite: (i) introducing all variables in order to use all the available information. (ii) Stepwise LDA which

performs a data dimensionality reduction procedure that should best classify the wine classes before introducing the variables. That was done by assigning samples to given wine qualities (classes) according to score rating determined by the corresponding wine experts. Then, we used the selection attribute method implemented in WEKA,³⁷ a comprehensive collection of data mining and machine learning tools. The selection attribute selection method searches the space of features subsets in order to find an optimal subset that has high correlation with the classes, that is, high predictive ability, but low intercorrelation between its features. To validate the predictions, we performed a repeated random subsampling cross-validation in which the data set is repeatedly split into training a validation sets. Here we tested 80, 66, and 50% splits as training sets. For every split an LDA model is created to fit the training set and validated with the testing set. The percentage of correct predictions is averaged over the splits and that gives the prediction power estimation.

Wine samples were stored in a volume of 3 mL in hermetic ambar vials. Before being closed with (crimped) silver caps, nitrogen was applied for 30 s and the samples were kept by -20°C .

RESULTS AND DISCUSSION

The starting point of our analysis (Figure 1a), was to determine whether or not the unrestricted metabolic profiles of wine contain sufficient information to allow the distinction of wines derived from different grape cultivars. For this purpose 216 independent red wines samples (all from independent bottles, derived directly from Chilean vineyards assuring the authenticity of the samples) covering four different grape types (72× Carmenère, 72× Cabernet

Table 1. Samples of Wine Cultivars

wine variety	year 2004	year 2005	year 2006	total
Carmenère (CM)	24	24	24	72
Cabernet Sauvignon (CS)	24	24	24	72
Merlot (ME)	12	12	12	36
Syrah (SY)	12	12	12	36

Wine bottles correspond to red wine produced from the cultivars Carmenère (CM), Cabernet Sauvignon (CS), Merlot (ME), and Syrah (SY), considering different vintages (2004, 2005, and 2006) and of different quality score rating previously determined by the corresponding wine expert from each vineyard.

Sauvignon, 36× Merlot, and 36× Syrah), were initially analyzed. These 216 samples included not only the above-mentioned four different cultivars but were also originated from different years of production, different vineyards and represented different quality score ratings (see Table 1).

Interestingly, it was immediately apparent, based on visual inspection of the crude total ion chromatograms, that the wine samples showed major differences depending on the grape cultivar they were derived from (Figure 1b and c). Up to 6400 detectable peaks in each ionization mode (negative and positive) were detectable (Figure 1d). A significant proportion of them (around 30%) were detectable exclusively in each variety. The highest number of exclusive mass traces was found for Cabernet Sauvignon samples, followed by Syrah, Carmenère, and Merlot

(Figure 2a and b). Even more interesting was the observation that only a minority (approximately 9%) of all peaks were shared between all four cultivars (Figure 2a and b). This observation clearly demonstrates the uniqueness of each wine cultivar derived from different grape types, a feature which has been praised since centuries though without any broad and objectively measured data set. When relating the accurate mass traces to databases (Experimental Section³¹) we were surprised to find that the majority of masses (62%) did not lead to a distinct annotation, suggesting that the majority of the compounds present in any wine sample have not yet been chemically ascertained.

As expected due to the diversity of the various chromatograms and the different number of mass traces present in the cultivars, hierarchical cluster analysis (HCA) separates these 216 samples preferentially according to their grape cultivar (Figure 3a and b). Interestingly, the same results, namely a cultivar specific separation of the wine samples was observed by considering only the 9% peaks common to each cultivar (Figure 3c and d), indicating that the HCA derived separation is not barely based on the presence/absence of cultivar specific mass traces, but also on quantitative differences within the molecules present in all cultivars. Essentially the same results were obtained when applying a principal component analysis (PCA), as shown in the Figure 3e and f, to the 216 samples. Even though the cultivar specific separation was a very interesting result, it also underlines

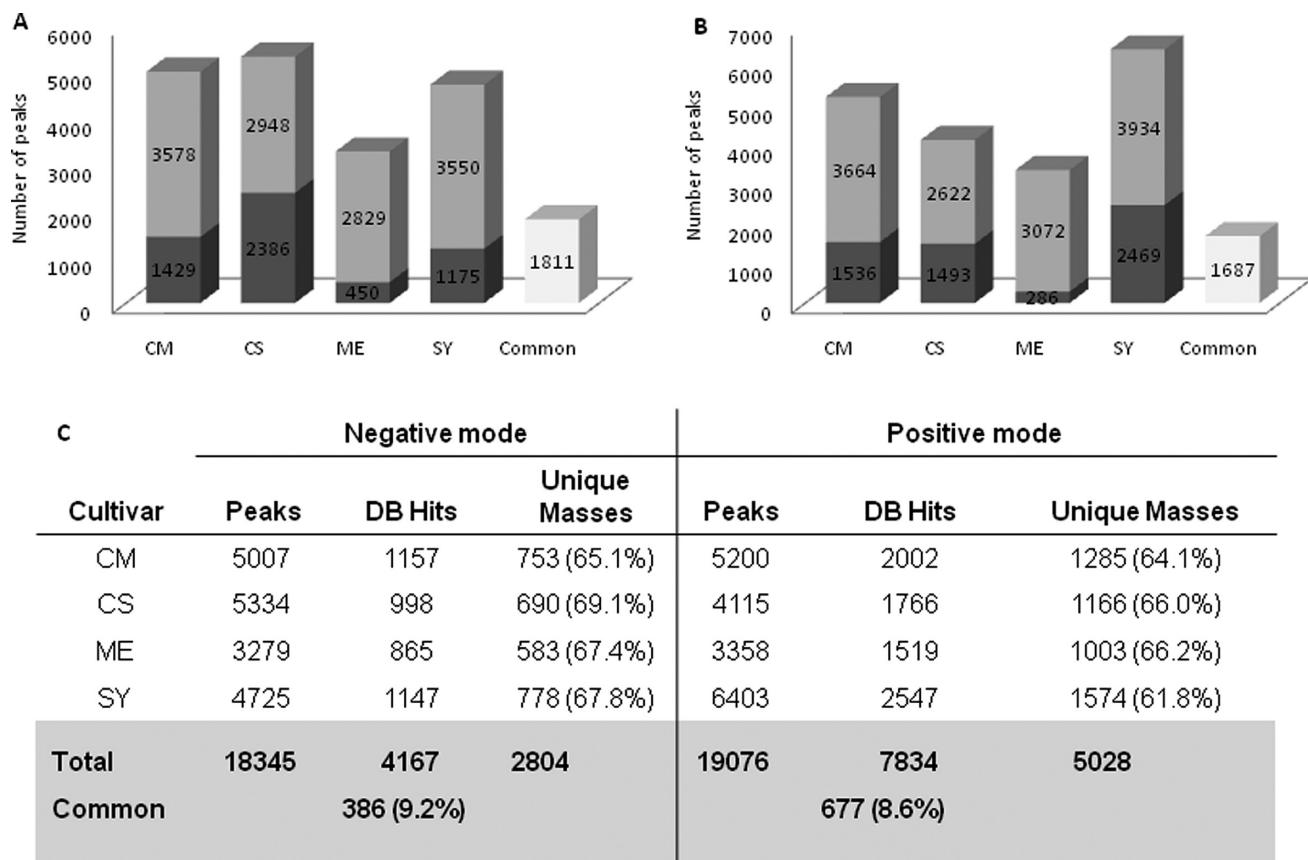


Figure 2. Number of peaks detected in Chilean red wine. (A), (negative mode), and (B) (positive mode), exclusive ions of each cultivar (CM: Carmenère; CS: Cabernet Sauvignon; ME: Merlot; and SY: Syrah) obtained after comparison with ions of the other three cultivars and removal of repeated ones among them (black boxes, cultivar specific peaks), Common represents the ions present in all four cultivars; (C), DB hits, number of chemical formulas retrieved after database searches; Unique Masses: masses leading to the database hits; Common: represents those masses from the total unique masses (2804 for negative and 5028 for positive modes) which are present in all four cultivars and match with hits in databases.

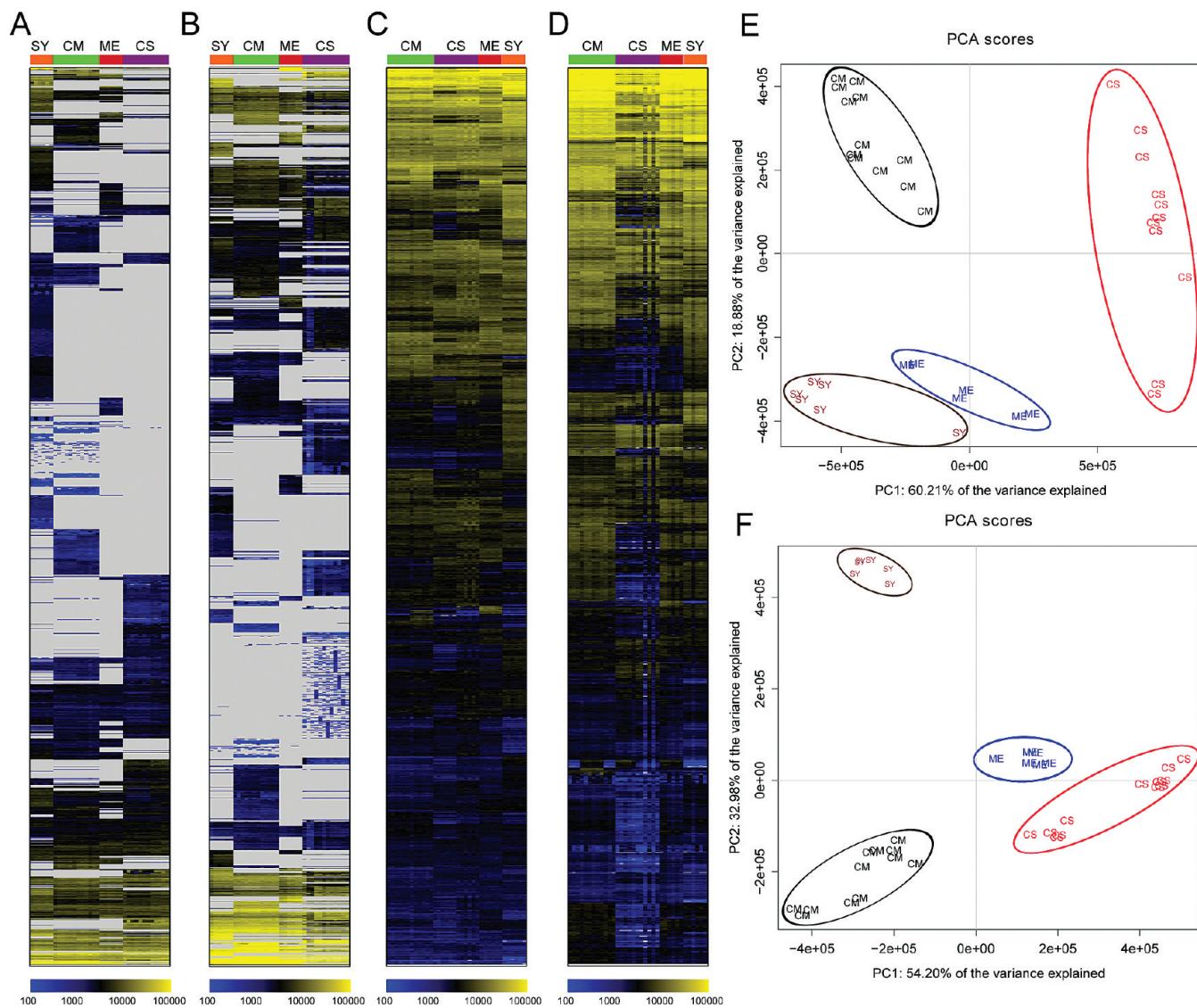


Figure 3. Metabolomic differences and statistical discrimination of wine cultivars. Heat map representation of 18 345 ions determined in negative (A) and 19 076 ions determined in positive (B) ionization modes in Carmenère (CM), Cabernet Sauvignon (CS), Merlot (ME), and Syrah (SY). (C), Heat map representation of common 1811 ions determined in negative and 1687 ions determined in positive (D) ionization mode in Carmenère (CM), Cabernet Sauvignon (CS), Merlot (ME), and Syrah (SY). Columns represent individual wine samples and the rows refer to distinct ions. Shades of yellow and shades of blue represent increase or decrease of a metabolite relative to the median metabolite expression in log10 scale, respectively. Missing values are represented in gray. The same ions used by the hierarchical clustering (C, D) were analyzed by principal component analysis (PCA) based on metabolomic alteration determined by negative (E) or positive (F) ions of Carmenère (CM), C. Sauvignon (CS), Merlot (ME), and Syrah (SY). To decrease complexity of the 216 samples, PCA was performed on average value of each group (36 samples) representing a value of each cultivar.

that this feature is clearly dominant over the other features (namely, vintage, origin and quality) contained in the 216 wine sample set.

A consequence of this observation is that discrimination of additional features requires a higher sample homogeneity (e.g., restricting the analysis to only one cultivar) or more sophisticated, or at least supervised statistical methods,^{32–34} relying on previous knowledge of the wine samples.

To discriminate other attributes different from wine cultivar, we restricted our second analysis to one type of wine, derived from only one cultivar (Cabernet Sauvignon). These samples originated from three distinct vineyards and from three consecutive years (2004, 2005, and 2006), resulting in six samples per year and per vineyard; therefore, totaling 54 bottles (Table 2).

In contrast to the chromatograms derived from the different cultivars from the initial study, the chromatograms derived solely from Cabernet Sauvignon samples were much more homogeneous. Still PCA examination of the metabolic profiles demonstrates striking differences between the samples from the different vintages resulting in a clear separation of the wines according to year (Figure 4a) respectively vineyard (Figure 4b). This result strongly confirms our assumption that removal of the most significant variable between the samples and therefore increasing the homogeneity of the samples leads to the efficient discrimination of additional, more subtle features like year or provenance of the wine.

Wine quality scoring is traditionally based on wine judges, winemakers, and technical staff, whose criteria are based on their

Table 2. Samples of Cabernet Sauvignon Samples Including Three Different Years, Vineyards and Quality Score Rating

vineyard	Wine score rating								
	low			medium			high		
	year 1	year 2	year 3	year 1	year 2	year 3	year 1	year 2	year 3
1	2004	2005	2006	2004	2005	2006	2004	2005	2006
2	2004	2005	2006	2004	2005	2006	2004	2005	2006
3	2004	2005	2006	2004	2005	2006	2004	2005	2006
total	18	18	18	18	18	18	18	18	162

All wine bottles correspond to red wine produced from cabernet Sauvignon by three different Chilean vineyards (1, 2, and 3). Three different harvest years (2004, 2005, and 2006) were considered as well as three different wine score ratings determined by the wine expert of each vineyard (low, medium, and high).

sensory evaluation. This sensory analysis includes measurement, interpretation, and understanding of human responses to the properties perceived by the senses such as sight, smell, and taste.^{4,35,36} We expected thus the discrimination of these subjective annotations to represent a challenge for every metabolome-based analysis. Therefore we decided to further challenge our analytical method to see if it also matches the requirements for quality score discrimination. For this purpose we added another 108 Cabernet Sauvignon samples, including next to the different vineyards and the different years also three different wine quality grading (high,

medium-, and low-quality, which had been determined by Chilean wine experts) (Table 2).

Interestingly, the application of PCA to this metabolomic data set, including the 162 samples did not result in a good separation according to quality. Our first assumption was that the comparison of different vineyards, years, and quality scores might complicate our unsupervised analysis. Therefore in a first step we again increased homogeneity of the samples by taking out two of the three variables. As shown in Figure 4c and d a very effective separation according to quality is achieved if only samples derived

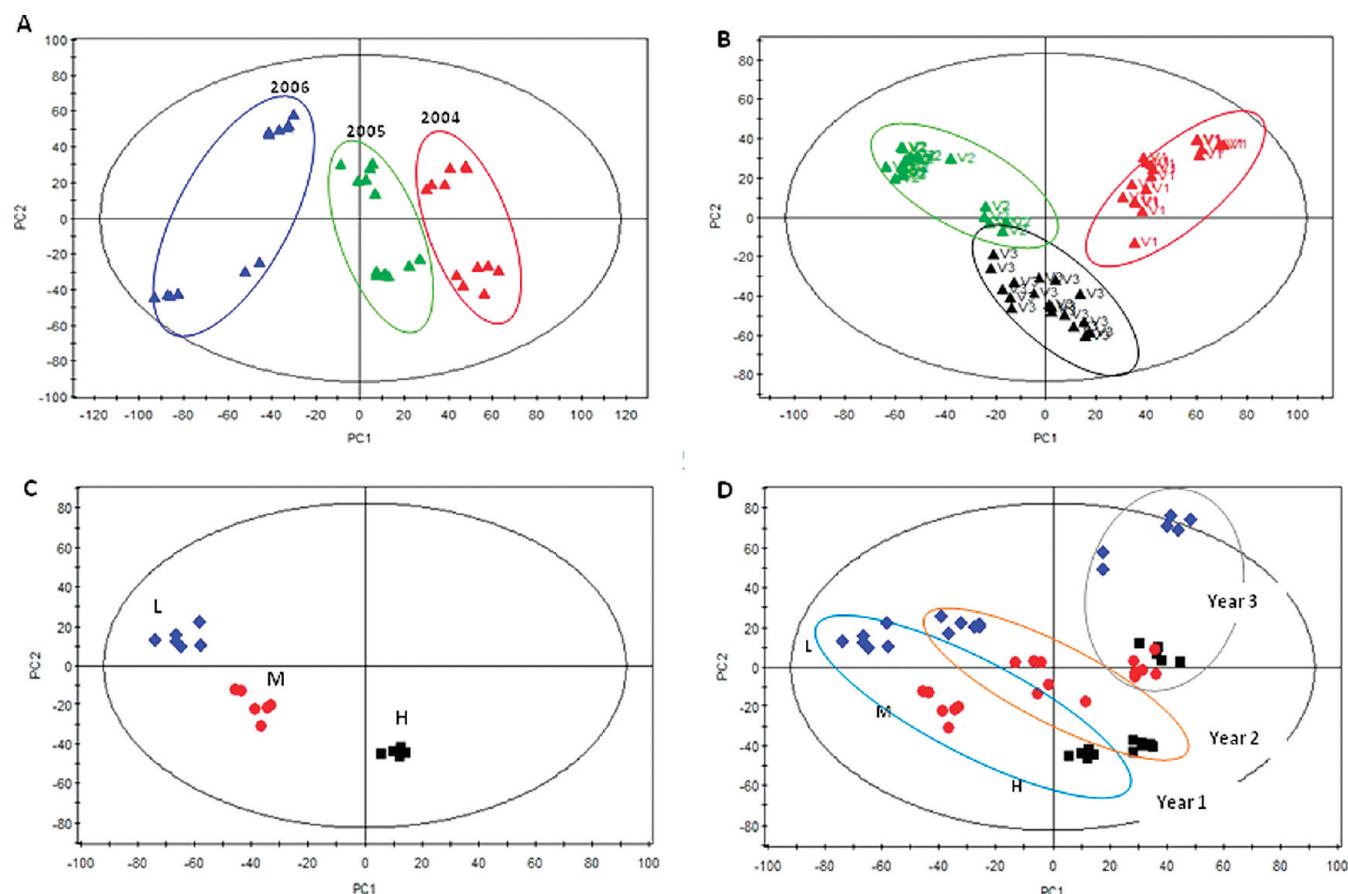


Figure 4. PCA analysis defined by the two first discriminant functions of the samples, measured in positive ionization modes. (A) Cabernet Sauvignon (CS) wines produced in 2004 (red triangle), 2005 (green triangle), and 2006 (blue triangle) and from bottle samples provided by different vineyards (V1, V2, and V3, representing different Chilean vineyards) (B) determined by positive ionization mode. (C) and (D) Principal component analysis of Chilean red wine ranged in three score ratings by wine experts: H, high; M, medium; and L, low. Wine samples, considering three different categories (H, M, and L) from a specific vineyard and a determined year, are clearly differentiated (C). PCA analysis considering wine samples classified in the three categories of a specific vineyard (arbitrary chosen) but considering three different harvest years are represented in (D) (total bottles 54).

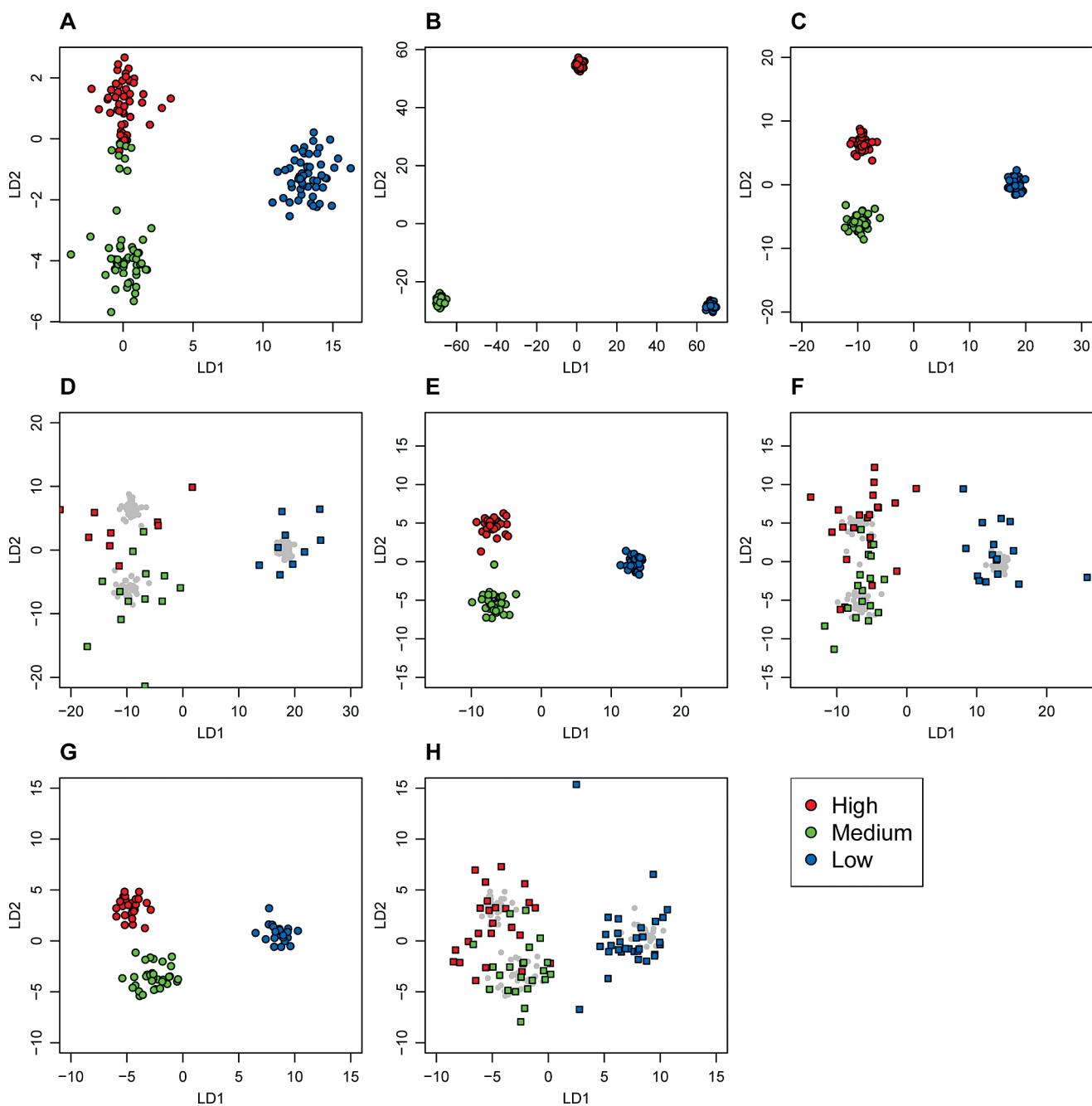


Figure 5. Discriminant analysis results of all the samples according to the type of wine score rating. (A) and (B) show the two first discriminant functions resulted from the stepwise LDA of the discriminant measured in negative (A) and positive (B) ionization modes. To check the predictive ability of LDA, we used leave-some-out cross validation. In this test, a number of samples are removed from the original data set (B, positive ionization mode). The classification model is rebuilt and the removed samples are classified in this new model. We tested several combinations of inputs: (C) and (D): 80% training set, 20% testing set, respectively; (E) and (F) 66% training set, 34% testing set, respectively; and (G) and (H): 50% training set, 50% testing set, respectively. Codification: high (red circle), medium (green circle), and low (blue circle) score rating.

from one year and one vineyard are compared to each other. This result is similar independent of the selected vineyards indicating, that sufficient differentiating information is contained in the metabolome data, but this is overwritten by stronger variance derived from other parameters (year and vineyard).

Taken together these results indicate that even though the information generated from the UPLC-FT-ICR MS-based wine metabolome analysis provides sufficient discrimination power and information to discriminate a multitude of features, these cannot be accessed in highly heterogenic data sets if only unsupervised

multivariate statistics are applied. These results may reflect two different independent observations: 1. Our method discriminates between the reproducible and therefore objective differences introduced by the use of different cultivars, the wine making process in different vineyards and the harvest time points chosen between different years and different vineyards. 2. The more subjective measure, namely quality, is not as well or instantaneously reflected in the metabolic profile, indicating that the sensory procedure applied by each wine expert to determine the score rating of a given wine might contain a certain degree of

variance, not fully covered by the metabolic pattern of the wine sample. This could be an additional explanation for the failure of our method to directly discriminate between some samples graded for instance with a medium score rating by one wine expert while being categorized as high or low by another wine expert and vice versa.

To this end we were therefore interested to ascertain whether a supervised statistical method would allow the discrimination of the different wine quality ratings, even in the presence of a complex data set containing not only the different quality scores but also different vintages and origins. Among the different supervised methods tested LDA (linear discriminant analysis) turned out to be very powerful. LDA was trained for classifying wine samples in the different score rating categories, considering as differentiating category only the wine quality (score rating), independently of vintage and origin (vineyard) (see material and methods).

A scatter plot of the two first discriminant functions (derived from all the variables measured in either negative or positive ionization modes) as well as the corresponding stepwise LDA, is shown in Figure 5a and b. Contrary to the unsupervised PCA results, a clear separation into three groups is readily obtained, when all 162 wine samples are grouped according to their quality score independently of vintage and provenance.

To test the robustness and predictive power of the LDA method, we used repeated random subsampling validation. This analysis showed that 80% of the training set samples are enough to get a correct classification (with a predictive power of more than 93%) for correct assignment of the quality score (Figure 5c and d). Even when only 66% of the samples were used to train the model, it still allows an efficient classification (with a predictive power of more than 89.1%) of the quality score (Figure 5e and f). Decreasing the number of samples of the training set thus obviously leads to loss in correct annotation, mostly affecting the correct classification of samples with high or medium quality (Figure 5h), whereas the samples with the lowest quality are still annotated with a high precision (>95%). These data suggest that differences between high and medium quality wines are less prominent on the metabolome level as compared to the respective differences toward low quality wines, again pointing to possible weaknesses in the commonly accepted, but subjective quality scoring system.

-
- (32) Lindon, J. C.; Holmes, E.; Nicholson, J. *Prog. Nucl. Magn. Reson. Spectrosc.* **2001**, *39*, 1–40.
 - (33) Trygg, J.; Holms, E.; Lundstedt, T. *J. Proteome Res.* **2007**, *6*, 469–479.
 - (34) Mahadevan, S.; Shah, S.; Marrie, T.; Slupsky, C. *Anal. Chem.* **2008**, *80*, 7562–7570.
 - (35) Atanasova, B.; Thomas-Danguin, T.; Chabanet, C.; Langlois, D.; Nicklaus, S.; Etievant, P. *Chem. Senses* **2005**, *30*, 209–217.
 - (36) Marshall, K.; Laing, D. G.; Jinks, A. L.; Hutchinson, I. *Chem. Senses* **2006**, *31*, 539–545.
 - (37) Witten, I. H.; Frank, E. In *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Series in Data Management Systems, 2nd ed.; Elsevier Inc.: New York, 2005.

Taken together the LDA model, based on UPLC-FT-ICR MS-based wine metabolome analysis, provides an excellent prediction and separation tool. If higher quality ratings with higher complexity are considered it has to be taken into account that these samples can only be efficiently scored and classified if sufficient well-annotated samples are available for the training set.

When applying the LDA model, developed on Cabernet Sauvignon wines, to other cultivars such as Carmenère, Merlot, or Syrah, no separation according to quality was achieved (data not shown). This observation therefore shows that quality in one wine is not necessarily reflected by the same chemical properties in a different cultivars. This observation is not really surprising given the unique pattern of wine quality associated with specific cultivars as always described by the wine industry. The results suggest that wine metabolome analysis allows the generation of enough and appropriate information to develop LDA- or alternative multivariate models for the efficient discrimination and/or classification of different wine features like cultivar, year or quality. However, due to the high complexity and diversity of the measured pattern these classification models should initially be developed specifically for each cultivar.

CONCLUSION

Unbiased metabolome analysis of wine, establishment of wine profiles (metabolic profiles), and data processing by both unsupervised and supervised multivariate procedures allow the digitalization of wine properties providing a novel approach for objective annotation of different wine attributes such as cultivar, year of production, vineyard, and quality score rating. Starting from these observations, the development of well annotated, cultivar specific databases, containing the metabolic profiles of several thousand scored wines samples might be a first step for the objective classification of wine samples from all over the world. Additionally, further validation of the underlying molecules, being responsible for the classification of the wine samples might as a next step lead the elucidation of commercially viable biomarkers for wine authentication purposes.

ACKNOWLEDGMENT

All wine samples used in this work were generously provided by different Chilean Vineyards through the Technological and business Consortium for the Vine and Wine Chile (VINNOVA). This work was supported in part by the Chilean Genome Initiative through Fondef-Conicyt (Project G07I1003) and Innova-Corfo (Project 06FC01IAC-141). A.C.-I. and P.G. contributed equally to this work.

Received for review November 23, 2009. Accepted February 19, 2010.

AC902678T