Breakthrough Technologies

A Liquid Chromatography-Mass Spectrometry-Based Metabolome Database for Tomato[1]

Sofia Moco*, Raoul J. Bino, Oscar Vorst, Harrie A. Verhoeven, Joost de Groot, Teris A. van Beek, Jacques Vervoort, and C.H. Ric de Vos

Laboratory of Biochemistry, Wageningen University, 6703 HA Wageningen, The Netherlands (S.M., J.V.); Plant Research International, 6700 AA Wageningen, The Netherlands (S.M., R.J.B., O.V., H.A.V., J.d.G., C.H.R.d.V.); Centre for BioSystems Genomics, 6700 AB Wageningen, The Netherlands (S.M., R.J.B., O.V., H.A.V., C.H.R.d.V.); Laboratory of Organic Chemistry, Wageningen University, 6703 HB Wageningen, The Netherlands (T.A.v.B.); and Laboratory of Plant Physiology, Wageningen University, 6703 BD Wageningen, The Netherlands (R.J.B.)

For the description of the metabolome of an organism, the development of common metabolite databases is of utmost importance. Here we present the Metabolome Tomato Database (MoTo DB), a metabolite database dedicated to liquid chromatography-mass spectrometry (LC-MS)- based metabolomics of tomato fruit (*Solanum lycopersicum*). A reproducible analytical approach consisting of reversed-phase LC coupled to quadrupole time-of-flight MS and photodiode array detection (PDA) was developed for large-scale detection and identification of mainly semipolar metabolites in plants and for the incorporation of the tomato fruit metabolite data into the MoTo DB. Chromatograms were processed using software tools for mass signal extraction and alignment, and intensity-dependent accurate mass calculation. The detected masses were assigned by matching their accurate mass signals with tomato compounds reported in literature and complemented, as much as possible, by PDA and MS/MS information, as well as by using reference compounds. Several novel compounds not previously reported for tomato fruit were identified in this manner and added to the database. The MoTo DB is available at http://appliedbioinformatics.wur.nl and contains all information so far assembled using this LC-PDA-quadrupole time-of-flight MS platform, including retention times, calculated accurate masses, PDA spectra, MS/MS fragments, and literature references. Unbiased metabolic profiling and comparison of peel and flesh tissues from tomato fruits validated the applicability of the MoTo DB, revealing that all flavonoids and α-tomatine were specifically present in the peel, while several other alkaloids and some particular phenylpropanoids were mainly present in the flesh tissue.

For understanding the dynamic behavior of a complex biological system, it is essential to follow, as unbiased as possible, its response to a conditional perturbation at the transcriptome, proteome, and metabolome levels. To study the dynamics of the metabolome, to analyze fluxes in metabolic pathways, and to decipher the biological roles of metabolites, the identification of the participating metabolites should be as unambiguous as possible. Metabolomics is defined as the analysis of all metabolites in an organism and concerns the simultaneous (multiparallel) measurement of all metabolites in a given biological system (Dixon and Strack, 2003). However, this is a technically challenging task, as no single analytical method is capable of extracting and detecting all metabolites at once due to the enormous chemical variety of metabolites and the large range of concentrations at which metabolites can be present. Therefore, the characterization of a complete metabolome requires different complementary analytical technologies. Currently, mass spectrometry (MS) is the most sensitive method enabling the detection of hundreds of compounds within single extracts.

Ideally, metabolome data should be incorporated into open access databases where information can be viewed, sorted, and matched. Different pathway resources are available that combine information from the omics technologies such as the Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg), MetaCyc (http://metacyc.org), or The Arabidopsis Information Resource (http://www.arabidopsis.org). Hitherto, research on plant metabolic profiling using chromatographic techniques coupled to MS technologies for database purposes has been accomplished by gas chromatography (GC)-MS analysis of extracts (Schauer et al., 2005; Tikunov et al., 2005). GC-MS entails high reproducibility in both chromatography and mass fragmentation patterns. This reproducibility enabled the development of common metabolite databases, e.g. GMD@CSB.DB (http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html) and the Fiehn-Library (http://fiehnlab.ucdavis.edu/compounds), that gather information mainly on primary metabolites.

Liquid chromatography (LC)-MS is the preferred technique for the separation and detection of the large and often unique group of semipolar secondary metabolites in plants. Specifically, high resolution accurate mass MS enables the detection of large numbers of parent ions present in a single extract and can provide valuable information on the chemical composition and thus the putative identity of large numbers of metabolites. Recently, accurate mass LC-MS was performed to detect secondary metabolites present in roots and leaves of Arabidopsis (*Arabidopsis thaliana*; von Roepenack-Lahaye et al., 2004), to study metabolic alterations in a light-hypersensitive mutant of tomato (*Solanum lycopersicum*; Bino et al., 2005), and to compare tubers of potato (*Solanum tuberosum*) of different genetic origin and developmental stages (Vorst et al., 2005). The variety of LC-MS systems, and the generally poorer retention time reproducibility of LC compared to GC, limits the establishment of a single optimized analytical procedure and hampers the comparison of LC-MS chromatograms between laboratories. Moreover, software tools able to transform automatically MS data into a list of (putative) plant metabolites, in particular for LC-MS, are not yet available. This implies that analyses of mass signal datasets are left to manual searches in the available chemical databases such as SciFinder, Pub-Chem, or Dictionary of Natural Products. To extend the applicability of LC-MS in plant metabolomics, efforts should be made in (1) the establishment of a routine and reproducible LC-MS method, (2) the annotation of the large numbers of mass signals detected, (3) the unambiguous identification of compounds, and (4) the development of a common reference database and searching tools for secondary metabolites in plants.

In this article we present an open access metabolite database for LC-MS, called Metabolome Tomato Database (MoTo DB), dedicated to tomato fruit. This database is based on literature information combined with experimental data derived from LC-MS-based metabolomics experiments. A reproducible and robust C_{18}-based reversed-phase LC-photodiode array detection (PDA)-electrospray ionization (ESI)-quadrupole time-of-flight (QTOF)-MS method was developed for the detection and putative identification of predominantly secondary metabolites of semipolar nature. The assignment of mass signals detected relies on the combination of the parameters: (1) accurate mass, (2) retention time, (3) UV/Vis spectral information, and (4) MS/MS fragmentation data. To demonstrate the applicability of the established LC-MS metabolomics platform including database searching, peel and flesh tissues from ripe tomato fruit were compared for differences in metabolic composition. Statistically significant differences in LC-QTOF MS profiles between the tissues were identified in an unbiased manner, and differential mass peaks were annotated by searching in the MoTo DB. Several compounds not previously reported in tomato were also identified and have been incorporated into the database. All available information in the MoTo DB can be searched at http://appliedbioinformatics.wur.nl.

RESULTS

Metabolites Present in Tomato Fruit According to Literature

First, a database was constructed based on literature research to include metabolites reported to be present in tomato fruit from both wild and cultivated varieties as well as transgenic tomato plants. Though some tomato varieties are known to contain anthocyanins in their fruit (Jones et al., 2003), so far, to our knowledge, there are no reports on the identification of this class of compounds in fruit tissue. Therefore, in our literature search we included reports on anthocyanin identification in seedlings of tomato. Names (common and International Union of Pure and Applied Chemistry [IUPAC]), Chemical Abstracts Service (CAS) registry number, molecular formula, monoisotopic accurate mass, published references, and other properties of each metabolite are systematized in this database. The database includes polar, semipolar, and apolar compounds. Because the procedure used by us for extraction, separation, and detection (see below) is biased toward compounds of semipolar nature, we expected mostly secondary metabolites like (poly)phenols, alkaloids, and derivatives thereof to be detected. Table I summarizes all (poly)phenolic compounds (48) and alkaloids (15) so far reported to be present in tomato fruit extracts, including compounds that have been identified only in fruits of transgenic tomato plants. Many compounds were assigned before MS technologies became available. The number of compounds identified by NMR is very limited.

Metabolite Extraction and LC-PDA-MS Analysis

A representative tomato fruit sample was obtained by combining fruits of 96 different tomato cultivars producing ripe red, orange-colored beef, round, or cherry type of fruits at different stages of ripening (Tikunov et al., 2005). In addition, some purple-skinned fruits were selected for analyses of anthocyanins, which is a class of tomato fruit compounds only occurring in specific varieties (Jones et al., 2003) or in transgenic plants (Mathews et al., 2003). Peel material was chosen as the starting material, as this tissue contains the highest levels of flavonoids (Muir et al., 2001), which represent an important class of secondary metabolites. The 75% methanol/water extract enabled separation by C_{18}-reversed-phase LC and detection by both PDA and MS of semipolar metabolites. Figure 1 shows an example of a chromatogram obtained upon LC-PDA-QTOF-MS analysis of 75% methanol/water extracts from tomato peel. These extracts were stable for several months at −20°C, as determined by comparing LC-PDA chromatograms. Only naringenin chalcone was observed to decay slowly into naringenin while standing in the autosampler (20°C) during a series of analyses (about 1.4 μg g^{-1} fresh weight h^{-1}).

To test the reproducibility of the LC system, chromatograms of the tomato fruit material that have been

Table I. *List of secondary metabolites identified in tomato fruit extracts according to literature*

Mol Form, Molecular formula; MM, monoisotopic molecular mass.

Compound	Mol Form	MM	Reference
p-Hydroxybenzoic acid	$C_7H_6O_3$	138.0317	Mattila and Kumpulainen (2002)
Salicylic acid	$C_7H_6O_3$	138.0317	Schmidtlein and Herrmann (1975), Petró-Turza (1987)
Cinnamic acid	$C_9H_8O_2$	148.0524	Petró-Turza (1987)
Protocatechuic acid	$C_7H_6O_4$	154.0266	Mattila and Kumpulainen (2002)[a]
m-Coumaric acid	$C_9H_8O_3$	164.0474	Hunt and Baker (1980)[a]
p-Coumaric acid	$C_9H_8O_3$	164.0473	Schmidtlein and Herrmann (1975)[a], Hunt and Baker (1980)[a], Petró-Turza (1987), Martinez-Valverde et al. (2002), Mattila and Kumpulainen (2002), Raffo et al. (2002), Le Gall et al. (2003a)[bc]
Vanillic acid	$C_8H_8O_4$	168.0423	Schmidtlein and Herrmann (1975), Mattila and Kumpulainen (2002)
Caffeic acid	$C_9H_8O_4$	180.0423	Schmidtlein and Herrmann (1975)[a], Hunt and Baker (1980)[a], Martinez-Valverde et al. (2002), Mattila and Kumpulainen (2002), Raffo et al. (2002), Sakakibara et al. (2003), Minoggio et al. (2003), Le Gall et al. (2003a)[bc]
Ferulic acid	$C_{10}H_{10}O_4$	194.0579	Schmidtlein and Herrmann (1975)[a], Hunt and Baker (1980)[a], Martinez-Valverde et al. (2002), Mattila and Kumpulainen (2002), Raffo et al. (2002), Minoggio et al. (2003)
Sinapic acid	$C_{11}H_{12}O_5$	224.0685	Schmidtlein and Herrmann (1975)[a]
Naringenin	$C_{15}H_{12}O_5$	272.0685	(Hunt and Baker, 1980)[a]; (Justesen et al., 1998)[a], (Martinez-Valverde et al., 2002)[a], (Raffo et al., 2002), (Minoggio et al., 2003)
Naringenin chalcone	$C_{15}H_{12}O_5$	272.0685	Hunt and Baker (1980)[a], Krause and Galensa (1992), Muir et al. (2001), Le Gall et al. (2003b)[b], Minoggio et al. (2003)
Kaempferol	$C_{15}H_{10}O_6$	286.0477	Stewart et al. (2000), Martinez-Valverde et al. (2002)[a], Tokusoglu et al. (2003)[a]
Quercetin	$C_{15}H_{10}O_7$	302.0427	Hertog et al. (1992), Crozier et al. (1997)[a], Justesen et al. (1998)[a], Stewart et al. (2000), Martinez-Valverde et al. (2002)[a], Raffo et al. (2002), Sakakibara et al. (2003), Tokusoglu et al. (2003)[a]
Myricetin	$C_{15}H_{10}O_8$	318.0376	Raffo et al. (2002), Sakakibara et al. (2003), Tokusoglu et al. (2003)[a]
p-Coumaric acid-*O*-*β*-D-glucoside	$C_{15}H_{18}O_8$	326.1002	Fleuriet and Macheix (1977), Reschke and Herrmann (1982)[a], Winter and Herrmann (1986)[c], Buta and Spaulding (1997)
p-Coumaroylquinic acid	$C_{16}H_{18}O_8$	338.1002	Fleuriet and Macheix (1977)
Caffeic acid-4-*O*-*β*-D-glucoside	$C_{15}H_{18}O_9$	342.0951	Fleuriet and Macheix (1977), Winter and Herrmann (1986)
Chlorogenic acid (3-*O*-caffeoylquinic acid)	$C_{16}H_{18}O_9$	354.0951	Fleuriet and Macheix (1977), Fleuriet and Macheix (1981), Winter and Herrmann (1986), Buta and Spaulding (1997), Martinez-Valverde et al. (2002), Mattila and Kumpulainen (2002), Raffo et al. (2002), Sakakibara et al. (2003), Minoggio et al. (2003), Le Gall et al. (2003a, 2003b)[bc]
4-*O*-Caffeoylquinic acid	$C_{16}H_{18}O_9$	354.0951	Winter and Herrmann (1986), Mattila and Kumpulainen (2002)
5-*O*-Caffeoylquinic acid	$C_{16}H_{18}O_9$	354.0951	Winter and Herrmann (1986)
Ferulic acid-*O*-*β*-D-glucoside	$C_{16}H_{20}O_9$	356.1107	Fleuriet and Macheix (1977), Reschke and Herrmann (1982), Winter and Herrmann (1986)
Feruloylquinic acid	$C_{17}H_{20}O_9$	368.1107	Fleuriet and Macheix (1977)
Tomatidine	$C_{27}H_{45}NO_2$	415.3450	Juvik et al. (1982),[a] Friedman et al. (1998)[a]
Tomatidenol	$C_{27}H_{43}NO_2$	413.3294	Juvik et al. (1982)[a], Friedman et al. (1994)[a], Friedman et al. (1997)[a], Friedman (2002)[a]
Naringenin-7-*O*-glucoside	$C_{21}H_{22}O_{10}$	434.1213	Hunt and Baker (1980), Le Gall et al. (2003a, 2003b)[bc]
Naringenin chalcone-glucoside	$C_{21}H_{22}O_{10}$	434.1213	Bino et al. (2005)
Astragalin	$C_{21}H_{20}O_{11}$	448.1006	Le Gall et al. (2003a, 2003b)[bc]
Dihydrokaempferol-7-*O*-hexoside and Dihydrokaempferol-?-*O*-hexoside	$C_{21}H_{22}O_{11}$	450.1162	Le Gall et al. (2003a, 2003b)[bc]
Isoquercitrin	$C_{21}H_{20}O_{12}$	464.0955	Muir et al. (2001)[b], Le Gall et al. (2003a, 2003b)[b]
Myricitrin	$C_{21}H_{20}O_{12}$	464.0955	Sakakibara et al. (2003)
Naringin	$C_{27}H_{32}O_{14}$	580.1792	Bovy et al. (2002)[abd]
Kaempferol-3-*O*-rutinoside	$C_{27}H_{30}O_{15}$	594.1585	Bovy et al. (2002)[bd], Le Gall et al. (2003b)[bc]
Kaempferol-3-7-di-*O*-glucoside	$C_{27}H_{30}O_{16}$	610.1534	Le Gall et al. (2003a, 2003b)[bc]
Rutin	$C_{27}H_{30}O_{16}$	610.1534	Fleuriet and Macheix (1977), Buta and Spaulding (1997), Stewart et al. (2000), Muir et al. (2001), Raffo et al. (2002); Le Gall et al. (2003a, 2003b)[bc], Minoggio et al. (2003)
Quercetin-3-*O*-trisaccharide	$C_{32}H_{38}O_{20}$	742.1956	Muir et al. (2001), Minoggio et al. (2003)
p-Coumaric acid-rutin conjugate	$C_{36}H_{36}O_{18}$	756.1902	Buta and Spaulding (1997)

(*Table continues on following page.*)

Table I. (*Continued from previous page*)

Compound	Mol Form	MM	Reference
Kaempferol-3-O-rutinoside-7-O-glucoside	$C_{33}H_{40}O_{20}$	756.2113	Le Gall et al. (2003a, 2003b)[bc]
Delphinidin-3-O-rutinoside-5-O-glucoside	$C_{33}H_{41}O_{21}^{+}$	773.2135	Mathews et al. (2003)[bd]
Petunidin-3-O-rutinoside-5-O-glucoside	$C_{34}H_{43}O_{21}^{+}$	787.2291	Mathews et al. (2003)[bd]
Malvidin-3-O-rutinoside-5-O-glucoside	$C_{35}H_{45}O_{21}^{+}$	801.2448	Mathews et al. (2003)[bd]
Delphinidin-3-O-(p-coumaroyl)rutinoside-5-O-glucoside	$C_{42}H_{47}O_{23}^{+}$	919.2503	Mathews et al. (2003)[bd]
Petunidin-3-O-(p-coumaroyl)-rutinoside-5-O-glucoside	$C_{43}H_{49}O_{23}^{+}$	933.2659	Bovy et al. (2002)[bd], Mathews et al. (2003)[bd]
Delphinidin-3-O-(caffeoyl)-rutinoside-5-O-glucoside	$C_{42}H_{47}O_{24}^{+}$	935.2452	Mathews et al. (2003)[bd]
Malvidin-3-O-(p-coumaroyl)-rutinoside-5-O-glucoside	$C_{44}H_{51}O_{23}^{+}$	947.2816	Bovy et al. (2002)[bd], Mathews et al. (2003)[bd]
Petunidin-3-(caffeoyl)-rutinoside-5-O-glucoside	$C_{43}H_{49}O_{24}^{+}$	949.2608	Bovy et al. (2002)[bd], Mathews et al. (2003)[bd]
Malvidin-3-(caffeoyl)-rutinoside-5-O-glucoside	$C_{44}H_{51}O_{24}^{+}$	963.2765	Mathews et al. (2003)[bd]
δ-Tomatine	$C_{33}H_{55}NO_{7}$	577.3979	Friedman et al. (1998)[a]
γ-Tomatine	$C_{39}H_{65}NO_{12}$	739.4507	Friedman et al. (1998)[a]
β-Tomatine	$C_{45}H_{75}NO_{17}$	901.5035	Friedman et al. (1998)[a]
Dehydrotomatine	$C_{50}H_{81}NO_{21}$	1,031.5301	Friedman et al. (1994), Kozukue and Friedman (2003)
α-Tomatine	$C_{50}H_{83}NO_{21}$	1,033.5458	Juvik et al. (1982), Willker and Leibfritz (1992)[c], Friedman et al. (1994), Yahara et al. (1996), Friedman et al. (1997), Friedman et al. (1998), Friedman (2002), Bianco et al. (2002), Kozukue and Friedman (2003)
Lycoperoside H	$C_{50}H_{83}NO_{22}$	1,049.5407	Yahara et al. (1996)[c], Yahara et al. (2004)[c]
Lycoperoside A	$C_{52}H_{85}NO_{23}$	1,091.5512	Yahara et al. (1996, 2004)[c]
Lycoperoside B	$C_{52}H_{85}NO_{23}$	1,091.5512	Yahara et al. (1996, 2004)[c]
Lycoperoside C	$C_{52}H_{85}NO_{23}$	1,091.5512	Yahara et al. (1996, 2004)[c]
Esculeoside B	$C_{56}H_{93}NO_{28}$	1,227.5884	Fujiwara et al. (2004)[c], Yahara et al. (2004)[c]
Esculeoside A	$C_{58}H_{95}NO_{29}$	1,269.5990	Fujiwara et al. (2003, 2004)[c], Yahara et al. (2004)[c], Yoshizaki et al. (2005)[c]
Lycoperoside F	$C_{58}H_{95}NO_{29}$	1,269.5990	Yahara et al. (2004)[c]
Lycoperoside G	$C_{58}H_{95}NO_{29}$	1,269.5990	Yahara et al. (2004)[c]

[a] Identified after hydrolysis.   [b] Identified in transgenic tomato plants.   [c] Identified using NMR data.   [d] Identified in seedlings.

analyzed over a period of 2 years (>100 samples) were manually compared for retention time shifts using some typical tomato compounds (Table II). Within a single series of analyses, the standard variation was very small (about 2 s) for all compounds tested. Between series of analyses over this time period, the maximum variation was 30 s, with a maximum retention time window of 1.1 min for naringenin chalcone. During this prolonged period, LC columns of different batches were used.

Comparison of Ionization Modes

Since compounds may preferentially ionize in either positive or negative mode in our LC system, which is based on a gradient of acetonitrile acidified with formic acid, we analyzed tomato extracts sequentially in both modes and compared the absolute mass signal intensities, expressed in peak heights, of the monoisotopic parent ions of some identified compounds. Phenolic acids and their carboxylic acid derivatives ionized better in negative ionization mode, while flavonoids gen-

erated higher signal intensities in positive ionization mode (Fig. 2). Nitrogen-containing compounds such as Phe and some alkaloids ionized better in positive mode, and were mainly detected as formic acid adducts in negative mode. These adducts were formed in the ionization source and were readily recognized in MS/MS mode from the loss of 46 D (formic acid). A loss of 18 D corresponding to a loss of water was also regularly observed in negative ionization mode.

Automatic Mass Alignment and Exact Mass Calculation

First, reproducibility of sample preparation and subsequent automated extraction and comparison of mass signal intensities, expressed as peak height using metAlign software (Bino et al., 2005; Vorst et al., 2005), was performed on a dataset obtained from LC-MS analysis of eight replicate extractions of tomato peel. The retention time correction used by the software to align all mass signals was, on average, 2.5 s, which is in accordance to the retention shift observed on manual inspection of the chromatograms (Table II). The overall
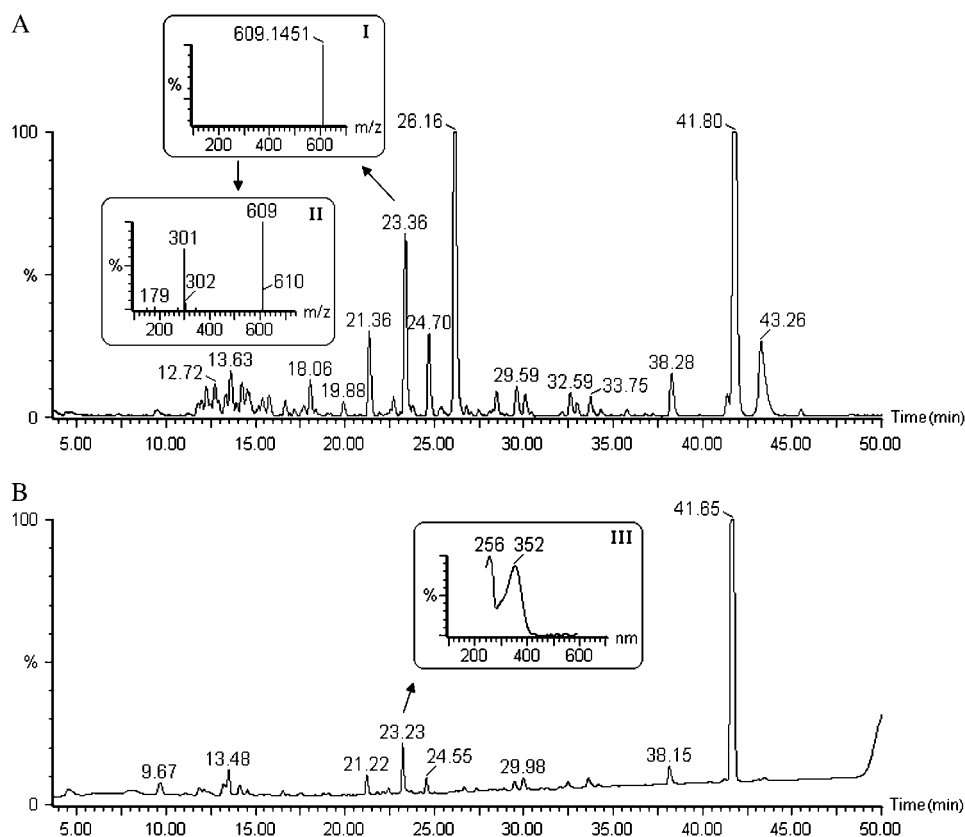
Figure 1. Typical chromatograms obtained from reversed-phase LC-PDA-ESI-QTOF-MS analysis of tomato peel extract. A, Total ion signal (QTOF MS). B, Absorbance signal (PDA). Retention times (in minutes) are indicated for the most intense peaks (difference between the two detectors is 0.15 min). Inserts in A show accurate mass (I) and MS/MS spectrum (II), and in B absorbance spectrum (III) obtained for the compound rutin eluting at 23.3 min.

variation in mass signal intensities between these replicate samples was <15%.

Automation of the calculation of the accurate mass of detected LC-MS signals was tested using a dataset of 44 tomato extracts obtained from both peel and flesh tissues analyzed in negative ionization mode. Upon metAlign-assisted data processing, 4,958 mass signals with signal-to-noise ratios >3 were extracted. It is known that exact mass measurements on QTOF instruments using lock mass correction provide the highest accuracy at analyte signal intensities that are similar to the lock mass signal (Colombo et al., 2004). To establish the dynamic range in signal intensity for producing high mass accuracy in our TOF MS, the deviation of manually measured mass (i.e. the mean of the three top scans of the extracted mass peak) from the theoretical mass was plotted against the parent mass signal intensity (ion counts at top scan) for some known tomato metabolites (Fig. 3). Typically, accurate mass measurements derived from peak intensities lower than the lock mass intensity resulted in a positive deviation from the real mass, while mass measurements from peak intensities higher than lock mass intensity resulted in a negative deviation. High mass accuracies (i.e. mass deviation less than 5 ppm) were observed within an analyte signal intensity window of 0.25 to 2.0 times the lock mass. Thus, to automatically calculate correct accurate masses for signals extracted and aligned by metAlign, a script called metAccure (O. Vorst, H.A. Verhoeven, C.H.R. de Vos, C.A. Maliepaard, and R.C.H.J. van Ham, unpublished data) was programmed to use only those scans with mass signal intensities within this intensity window. In this way, appropriate accurate masses were automatically obtained for 479 (about 10%) of the total mass signals detected in ESI-negative mode, in which isotopes, adducts, and fragments are included. This number indicates that for the majority of extracted mass signals, though having a chromatographically relevant signal-to-noise ratio of at least 3, the intensities in the

Table II. *Retention time shifts observed during LC-QTOF-MS analysis of tomato fruit*

Ret (min), Retention time, in minutes; Av, average; StDev, standard deviation; Wd, retention time window.

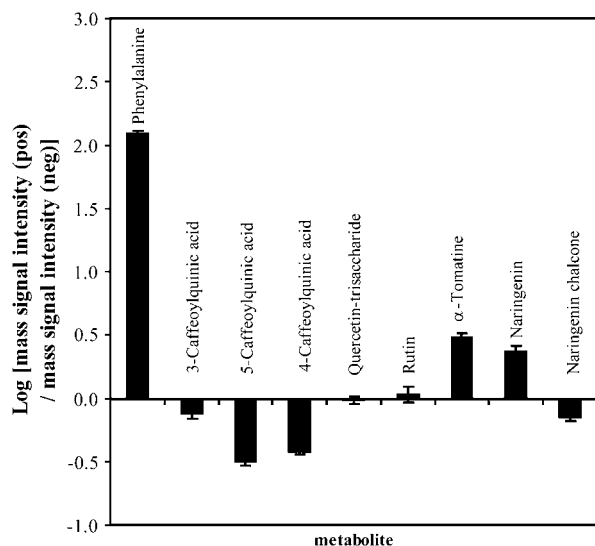| Ret Metabolite	Chlorogenic Acid			Rutin			Naringenin Chalcone		
	Av	StDev	Wd	Av	StDev	Wd	Av	StDev	Wd
min									
Within series (*n* = 13)	14.42	0.03	0.09	23.40	0.04	0.13	41.81	0.03	0.11
In-between series (*n* = 6)	14.92	0.33	0.79	23.85	0.50	0.99	42.26	0.50	1.12

Figure 2. Peak intensity ratios, in logarithmic scale, of mass signals (peak height) obtained in positive and negative ionization modes for some metabolites found in tomato peel extracts.

samples analyzed were too low to estimate properly their accurate mass, either by automated calculation through metAccure or by manual calculation.

Identification of Tomato Metabolites

The identification of compounds reported to be present in tomato fruit was done using two approaches. First, 19 available standard compounds (see "Materials and Methods") were injected and compared for retention time, accurate mass, and UV/Vis spectra with LC peaks detected in the extracts from the pooled peel material of the 96 tomato cultivars. In this way, chlorogenic acid (i.e. 3-caffeoylquinic acid), rutin, kaempferol-rutinoside, naringenin, naringenin chalcone, and α-tomatine were identified. Second, the chromatograms from the 44 LC-MS data sets were checked for the presence of accurate masses, as calculated by metAccure, corresponding to metabolites that were expected to be detected with our system (Table I). The accurate mass hits were subsequently combined with PDA and MS/MS fragmentation data for further identification and confirmation of metabolites. As an example, data of known tomato metabolites observed in extracts of the pooled peel material of the 96 tomato cultivars, derived by LC-PDA-MS and MS/MS analyses in negative mode, are listed in Table III. In an analogous way, the presence of anthocyanins was confirmed by LC-PDA-QTOF-MS/MS analysis (positive mode) in peel extracts from purple-skin tomato fruits (data not shown). Using this primarily accurate mass-directed targeted approach, about 41% (25 compounds) of the metabolites cited in Table I were identified in both tomato peel samples. In addition, caffeic acid, ferulic acid, p-coumaric acid, quercetin, and kaempferol aglycones could be detected but only

after acid hydrolysis of the extract. All experimental LC-MS information gathered for these metabolites, including retention time window, accurate mass, PDA spectral information, and MS/MS data generated at different collision energies were added to the MoTo DB.

Database Building

The data from Table I were used as a foundation upon which to initiate the tomato fruit LC-MS database. From the molecular formula, the accurate mass of each component was calculated using the "Isotopic compositions of the elements 1997" list (Rosman and Taylor, 1998) for accurate mass assignments. The observed mass, together with a mass accuracy setting, is the main search entry for this database (Fig. 4). A choice on the entry form is provided to enable ionization-specific correction of mass spectrometer data, to submit the proper mass value of the uncharged molecule to the database. Mass accuracy can be set from 1 to 1,000 ppm, thus enabling the matching of data from detectors generating masses with either low or high accuracy. All other properties of the compounds are stored in a table, which can be accessed from the hit list after mass searching. Each hit suggests either a metabolite previously found in literature and validated by experimental data (Table III) or a novel
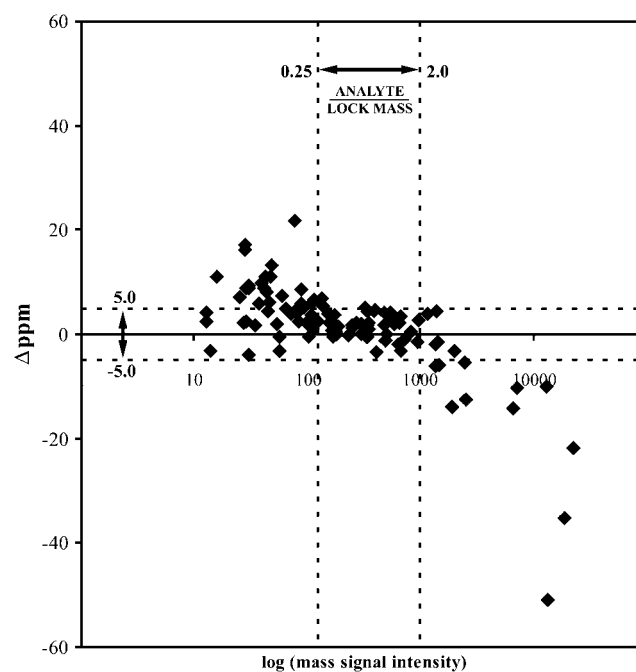

Figure 3. Difference between observed and theoretical monoisotopic masses, calculated as Δppm (y axis), as a function of the parent ion signal intensity, expressed as ion counts/scan at center of peak (x axis, $\log_{10}$-transformed data) for some identified compounds in tomato peel extracts. Threshold levels for mass accuracies between $+5$ and -5 ppm, and for analyte mass signal intensities between 0.25 and 2.0 times the lock mass signal intensity are indicated with dotted lines.

Table III. *Metabolites that have previously been reported in literature, identified by LC-PDA-ESI-QTOF-MS/MS (negative ionization mode) in tomato peel extracts*

Ret (min), Retention time, in minutes; Av, average; StDev, standard deviation; Av m/z, average found mass signal; UV/Vis, absorbance maximums in the UV/Vis range; Mol Form, molecular formula of the metabolite; Theo. Mass, theoretical monoisotopic mass calculated for the ion (M-H)⁻; Mean Δ (ppm), deviation between the averages of found accurate mass and real accurate mass, in ppm; Putative ID, putative identification of metabolite; () FA, formic acid adduct; −, data not found; (S), identification confirmed by the standard compound; I, II, III, IV, V, and VI, different isomers (only one reported in literature).

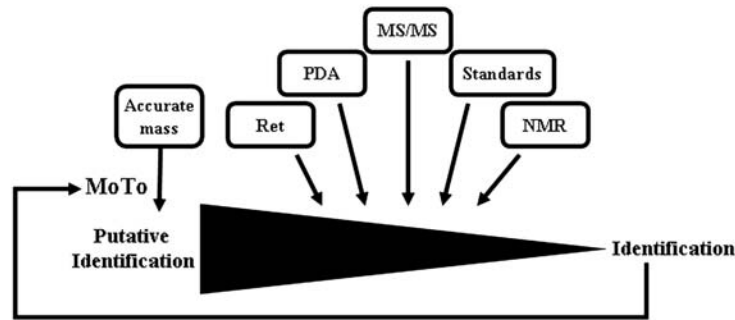| Ret Av	Ret StDev	Av m/z	UV/Vis	MS/MS Fragments	Mol Form	Theo. Mass	Mean Δ	Putative ID
min							*ppm*	
9.45	0.09	341.0883	–	179, 135	$C_{15}H_{18}O_9$	341.0878	1.52	Caffeic acid-hexose I
9.75	0.08	325.0930	294sh, 313	163	$C_{15}H_{18}O_8$	325.0929	0.25	Coumaric acid-hexose I
10.32	0.08	341.0883	310	179, 161, 135	$C_{15}H_{18}O_9$	341.0878	1.58	Caffeic acid-hexose II
11.35	0.08	341.0883	302sh, 318	281, 251, 233, 221, 179, 161, 135	$C_{15}H_{18}O_9$	341.0878	1.53	Caffeic acid-hexose III
12.08	0.06	355.1036	290sh, 313	193, 177, 145	$C_{16}H_{20}O_9$	355.1035	0.31	Ferulic acid-hexose I
12.58	0.07	341.0883	–	181, 179, 137, 135	$C_{15}H_{18}O_9$	341.0878	1.49	Caffeic acid-hexose IV
13.32	0.05	341.0883	–	281, 221, 181, 179, 161, 137, 135	$C_{15}H_{18}O_9$	341.0878	1.39	Caffeic acid-hexose V
13.43	0.07	353.0878	300sh, 327	191, 173, 127	$C_{16}H_{18}O_9$	353.0878	0.01	3-Caffeoylquinic acid
13.71	0.07	325.0929	285	163, 119	$C_{15}H_{18}O_8$	325.0929	0.05	Coumaric acid-hexose II
14.41	0.10	353.0878	295sh, 327	179, 173	$C_{16}H_{18}O_9$	353.0878	−0.08	5-Caffeoylquinic acid (S)
15.90	0.05	355.1036	–	193, 175, 160	$C_{16}H_{20}O_9$	355.1035	0.42	Ferulic acid-hexose II
15.98	0.06	341.0886	–	179	$C_{15}H_{18}O_9$	341.0878	2.26	Caffeic acid-hexose VI
16.76	0.07	353.0880	323	191, 173, 161, 127	$C_{16}H_{18}O_9$	353.0878	0.49	4-Caffeoylquinic acid
19.53	0.25	1,272.5901	–	1,227, 1,095, 1,065, 933, 866, 770	$C_{57}H_{95}NO_{30}$	1,272.5866	2.75	(Esculeoside B) FA
21.42	0.04	741.1870	256, 299sh, 351	301, 271, 255	$C_{32}H_{38}O_{20}$	741.1884	−1.82	Quercetin-hexose-deoxyhexose-pentose
22.83	0.06	1,314.6001	–	1,269, 1,137, 1,107, 974, 770, 752	$C_{59}H_{97}NO_{31}$	1,314.5972	2.21	(Lycoperoside G) FA or (Lycoperoside F) FA or (Esculeoside A) FA I
23.43	0.04	609.1451	256, 299sh, 355	301, 271, 255	$C_{27}H_{30}O_{16}$	609.1461	−1.59	Quercetin-Glc-rhamnose (S)
25.48	0.16	1,314.6005	–	1,269, 1,137, 1,107, 975, 908, 866, 812, 770, 752, 275, 179, 161, 149, 143, 125, 113	$C_{59}H_{97}NO_{31}$	1,314.5972	2.54	(Lycoperoside G) FA or (Lycoperoside F) FA or (Esculeoside A) FA II
26.37	0.21	1,314.6021	–	1,270, 1,138, 1,108, 976, 909, 813, 753, 179, 161, 143, 125, 113	$C_{59}H_{97}NO_{31}$	1,314.5972	3.74	(Lycoperoside G) FA or (Lycoperoside F) FA or (Esculeoside A) FA III
26.41	0.03	593.1505	368	285	$C_{27}H_{30}O_{15}$	593.1512	−1.09	Kaempferol-Glc-rhamnose (S)
26.44	0.39	1,094.5382	–	1,049	$C_{51}H_{85}NO_{24}$	1,094.5389	−0.59	(Lycoperoside H) FA
32.46	0.37	1,078.5463	–	1,033, 871, 738, 576, 161, 143	$C_{51}H_{85}NO_{23}$	1,078.5440	2.14	(α-Tomatine) FA (S)
32.59	0.22	1,136.5539	–	1,091, 958, 928, 796, 635, 149, 143, 113	$C_{53}H_{87}NO_{25}$	1,136.5494	3.91	(Lycoperoside C) FA or (Lycoperoside B) FA or (Lycoperoside A) FA₃
32.65	0.02	433.1135	315sh, 368	271, 151	$C_{21}H_{22}O_{10}$	433.1140	−1.21	Naringenin chalcone-hexose I
41.43	0.05	271.0617	288, 303sh	151,119,107	$C_{15}H_{12}O_5$	271.0612	1.84	Naringenin (S)
41.86	0.05	271.0615	365	151, 119, 107	$C_{15}H_{12}O_5$	271.0612	1.15	Naringenin chalcone (S)

compound (Table IV). Links with the PubChem and MedLine databases are available for extended, external searches on particular or related components. The information for each compound includes molecular formula, molecular mass, CAS number, IUPAC name, and analytical properties such as retention time, MS/MS fragments, and UV/Vis absorbance maxima, when available. Literature references related to the occurrence in tomato fruit are also listed. Since our aim is to provide a compound database with data from literature and/or experimental MS/MS data, we did not include unknown or novel compounds that have not been validated.

Comparison of Metabolic Profiles of Peel and Flesh Tissues

The applicability of the LC-MS platform and metabolite database to automatically extract and annotate (differentially accumulating) mass signals was tested with red, ripe fruits of tomato cultivar Money Maker. Since we are interested in the differential distribution

Figure 4. A, Strategy applied for data analysis and identification of metabolites in tomato fruit, using LC-PDA-QTOF MS. Key entry into the database is the (intensity-corrected) accurate mass. B, Screenshot from the MoTo database query frame. Detected masses can be filled in (in this example *m/z* 609 in negative-ionization mode) and searched against the database at user-defined mass accuracy (first frame). If at least one mass hit is found in the database, the elemental compositions, deviations from accurate masses, and IUPAC names of the corresponding metabolites are indicated, as well as links to PubChem, if applicable, and our own experimental data (second frame). The last frame shows the experimental and literature information available for the selected compound.

of metabolites and their biochemical pathways between tomato fruit tissues, peel and flesh material was separated from whole ripe fruits and analyzed by LC-PDA-ESI-QTOF-MS in both positive and negative ion modes.

After automatic peak extraction and alignment of samples per ionization mode using metAlign, 2,944 mass signals (signal-to-noise ratio >3) were obtained in negative mode and 4,059 in positive mode. Since both tissues had similar water content (i.e. flesh: 94%, peel: 93%; $n = 8$; determined by freeze drying), the

intensities of their mass signals were directly comparable. For each aligned mass peak, the extracts from both tissues were compared for significant differences in signal intensity (based on eight extraction repetitions) using the Student's *t* test tool within metAlign. As expected, the mass profiles of these fruit tissues were markedly different. About 38% of the total of mass signals detected were significantly ≥1.5-fold higher in the peel extracts than in the flesh extracts (1,095 signals for negative mode and 1,566 for positive mode), and about 25% were higher in flesh than in peel

Table IV. *Novel metabolites identified or putatively assigned by LC-PDA-ESI-QTOF-MS/MS in tomato fruit extracts (abbreviations as in Table III)*

Ret		Av *m/z*	UV/Vis	MS/MS Fragments	Mol Form	Theo. Mass	Mean Δ	Putative ID
Av	StDev							
min							*ppm*	
4.74	0.05	299.0771	251	137	$C_{13}H_{16}O_8$	299.0772	−0.48	Hydroxybenzoic acid-hexose
7.42	0.07	380.1558	–	146	$C_{15}H_{27}NO_{10}$	380.1562	−1.11	Pantothenic acid-hexose
12.99	0.05	431.1557	–	269, 161, 143, 125, 119, 113, 101	$C_{19}H_{28}O_{11}$	431.1559	−0.43	Benzyl alcohol-dihexose
14.76	0.05	771.1989	263sh, 351	609, 463, 301	$C_{33}H_{40}O_{21}$	771.1989	−0.01	Quercetin-dihexose-deoxyhexose
15.47	0.06	595.1665	–	475, 385, 355	$C_{27}H_{32}O_{15}$	595.1668	−0.51	Naringenin chalcone-dihexose or Naringenin-dihexose
15.82	0.04	401.1452	–	293, 269, 233, 191, 161, 149, 131, 125, 101	$C_{18}H_{26}O_{10}$	401.1453	−0.37	Benzyl alcohol-hexose-pentose
24.77	0.15	1,312.5872	–	1,266, 1,135, 1,105	$C_{59}H_{95}NO_{31}$	1,312.5815	4.33	(Dehydrolycoperoside G) FA or (Dehydrolycoperoside F) FA or (Dehydroesculeoside A) FA
27.05	0.12	515.1193	301sh, 323	353, 335, 191, 179, 173	$C_{25}H_{24}O_{12}$	515.1195	−0.45	Dicaffeoylquinic acid I
27.60	0.07	515.1191	301sh, 323	353, 191, 179	$C_{25}H_{24}O_{12}$	515.1195	−0.72	Dicaffeoylquinic acid II
29.71	0.07	515.1188	301sh, 327	353, 299, 203, 191, 179, 173, 135	$C_{25}H_{24}O_{12}$	515.1195	−1.40	Dicaffeoylquinic acid III
30.11	0.04	887.2246	256, 301sh, 323	741, 723, 301, 271, 255, 179	$C_{41}H_{44}O_{22}$	887.2251	−0.57	Quercetin-hexose-deoxyhexose-pentose-*p*-coumaric acid
32.16	0.03	433.1137	307sh, 360	271, 151	$C_{21}H_{22}O_{10}$	433.1140	−0.84	Naringenin chalcone-hexose II
38.40	0.08	677.1503	301sh, 327	515	$C_{34}H_{30}O_{15}$	677.1512	−1.29	Tricaffeoylquinic acid I
39.78	0.11	677.1493	292sh, 325	515, 353, 335, 179, 173	$C_{34}H_{30}O_{15}$	677.1512	−2.82	Tricaffeoylquinic acid II

(794 for negative mode and 880 for positive mode). Chromatographic mass peaks detected in negative ionization mode that were significantly different between the extracts from both tissues are visualized in Figure 5. Subsequent metAccure-assisted accurate mass calculation of the differential mass peaks and searching for analogous masses in the MoTo DB indicated that flavonoids and derivatives thereof and α-tomatine were mainly occurring in the peel extracts. On the other hand, some phenylpropanoids (h, 52-fold; i, 2-fold) as well as glycosylated steroids such as glycosylated spirosolanols (j, 130-fold) were significantly higher in the flesh extracts. An intense mass signal, k, was solely detected in the extracts from flesh tissue and could be identified as the parent ion of a hydroxyfurostanol tetrahexose (e.g. tomatoside A) from the accurate mass observed ([M-H]$^-$ = 1,081.5442, $C_{51}H_{85}O_{24}{}^-$, 1.0 ppm difference from theoretical mass) and its MS/MS fragmentation pattern.

DISCUSSION

Metabolomics is developing as an important functional genomics tool. Technical improvements in the large-scale determination of metabolites in complex plant tissues and dissemination of metabolomics research data are essential (Sumner et al., 2003; Bino et al.,

2004). A major challenge is to construct consolidated metabolite libraries and to develop metabolite-specific data management systems. Here we set out to establish a reproducible LC-PDA-MS-based metabolomics platform including a LC-MS metabolite database and mass-directed searching tools for a commonly used plant material, i.e. tomato fruit.

An in-depth literature study was performed to obtain as much information as possible on metabolites previously detected in tomato fruits. Because tomato is an important crop, numerous analytical studies aimed at identifying its constituents have been performed. However, a number of problems arise when building such a database from the literature. First, finding the exact identity of a specific natural compound can be troublesome since common names or non-IUPAC nomenclatures are often used. Second, studies performed without MS or NMR technologies might lead to questioning the validity of at least some of the assigned compounds. Third, it is known that using harsh conditions during sample preparation may produce artifacts, which can result in the correct identification, but of a compound not occurring in the original biological sample. For instance, it has long been thought that the flavanone naringenin instead of naringenin chalcone was the main tomato flavonoid (Krause and Galensa, 1992). This is probably due to unforeseen cyclization of

Figure 5. Unbiased LC-QTOF MS-based comparative profiling of aqueous-methanol extracts from peel and flesh tissues from ripe tomato fruit (var. Money-maker). Mass chromatograms (*m/z* 100–1,500) were acquired in ESI-negative mode. Retention times (in minutes) and nominal masses of the most intense signals are indicated in the chromatograms (plotted as base peak intensities [BPI], from 4–50 min). A, Representative original chromatogram of peel tissue. B, Representative original chromatogram of flesh tissue. C, Differential chromatogram for metabolites that are significantly ($P < 0.05$; $n = 8$ extracts) at least 1.5-fold higher in extracts from peel compared to flesh tissue (peaks pointing upwards) or higher in extracts from flesh compared to peel tissue (peaks pointing downwards). a, Coumaric acid-hexose II; b, quercetin-hexose-deoxyhexose-pentose; c, rutin; d, kaempferol-hexose-deoxyhexose-pentose or querce-tin-dideoxyhexose-pentose; e, α-tomatine; f, narin-genin; g, naringenin chalcone; h, caffeic acid-hexose II; i, 3-caffeoylquinic acid; j, spirosolanol-trihexose; and k, hydroxyfurastanol tetrahexoside.
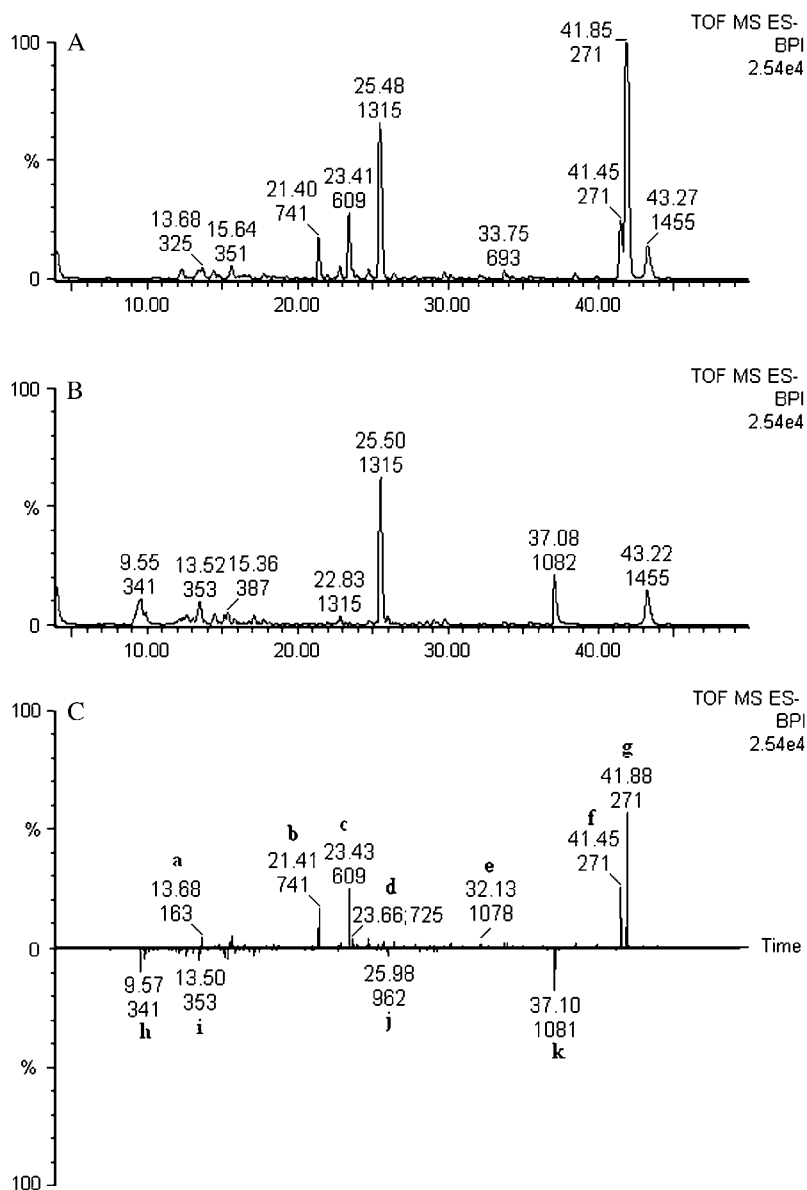
the chalcone to the corresponding flavanone during sample preparation and compound isolation. Likewise, some of the metabolites reported in literature have been identified after an enzymatic or chemical hydrolysis step. In the nonhydrolyzed tomato peel extract we exclusively found a range of glycosylated forms of caffeic acid, coumaric acids, and the flavonols quercetin and kaempferol, while the corresponding aglycones were only detectable after acid hydrolysis of the same sample.

The amount of information obtained by a single LC-QTOF MS analysis can be extensive and the use of dedicated software for data processing and comparison is crucial. The extraction of relevant mass signals and the subsequent alignment of chromatograms were performed using metAlign (Vorst et al., 2005). An average of 2 s variation within series of analyses and

30 s between analyses over a 2-year time period is an indication of high chromatographic reproducibility. These retention time shifts are sufficiently low to align correctly and thus compare samples when analyzed under the same chromatographic conditions. Variation in metabolite retention is a known and common obstacle in LC and thus important to take into account when searching LC-MS-based databases for comparable masses. Representative retention times and retention indexes of unknown mass peaks relative to tomato key compounds, such as rutin, chlorogenic acid, and naringenin, can be of use when comparing data generated by different LC systems or with a different type of C_{18}-reversed-phase column.

MetAccure (O. Vorst, H.A. Verhoeven, C.H.R. de Vos, C.A. Maliepaard, and R.C.H.J. van Ham, unpublished data) is an important tool for automated accurate mass

calculation of all aligned mass signals from the metAlign output. Within a specific range of mass signal intensities (depending on the specificities of the TOF MS and lock mass intensity used), the metAccure-assisted accurate mass calculations enabled the assignment of compounds. By calculating the average of all detected accurate masses of a certain aligned mass peak over all samples analyzed (taking into account only those scans with the correct range of ion intensities), high mass accuracies were obtained, i.e., frequently within 1 ppm and, in all cases, within 4 ppm deviation from the predicted mass (Table III). Apparently, this high mass accuracy was consistent over the entire mass range analyzed (mass-to-charge ratio [m/z] 100–1,500; accuracies better than 3 ppm were obtained for metabolites at both low [e.g. 271.0615 for naringenin chalcone] and high m/z values [e.g. 1,314.6005 for the formic acid adducts of the possible isomers lycoperoside G or F or esculeoside A]. With the QTOF instrument used, the metAccure script was able to generate appropriate accurate masses for about 10% of the total mass peaks detected in ESI-negative mode. Evidently, this percentage is highly dependent on the dynamic range of accurate mass measurements of the mass spectrometer used, as well as on the concentrations of each metabolite in the samples analyzed. By changing the lock mass-to-analyte ratio in successive analyses of the same sample it should be possible, in principle, to obtain accurate mass data for a wider range of amplitudes, leading to an expansion of the dynamic range.

The identification of compounds, in particular secondary metabolites, through a metabolomic profiling approach encounters some major difficulties. First, the number of commercially available standards of secondary metabolites reported to be present in a specific plant species or tissue is low. Second, in an automated online separation, PDA detection, MS measurement, and/or MS/MS fragmentation of mass signals, it is difficult to meet optimized levels for all eluting compounds. Due to overlapping compounds, low intensity mass signals, or difficulties in the isolation of the mass signal for MS/MS fragmentation, the extraction of usable information for identification purposes can be complicated. Third, the lack of dedicated software and databases that integrate spectroscopic and MS data limits the identification procedure to a manual level. Nevertheless, by these means 43 metabolites could be readily assigned in the tomato fruit extract (Tables III and IV), leaving more to be identified. The total number of compounds detectable by our LC-MS system is difficult to calculate due to the presence of mass signals from isotopes, adducts, and unintended in-source fragmentation. Using the strategy demonstrated in this study, the assignment of compounds lies on the integration of different sources of information (accurate mass, retention time, fragmentation pattern, and UV/Vis spectra). In addition to experimental data, previous findings and biochemical evidence can complement certain putative assignments.

In the MoTo DB we established searching tools to link an observed mass in LC-MS chromatograms to the putative tomato metabolite, through calculating the exact monoisotopic mass of each metabolite for both positive and negative ionization modes. Identifications can be validated using the retention time intervals, PDA spectra, and MS/MS data so far available. The link with external databases allows searching for similar molecules from other sources.

Some compounds reported in literature appear to occur more than once in our chromatograms, e.g. p-coumaroylhexoside, caffeoylhexoside, and naringenin chalcone-hexoside (Table III). Apparently, these metabolites can exist as different constitutional isomers in tomato fruit. The position and/or nature of the sugar substitution can influence the polarity and therefore the retention time of the compound. From the literature it is often unclear which particular isomer is mentioned. Three chromatographic peaks corresponding to caffeoylquinic acids were found. According to previous studies with comparable analytical systems (Clifford et al., 2003), the order of elution is likely 5-caffeoylquinic acid, followed by 3-caffeoylquinic acid, and then 4-caffeoylquinic acid (Table III).

Applying the same data analysis strategy, novel derivatives of phenolic acids and flavonoids were putatively assigned and information on the level of their identification are presented (Table IV). Dicaffeoylquinic acid (three isomers) and tricaffeoylquinic acid (two isomers) were identified in tomato, and novel glycosides of naringenin, naringenin chalcone, and quercetin were detected. The chromatographic separation of several isomers of coumaroyl- and caffeoylhexosides, of which only one has previously been described, also indicates the high resolution power of our LC-MS set up. MS/MS fragmentation can sometimes distinguish between constitutional isomers, however in most cases other approaches such as NMR will have to be performed to unravel the complete and exact structure of novel compounds. These NMR studies are part of our future activities in tomato metabolomics. Ideally, the combination of LC/MS/NMR should be performed for the unambiguous structure elucidation of metabolites (Exarchou et al., 2003; Sumner et al., 2003; Wolfender et al., 2003). Organizing all such analytical data into a single database will facilitate the identification of compounds and will further improve the quality and quantity of compound annotation through database searching.

By making use of the MoTo DB and the LC-PDA-MS platforms established, extracts from two tissues in tomato fruit, peel and flesh, were compared for relative differences in LC-MS signals in an untargeted manner (Fig. 5). As was expected from previous experiments (e.g. Muir et al., 2001; Bovy et al., 2002) most of the flavonoid species and their glycosides were detected in the extracts of peel tissue, while in the flesh extracts these compounds were hardly or not detectable at all. The specific accumulation of flavonoids in peel is in accordance with the idea that these

compounds play a role in the protection against stress, for example by UV light (Winkel-Shirley, 2002). On the other hand, by using this untargeted approach it became clear that tomato flesh contains markedly higher amounts of, among many still unknown metabolites, specific phenolic compounds such as caffeoylhexose II and 3-caffeoylquinic acid, as well as glycosylated alkaloids of the spirosolanol type. A compound uniquely present in the extracts from flesh tissue was identified as a hydroxyfurostanol tetrahexose, which might correspond to tomatoside A (Schelochkova et al., 1980). This molecule has a brassinosteroid-like structure and is structurally related to spirosolanes. Recently, highly active biosynthesis of brassinosteroids has been found in developing tomato fruits (Montoya et al., 2005). As yet, neither the biological functions nor the mechanisms underlying the specific accumulation of these phenolic acids and glycosylated spirosolanols in the flesh of the fruit are known. Clearly, further research into the differential distribution of (secondary) metabolites between peel and flesh tissues of tomato fruit, by analyzing these tissues from fruits from several cultivars, may provide novel information on tissue-specific regulation of biochemical pathways.

CONCLUSION

The maturation of metabolomics as the next cornerstone of functional genomics ultimately depends on the establishment of databases (Sumner et al., 2003; Bino et al., 2004). However, at the moment there are no effective database tools to query and/or comprehensively mine LC-MS-based plant metabolomics data through automated database search engines. The generation of such tools depends on the availability of metabolite databases that can be trusted and for which the source of data and its history are maintained and made publicly accessible. Here we present the first step to implement such an open access metabolite database, the MoTo DB dedicated to tomato, which intends to systematize metabolite LC-MS, MS/MS, and absorbance spectra information for common knowledge. The next step is to utilize the validated metabolomic information to study the dynamics of the metabolome, to elucidate mutants and gene functions based on differential metabolic profiles, and to decipher the biological relevance of each metabolite. The combination of information from other omics technologies can lead to a wider view on the systems biology of the plant studied. As a result, the integration of databases from these different disciplines will be inevitable.

MATERIALS AND METHODS

Plant Material

A large pool of tomato (*Lycopersicum esculentum*, now *Solanum lycopersicum*) fruit material was prepared by combining fruits from turning, pink, and red ripe stages of development of 96 different tomato cultivars representing the three major types of tomato fruits (i.e. cherry, Dutch beef, and normal round tomatoes). These plants were grown in an environmentally controlled green-

house located in Wageningen, The Netherlands, during the summer and autumn of 2003. Plants were grown in rock wool plugs connected to an automatic irrigation system comparable to standard commercial cultivation conditions. For analysis of anthocyanins, purple-colored fruits from offspring of a crossing of two natural mutants, $Af \times hp-2^{j}$ (van Tuinen et al., 2005), were harvested at the ripe stage of development. Peel (about 2 mm thickness) was removed from fruits, ground into a fine powder in liquid nitrogen, and stored at $-80°C$ until further analysis. For metabolite profile comparison of peel and flesh, red ripe fruits of cultivar Money Maker were used of which peel (2 mm thickness) and flesh (rest of fruit) were separated and used as described.

Extraction

Of the frozen tomato powder, 0.5 g fresh weight was weighed and extracted with 1.5 mL pure methanol (final methanol concentration in the extract approximately 75%). Hydrolyzed extracts were prepared by sequentially adding 1 mL of 0.1% tert-butylhydroquinone in methanol solution and 0.4 mL of HCl 6 M to 0.6 g fresh weight tomato material, shaking in a water bath at 90°C to 95°C for 1 h, and adding 2 mL of methanol (Bovy et al., 2002). All samples were sonicated for 15 min, filtered through a 0.2 μm inorganic membrane filter (Anotop 10, Whatman), and analyzed.

Chemicals

Standard compounds *p*-coumaric acid, protocatechuic acid, salicylic acid, caffeic acid, ferulic acid, cinnamic acid, myricetin, and naringenin were purchased from ICN; *p*-hydroxybenzoic acid, chlorogenic acid quercetin, Phe, sinapic acid, and α-tomatine from Sigma; vanillic acid and rutin (quercetin-3-*O*-rutinoside) from Acros; naringenin chalcone from Apin Chemicals, kaempferol and kaempferol-3-*O*-rutinoside from Extrasynthese; and tert-butylhydroquinone from Aldrich. Acetonitrile HPLC supragradient and methanol absolute HPLC supragradient were obtained from Biosolve. Formic acid for synthesis 98% to 100% was from Merck-Schuchardt, HCl 37% for analysis from Acros, and ultrapure water was obtained from an Elga Maxima purification unit (Bucks). Leucine enkaphaline was purchased from Sigma.

Chromatographic Conditions

HPLC was carried out using a Waters Alliance 2795 HT system with a column oven. For chromatographic separation, a Luna C18(2) precolumn (2.0 × 4 mm) and analytical column (2.0 × 150 mm, 100 Å, particle size 3 μm) from Phenomenex were used. Five microliters of sample was injected into the system for LC-PDA-MS analysis. Degassed solutions of formic acid:ultrapure water (1:10³, v/v; eluent A) and formic acid:acetonitrile (1:10³, v/v; eluent B) were pumped at 0.19 mL min⁻¹ into the HPLC system. The gradient applied started at 5% B and increased linearly to 35% B in 45 min. Then, for 15 min the column was washed and equilibrated before the next injection. The column temperature was kept at 40°C and the samples at 20°C. The room temperature was maintained at 20°C.

Detection of Metabolites by PDA and MS

The HPLC system was connected online to a Waters 2996 PDA detector, set to acquire data every second from 240 to 600 nm with a resolution of 4.8 nm, and subsequently to a QTOF Ultima V4.00.00 mass spectrometer (Waters-Corporation, MS technologies). An ESI source working either in positive or negative ion mode was used for all MS analyses. Before each series of analyses, the mass spectrometer was calibrated using phosphoric acid:acetonitrile:water (1:10³:10³, v/v/v) solution. Capillary voltage, collision energy, and desolvation temperature were optimized to obtain a series of phosphoric acid clusters suitable for calibration between m/z 80 and 1,500. During sample analyses, the capillary voltage was set to 2.75 kV and the cone at 35 V. Source and desolvation temperatures were set to 120°C and 250°C, respectively. Cone gas and desolvation gas flows were 50 and 500 Lh⁻¹, respectively. In the positive ion mode, the collision energy was 5 eV while in the negative ion mode it was 10 eV. Resolution was set at 10,000 and during calibration the MS parameters were adjusted to achieve such a resolution.

TOF-MS data were acquired in centroid mode. During LC-MS analyses scan durations of 0.9 s and an interscan time of 0.1 s were used. For LC-MS/MS measurements, 10 μL of sample was injected into the system and MS/MS measurements were made with 0.40 s of scan duration and 0.10 s of interscan

delay with increasing collision energies according to the following program: 5 (ESI positive) or 10 (ESI negative), 15, 30, and 50 eV.

A lockspray source was equipped with the mass spectrometer allowing online mass correction to obtain high mass accuracy of analytes. Leucine enkephalin, $[M+H]^+ = 556.2766$ and $[M-H]^- = 554.2620$, was used as a lock mass, being continuously sprayed into a second ESI source using an LKB Bromma 2150 HPLC pump, and sampled every 10 s, producing an average intensity of 500 counts/scan in centroid mode (approximately 100 count/scan in continuum mode).

Data Analysis and Alignment

Acquisition of LC-PDA-MS data was performed under MassLynx 4.0 (Waters). MassLynx was used for visualization and manual processing of LC-PDA-MS/MS data. Mass data were automatically processed by metAlign version 1.0 (www.metalign.nl). MetAlign transforms accurate masses into nominal masses to shorten the calculation time and minimize the number of mass bins. Baseline and noise calculations were performed from scan number 225 to 2,475, corresponding to retention times 4.0 min to 49.3 min. The maximum amplitude was set to 15,000 and peaks below three times the local noise were discarded. The .csv file output containing nominal mass peak intensity data (peak heights, i.e. ion counts/scan at the center of the peak) at aligned retention times (scans) over all samples processed was used for further data processing. A script called metAccure was used for the calculation of accurate masses from the metAlign-extracted peaks. MetAccure calculates the accurate mass, using only those scans in which signal intensities are within a user-defined window relative to the lock mass intensity of each mass signal using the .csv files containing retention time alignments, originating from metAlign analysis, in combination with the original data in NetCDF format, created from MassLynx.raw files by Dbridge (O. Vorst, H.A. Verhoeven, C.H.R. de Vos, C.A. Maliepaard, and R.C.H.J. van Ham, unpublished data). Comparison of extracts from peel and flesh tissues for significant differences in intensity of each aligned mass signal was made using the t-student statistical tool within metAlign (level of significance set at 0.05). The settings for baseline corrections and signal alignment were analogous to those described above.

Annotation of Metabolites

Datasets obtained after metAlign and metAccure treatment were analyzed as (retention time × accurate mass × peak intensity) matrixes for metabolite identification. $[M+H]^+$ and $[M-H]^-$ values were calculated for metabolites present in Table I and used for sorting with the matrixes. Data collected during the first 4.0 min of chromatography were discarded. Novel metabolites were identified by calculating the elemental composition from accurate mass measurements using the MassLynx software. The tolerance was set at 5 ppm, taking into account the correct analyte-lock mass signal ratio. For an observed accurate mass, a list of possible molecular formulas was obtained, selected for the presence of C, H, O, and N. In addition, raw datasets were checked manually in MassLynx for retention time, UV/Vis spectra, and QTOF-MS/MS fragmentation patterns for chromatographically separated peaks, complementing the accurate mass-based elemental formulas. The combination of accurate mass data, retention time (as an indication of polarity), UV/Vis spectra, and MS/MS data allowed a putative identification of metabolites. Best matches were searched in the Dictionary of Natural Products and SciFinder databases for possible structures. The putative identifications were confirmed by published data and with standard compounds, if commercially available.

MoTo DB Buildup

Based on available literature information about compounds identified in tomato, information acquired from LC-PDA-MS analysis of tomato fruit was used to validate each metabolite: (1) a retention time; (2) accurate mass in the form of monoisotopic mass (neutral) and in the ion forms $(M+H)^+$ and $(M-H)^-$; (3) elemental compositions; (4) MS/MS fragments; and (5) maximum absorbance peaks in UV/Vis. Given a found mass and a Δppm (or ΔmD) that is set by the user, the database can find possible matches. Formic acid, if detected, was also included in the database. The database is implemented in MySQL and running on a Linux cluster.

ACKNOWLEDGMENTS

LITERATURE CITED

Bianco G, Schmitt-Kopplin P, De Benedetto G, Kettrup A, Cataldi TR (2002) Determination of glycoalkaloids and relative aglycones by non-aqueous capillary electrophoresis coupled with electrospray ionization-ion trap mass spectrometry. Electrophoresis **23:** 2904–2912

Bino RJ, de Vos CHR, Lieberman M, Hall RD, Bovy A, Jonker HH, Tikunov Y, Lommen A, Moco S, Levin I (2005) The light-hyperresponsive *high pigment-2^dg* mutation of tomato: alterations in the fruit metabolome. New Phytol **166:** 427–438

Bino RJ, Hall RD, Fiehn O, Kopka J, Saito K, Draper J, Nikolau BJ, Mendes P, Roessner-Tunali U, Beale MH, et al (2004) Potential of metabolomics as a functional genomics tool. Trends Plant Sci **9:** 418–425

Bovy A, de Vos CHR, Kemper M, Schijlen E, Almenar Pertejo M, Muir S, Collins G, Robinson S, Verhoeyen M, Hughes S, et al (2002) High-flavonol tomatoes resulting from the heterologous expression of the maize transcription factor genes LC and C1. Plant Cell **14:** 2509–2526

Buta JG, Spaulding DW (1997) Endogenous levels of phenolics in tomato fruit during growth and maturation. J Plant Growth Regul **16:** 43–46

Clifford MN, Johnston KL, Knight S, Kuhnert N (2003) Hierarchical scheme for LC-MS^n identification of chlorogenic acids. J Agric Food Chem **51:** 2900–2911

Colombo M, Sirtori FR, Rizzo V (2004) A fully automated method for accurate mass determination using high-performance liquid chromatography with a quadrupole/orthogonal acceleration time-of-flight mass spectrometer. Rapid Commun Mass Spectrom **18:** 511–517

Crozier A, Lean MEJ, McDonald MS, Black C (1997) Quantitative analysis of the flavonoid content of commercial tomatoes, onions, lettuce, and celery. J Agric Food Chem **45:** 590–595

Dixon RA, Strack D (2003) Phytochemistry meets genome analysis, and beyond. Phytochemistry **62:** 815–816

Exarchou V, Godejohann M, van Beek TA, Gerothanassis IP, 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. Anal Chem **75:** 6288–6294

Fleuriet A, Macheix JJ (1977) Effect des blessures sur les composés phénoliques des fruits de tomates «cerise» (*Lycopersicum esculentum* var. cerasiforme). Physiol Veg **15:** 239–250

Fleuriet A, Macheix J-J (1981) Quinyl esters and glucose derivatives of hydroxycinnamic acids during growth and ripening of tomato fruit. Phytochemistry **20:** 667–671

Friedman M (2002) Tomato glycoalkaloids: role in the plant and in the diet. J Agric Food Chem **50:** 5751–5780

Friedman M, Kozukue N, Harden LA (1997) Structure of the tomato glycoalkaloid tomatidenol-3-beta-lycotetraose (dehydrotomatine). J Agric Food Chem **45:** 1541–1547

Friedman M, Kozukue N, Harden LA (1998) Preparation and characterization of acid hydrolysis products of the tomato glycoalkaloid alpha-tomatine. J Agric Food Chem **46:** 2096–2101

Friedman M, Levin CE, Mcdonald GM (1994) α-Tomatine determination in tomatoes by HPLC using pulsed amperometric detection. J Agric Food Chem **42:** 1959–1964

Fujiwara Y, Takaki A, Uehara Y, Ikeda T, Okawa M, Yamauchi K, Ono M, Yoshimitsu H, Nohara T (2004) Tomato steroidal alkaloid glycosides, esculeosides A and B, from ripe fruits. Tetrahedron **60:** 4915–4920

Fujiwara Y, Yahara S, Ikeda T, Ono M, Nohara T (2003) Cytotoxic major saponin from tomato fruits. Chem Pharm Bull (Tokyo) **51:** 234–235

Hertog MGL, Hollman PCH, Katan MB (1992) Content of potentially anticarcinogenic flavonoids of 28 vegetables and 9 fruits commonly consumed in the Netherlands. J Agric Food Chem **40:** 2379–2383

Hunt GM, Baker EA (1980) Phenolic constituents of tomato fruit cuticles. Phytochemistry **19:** 1415–1419

Jones CM, Mes P, Myers JR (2003) Characterization and inheritance of the *Anthocyanin fruit* (*Aft*) tomato. J Hered **94:** 449–456

Justesen U, Knuthsen P, Leth T (1998) Quantitative analysis of flavonols, flavones, and flavanones in fruits, vegetables and beverages by high-performance liquid chromatography with photo-diode array and mass spectrometric detection. J Chromatogr A **799:** 101–110

Juvik JA, Stevens MA, Rick CM (1982) Survey of the genus Lycopersicon for variability in alpha-tomatine content. HortScience **17:** 764–766

Kozukue N, Friedman M (2003) Tomatine, chlorophyll, beta-carotene and lycopene content in tomatoes during growth and maturation. J Sci Food Agric **83:** 195–200

Krause M, Galensa R (1992) Determination of naringenin and naringenin-chalcone in tomato skins by reversed phase HPLC after solid-phase extraction. Z Lebensm Unters Forsch **194:** 29–32

Le Gall G, Colquhoun IJ, Davis AL, Collins GJ, Verhoeyen ME (2003a) Metabolite profiling of tomato (*Lycopersicon esculentum*) using ^{1}H NMR spectroscopy as a tool to detect potential unintended effects following a genetic modification. J Agric Food Chem **51:** 2447–2456

Le Gall G, DuPont MS, Mellon FA, Davis AL, Collins GJ, Verhoeyen ME, Colquhoun IJ (2003b) Characterization and content of flavonoid glycosides in genetically modified tomato (*Lycopersicon esculentum*) fruits. J Agric Food Chem **51:** 2438–2446

Martinez-Valverde I, Periago MJ, Provan G, Chesson A (2002) Phenolic compounds, lycopene and antioxidant activity in commercial varieties of tomato (*Lycopersicum esculentum*). J Sci Food Agric **82:** 323–330

Mathews H, Clendennen SK, Caldwell CG, Liu XL, Connors K, Matheis N, Schuster DK, Menasco DJ, Wagoner W, Lightner J, et al (2003) Activation tagging in tomato identifies a transcriptional regulator of anthocyanin biosynthesis, modification, and transport. Plant Cell **15:** 1689–1703

Mattila P, Kumpulainen J (2002) Determination of free and total phenolic acids in plant-derived foods by HPLC with diode-array detection. J Agric Food Chem **50:** 3660–3667

Minoggio M, Bramati L, Simonetti P, Gardana C, Iemoli L, Santangelo E, Mauri PL, Spigno P, Soressi GP, Pietta PG (2003) Polyphenol pattern and antioxidant activity of different tomato lines and cultivars. Ann Nutr Metab **47:** 64–69

Montoya T, Nomura T, Yokota T, Farrar K, Harrison K, Jones JG, Kaneta T, Kamiya Y, Szekeres M, Bishop GJ (2005) Patterns of *Dwarf* expression and brassinosteroid accumulation in tomato reveal the importance of brassinosteroid synthesis during fruit development. Plant J **42:** 262–269

Muir SR, Collins GJ, Robinson S, Hughes S, Bovy A, De Vos CHR, van Tunen AJ, Verhoeyen ME (2001) Overexpression of petunia chalcone isomerase in tomato results in fruit containing increased levels of flavonols. Nat Biotechnol **19:** 470–474

Petró-Turza M (1987) Flavor of tomato and tomato products. Food Rev Int **2:** 309–351

Raffo A, Leonardi C, Fogliano V, Ambrosino P, Salucci M, Gennaro L, Bugianesi R, Giuffrida F, Quaglia G (2002) Nutritional value of cherry tomatoes (*Lycopersicon esculentum* cv. Naomi F1) harvested at different ripening stages. J Agric Food Chem **50:** 6550–6556

Reschke A, Herrmann K (1982) Vorkommen von 1-O-hydroxycinnamyl-β-D-glucosen im gemüse. 1. Phenolcarbonsäure-verbindungen des gemüses. Z Lebensm-Unters-Forsch **174:** 5–8

Rosman KJR, Taylor PDP (1998) Isotopic compositions of the elements 1997. Pure Appl Chem **70:** 217–235

Sakakibara H, Honda Y, Nakagawa S, Ashida H, Kanazawa K (2003) Simultaneous determination of all polyphenols in vegetables, fruits, and teas. J Agric Food Chem **51:** 571–581

Schauer N, Steinhauser D, Strelkov S, Schomburg D, Allison G, Moritz T, Lundgren K, Roessner-Tunali U, Forbes MG, Willmitzer L, et al (2005) GC-MS libraries for the rapid identification of metabolites in complex biological samples. FEBS Lett **579:** 1332–1337

Schmidtlein H, Herrmann K (1975) Über die phenolsäuren des gemüses. II. Hydroxyzimtsäuren und hydroxybenzoesäuren der frucht- und samengemüsearten. Z Lebensm Unters Forsch **159:** 213–218

Schelochkova AP, Vollerner JS, Koshoev KK (1980) Tomatoside A from *Licopersicum esculentum* seeds. Khim Prir Soedin **4:** 533–540

Stewart AJ, Bozonnet S, Mullen W, Jenkins GI, Lean MEJ, Crozier A (2000) Occurrence of flavonols in tomatoes and tomato-based products. J Agric Food Chem **48:** 2663–2669

Sumner LW, Mendes P, Dixon RA (2003) Plant metabolomics: large-scale phytochemistry in the functional genomics era. Phytochemistry **62:** 817–836

Tikunov Y, Lommen A, de Vos CHR, Verhoeven HA, Bino RJ, Hall RD, Bovy AG (2005) A novel approach for nontargeted data analysis for metabolomics: large-scale profiling of tomato fruit volatiles. Plant Physiol **139:** 1125–1137

Tokusoglu O, Unal MK, Yildirim Z (2003) HPLC-UV and GC-MS characterization of the flavonol aglycons quercetin, kaempferol, and myricetin in tomato pastes and other tomato-based products. Acta Chromatogr **13:** 196–207

van Tuinen A, de Vos CHR, Hall RD, van der Plas LHW, Bowler C, Bino RJ (2005) Use of metabolomics for development of tomato mutants with enhanced nutritional value by exploiting natural non-GMO light-hyperresponsive mutants. *In* P Jaiwal, ed, Plant Genetic Engineering: Improvement of the Nutritional and the Therapeutic Qualities of Plants. Agritech Publications/Agricell Report, Shrub Oak, NY

von Roepenack-Lahaye E, Degenkolb T, Zerjeski M, Franz M, Roth U, Wessjohann L, Schmidt J, Scheel D, Clemens S (2004) Profiling of Arabidopsis secondary metabolites by capillary liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry. Plant Physiol **134:** 548–559

Vorst O, de Vos CHR, Lommen A, Staps RV, Visser RGF, Bino RJ, Hall RD (2005) A non-directed approach to the differential analysis of multiple LC-MS derived metabolic profiles. Metabolomics **1:** 169–180

Willker W, Leibfritz D (1992) Complete assignment and conformational studies of tomatine and tomatidine. Magn Reson Chem **30:** 645–650

Winkel-Shirley B (2002) Biosynthesis of flavonoids and effects of stress. Curr Opin Plant Biol **5:** 218–223

Winter M, Herrmann K (1986) Esters and glucosides of hydroxycinnamic acids in vegetables. J Agric Food Chem **34:** 616–620

Wolfender JL, Ndjoko K, Hostettmann K (2003) Liquid chromatography with ultraviolet absorbance-mass spectrometric detection and with nuclear magnetic resonance spectroscopy: a powerful combination for the on-line structural investigation of plant metabolites. J Chromatogr A **1000:** 437–455

Yahara S, Uda N, Nohara T (1996) Lycoperosides A-C, three stereoisomeric 23-acetoxyspirosolan-3 beta-ol beta-lycotetraosides from Lycopersicon esculentum. Phytochemistry **42:** 169–172

Yahara S, Uda N, Yoshio E, Yae E (2004) Steroidal alkaloid glycosides from tomato (*Lycopersicon esculentum*). J Nat Prod **67:** 500–502

Yoshizaki M, Matsushita S, Fujiwara Y, Ikeda T, Ono M, Nohara T (2005) Tomato new sapogenols, isoesculeogenin A and esculeogenin B. Chem Pharm Bull (Tokyo) **53:** 839–840