

Intra- and inter-metabolite correlation spectroscopy of tomato metabolomics data obtained by liquid chromatography-mass spectrometry and nuclear magnetic resonance

Sofia Moco · Jenny Forshed · Ric C. H. De Vos ·
Raoul J. Bino · Jacques Vervoort

Received: 13 December 2007 / Accepted: 9 May 2008 / Published online: 26 May 2008
© The Author(s) 2008

Abstract Nuclear magnetic resonance (NMR) and liquid chromatography-mass spectrometry (LCMS) are frequently used as technological platforms for metabolomics applications. In this study, the metabolic profiles of ripe fruits from 50 different tomato cultivars, including beef, cherry and round types, were recorded by both ^1H NMR and accurate mass LC-quadrupole time-of-flight (QTOF) MS. Different analytical selectivities were found for these both profiling techniques. In fact, NMR and LCMS provided complementary data, as the metabolites detected belong to essentially different metabolic pathways. Yet, upon unsupervised multivariate analysis, both NMR and LCMS datasets revealed a clear segregation of, on the one hand, the cherry tomatoes and, on the other hand, the beef and

round tomatoes. Intra-method (NMR–NMR, LCMS–LCMS) and inter-method (NMR–LCMS) correlation analyses were performed enabling the annotation of metabolites from highly correlating metabolite signals. Signals belonging to the same metabolite or to chemically related metabolites are among the highest correlations found. Inter-method correlation analysis produced highly informative and complementary information for the identification of metabolites, even in the case of low abundant NMR signals. The applied approach appears to be a promising strategy in extending the analytical capacities of these metabolomics techniques with regard to the discovery and identification of biomarkers and yet unknown metabolites.

Electronic supplementary material The online version of this article (doi:10.1007/s11306-008-0112-8) contains supplementary material, which is available to authorized users.

S. Moco (✉) · J. Vervoort
Laboratory of Biochemistry, Wageningen University,
Dreijenlaan 3, 6703 HA Wageningen, The Netherlands
e-mail: sofia.moco@wur.nl

S. Moco · R. C. H. De Vos · R. J. Bino
Plant Research International, P.O. Box 16,
6700 AA Wageningen, The Netherlands

S. Moco · R. C. H. De Vos · R. J. Bino · J. Vervoort
Centre for BioSystems Genomics, P.O. Box 98,
6700 AB Wageningen, The Netherlands

J. Forshed
Karolinska Biomics Center, Karolinska Institutet, Karolinska
University Hospital Solna Z5:02, 17 176 Stockholm, Sweden

R. J. Bino
Laboratory of Plant Physiology, Wageningen University,
Arboretumlaan 4, 6703 BD Wageningen, The Netherlands

Keywords Metabolomics · Liquid chromatography-mass spectrometry · Nuclear magnetic resonance · Correlations · Principal component analysis

1 Introduction

Analytical methods such as NMR and MS provide information about the chemical composition of biological samples, at the molecular level. In plant metabolomics, these two technologies are commonly used as independent approaches for profiling plant systems (Le Gall et al. 2003; Moco et al. 2006a; Tikunov et al. 2005; Ward et al. 2003). The (online) combination of NMR and MS technologies (Jaroszewski 2005; Seger and Sturm 2007; Tatsis 2007; Wolfender et al. 2005), in particular through LC-solid phase extraction (SPE)-NMR-MS (Exarchou et al. 2003) has been used for the efficient detection, separation, isolation and unequivocal structural elucidation of metabolites from plant origin. This analytical approach seems therefore

especially useful in the identification of biomarkers discovered in high-throughput LCMS and ^1H -NMR metabolomics studies.

LCMS and NMR are distinct analytical techniques, concerning detection and sensitivity. On the one hand, LCMS is a fast and sensitive technique. However, the separation of metabolites is dependent on the chromatographic column used, the detection is dictated by the ionization aptitude of the analytes and the molecular elucidation has some intrinsic limitations, such as the resolution of isomers. On the other hand, NMR is indiscriminative towards matrix properties, given that the analytes are soluble. NMR is a highly selective technique for distinguishing molecular structures, but has a lower sensitivity compared to MS (Moco et al. 2007a).

The statistical combination of metabolomics data from profiling equivalent samples by NMR and LCMS opens opportunities to relate spectrometric and spectroscopic properties for single metabolites. These statistical strategies have been applied to large-scale analytical analyses of urine (Crockford et al. 2006; Forshed et al. 2007a, b). Beyond the analytical acquisition of data, the interpretation of metabolomics data is highly dependent on the performed data analysis. Due to the production of large datasets and complexity of untargeted metabolic fingerprinting techniques, appropriate conclusions arise only after pre-processing and statistical validation of metabolomics datasets. Multivariate analyses methods are useful in discriminating information, dealing with the redundancy often present in metabolomics data (Trygg et al. 2007).

One of the statistical methods to establish relationships between metabolite signals belonging to a biological system is correlation analysis. The presence of high correlations in metabolomics data can be attributed to several phenomena which might not be directly related to pathway proximity (Camacho et al. 2005; Steuer 2006). It has been proposed that the highest correlations belong to either an equilibrium situation (positive correlations) or to a mass conservation situation in which there is a moiety-conserved cycle (negative correlations) (Camacho et al. 2005; Steuer 2006).

Both MS (Fraser et al. 2007; Moco et al. 2006a; Schauer et al. 2005; Tikunov et al. 2005) and NMR (Le Gall et al. 2003; Mattoo et al. 2006; Mounet et al. 2007) have previously been used for the profiling of metabolites present in fruits of tomato (*Solanum lycopersicum*). In the present study, we used both ^1H NMR and LC-QTOF-MS techniques to record the metabolic profiles of 50 different tomato cultivars. The semi-polar metabolite content of the fruits was captured by using methanol as extraction solvent. These extracts were successively analysed by ^1H NMR and LCMS, and relationships between the various signals obtained by each method were studied

by performing correlation analyses within the datasets separately. Additionally, the correlation between LCMS and NMR signals across the 50 samples was investigated. Our strategy provided new insight into the complementarity and coincidence of LCMS and NMR as metabolite profiling technologies and as molecular elucidators, applied to the assignment of metabolites in tomato fruit.

2 Materials and methods

2.1 Plant material

Fruits from 50 different cultivars of tomato (*Solanum lycopersicum*), at the ripe stage of development, were obtained from a series of 94 different genotypes grown simultaneously in greenhouses in Wageningen (Tikunov et al. 2005). For each cultivar, a selection of red ripe fruits (12 for round and beef tomatoes and 18 for cherry tomatoes) was pooled to make a representative fruit sample. The fruits were chopped into small pieces and immediately frozen in liquid nitrogen. The frozen material was ground to a fine powder and stored at -80°C before further analysis. Based on the results from analyses of natural volatile compounds in these tomatoes (Tikunov et al. 2005), a sub-collection of 50 cultivars, corresponding to the most contrasting metabolite profiles, was chosen for the analyses performed in this study. These 50 different cultivars included 17 cherry, 26 round and 7 beef type of tomato fruits.

2.2 Chemicals

The standard compounds L-tryptophan (CAS number 73-22-3), D-(+)-glucose (CAS number 50-99-7) and citric acid (CAS number 77-92-9) were purchased from Merck (Damstadt, Germany), rutin (CAS number 153-18-4) from Aldrich (Steinheim, Germany) and chlorogenic acid (CAS number 327-97-9) and α -tomatine (CAS number 17406-45-0) from Sigma (St. Louis, USA). Methanol- d_4 ($\text{H}_2\text{O} + \text{D}_2\text{O} < 0.03\%$) was purchased from Euriso-top (Gif-Sur-Yvette, France) and protonated acetonitrile HPLC supra gradient quality was obtained from Biosolve (Valkenswaard, The Netherlands). Formic acid for synthesis, 98–100%, was purchased from Merck-Schuchardt (Hohenbrunn, Germany). Ultra pure water was obtained from an Elga Maxima purification unit (Bucks, UK).

2.3 Sample preparation for NMR and LCMS analysis

About 0.3 g of fresh weight of tomato fruit powder was freeze-dried immediately before proceeding with NMR and LCMS analyses. To the dried powder, 1.2 ml of methanol- d_4 was added as extraction solvent. The extracts were

sonicated for 15 min, followed by a 5 min-centrifugation (3,000g) step. After filtration of the supernatants through a 0.2 µm inorganic membrane filter (Anotop 10 Whatman, Maidstone, England), exactly 600 µl of tomato extract were transferred to dry 5 mm NMR tubes and taken to NMR analysis. After NMR analysis, the methanol-d₄ tomato extracts were diluted to 25% (v/v) ultra pure water. The diluted extracts were sonicated, centrifuged and filtered before LCMS analysis. In between analyses, the extracts were kept at 4°C.

Standard compounds were dissolved separately in methanol-d₄, to obtain a dilution series of six different concentrations (between about 4 and 130 mg/l). These samples were taken to NMR analyses and later prepared for LCMS, as described above.

2.4 NMR analysis

¹H NMR measurements were carried out in a 500 MHz Bruker AMX NMR spectrometer, proton frequency 500.137 MHz, equipped with a 5 mm TXI probe. A zg pulse sequence was used for the acquisition of ¹H NMR spectra. All measurements were performed at 298 K, containing 1,536 scans with 4 initial dummy scans. The receiver gain was set to 512 and the acquisition time to 2.23 s of 32,768 data points and spectral width of 14.7018 ppm. A 45° pulse was given with a delay of 1.5 s. In total, each measurement took 1 h 35 min 56 s of acquisition time. Data acquisition was done under the control of Bruker XWIN-NMR version 2.1. The datasets were Fourier-transformed, corrected for phase, calibrated for the chemical shift axis towards the resonance of the methanol signal ($\delta = 3.31$ ppm) and baseline corrected.

2.5 LCMS analysis

The extracts were analysed by reversed phase LCMS for their metabolite contents, following the protocol described previously (Moco et al. 2006a). The LC-QTOF-MS analyses were carried out in electrospray (ESI) negative mode. In brief, a Waters Alliance 2795 HT system equipped with a Luna C18(2) pre-column (2.0 × 4 mm) and analytical column (2.0 × 150 mm, 100 Å, particle size 3 µm) from Phenomenex (Torrance, CA, USA) were used for chromatographic separation. The HPLC system was connected online to a Waters 2996 PDA detector and subsequently to a QTOF Ultima V4.00.00 mass spectrometer (Waters-Corporation, MS technologies, Manchester, UK). Mass range and scan rate were set to record m/z 100–1,500/s.

2.6 Data pre-processing

Acquisition, visualization and manual processing of LCMS data were performed under MassLynx 4.0 (Waters). Mass

data were automatically processed by metAlign version 1.0 (<http://www.metalalign.nl>). Baseline correction and noise calculations, which are integrated within this software, were performed from scan number 75 to 2,550, corresponding to retention times 1.5 min (before injection peak) to 50.1 min (before column washing) and the maximum amplitude was set to 30,000. More details about the settings of metAlign can be found elsewhere (De Vos et al. 2007). Mass signals below 50 units of intensity (average through the samples) were discarded, reducing the size of the matrix to 12%.

Processing and visualization of ¹H NMR data was done using Bruker TopSpin version 2.0. (Germany). The matrix of chemical shift amplitudes across all tomato samples was calculated by using a bucket analysis within the AMIX software (Bruker, Germany). The signals were integrated by sum of intensities and normalized to total spectral intensity between samples. A bucket width of 0.01 ppm was applied and the following spectral intervals were excluded: 4.71–5.05 ppm (water signal) and 2.29–3.32 ppm (methanol signal).

The NMR dataset was normalized to total signal intensity per sample and was not further scaled. In contrast, the LCMS data were not normalized to total intensities; instead the intensity of each mass signal was separately scaled over all samples, so that all variables had unit variance.

2.7 Data analysis

The data matrixes of NMR and LCMS were subjected to multivariate and statistical analyses using MATLAB, version 7.10 (MathWorks Inc). Principal component analyses (PCA) were performed on the NMR and LCMS data sets separately. Correlation analyses were performed within NMR signals and LCMS signals, separately, as well as between NMR and LCMS signals. For the correlation analyses, the Pearson correlation coefficient (*corrcoef*) was used. Equation 1 was applied in the calculation of NMR and LCMS correlations, in which *cov* is the covariance matrix of NMR and LCMS variables and σ_{NMR} and σ_{LCMS} are the standard deviations of NMR and LCMS, respectively (including the values of one variable in all samples). The confidence level in all correlations was >99.99%.

$$\text{corrcoef} = \frac{\text{cov}(\text{NMR}, \text{LCMS})}{\sigma_{\text{NMR}} \cdot \sigma_{\text{LCMS}}} \quad (1)$$

Using Mahalanobis distance calculation, spurious or “false” correlations were identified in the correlations. These were correlations where only one sample gave rise to correlation, while the rest of the samples were randomly ordered. The Mahalanobis distance from the deviating sample to the sample group were taken from the correlation plots in which one axis represents the values from one

variable in all samples, and the other axis represent another variable in all samples, Eq. 2.

$$\text{Mahalanobis distance} = \sqrt{\sum_{i=1}^p \left(\frac{x_i - y_i}{\sigma_i} \right)^2}, \quad (2)$$

where the elements of \mathbf{x} are denoted by x_i , the elements of vector \mathbf{y} are denoted by y_i and σ_i is the standard deviation of x_i over all samples. Each sample at a time was regarded as \mathbf{y} , and the distance to the centre (mean) of the remaining samples in \mathbf{x} were measured. One or more Mahalanobis distances >50 in a given correlation plot revealed the presence of a “false” correlation, therefore this correlation was discarded.

Two samples were found to be outliers in the LCMS and NMR data. These were discarded based on three criteria: (1) distinct segregation from the other samples in the PCA plots, (2) abnormal characteristics of the raw data (such as evident differences in the baseline or noise) and (3) outlier-dependent correlations within and between LCMS and NMR data. The correlations presented here were subsequently calculated omitting these outliers.

The visualization of the NMR and LCMS positive correlations was obtained using the open source bioinformatics software Cytoscape version 2.5.2 (<http://www.cytoscape.org>) which enables the visualization of molecular interaction networks. The layout degree sorted circle was applied to the NMR–LCMS positive correlation data ($r \geq 0.8$). This layout algorithm sorts nodes in a circle by degree of the nodes.

For the standard compounds, the coefficients of correlation of the LCMS signal intensities (obtained by metAlign) and NMR resonance intensities (obtained by manual signal integration using Bruker TopSpin) were obtained by the least squares linear regression, using the Statistical Analysis Tools of Microsoft Excel 2003.

3 Results and discussion

3.1 ^1H NMR analyses

^1H NMR profiling was performed on methanol- d_4 extracts of 50 different cultivars of tomato fruit. The NMR spectra appeared as a complex ensemble of resonances, indicating the presence of a multitude of metabolites. In addition, given the presence of resonances throughout the whole spectral width (0–10 ppm), variable chemical features ranging from aliphatic to aromatic groups were detected in the extracts. In particular, intense signals in the sugar region, 3–6 ppm, were observed indicating the presence of glycosylated metabolites and free sugars (Fig. 1). From the visual comparison of the spectra, there is an obvious

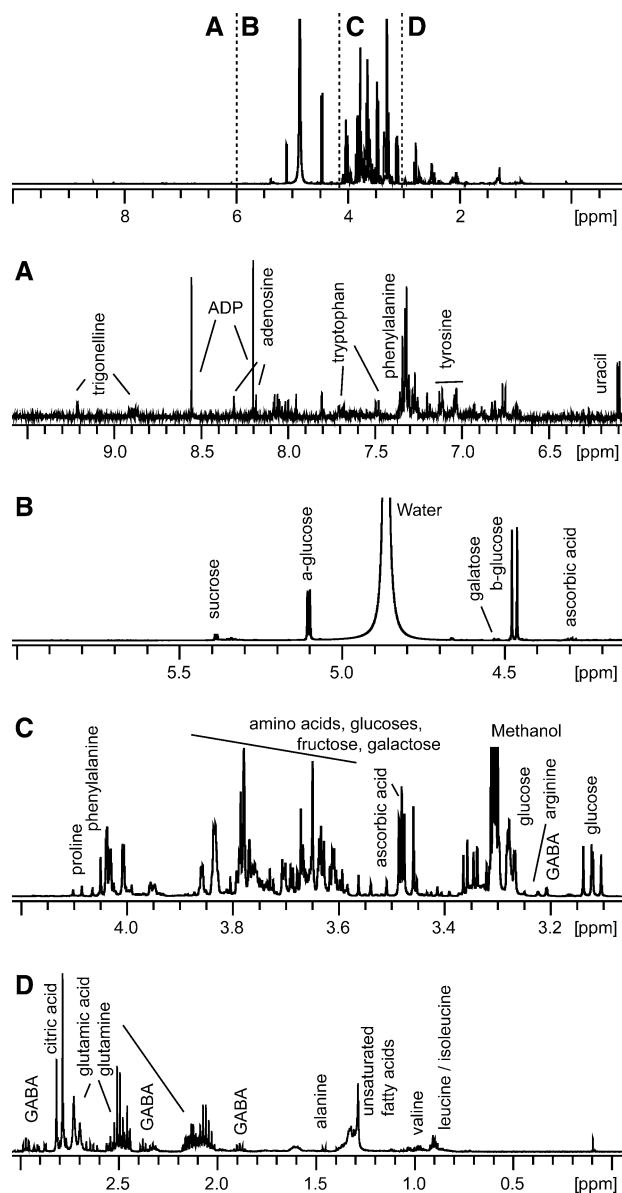


Fig. 1 NMR spectrum of a cherry tomato fruit cultivar (top frame), indicating distinct regions: (a) aromatic region, (b, c) sugar region and (d) amino acid region. Metabolites are indicated next to the respective resonances: GABA = γ -aminobutyric acid; ADP = adenosine diphosphate

similarity between all spectra, suggesting similar metabolic profiles between the different tomato cultivars.

The assignment of NMR resonances to metabolites was based on previously reported findings for tomato fruit samples (Le Gall et al. 2003; Mattoo et al. 2006; Mounet et al. 2007; Sobolev et al. 2003) and NMR-based databases: the Spectral Database for Organic Compounds (SDBS; http://www.aist.go.jp/RIODB/SDBS/cgi-bin/cre_index.cgi), the Flavonoid Database (Moco et al. 2006b), the Human Metabolome Database (HMDB; <http://www.hmdb.ca>) and the Biological

Magnetic Resonance Data Bank (BMRB; <http://www.bmrwisc.edu>).

The analysis of (non-fractionated) tomato fruit extracts by ^1H NMR (Fig. 1) allowed the detection of essentially primary (polar) metabolites such as sugars, amino acids, organic acids and nucleotides, and the abundance of the corresponding resonances indicated the high natural concentration of these metabolites in the fruits. The relatively low abundance of secondary metabolites and the large amount of resonances in the spectrum (as a consequence of the presence of highly abundant metabolites and resonance overlap) made the detection of secondary metabolites, such as phenolic acids, flavonoids and alkaloids, more difficult as compared to the detection of the primary metabolites.

Based on these NMR profiles, the relative differences between the various tomato varieties were visualised by plotting the scores of a principal component analysis (PCA) (Fig. 2). On PC1 (largest variation of all variables) a clear segregation between, on the one hand, the cherry tomatoes and, on the other hand, the beef and round tomatoes was found. The segregation between these two groups was mainly the result of spectral differences in the sugar region, as was concluded from the loading variables of the PCA model (data not shown). Beef and round tomato types were not separated by PC1, implying relatively similarity of the metabolic profiles in the sugar region.

3.2 LCMS analyses

The same extracts used for ^1H NMR were taken for LCMS profiling, after aqueous dilution to 25% (v/v). The obtained metabolic profiles (Fig. 3) were analogous to the ones described before (Moco et al. 2006a, 2007b). The assignment of metabolites was based on the previously published

LCMS-database of tomato fruit metabolites (MoToDB; Moco et al. 2006a, 2007b).

Using C18-reversed phase-LCMS, essentially semi-polar metabolites, such as phenolic acids, alkaloids and flavonoids, were detected. The polar metabolites (i.e. more polar than the stationary phase of the column used for the chromatographic separation), which include sugars, organic acids, most amino acids and nucleotides, elute as large overlapping chromatographic signals before 4 min of retention time. In the PCA scores plot (Fig. 4) the cherry tomatoes were clearly separated from the round and beef tomatoes. The tendency shown in this LCMS-based PCA plot is, in a large extent, analogous to the plot obtained from the NMR analyses (Fig. 2). The similarity between both plots is remarkable given the observed divergence in metabolites detected (compare annotations in Figs. 1 and 3). Therefore, it can be proposed that the metabolic differences, which separate the cherry tomatoes from the round and beef types, belong to both primary (mostly detected by NMR) and secondary (mostly detected by LCMS) metabolism.

3.3 NMR and LCMS dose-responses of standard compounds

Six standard compounds (glucose, citric acid, tryptophan, chlorogenic acid, rutin and tomatine) at different concentrations were analysed by both NMR and LCMS for the assessment of instrumental dose-response relationships (data not shown). As was expected, with both methods an increase in signal height (for MS mass signals) or area (for NMR resonances) was registered with increasing concentrations of the standard compound. However, the linearity range of the instrumental signals along the compound concentrations proved differently between the two techniques. For NMR, the instrumental response was linear for all resonances and for all metabolites tested. As a consequence, the correlations between different resonances of the same metabolite were linear as well (the obtained correlation coefficients, r , were on average 1.00). It is well-known that modern NMR spectrometers have a wide dynamic range (up to 22 bits—more than one million—in the latest hardware configurations), demonstrating the power of NMR as a quantitative technique. On the other hand, as to be expected, MS is more sensitive than NMR which was made clear from the fact that its dynamic range was shifted towards lower absolute concentration ranges as compared to the NMR. The QTOF-MS Ultima instrument used in this study has, however, limitations at the higher concentrations, resulting in a saturation of the response, as reported previously (Chernushevich et al. 2001; Moco et al. 2006a). New types of MS instruments have improved hardware configurations that enable detection in a larger

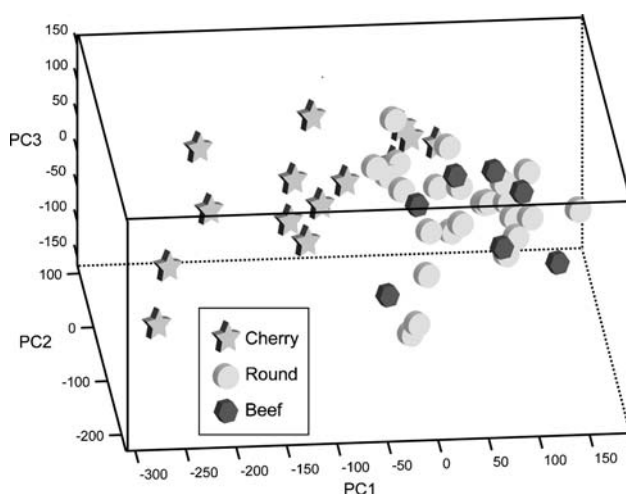


Fig. 2 PCA plot of normalized and mean-centred bucket NMR data of fruits from tomato cultivars (explained variance by PC1 98.5%, PC2 0.6% and PC3 0.3%)

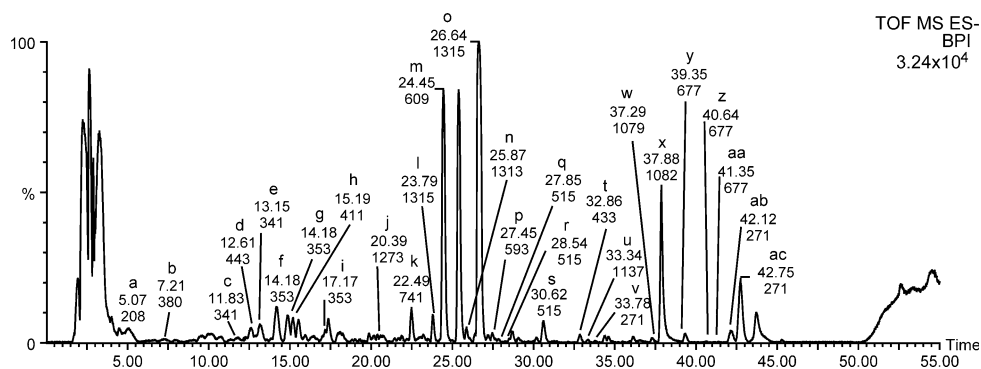


Fig. 3 Negative ion mode electrospray ionization (ESI⁻)-LCMS chromatogram of fruits from a cherry tomato cultivar. For the following metabolites, the retention time (min) and the detected mass are indicated: *a*, (phenylalanine)FA; *b*, zeatin hexose; *c*, *e*, caffeic acid hexose; *d*, dehydrophaseic acid hexose; *f*, *g*, *i*, caffeoylquinic acid; *h*, (iso)pentyl dihexose; *j*, (esculeoside B)FA; *k*, quercetin-hexose-deoxyhexose-pentose; *l*, *o*, (lycoperoside F)FA or (lycoperoside G)FA or(esculeoside

A)FA; *m*, rutin; *n*, (dehydrolycoperoside F)FA or (dehydrolycoperoside G)FA or(dehydroesculeoside A)FA; *p*, kaempferol-3-*O* rutinose; *q*, *r*, *s*, dicaffeoylquinic acid; *t*, *v*, naringenin chalcone-hexose; *u*, (lycoperoside A)FA or (lycoperoside B)FA or (lycoperoside C)FA; *w*, α -tomatin; *x*, tomatoside A; *y*, *z*, *aa*, tricaffeoylquinic acid; *ab*, naringenin; *ac*, naringenin chalcone

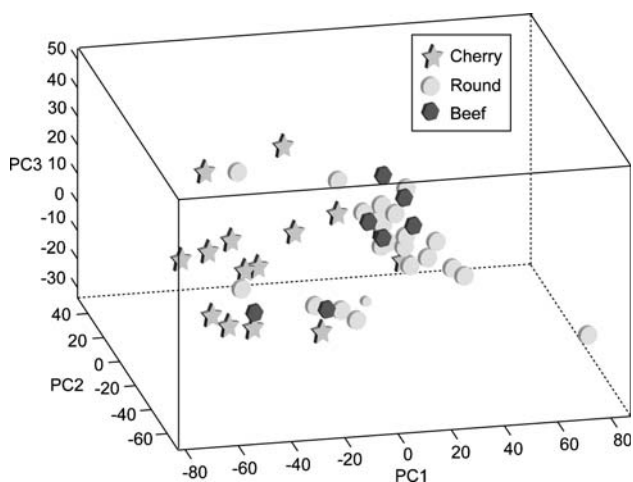


Fig. 4 PCA of mean-centred unit-variance scaled LCMS data of tomato fruit cultivars (explained variance by PC1 90.0%, PC2 4.7% and PC3 1.2%)

dynamic range. With the experimental conditions used in our LCMS analyses, the detector response was linear for parent ion intensities up to about 20,000 counts per scan, corresponding to about 25 μ M of the reference compounds. Above this maximum, there was no longer a linear response with increasing concentrations. At such relative high concentrations, better linearity was obtained for the 2nd and 3rd isotope signals.

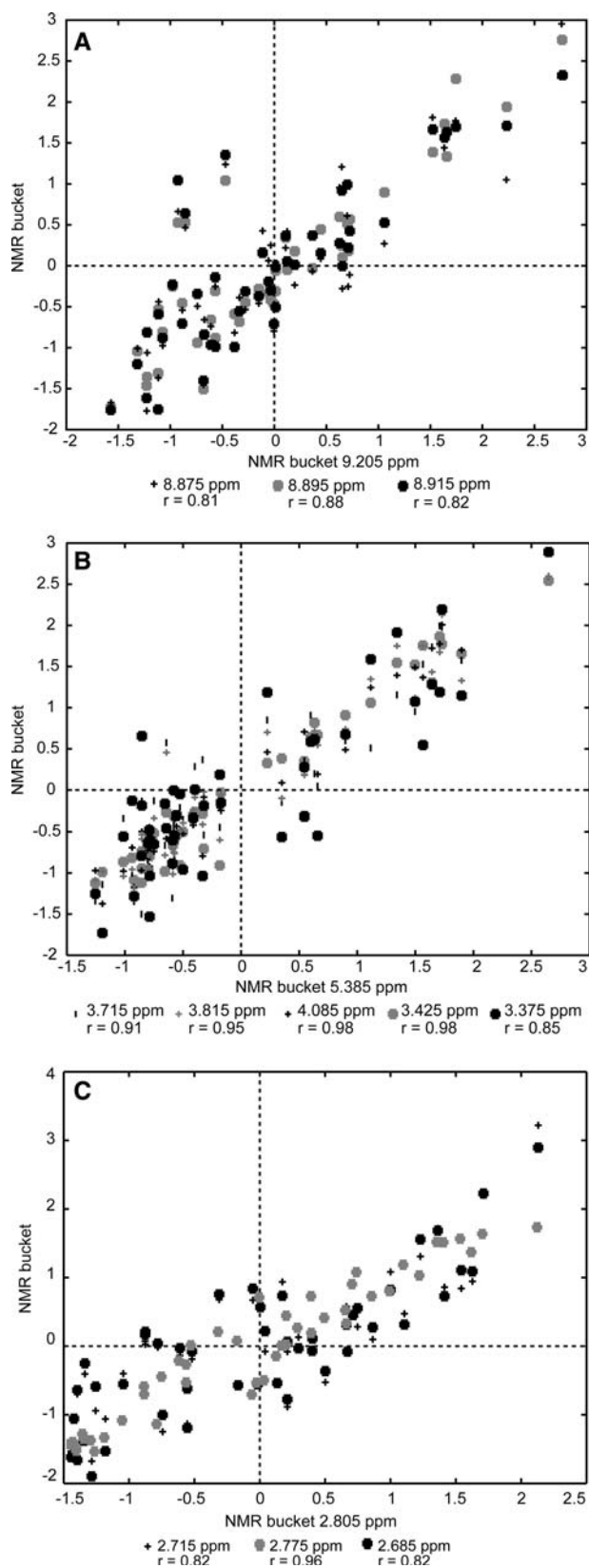
3.4 NMR–NMR signal correlations across samples

Correlation analyses allow the assessment of the degree of linear association between two variables. Correlations can range from the value -1 to $+1$. Negative correlations are obtained for $r < 0$, which indicate linear negative associations between two variables, while positive correlations

are obtained for $r > 0$, indicating positive linear associations. The lack of any association between two variables is originated by $r = 0$. In this study, the similarity in behaviour across the various tomato samples was assessed in terms of signal intensity. For example, the intensity of signal *x*, present in all tomato cultivars, is consistently related to the intensity of signal *y*. These associations between signals are given by their correlation coefficients.

To identify significantly correlating NMR-signals in the data matrix of tomato cultivars, correlation analysis was carried out on the intensities of the different NMR signals (buckets) over all tomato samples, after normalization towards total signal intensities. In these NMR–NMR correlation studies we firstly focussed on the subsection of the matrix with correlation coefficients $|r| \geq 0.8$. This level is considered to be sufficiently high to allow identification of robust associations between variables. From a dataset of 1,008 NMR-buckets, more than 30,000 correlations were found, corresponding to 3.0% of the mathematical possibilities, in which all buckets participated at least once in a correlation pair. In principle, NMR signals belonging to the same compound should show the highest correlations with each other across samples. As examples, NMR–NMR correlations of signals from trigonelline, sucrose and citric acid in the tomato samples are shown in detail in Fig. 5.

The alkaloid trigonelline produced very low signal intensities in the NMR spectra of tomato fruits (see Fig. 1). Due to high deshielding of some of the trigonelline protons, this compound has resonances in a relatively empty region of the NMR spectrum (e.g. at 9.2 and 8.9 ppm). The resonances belonging to trigonelline exhibited high linear correlations ($r \geq 0.81$) between each other across all samples, as shown in Fig. 5a. In fact, this correlation analysis enabled the assignment of signals that otherwise



would have been masked by the high complexity of the tomato extracts and the relative low concentration of trigonelline. Using this NMR–NMR signal correlation

◀ **Fig. 5** Selected normalized signal intensities of NMR buckets of ^1H NMR spectra of tomato cultivars belonging to resonances of the same compound: (a) trigonelline, (b) sucrose and (c) citric acid. The correlation coefficients (r) are displayed for each pair of NMR buckets (x axis, y axis)

approach, a full ^1H NMR characterization of trigonelline was achieved (Fig. S.1, Table S.1).

Sucrose is an abundant compound in tomato fruits, which was evident from the high and characteristic signal at 5.385 ppm in the ^1H NMR spectra (Fig. 1). This disaccharide has a complex NMR spectrum, in particular in the 3–4.5 ppm region which overlaps with the resonances of other free sugars and sugar moieties. Using NMR–NMR correlations across the 50 tomato samples, it was possible to identify other resonances belonging to sucrose (Fig. 5b).

Citric acid was also detected in tomato fruits and its resonance signals appear around 2.8 ppm. The four resonances of citric acid were readily identified from their strong correlations ($r \geq 0.82$) (Fig. 5c).

Given the complexity of the ^1H -NMR spectra of crude tomato fruit extracts, the overlap of signals is a difficulty in the assignment of endogenous metabolites. The identification of NMR resonances belonging to the same compound was facilitated by correlation analyses, enabling the assignment of signals from the same compound within complex mixtures. As the pre-processing of NMR data was performed by bucketing, there was an associated loss in spectral resolution, as well as propagation of resonances through more than one bucket along the samples. However, as there are unavoidable shifts between NMR spectra (even under the best controlled experimental conditions), a direct comparison of all signals is currently impossible, and a pre-processing approach such as bucketing is mandatory.

3.5 LCMS–LCMS signal correlations across samples

Analogously to the NMR dataset, the LCMS dataset of the tomato cultivars was statistically analysed for correlations between LCMS signals across the samples. The pre-processed data matrix contained 3,374 mass signals, aligned by retention time and m/z . More than 130,000 significant strong correlations ($|r| > 0.80$) were obtained, corresponding to 1.2% of the mathematical possibilities, in which all m/z signals participated at least once in a correlation pair. The highest positive correlations ($r \geq 0.96$) were found for mass signals belonging to the same metabolite, such as adducts and fragments. As an example, three compounds were analysed in more detail: sucrose (341 m/z), phenylalanine (164 m/z), and caffeoylquinic acid (353 m/z , retention time 14.9 min) (Fig. 6).

Sucrose was detected both by NMR and LCMS. Sucrose has a short elution time from the LC column: between 2.1

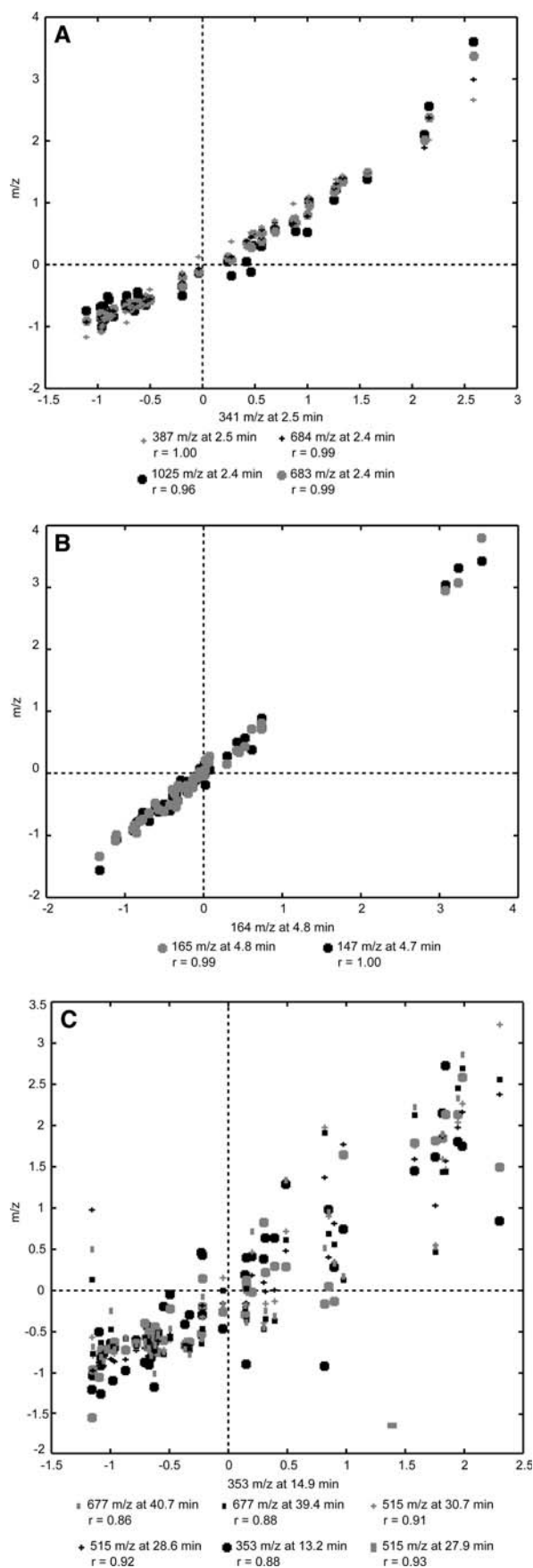


Fig. 6 Selected signal heights of LCMS signals (m/z) with high correlation coefficients (r) to the compounds: (a) sucrose (341 m/z)—387 m/z = [sucrose + HCOOH - H]⁻, 683 m/z = [sucrose + sucrose - H]⁻, 684 m/z = 2nd isotope of [sucrose + sucrose - H]⁻, 1,025 m/z [sucrose + sucrose + sucrose - H]⁻; (b) phenylalanine (164 m/z)—165 m/z = 2nd isotope of [phenylalanine - H]⁻, 147 m/z = [phenylalanine - NH₃ - H]⁻ and (c) caffeoylquinic acid II (353 m/z at 14.9 min)—677 m/z at 40.7 min = [tricafeoylquinic acid II - H]⁻, 677 m/z at 39.4 min = [tricafeoylquinic acid I - H]⁻, 515 m/z at 30.7 min = [dicaffeoylquinic acid III - H]⁻, 515 m/z at 28.6 min = [dicaffeoylquinic acid II - H]⁻, 515 m/z at 27.9 min = [dicaffeoylquinic acid I - H]⁻, 353 m/z at 13.2 min = [caffeoylquinic acid I - H]⁻. The correlation coefficients (r) are displayed for each pair of LCMS signals (x axis, y axis)

and 2.8 min, with a large overlap with other eluting polar metabolites. Parent ion adducts (dimer and trimer), isotopes and formic acid adducts of sucrose exhibited correlations with the molecular ion of sucrose (341 m/z), with $r > 0.96$ (Fig. 6a).

Also the amino acid phenylalanine is detectable by both NMR and LCMS in tomato fruit extracts. Very high correlations ($r > 0.99$) were found between isotopes and a fragment of phenylalanine and the molecular ion of phenylalanine (164 m/z) (Fig. 6b).

A series of caffeoylquinic acid derivatives have been putatively assigned in tomato: three caffeoylquinic acid isomers, three dicaffeoylquinic acid isomers and three tricafeoylquinic acid isomers (Moco et al. 2006a). For the secondly eluting isomer of caffeoylquinic acid (m/z 353 at retention time 14.9 min, assigned as 5-caffeoylquinic acid (Moco et al. 2006a), not only signals related to this specific isomer such as isotopes, adducts, parent ion adducts were found to be highly correlated ($r > 0.98$), but interestingly also two signals derived from other caffeoylquinic acids derivatives, including dicaffeoylquinic acids and tricafeoylquinic acids (Fig. 6c, Table S.2). Thus, this correlation analysis enabled us to verify the presence of other caffeoylquinic acid derivatives in tomato fruit: clearly indicating the potential of this strategy to identify biochemically related compounds within complex mixtures.

High positive ($r > 0.8$) LCMS–LCMS correlations were found not only between signals belonging to the same metabolite, but also among chemically related metabolites, such as the isomer series of (di and tri)caffeoylquinic acids. This observation suggests the presence of equilibriums and/or asymmetric control between the various (di and tri)caffeoylquinic acids (Camacho et al. 2005; Steuer 2006).

3.6 NMR–LCMS signal correlations across biological samples

NMR and LCMS datasets acquired for the different tomato cultivars were further analysed using an inter-method correlation analysis. In order to achieve this, each LCMS

signal (m/z value at a certain retention time) was directly compared to each chemical shift present in the NMR buckets, producing a $3,374 \times 1,008$ correlation matrix. This correlation matrix can be visualized as a heat-map. As an example, a section of the obtained NMR–LCMS correlation map, for $|r| \geq 0.8$, is shown in Fig. 7. It can be seen that resonances at 3.3 and 5.3 ppm had a high positive correlation to numerous m/z signals (Fig. 3), suggesting that these LCMS signals are originated from glycosylated metabolites. In fact, at this retention time region (23.12–27.55 min), metabolites such as glycosylated flavonoids (quercetin and kaempferol glycosides) and alkaloids of the lycoperside-type are present (see Fig. 3).

The performed NMR–LCMS correlation analysis led to 514 significant positive and 47 significant negative correlations for $|r| \geq 0.8$, corresponding to 0.16‰ of the mathematical possibilities, in which 6% of the signals (LCMS m/z 's and NMR buckets) participated at least once in a correlation pair. To visualize the connectivities between these highly correlated NMR buckets and LCMS signals, a network was constructed (Fig. 8). This network has 246 nodes (total number of unique NMR buckets and LCMS signals) and 514 edges (total number of positive correlations). The degree of connectivity between NMR buckets and LCMS signals is high, as only a few isolated clusters were observed within this network.

The largest cluster (Fig. 8a) comprehends more than 50% of the obtained nodes. The represented NMR buckets belong to sugar resonances (between 3.4 and 5.4 ppm), except for one bucket (at 8.205 ppm). This latter bucket is tentatively assigned to formic acid and establishes high correlations to a range of LCMS signals. Mostly sugars (free and bound) are present in the LCMS signals within this cluster, as seen by the abundance of signals related to

polar metabolites such as free sugars (retention times lower than 4 min) and glycosylated alkaloids and saponins corresponding to high m/z values. In fact, the NMR buckets 5.375, 5.385 and 5.395 ppm, which correspond to the anomeric protons of sugar moieties, correlated to 75% of the total of LCMS signals within this cluster, suggesting the abundance of glycosylated metabolites.

The second largest cluster (Fig. 8b) corresponds to mainly naringenin or naringenin chalcone derivatives. Tomato fruit contains endogenous naringenin chalcone which after extraction slowly undergoes chemical cyclization into naringenin (Moco et al. 2006a, b). Therefore the presence of naringenin is linked to the presence of naringenin chalcone. This cluster B contains various NMR buckets representing the ^1H NMR chemical shifts of both naringenin and naringenin chalcone (Table 1), as well as LCMS signals belonging to these two metabolites, including molecular ions, isotopes (up to the 4th), adducts, dimers (with Na-formic acid), fragments and isotopes of fragments, as well as hexose-substituted species. In total, eight different naringenin or naringenin chalcone derivatives could be identified by means of this NMR–LCMS correlation analysis. In fact, the intensity of the NMR resonances belonging to naringenin is rather low, which would make identification of this metabolite in tomato by only ^1H -NMR profiling very difficult. Within cluster B are also NMR–LCMS connectivities related to signals from two kaempferol glycosides: kaempferol-hexose-deoxyhexose-pentose ($[\text{M}-\text{H}]^- = 725$ m/z , and kaempferol-3-rutinoside ($[\text{M}-\text{H}]^- = 593$ m/z , 2nd isotope of $[\text{M}-\text{H}]^- = 594$ m/z and Na-formic acid adduct = 661 m/z). It is well known that naringenin, naringenin chalcone and kaempferol are derived from the same flavonoid biosynthetic pathway (see e.g. KEGG PATHWAY Database; <http://www.genome.jp/kegg/pathway.html>), so the high correlations between LCMS and

Fig. 7 Section of the correlation map of NMR–LCMS for tomato fruit: [3.005–6.995 ppm (NMR); 2,900–3,100 (LCMS peak number) = (23.12–27.55 min)]. Horizontally, the NMR buckets (ppm) of the samples are overlaid and vertically the LCMS peak numbers (peak numbers increase with retention time and m/z) of the samples are overlaid. In the central frame, the correlation coefficients, r , for NMR–LCMS correlations are displayed as a blue-red heat map

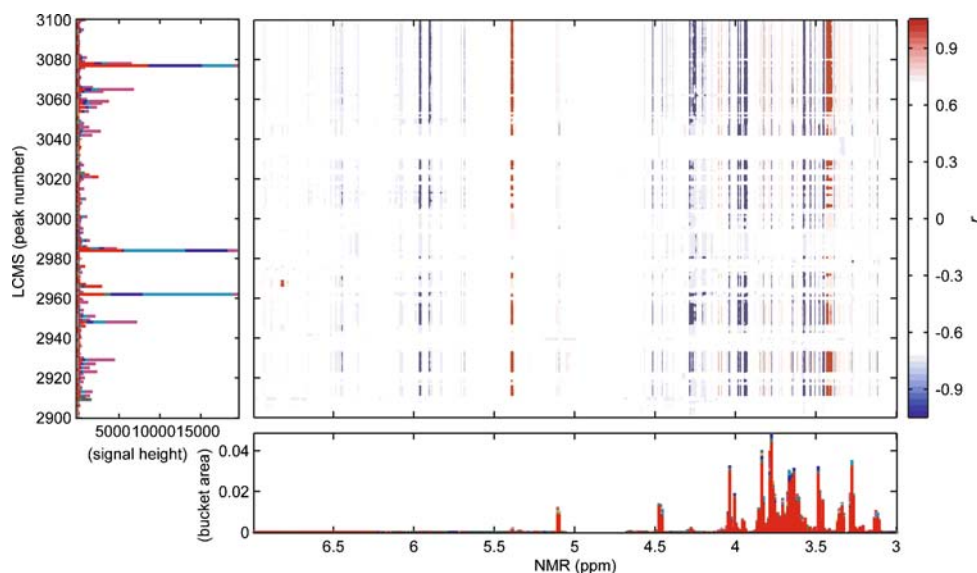


Fig. 8 Network structures shown in (a–g), obtained from the NMR–LCMS correlations of tomato fruit data, for $r \geq 0.8$. Correlations between NMR buckets (in ppm; as black circles) and LCMS signals (as m/z at a specific retention time, in min; as white circles) are linked by a grey line

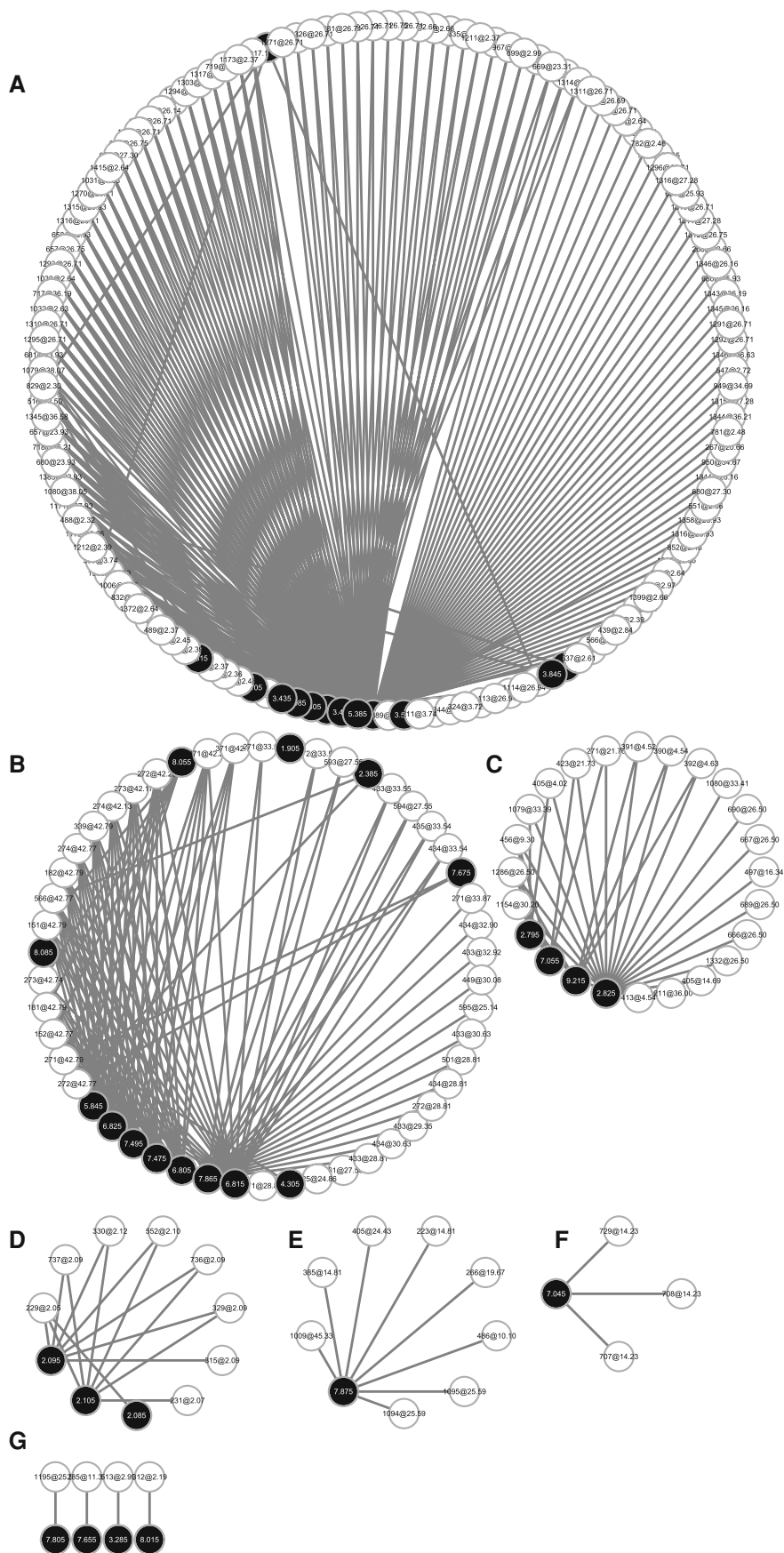
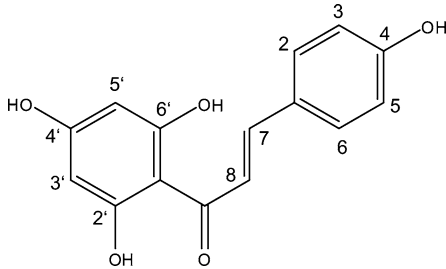
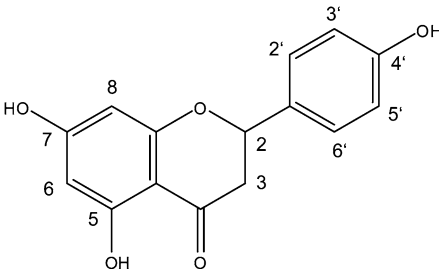


Table 1 Putative assignment of naringenin and naringenin chalcone from the NMR buckets obtained by NMR–LCMS correlation analysis and network representation

NMR bucket (ppm)	Naringenin chalcone	Naringenin
		
5.845	H3'/H5'	H8 or H6
6.805	H5/H3	H3'/H5'
6.815		
6.825		
7.475	H2/H6	H2'/H6'
7.495		
7.675	H7	–
7.865	H2/H6 with 4-conjugation (glycolsylation?)	H2'/H6' with 4'-conjugation (glycosylation?)
8.055	H8	–
8.085		–

NMR signals of these flavonoids are a good indication of the potency of this approach to complement metabolomics data.

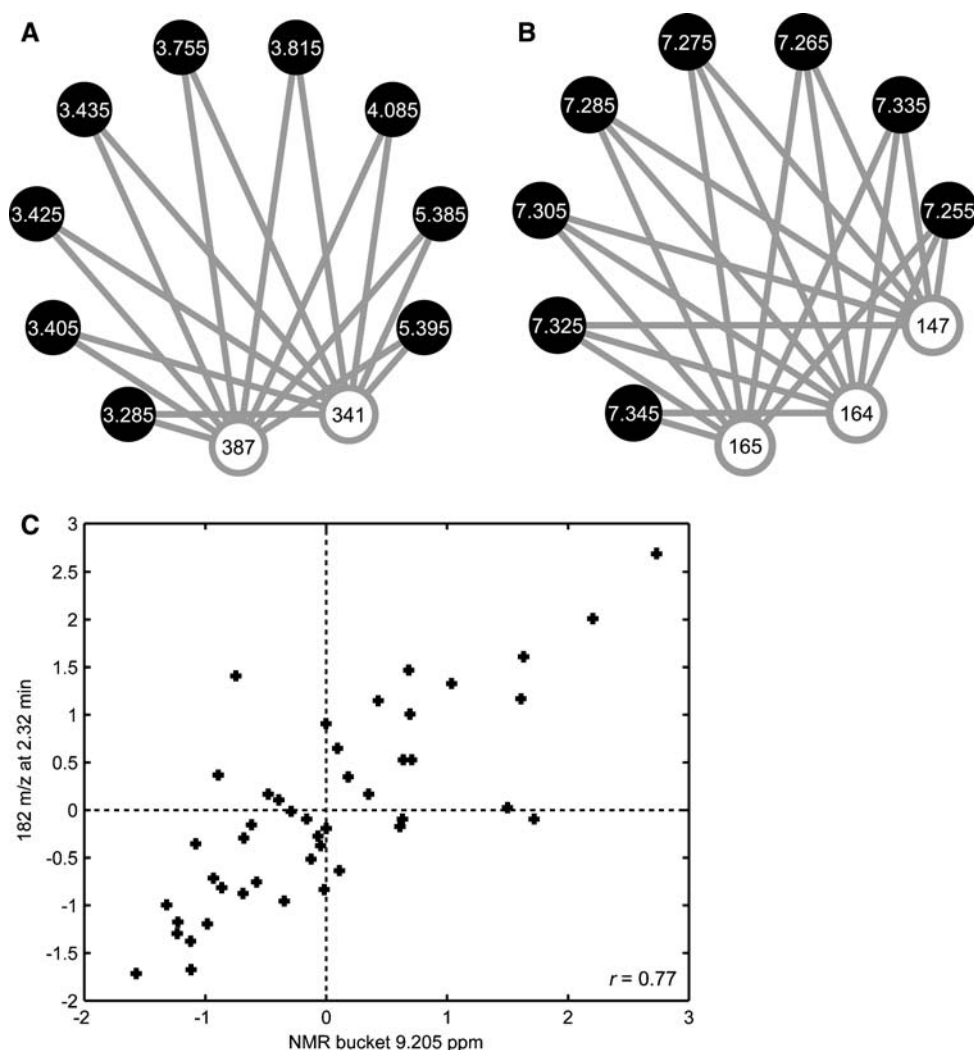
The other structures observed within this NMR–LCMS correlation network (Fig. 8c–g) correspond to metabolites with either low NMR–LCMS overlap or to NMR regions which were found to be more prone for chemical shift shifts. A logical chemical dependency between NMR buckets and LCMS signals could be identified in none of these clusters. These structures may therefore be related to chemical properties of yet unknown metabolites, or be a consequence of statistical coincidence.

Upon lowering the NMR–LCMS correlation threshold from $|r| \geq 0.8$ to $|r| \geq 0.7$, the number of significant correlations increased considerably: from 514 to 5,258 positive correlations and from 47 to 1,638 negative correlations. Within the dataset of $0.7 \leq |r| < 0.8$, several metabolites with high intra-method signal correlations (Figs. 5, 6) were also found to correlate in this inter-method approach, such as sucrose (Fig. 9a), phenylalanine (Fig. 9b) and trigonelline (Fig. 9c).

The observed correlations within and between NMR and LCMS variables can be the result of different interactions between the variables. They may be the effect of the presence of different signals derived from exactly the same metabolite (e.g. Figs. 5, 6a, b, 9), from biochemically related metabolites (Fig. 6c), or from unrelated metabolites that have a similar abundance pattern across the samples.

From the present correlation analyses, it can be concluded that the chemical overlap between NMR and LCMS signals obtained by our tomato metabolic profiling techniques was significant. The percentage of overlap between NMR and LCMS signals was roughly 6% (with $r \geq 0.8$). However, it is difficult to infer the number of coincident metabolites captured by both NMR and LCMS. Nevertheless, several adaptations or/and improvements in the hardware, analytical set up, data pre-processing and statistical analyses strategies may help to increase the quality and number of intra-method correlated signals obtained by NMR- and LCMS-based metabolic profiling approaches. Firstly, the investigation of sample fractionation protocols can contribute to diminish NMR-spectral complexity and inherent signal overlap. Secondly, optimization of the nature of interaction and polarity range of the chromatography used in the LCMS can be explored. For example, the use of hydrophilic columns, which are more suitable for chromatographic separation of polar metabolites also detected by NMR (such as amino acids, organic acids, sugars) than the reverse-phase column used in the present study, may result in an increased chemical overlap. Thirdly, an increase in the response linearity of MS at high metabolite abundance, resulting in a better linear correlation with the less sensitive NMR, can be obtained by using wider-dynamic range MS instruments, such as modern TOF-MS machines (e.g. equipped with an analog-to-digital detector) or Fourier Transform (FT)-MS. The improved

Fig. 9 Examples of NMR–LCMS correlations for specific metabolites represented as networks (a, b; for colour coding see Fig. 8) or as a graph (c): (a) sucrose (NMR–LCMS correlations with $0.70 < r < 0.78$); (b) phenylalanine (NMR–LCMS correlations with $0.70 < r < 0.78$) and (c) trigonelline (NMR–LCMS correlation with $r = 0.77$)



mass accuracy of FTMS instruments can also facilitate the identification of metabolite signals observed in NMR–LCMS correlations. In addition, the combination of data obtained by negative and positive ionization modes can be fruitful to increase the overlap of LCMS with NMR. Fourthly, developments on the quality of NMR data, such as obtaining flat baselines, very precise temperature control and increase in sensitivity (by making use of cryogenic probes, higher magnetic field instruments or other type of NMR measurements), can be advantageous. Fifthly, the development of improved deconvolution software for NMR spectra, which can be used for batch-wise pre-processing of NMR spectra from complex mixtures, to avoid the usage of bucketing methods which result in a reduction of spectral resolution and inherent overlap of frequency signals from different metabolites. Sixthly, the development of algorithms that can correct for non-linearity in instrumental responses can contribute to the quality of the inter-method correlations. The implementation of one or more of these items can lead to improvements in both the

quantity and quality of the NMR–LCMS correlations and may result in a wider application of this promising strategy in metabolomics studies.

4 Concluding remarks

The application of intra- and inter-method correlation analysis of untargeted metabolomics data, obtained by LCMS- and ^1H -NMR-based profiling of exactly the same biological material, appears to be a powerful strategy in discovering and linking metabolite information, which can be used in biomarker discovery and annotation, and in the analysis of metabolic pathways. The percentage of obtained intra-correlations, in both cases $\leq 3\%$, indicates the presence of a complex control among the metabolites, for the amount of low correlations ($|r| < 0.6$) prevails.

Despite the different analytical selectivities and sensitivities of the NMR and LCMS used in the present study, the fusion of metabolomics data through NMR–LCMS

inter-correlation analysis generated a substantial number of significantly correlating metabolite signals. Based on these correlating signals, the identification of the underlying metabolites was facilitated and relationships between chemically-related metabolites could be observed, proving the validity of this approach. Precisely due the analytical differences inherent to NMR and LCMS, a larger section of the metabolome could be captured, as together, NMR and LCMS, contributed to the detection of a larger and wider range of chemically diverse metabolites. Hence, we can conclude that intra- and inter-method correlation analyses have potential applications in the interpretation of metabolomics data obtained by $^1\text{H-NMR}$ and LCMS.

Acknowledgements The authors thank Pim Lindhout and Arnaud Bovy for the assistance on the selection of tomato cultivars, and Oscar Vorst, Arjen Lommen, Lutgarde Buydens, Peter Neidig and the reviewers for valuable comments and reflections on the fusion of NMR and LCMS data. This work was financially supported by the EU RTD project “Capillary NMR”, a European Community-Access to Research Infrastructure action of the Improving Human Potential Program, Contract HPRI-CT-1999-00085, contract HPRI-CT-1999-50018, the research programme of the Centre of BioSystems Genomics (CBSG) which is a part of The Netherlands Genomics Initiative/Netherlands Organization for Scientific Research, the EU project “META-PHOR”, contract number FOOD-CT-2006-036220 and the Wageningen NMR Centre.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

- Camacho, D., de la Fuente, A., & Mendes, P. (2005). The origin of correlations in metabolomics data. *Metabolomics*, *1*, 53–63. doi:10.1007/s11306-005-1107-3.
- Chernushevich, I. V., Loboda, A. V., & Thomson, B. A. (2001). An introduction to quadrupole-time-of-flight mass spectrometry. *Journal of Mass Spectrometry*, *36*, 849–865. doi:10.1002/jms.207.
- Crockford, D. J., Holmes, E., Lindon, J. C., Plumb, R. S., Zirah, S., Bruce, S. J., Rainville, P., Stumpf, C. L., & Nicholson, J. K. (2006). Statistical heterospectroscopy, an approach to the integrated analysis of NMR and UPLC-MS data sets: Application in metabonomic toxicology studies. *Analytical Chemistry*, *78*, 363–371. doi:10.1021/ac051444m.
- De Vos, R. C. H., Moco, S., Lommen, A., Keurentjes, J. J. B., Bino, R. J., & Hall, R. D. (2007). Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. *Nature Protocols*, *2*, 778–791. doi:10.1038/nprot.2007.95.
- Exarchou, V., Godejohann, M., van Beek, T. A., Gerotherassis, I. P., & Vervoort, J. (2003). LC-UV-solid-phase extraction-NMR-MS combined with a cryogenic flow probe and its application to the identification of compounds present in Greek oregano. *Analytical Chemistry*, *75*, 6288–6294. doi:10.1021/ac0347819.
- Forshed, J., Idborg, H., & Jacobsson, S. P. (2007a). Evaluation of different techniques for data fusion of LC/MS and $^1\text{H-NMR}$. *Chemometrics and Intelligent Laboratory Systems*, *85*, 102–109. doi:10.1016/j.chemolab.2006.05.002.
- Forshed, J., Stolt, R., Idborg, H., & Jacobsson, S. P. (2007b). Enhanced multivariate analysis by correlation scaling and fusion of LC/MS and $^1\text{H-NMR}$ data. *Chemometrics and Intelligent Laboratory Systems*, *85*, 179–185. doi:10.1016/j.chemolab.2006.06.012.
- Fraser, P. D., Enfissi, E. M. A., Goodfellow, M., Eguchi, T., & Bramley, P. M. (2007). Metabolite profiling of plant carotenoids using the matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *The Plant Journal*, *49*, 552–564. doi:10.1111/j.1365-3113X.2006.02949.x.
- Jaroszewski, J. W. (2005). Hyphenated NMR methods in natural products research, part 2: HPLC-SPE-NMR and other new trends in NMR hyphenation. *Planta Medica*, *71*, 795–802. doi:10.1055/s-2005-873114.
- Le Gall, G., Colquhoun, I. J., Davis, A. L., Collins, G. J., & Verhoeven, M. E. (2003). Metabolite profiling of tomato (*Lycopersicon esculentum*) using $^1\text{H-NMR}$ spectroscopy as a tool to detect potential unintended effects following a genetic modification. *Journal of Agricultural and Food Chemistry*, *51*, 2447–2456. doi:10.1021/jf0259967.
- Mattoo, A. K., Sobolev, A. P., Neelam, A., Goyal, R. K., Handa, A. K., & Segre, A. L. (2006). Nuclear magnetic resonance spectroscopy-based metabolite profiling of transgenic tomato fruit engineered to accumulate spermidine and spermine reveals enhanced anabolic and nitrogen-carbon interactions. *Plant Physiology*, *142*, 1759–1770. doi:10.1104/pp.106.084400.
- Moco, S., Bino, R. J., Vorst, O., Verhoeven, H. A., de Groot, J., van Beek, T. A., Vervoort, J., & De Vos, R. C. H. (2006a). A liquid chromatography-mass spectrometry-based metabolome database for tomato. *Plant Physiology*, *141*, 1205–1218. doi:10.1104/pp.106.078428.
- Moco, S., Tseng, L. H., Spraul, M., Chen, Z., & Vervoort, J. (2006b). Building-up a comprehensive database of flavonoids based on nuclear magnetic resonance data. *Chromatographia*, *9/10*, 503–508. doi:10.1365/s10337-006-0077-6.
- Moco, S., Bino, R., De Vos, R. C. H., & Vervoort, J. (2007a). Metabolomics technologies and metabolite identification. *TrAC Trends in Analytical Chemistry*, *26*, 855–866. doi:10.1016/j.trac.2007.08.003.
- Moco, S., Capanoglu, E., Tikunov, Y., Bino, R. J., Boyacioglu, D., Hall, R. D., Vervoort, J., & De Vos, R. C. H. (2007b). Tissue specialization at the metabolite level is perceived during the development of tomato fruit. *Journal of Experimental Botany*, *58*, 4131–4146. doi:10.1093/jxb/erm271.
- Mounet, F., Lemaire-Chamley, M., Maucourt, M., Cabasson, C., Giraudel, J. L., Deborde, C., Lessire, R., Gallusci, P., Bertrand, A., Gaudillere, M., Rothan, C., Rolin, D., & Moing, A. (2007). Quantitative metabolic profiles of tomato flesh and seeds during fruit development: Complementary analysis with ANN and PCA. *Metabolomics*, *3*, 273–288. doi:10.1007/s11306-007-0059-1.
- Schauer, N., Zamir, D., & Fernie, A. R. (2005). Metabolic profiling of leaves and fruit of wild species tomato: A survey of the Solanum lycopersicum complex. *Journal of Experimental Botany*, *56*, 297–307. doi:10.1093/jxb/eri057.
- Seger, C., & Sturm, S. (2007). Analytical aspects of plant metabolite profiling platforms: Current standings and future aims. *Journal of Proteome Research*, *6*, 480–497. doi:10.1021/pr0604716.
- Sobolev, A. P., Segre, A., & Lamanna, R. (2003). Proton high-field NMR study of tomato juice. *Magnetic Resonance in Chemistry*, *41*, 237–245. doi:10.1002/mrc.1176.
- Steuer, R. (2006). On the analysis and interpretation of correlations in metabolomic data. *Briefings in Bioinformatics*, *7*, 151–158. doi:10.1093/bib/bbl009.
- Tatsis, E. (2007). Identification of the major constituents of *Hypericum perforatum* by LC/SPE/NMR and/or LC/MS. *Phytochemistry*, *68*, 383–393. doi:10.1016/j.phytochem.2006.11.026.

- Tikunov, Y., Lommen, A., De Vos, R. C. H., Verhoeven, H. A., Bino, R. J., Hall, R. D., & Bovy, A. G. (2005). A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. *Plant Physiology*, *139*, 1125–1137. doi:[10.1104/pp.105.068130](https://doi.org/10.1104/pp.105.068130).
- Trygg, J., Holmes, E., & Lundstedt, T. (2007). Chemometrics in metabonomics. *Journal of Proteome Research*, *6*, 469–479. doi:[10.1021/pr060594q](https://doi.org/10.1021/pr060594q).
- Ward, J. L., Harris, C., Lewis, J., & Beale, M. H. (2003). Assessment of H-1 NMR spectroscopy and multivariate analysis as a technique for metabolite fingerprinting of *Arabidopsis thaliana*. *Phytochemistry*, *62*, 949–957. doi:[10.1016/S0031-9422\(02\)00705-7](https://doi.org/10.1016/S0031-9422(02)00705-7).
- Wolfender, J. L., Queiroz, E. F., & Hostettmann, K. (2005). Phytochemistry in the microgram domain—a LC-NMR perspective. *Magnetic Resonance in Chemistry*, *43*, 697–709. doi:[10.1002/mrc.1631](https://doi.org/10.1002/mrc.1631).