Journal of Chromatography A, 1217 (2010) 6718–6723

Contents lists available at ScienceDirect

Journal of Chromatography A

journal homepage: www.elsevier.com/locate/chroma

# Target list building for volatile metabolite profiling of fruit

Aniko Kende*, David Portwood, Adam Senior, Mark Earll, Elek Bolygo, Mark Seymour

*Analytical Sciences, Syngenta Jealott's Hill International Research Centre, Jealott's Hill, Bracknell RG42 6EY, Berkshire, United Kingdom*

ARTICLE INFO

*Article history:*
Available online 8 June 2010

*Keywords:*
Volatile analysis
Melon aroma
Dynamic headspace
Gas chromatography–mass spectrometry
Spectral alignment
Multivariate statistics

ABSTRACT

Fruit flavour is the combination of numerous biochemicals: sugars for sweetness, acids for sourness and volatile metabolites for aroma. The objective of this study was to establish a method to develop a target list of statistically relevant compounds for the characterization of melon from non-targeted data, while preserving the profile information. Five different varieties were sampled (sampling 12 biological replicates from 12 plants) using dynamic headspace extraction, then analysed by gas chromatography–mass spectrometry in full scan mode. Using Metalign and SIMCA-P software the raw data was spectrally aligned and then subjected to principal component analysis (PCA). The principal component analysis plot showed good separation of the five varieties based on their full scan GC–MS profile. Mass spectral data points responsible for the differences between varieties were highlighted by further statistical analysis. The mass spectra were then reconstructed and the corresponding chemicals identified using library search or reference standards were available to create a new target component list. To validate the new target list, the initial data set was re-processed using the targeted approach and the results subjected again to principal component analysis. The two representations showed excellent agreement on the separation of the five varieties. The new target list obtained from this study can be applied to differentiate and characterize the volatile profile of melon varieties using a list of statistically significant compounds.



## 1. Introduction

Secondary metabolites of fruits include a large number of volatile compounds which contribute to the flavour and thus have a strong impact on customer response to different varieties [1].

There are various methods for measuring the volatile metabolites of plant material. These methods usually consist of a sample preparation technique to extract the volatile metabolites from the tissue, such as distillation [2–4], static [2] or dynamic headspace extraction [5–9], solid phase microextraction [2,4,5,10], direct extraction [11] or supercritical fluid extraction [2,4] and most typically gas chromatographic separation often using mass spectrometry as a detector of choice to allow metabolite identification [2–11].

The volatile profile data generated can be interrogated using different approaches depending on the focus of the study. Multivariate statistical tools can reveal profile differences using GC–MS data, without the need of previous peak identification [12,13]. For this however spectral alignment of the data is crucial. Metabolomic studies involving mass spectrometry have the dual problem of shifts in retention times between data files and also small shifts in the mass axis within spectra. Therefore, the data files need to be aligned before comparing against each other. Many alignment strategies have been described in the literature. In principle, there are two types of data alignment. In the first approach, data files are aligned against a master chromatogram. In the second approach, mass spectra of individual components within the data files are matched to those in a spectral database.

Nielsen et al. [14,15] published a correlation optimised warping (COW) algorithm that aligned two-dimensional chromatographic data without prior peak picking. The alignment of gas chromatography mass spectrometry data based on peak lists (MSFACTs) was described by Duran et al. [16]. Due to the highly complex nature of the data, it was important to include the mass spectrometric dimension in the data alignment as well as chromatography data. Most of the recently described alignment algorithms identify first marker peaks that can be used for alignment or they dissect the mass spectra into individual $m/z$ traces (mass chromatograms) and shift the time axis of the chromatograms until all peaks are aligned. The following enumeration lists a few published methods, with no claim for completeness: MarkerLynx (Waters Corporation, Milford, MA [17]), Metalign [13,18], XCMS [19,20], MZMine [21], MET-IDEA [22], EQUEST [23].

Once the profile differences are revealed, studies often focus on the chemicals underlying the observed differences and in this case compound identification becomes necessary [9,13]. Due to

* Corresponding author. Tel.: +44 1344414396.
*E-mail address:* aniko.kende@syngenta.com (A. Kende).

0021-9673/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2010.05.030

the complex chromatograms typically obtained for fruit volatiles even the most thorough compound identification cannot guarantee that all compounds which contribute to profile differences are included, risking information loss during data reduction. Moreover comprehensive compound identification may also include numerous compounds without any significant contribution to the profile differences.

The objective of this study was to establish a method to develop a target list of volatile compounds derived from a non-targeted analysis which have a significant contribution to the profile differences of the studied melon varieties. We used spectrally aligned chromatographic data and multivariate statistics to achieve this and then validated the target list against the non-targeted results to prove that no major loss of information occurs in spite of the data reduction.

## 2. Materials and methods

### 2.1. Sample preparation

Internal standard 1,4-dichlorobenzene was purchased from Sigma–Aldrich (Gillingham, UK). It was diluted with methanol to give 500 μg/mL ISTD solution. To each sample 5 μL of the ISTD solution was added giving a theoretical maximum trapped amount of 2500 ng.

Five melon varieties were grown and the fruits were harvested at full maturity. Fruits from 12 plants of each variety were sampled in replicates where fruit size allowed using dynamic headspace extraction. From each fruit 12 cores were cut and weighed into a 500 mL Duran bottle. The internal standard was introduced into its container and then the bottle was closed with a specially transformed screw cap (Fig. 1). The sample bottle was then attached to the extraction manifold, left to equilibrate to 42 °C, then subjected to dynamic headspace extraction. Through the inlet line the headspace was purged by nitrogen at 50 mL/min for 35 min, while volatile components were trapped on the absorbent tube attached to the outlet line. The absorbent tubes were purchased prefilled from Markes Int. (Llantrisant, UK) containing 200 mg Tenax TA and 150 mg Carbograph 1TD. After extraction the absorbent tubes were capped off with gas-tight brass fittings until analysis.

**Fig. 1.** Sampling bottle with specially transformed cap.

### 2.2. Gas chromatography–mass spectrometry

Samples were analysed using a Unity/Ultra thermo desorption system attached to an Agilent 6890-5973 GC–MS. Conditions of the thermal desorption step were the following: tubes were desorbed

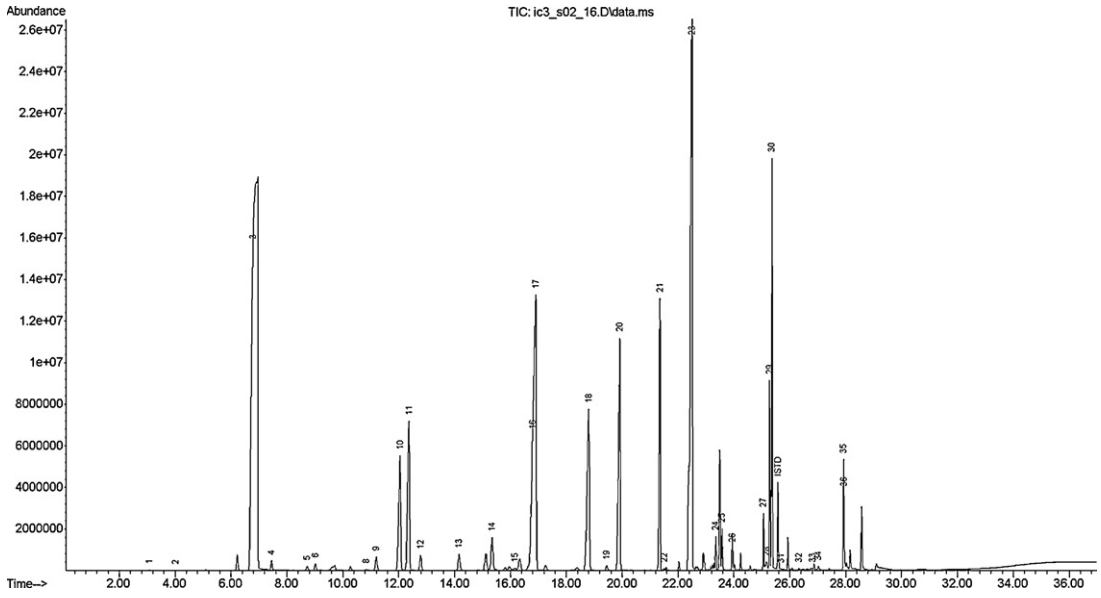


**Fig. 2.** Typical melon chromatogram.

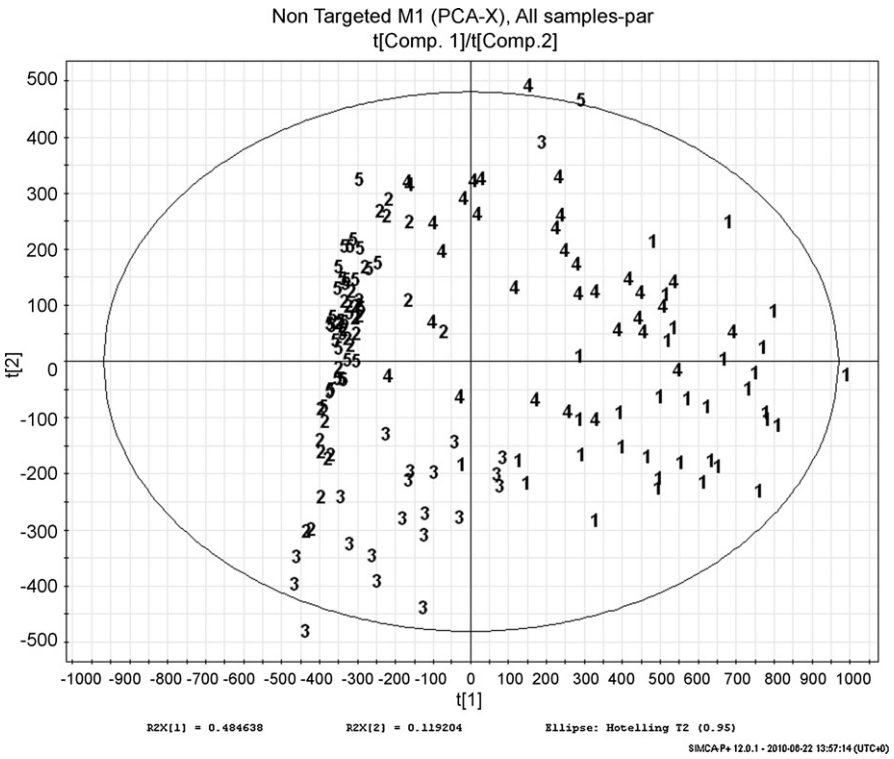


**Fig. 3.** PCA scores plot of the non-targeted approach.

for 4 min at 280 °C while the trap was held at −10 °C using 168:1 split ratio. After the trapping step the trap was heated to 300 °C at 100 °C/s to transfer the volatile components on the GC column. The GC column was VF-624 ms 20 m × 0.15 mm × 0.84 μm (Varian, Palo Alto, USA). The carrier gas was He 5.0 at 0.6 mL/min. The oven program started at 35 °C held for 4 min then ramped at 2 °C/min to 65 °C, ramped at 15 °C/min to 290 °C then held for 3 min. The data acquisition on the mass spectrometer was done in full scan mode in the range 35–400 amu, with a scan rate of 3.89 scan/s, at 230 °C ion source temperature and 150 °C quadruple temperature. MSD Chemstation D.02.00.275 was used as data acquisition and processing system. A typical melon chromatogram is shown in Fig. 2.

### 2.3. Spectral alignment and multivariate statistics

Metalign software was used for spectral alignment of the three-dimensional GC–MS data.

SIMCA-P Multivariate analysis software version 11 from Umetrics AB was used for principal components analysis (PCA) and partial least squares discriminant (PLS-DA) analysis. PCA is a data reduction and visualisation method which seeks to find the main underlying trends in a dataset and present these in simple graphical form. The scores plot shows the relationships present in the observations and the loadings plot shows the relationships between the variables. PCA is an unsupervised method which reflects patterns in datasets with no prior knowledge. In contrast PLS-DA is a supervised method where class membership is assigned before the analysis and a maximum separation projection of the data is made. PLS-DA utilises PLS regression with a binary $Y$ variable representing class assignment and may be thought of as a multivariate equivalent of linear discriminant analysis. In this study the PLS-DA model is used diagnostically, where the regression coefficients highlight which variables are responsible for class separation.

The quality of a multivariate model is measured in terms of both the variance explained and the predictive variance determined by cross-validation (CV). The variance explained represents the fit of the data determined by analysing the amount of variation left in the model residuals. It is summarised by the $R^2$ parameter. The $Q^2$ parameter summarises the predictive variation and is calculated by a "leave many out" cross-validation method. By default in the SIMCA-P software one-seventh of the data is removed and the model re-calculated. The left out data is then predicted by the reduced model and the predictions compared to the actual values. This is completed seven times so that every observation is left out of the model at least once and the total predicted residuals are calculated. From this the $Q^2$ is calculated which is used both as an estimate of the predictive merit of a model and to determine the optimal number of components. Taking too many components may entrain noise (random variation) into the model and is avoided by stopping calculation of the next component if the $Q^2$ starts to decrease.

In addition PLS-DA models may be subjected to a permutation test where the class memberships are deliberately scrambled in order to check that the model could not have arisen by chance. This is a necessary step with "omics" datasets where the large numbers of variables and low number of observations mean that spurious patterns with little or no predictive merit may arise by chance. For an in depth discussion of multivariate model validation the reader is encouraged to read Ref. [24].

## 3. Results and discussion

### 3.1. Non-targeted profile analysis with PCA

The raw data files were first spectrally aligned using Metalign and then a scan number–fragment–intensity (non-targeted) matrix was extracted from all data files. The data was then loaded into SIMCA-P, where it was subjected to principal component analysis in

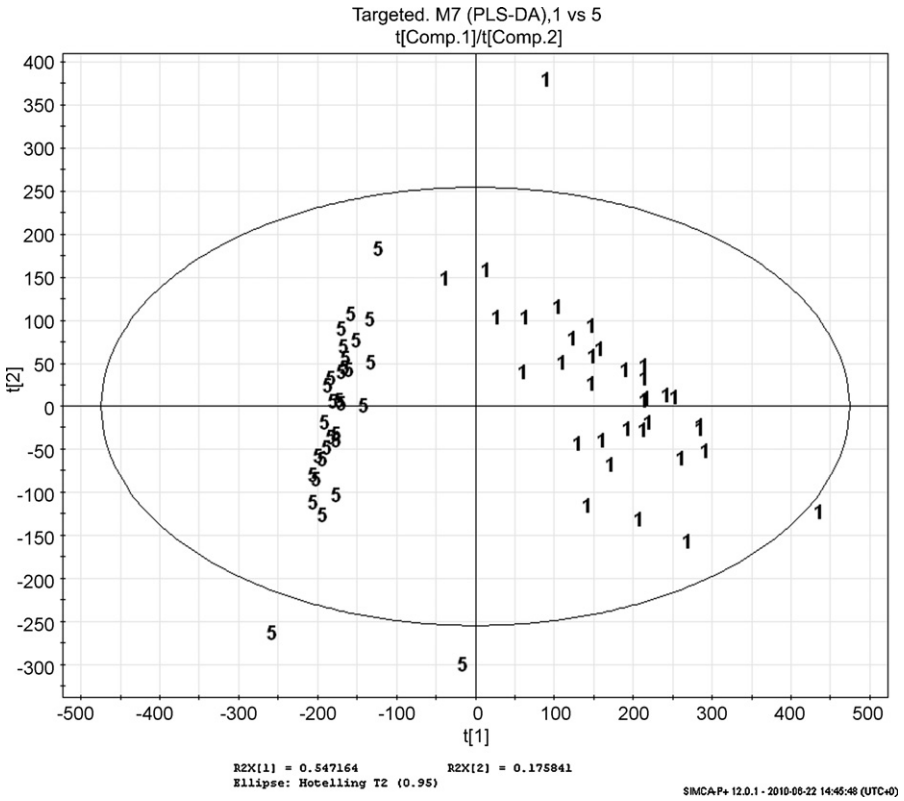


**Fig. 4.** PLS-DA scores plot of varieties 1 and 5.

**Table 1**
List of analytes.

| Nr. | Compounds | Target ion [*m*/*z*] | RT [min] | ID | NIST match |
|---|---|---|---|---|---|
| 1 | Ethyl alcohol | 45 | 3.06 | T | 900 |
| 2 | Methyl acetate | 43 | 3.92 | T | 948 |
| 3 | Ethyl acetate | 43 | 6.61 | std | |
| 4 | Methyl propanoate | 57 | 7.26 | T | 944 |
| 5 | Isobutanol | 43 | 8.54 | T | 925 |
| 6 | Isopropyl acetate | 43 | 8.81 | T | 941 |
| 7 | Acetic acid | 43 | 10.59 | T | |
| 8 | 1-Butanol | 56 | 10.55 | T | 826 |
| 9 | S-methyl thiolacetate | 43 | 10.88 | std | |
| 10 | Ethyl propanoate | 57 | 11.68 | T | 952 |
| 11 | Propyl acetate | 43 | 11.99 | std | |
| 12 | Methyl butyrate | 74 | 12.45 | std | |
| 13 | Dimethyl disulfide | 94 | 13.81 | T | 965 |
| 14 | Ethyl isobutyrate | 43 | 14.99 | T | 880 |
| 15 | 2-Methyl-1-butanol | 57 | 15.78 | std | |
| 16 | Methyl 2-methylbutyrate | 88 | 16.42 | std | |
| 17 | Isobutyl acetate | 43 | 16.51 | std | |
| 18 | Ethyl butyrate | 71 | 18.36 | std | |
| 19 | Propyl propanoate | 57 | 19.13 | T | 921 |
| 20 | Butyl acetate | 43 | 19.57 | std | |
| 21 | Ethyl 2-methylbutyrate | 57 | 21.14 | std | |
| 22 | Ethyl 3-methylbutyrate | 88 | 21.32 | T | 828 |
| 23 | 2-Methylbutyl acetate | 43 | 22.30 | std | |
| 24 | n-Pentylacetate | 43 | 23.22 | std | |
| 25 | Methyl hexanoate | 43 | 23.44 | T | 817 |
| 26 | Ethyl 2-methyl-2-butenoate | 55 | 23.87 | std | |
| 27 | Ethyl hexanoate | 88 | 24.95 | std | |
| 28 | Ethyl 2-methylthioacetate | 61 | 25.05 | std | |
| 29 | (Z)-3-Hexenyl acetate | 67 | 25.16 | std | |
| 30 | Hexyl acetate | 43 | 25.24 | std | |
| 31 | Eucalyptol | 108 | 25.57 | T | 671 |
| 32 | 2,3-Butanediol diacetate | 43 | 26.22 | std | |
| 33 | Ethyl sorbate | 67 | 26.71 | std | |
| 34 | Ethyl 3-metylthiopropanoate | 74 | 26.93 | std | |
| 35 | Benzyl acetate | 108 | 27.81 | std | |
| 36 | Ethyl octanoate | 88 | 27.84 | std | |
| ISTD | 1,4-Dichlorobenzene | 146 | 25.45 | | |

T: tentative identification based on Library search, with NIST match factor; std: identified using a reference standard.

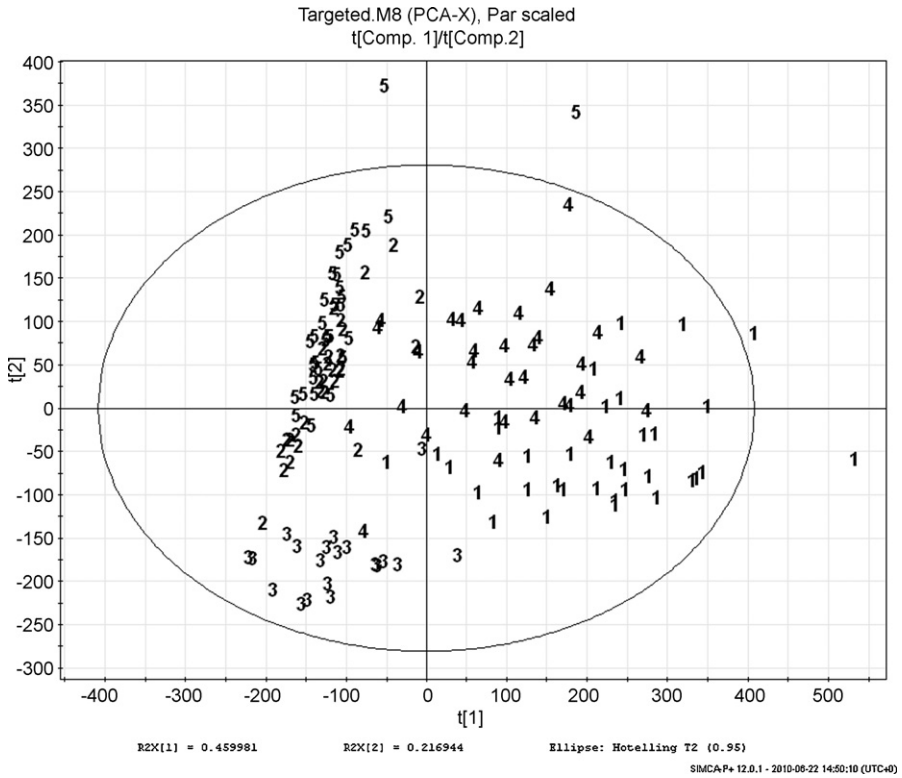


**Fig. 5.** PCA scores plot of the targeted results.

unit variance and Pareto scaling. Unit variance scaling (UV) gives all variables equal influence on the model ensuring small but interesting correlations are not overlooked. A disadvantage of this scaling is exaggeration of experimental noise from small components or baseline effects. Pareto scaling is a compromise between UV scaling and no scaling where medium scale features are up-weighted and large features down-weighted without amplifying the low-level noise. Pareto allows a consideration of the magnitude of the metabolites present. In this case the scaling had very little effect on the clustering in scores-space and for further analysis the Pareto

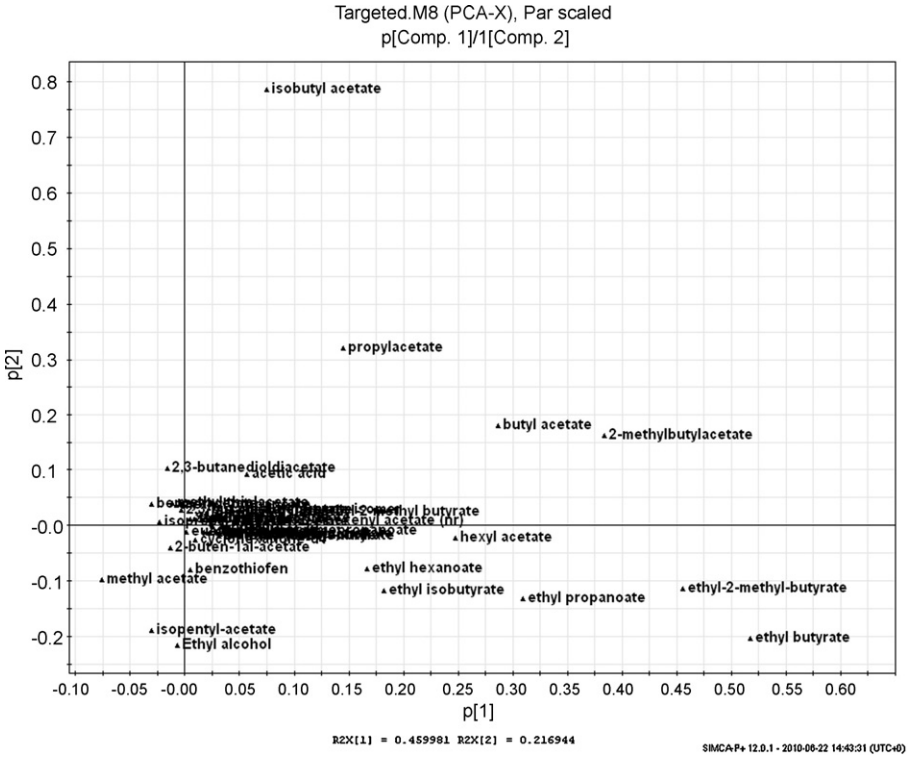


**Fig. 6.** PCA loadings plot of the targeted results.

scaling was chosen. The model had good fit and predictive values: $R^2 = 0.698$, $Q^2 = 0.615$. The $Q^2$ was calculated by the standard 7-fold cross-validation available in the software. The close agreement between $R^2$ and $Q^2$ indicates a stable model with high likelihood of predictive performance. The scores plot of the non-targeted data showed good separation of the five varieties (Fig. 3).

### 3.2. Target list building using PLS discriminant analysis

The data was then subjected to partial least squares discriminant analysis (PLS-DA) to identify scan number–fragment data pairs which contribute most to the observed separation of the varieties. The resultant score plots show the degree whose varieties are separated and the coefficient plots show the magnitude, sign and confidence of the contribution of each variable to the separation (Fig. 4).

The data pairs identified as important were then re-grouped based on the scan numbers and the corresponding peaks were located in the original chromatograms. Tentative peak identification was carried out using NIST library search, some of which was later confirmed by the use of reference standards. The list of melon volatiles is presented in Table 1.

### 3.3. Targeted profile analysis and validation

Once the list of components was established, a targeted data processing method was created. The processing method was then applied to all chromatograms to obtain a sample–relative analyte concentration (targeted) data matrix. The next step was to validate that the target list based data reduction does not result in loss of information. The targeted results were subjected to PCA. The model obtained had again good fit and predictive values: $R^2 = 0.738$, $Q^2 = 0.484$. The scores plot of the targeted results showed an excellent agreement with the scores plot of the non-targeted results (Fig. 5). The volatile compounds contributing most to the profile differences of the varieties were presented by the loadings plot (Fig. 6). $C_5$–$C_8$ esters were identified as most significant of the target compounds.

## 4. Conclusions

A systematic method is described for developing a target list of statistically significant metabolites from a non-targeted metabolomic dataset. Volatile components of fruits and vegetables were determined using TD GC–MS in full scan. The data from non-targeted metabolomic analysis of the samples was used to identify the list of components which contribute to the statistically significant differences between the profiles of the studied varieties. The obtained target list was validated when the data was re-processed in targeted mode and subjected to PCA. The PCA plot obtained using the revised target list showed excellent agreement with that obtained from the non-targeted analysis. The validated target list can be applied to characterize the volatile profile of related melon varieties with component identification.

## References

[1] R. Harker, N.Z. Kiwifruit J. 166 (2004) 5.
[2] E.E. Stashenko, et al., J. Chromatogr. A 1025 (2004) 105.
[3] J.G.S. Maia, et al., J. Food Compos. Anal. 21 (2008) 574.
[4] C. Nunes, et al., Food Chem. 111 (2008) 897.
[5] C. Bicchi, et al., J. Chromatogr. A 1184 (2008) 220.
[6] S. Kreutzmann, et al., Dev. Food Sci. 43 (2006) 505.
[7] J.P. Mattheis, et al., Phytochemistry 31 (3) (1992) 771.
[8] N. Narain, et al., Food Chem. 102 (3) (2007) 726.
[9] L. Rosillo, et al., J. Chromatogr. A 847 (1999) 155.
[10] J.M. Obando-Ulloa, et al., Food Chem. 118 (3) (2010) 815.
[11] S. Elss, et al., LWT—Food Sci. Technol. 38 (2005) 263.
[12] A.Z. Berna, et al., Postharvest Biol. Technol. 46 (2007) 230.
[13] Y.M. Tikunov, et al., Plant Physiol. 139 (2005) 1125.
[14] V. Nielsen, et al., J. Chromatogr. A 805 (1998) 17.
[15] V. Nielsen, et al., Anal. Chem. 71 (1999) 727.
[16] A.L. Duran, et al., Bioinformatics 19 (17) (2003) 2283.
[17] R.E. Williams, et al., Toxicology 207 (2005) 179.
[18] O. Vorst, et al., Metabolomics 1 (2) (2005) 169.
[19] C.A. Smith, et al., Anal. Chem. 78 (2006) 779.
[20] A. Nordström, et al., Anal. Chem. 78 (2006) 3289.
[21] M. Katajamaa, et al., Bioinformatics 6 (2005) 179.
[22] C.D. Broeckling, et al., Anal. Chem. 78 (2006) 4334.
[23] J. van der Greef, et al., Adv. Mass Spectrom. 16 (2004) 145.
[24] L. Eriksson, E. Johansson, N. Kettaneh-Wold, J. Trygg, C. Wikström, S. Wold, Multi- and Megavariate Data Analysis Part I and Part II, 2nd ed., Umetrics, Sweden, 2006.