A Novel Approach for Nontargeted Data Analysis for Metabolomics. Large-Scale Profiling of Tomato Fruit Volatiles[1][w]

Yury Tikunov, Arjen Lommen, C.H. Ric de Vos, Harrie A. Verhoeven, Raoul J. Bino, Robert D. Hall, and Arnaud G. Bovy*

Centre for BioSystems Genomics, 6700 AB Wageningen, The Netherlands (Y.T., A.L., C.H.R.d.V., H.A.V., R.J.B., R.D.H., A.G.B.); Plant Research International, 6700 AA Wageningen, The Netherlands (Y.T., C.H.R.d.V., H.A.V., R.J.B., R.D.H., A.G.B.); RIKILT, Institute for Food Safety, 6700 AE Wageningen, The Netherlands (A.L.); and Laboratory for Plant Physiology, Wageningen University, 6703 BD Wageningen, The Netherlands (R.J.B.)

To take full advantage of the power of functional genomics technologies and in particular those for metabolomics, both the analytical approach and the strategy chosen for data analysis need to be as unbiased and comprehensive as possible. Existing approaches to analyze metabolomic data still do not allow a fast and unbiased comparative analysis of the metabolic composition of the hundreds of genotypes that are often the target of modern investigations. We have now developed a novel strategy to analyze such metabolomic data. This approach consists of (1) full mass spectral alignment of gas chromatography (GC)-mass spectrometry (MS) metabolic profiles using the MetAlign software package, (2) followed by multivariate comparative analysis of metabolic phenotypes at the level of individual molecular fragments, and (3) multivariate mass spectral reconstruction, a method allowing metabolite discrimination, recognition, and identification. This approach has allowed a fast and unbiased comparative multivariate analysis of the volatile metabolite composition of ripe fruits of 94 tomato (*Lycopersicon esculentum* Mill.) genotypes, based on intensity patterns of >20,000 individual molecular fragments throughout 198 GC-MS datasets. Variation in metabolite composition, both between- and within-fruit types, was found and the discriminative metabolites were revealed. In the entire genotype set, a total of 322 different compounds could be distinguished using multivariate mass spectral reconstruction. A hierarchical cluster analysis of these metabolites resulted in clustering of structurally related metabolites derived from the same biochemical precursors. The approach chosen will further enhance the comprehensiveness of GC-MS-based metabolomics approaches and will therefore prove a useful addition to nontargeted functional genomics research.

Functional genomics technologies designed to assess gene activity (transcriptomics) and protein accumulation (proteomics) are now well established in the quest to link gene to function (Holtorf et al., 2002). Subsequently, metabolomics approaches have been forwarded as a means to link the functional biochemical phenotype to other functional genomics data (Weckwerth and Fiehn, 2002; Sumner et al., 2003; Bino et al., 2004; Hall et al., 2005). Like transcriptomics and proteomics, metabolomics involves two main components: instrumental analysis (analytical) and data analysis (bioinformatics). Both topics need to be as comprehensive as possible for true, broad, metabolic profiling and comparative analysis of the biochemical status of living organisms. Several analytical methods for metabolomics have already been reported using model plants in genomic studies (Fiehn et al., 2000a, 2000b; Roessner et al., 2000, 2001; Huhman and Sumner, 2002; Tolstikov and Fiehn, 2002; Roessner-Tunali et al., 2003; Kopka et al., 2004; Desbrosses et al., 2005). A significant number of these studies have, however, been dedicated to metabolic profiling specifically of the nonvolatile compounds involved in primary plant metabolism using gas chromatography (GC) coupled to mass spectrometry (MS). Another significant part of the plant metabolome, comprising the volatile metabolites, is of a particular interest, since they play an important role in fundamental processes such as signaling mechanisms and interorganism interactions (Shulaev et al., 1997; Seskar et al., 1998; Ozawa et al., 2000; Arimura et al., 2002; Liechti and Farmer, 2002; Dicke et al., 2003; Dudareva et al., 2004; Engelberth et al., 2004; Ryu et al., 2004). In addition, these components are also of great agronomic importance as volatile metabolites are major determinants of food and flower quality in terms of flavor and fragrance (Buttery and Ling, 1993; Baldwin et al., 2000, 2004; Yilmaz et al., 2001; Tandon et al., 2003; Krumbein et al., 2004; Simkin et al., 2004; Ruiz et al., 2005).

Solid phase microextraction (SPME-GC-MS) is an analytical approach that is suitable for metabolomics studies of volatiles since it is renowned for its high sensitivity, reproducibility, and robustness (Yang and

[1] This work was supported by the research program of the Centre of BioSystems Genomics, which is part of the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research.

* Corresponding author; e-mail arnaud.bovy@wur.nl; fax 31–317–418094.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Arnaud G. Bovy (arnaud.bovy@wur.nl).

[w] The online version of this article contains Web-only data.
www.plantphysiol.org/cgi/doi/10.1104/pp.105.068130.

Peppard, 1994; Matich et al., 1996; Verhoeven et al., 1997; Song et al., 1997, 1998; Augusto et al., 2000; Verdonk et al., 2003). GC-MS-based approaches utilize gas chromatographic separation of metabolites extracted from plant material and, in the case of SPME, the volatiles are first extracted from the headspace above the plant material using a specially designed adsorbant fiber (Fig. 1A). Subsequently, separated metabolites are fragmented to charged molecular fragments—ions—that are then detected in the mass spectrometer. Each metabolite produces a unique spectrum of molecular fragments with specific masses and a fixed relative abundance. This unique fingerprint can therefore be used for metabolite recognition and identification.

Hundreds of different metabolites can be detected in crude plant extracts using GC-MS. This is, however, just a small fraction of the more than 10,000 metabolites that have been described in plants (Fiehn et al., 2000b). However, even this limited amount of biochemical information cannot be fully subjected to a comparative metabolomic analysis when conventional strategies are used. Such strategies, in general, consist of three consecutive steps (Fig. 1B). First, metabolites must be recognized and (or) identified from the tens of thousands of molecular fragments that constitute a typical GC-MS profile (Fig. 1B-a). Second, quantitative values (often relative) of the identified metabolites are aligned throughout all the metabolic profiles of the genotypes (Fig. 1B-b) in order to perform the third step, comparative analyses of their metabolic phenotypes using multivariate exploratory techniques dedicated for metabolomics (e.g. hierarchical cluster analysis [HCAi], principal component analysis [PCA], self-organizing maps, etc.; Fig. 1B-c). The metabolic data can then also be linked to other data derived using other functional genomics technologies. The comprehensiveness of this strategy therefore depends on the number of metabolites that can be identified in the samples to be compared. However, the extreme complexity of the plant metabolome already generates a bottleneck at the first step in this algorithm: Despite using chromatographic separation, metabolites still coelute prior to being subjected to MS. Consequently, this coelution results in overlapping of the unique fragmentation patterns. In addition to the problem of coelution, a high variability in metabolite quantity within large numbers of biological samples further complicates metabolite identification and thus limits the entire analysis to a metabolite subset that includes only those compounds that can be reliably identified throughout all genotypes. Many other possible metabolic differences may then be overlooked. To overcome these limitations and make the metabolomic data analysis truly comprehensive and unbiased, we offer a novel strategy for data analysis (Fig. 1C). This strategy is based on a fully automated alignment of metabolic profiles at the level of individual molecular fragments without prior assignment to the chemical structures of the metabolites they represent. Subsequently, a multivariate comparative analysis of individual metabolic profiles is performed, which is based on all chemical information derived by an analytical approach. Although this strategy initially removes the need for prior metabolite identification, this is eventually still required in order to put a biological meaning to the differences found. To relate the thousands of molecular fragments normally constituting a chromatogram to their parental metabolites, a novel approach, multivariate mass spectra reconstruction (MMSR), has been developed. Using MMSR, clusters of related metabolite fragments can be recognized and the corresponding metabolites subsequently identified.

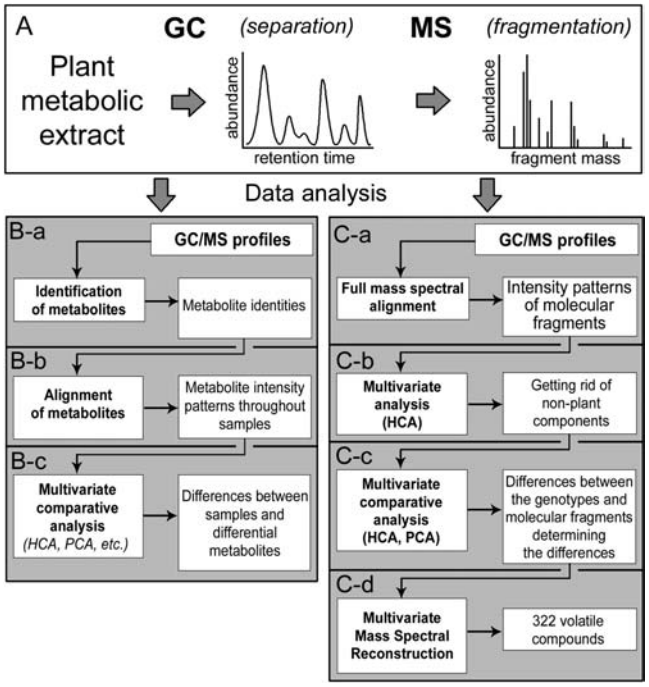

Figure 1. GC-MS-based metabolomics. A, Analytical approach used. B, Conventional approach. C, Alternative, unbiased approach to GC-MS data analysis.

The entire strategy of data analysis is universal for many kinds of mass spectral data and exceeds approaches of unbiased metabolomic data analysis in terms of resolution and comprehensiveness (Nielsen et al., 1998; Fraga et al., 2001; Johnson et al., 2003; Jonsson et al., 2004; Wiener et al., 2004; Willse et al., 2005). Also, it uses widely available software tools and simple basic statistical procedures, both of which make it useful for a wide range of studies in the fields of biochemistry, physiology, functional genomics, and systems biology.

The strategy was used for a comparative multivariate analysis of a set of 94 contrasting tomato (*Lycopersicon esculentum* Mill.) genotypes covering the variation in the germplasm of commercial tomato varieties. The analysis was based on the profiles of all volatiles that could be detected by the analytical method used (SPME-GC-MS) and revealed a total of 322 different compounds in the entire genotype set. This covers approximately 80% of the more than 400

tomato volatile compounds, which have been detected in tomato fruit using different analytical methods (for review, see Petro-Turza, 1987).

RESULTS

Automated Sequential Headspace SPME-GC-MS: Method Development

In order to produce and release volatiles, tomato material (e.g. juice or pulp) is usually incubated for a fixed period, during which essential enzymes such as lipoxygenase and hydroperoxide lyases are allowed to remain active. This is followed by the addition of concentrated $CaCl_2$ to stop enzyme activity and to drive the volatiles into the headspace (Bezman et al., 2003; Verdonk et al., 2003). To test this method for its suitability for effective, prolonged, sequential automated analysis of tomato samples, the SDs of the 15 major tomato volatiles (Baldwin et al., 2000), measured sequentially in four sample replications, each after 3-h intervals, were calculated (Table I). The addition of $CaCl_2$ alone resulted in large variations in metabolite abundance between replicate analyses (average % SD = 41%; Table I). However, a marked improvement in reproducibility (average % SD = 9%) was achieved by the addition of NaOH/EDTA solution, which was chosen for its effectiveness compared to a number of alternative buffers tested (data not shown). In combination with subsequent $CaCl_2$-induced enzyme inactivation, this procedure resulted in sufficient stability and reproducibility over a 12-h period. On average, the biological variation between the genotypes was then approximately 10 times the analytical variation. To estimate the metabolic variation that can be observed within a genotype, samples of five individual fruits of the same genotype were analyzed. The fruit-to-fruit variation, which, in fact, included the analytical variation, observed for the 15 volatiles ranged from 8% to 35% SD. For all metabolites, the fruit-to-fruit variation was significantly less than the biological variation between genotypes, according to % SD and range between lowest and highest value (Table I).

In total, 94 tomato fruit samples, in duplicate, were profiled for volatile metabolites. Consequently, including the daily external reference samples, 198 GC-MS datasets were obtained in this tomato volatile study.

A Stepwise Approach for Nontargeted Data Analysis

Step 1 (Fig. 1C-a) is as follows. The entire 198-sample GC-MS dataset was analyzed using the dedicated MetAlign software package. After automated baseline correction, intensities of approximately 20,000 molecular fragments with corresponding retention times were aligned throughout 198 GC-MS profiles by MetAlign.

Table I. *Biological and analytical variation of the tomato volatile metabolites*

For the analysis, a mix of tomato samples was made and separate aliquots were measured after 0, 4, 8, and 12 h. Using these four measurements, % SD (presented as the % of total value) was calculated for the sole use of $CaCl_2$ (second column) and for the combination NaOH/EDTA + $CaCl_2$ (third column). For the analysis of biological variation within genotype (fourth column), five individual fruits of the same genotype were profiled for volatiles, and % SD for these five replicates was calculated. Biological variation between genotypes (fifth column) was calculated as % SD of means of all 94 tomato samples when NaOH/EDTA + $CaCl_2$ procedure was used. The maximal relative fruit-to-fruit variation as well as the maximal variation between all 94 genotypes was calculated as the ratio of maximal and minimal relative values of the 15 volatiles across the five fruits and the 94 genotype samples, respectively. It is given in parenthesis as fold difference (fourth and fifth columns).

Metabolites	Analytical Variation, % SD		Biological Variation within Genotype n=5, % SD	Biological Variation between Genotypes n=94, % SD
	$CaCl_2$	NaOH/EDTA + $CaCl_2$		
1-Penten-3-one	19	7	15 (1.4)	45 (6)
2-Isobutylthiazole	39	4	13 (1.4)	64 (47)
2-Methylbutanal	88	11	34 (2.7)	60 (16)
2-Methylbutanol	38	8	35 (2.5)	76 (39)
3-Methylbutanol	37	7	28 (2.2)	74 (29)
6-Methyl-5-hepten-2-one	33	4	17 (1.5)	37 (7)
β-Ionone	51	17	10 (1.3)	62 (14)
E-2-Heptenel	38	8	8 (1.2)	39 (10)
E-2-Hexenal	34	5	25 (1.8)	37 (6)
Hexanal	17	3	10 (1.3)	29 (8)
Methyl salicylate	49	9	21 (1.6)	173 (656)
Phenylacetaldehyde	42	6	23 (1.8)	162 (401)
Phenylethanol	48	4	19 (1.5)	198 (686)
Z-3-Hexenal	40	23	23 (1.7)	28 (7)
Z-3-Hexenol	39	13	17 (1.5)	103 (28)
Average	41	9	18	79

Step 2 (Fig. 1C-b) is as follows. A common problem in SPME-GC-MS analyses is the production of molecular fragments originating from contaminants coming from the SPME fiber material. These molecular fragments had a typical pattern of occurrence throughout the samples, which was very different from the plant-derived molecular fragments. These could therefore be efficiently recognized by means of, for example, HCA (Fig. 2). The cluster of molecular fragments, which was clearly separate from the other clusters (Fig. 2), had mass and retention time characteristics that were identical to those of nonplant compounds identified in blank injections. Therefore, this entire group of nonplant molecular fragments, which were highly correlated to the contaminant-specific fragments (such as *m/z* 207, 267, 355, etc.) related to a number of polysiloxanes, could readily be excluded from the dataset before further analysis. This is an essential prerequisite before effective comparison of the plant-specific data can be made.

Step 3 (Fig. 1C-c) is as follows. The data matrix cleaned of the fiber contaminants was subjected to a multivariate comparative analysis. First, HCA of the 94 tomato genotypes was performed using the Pearson correlation between means of genotype analytical replicates. The HCA revealed a high correlation between the reference samples, which were analyzed daily during the entire experiment in order to monitor the stability of the analytical system (Fig. 3A). The cherry genotypes formed a distinct cluster, clearly separated from the round and beef varieties. The latter two tomato types could not be separated into distinct groups. One cherry genotype could be regarded as intermediate by its volatile composition, due to its location at the very edge of the round-beef cluster.

PCA revealed two major types of metabolic differences within the 94 tomato genotypes (Fig. 3B). First, in accordance with HCA, PCA showed a clear between-fruit-type variation, separating the cherry tomatoes, on the one hand, from the round and beef tomatoes, on the other hand (vector 1). In addition, PCA revealed a clear within-type variation in metabolite content, separating the 94 tomato cultivars into two groups independent of fruit type (vector 2). The daily replicated reference samples are located in the middle of both vectors of genotype differentiation. This is logical, since the reference sample was created by pooling of fruit material of several genotypes of each fruit type. The molecular fragments determining both the between- and within-type variations could be found by projection of the genotype differentiation vectors onto the PCA plot showing the distribution of the molecular fragments (Fig. 3C).

Step 4 (Fig. 1C-d) is as follows. A novel MMSR strategy was developed to reconstruct chemical structures of metabolites from the molecular fragment information of GC-MS profiles and subsequently to discover a biochemical meaning of the metabolic differences found.

The approach is based on two points. First, since fragmentation of a metabolite by the mass spectrometer occurs after chromatographic separation, molecular fragments derived from the same metabolite will appear within a peak of a certain width at a certain retention time in a chromatogram. Second, the relative ratio between intensities of molecular fragments derived from the same metabolite is constant. Therefore, the expression patterns of these molecular fragments must be identical throughout a set of variable metabolic profiles and hence must be highly correlated to each other. Based on these points, a metabolite may be defined as a group of highly correlated molecular fragments situated within a certain retention time window. Proceeding from this definition, all of the 20,000 molecular fragments were subjected to HCA by calculating the Pearson correlation between their intensity patterns throughout the GC-MS profiles of all the tomato genotypes analyzed. HCA resulted in clustering of molecular fragments showing identical or highly similar patterns of intensities throughout all GC-MS datasets (Fig. 4A). Those molecular fragments, which clustered together with a Pearson correlation coefficient equal to or higher than 0.8 and were

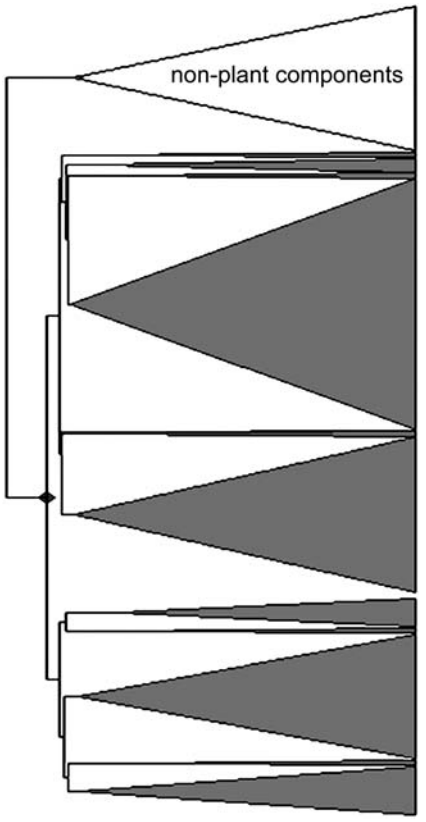

Figure 2. HCA of >20,000 molecular fragments based on their expression patterns throughout 198 GC-MS profiles. To simplify the view, only the highest branches of the dendrogram are displayed, showing the main groups of compounds as triangles. This procedure produced a dendrogram revealing a distinct cluster of nonplant components, comprising molecular fragments derived from constituents of the SPME fiber material that could then be readily removed from the dataset prior to further analysis.

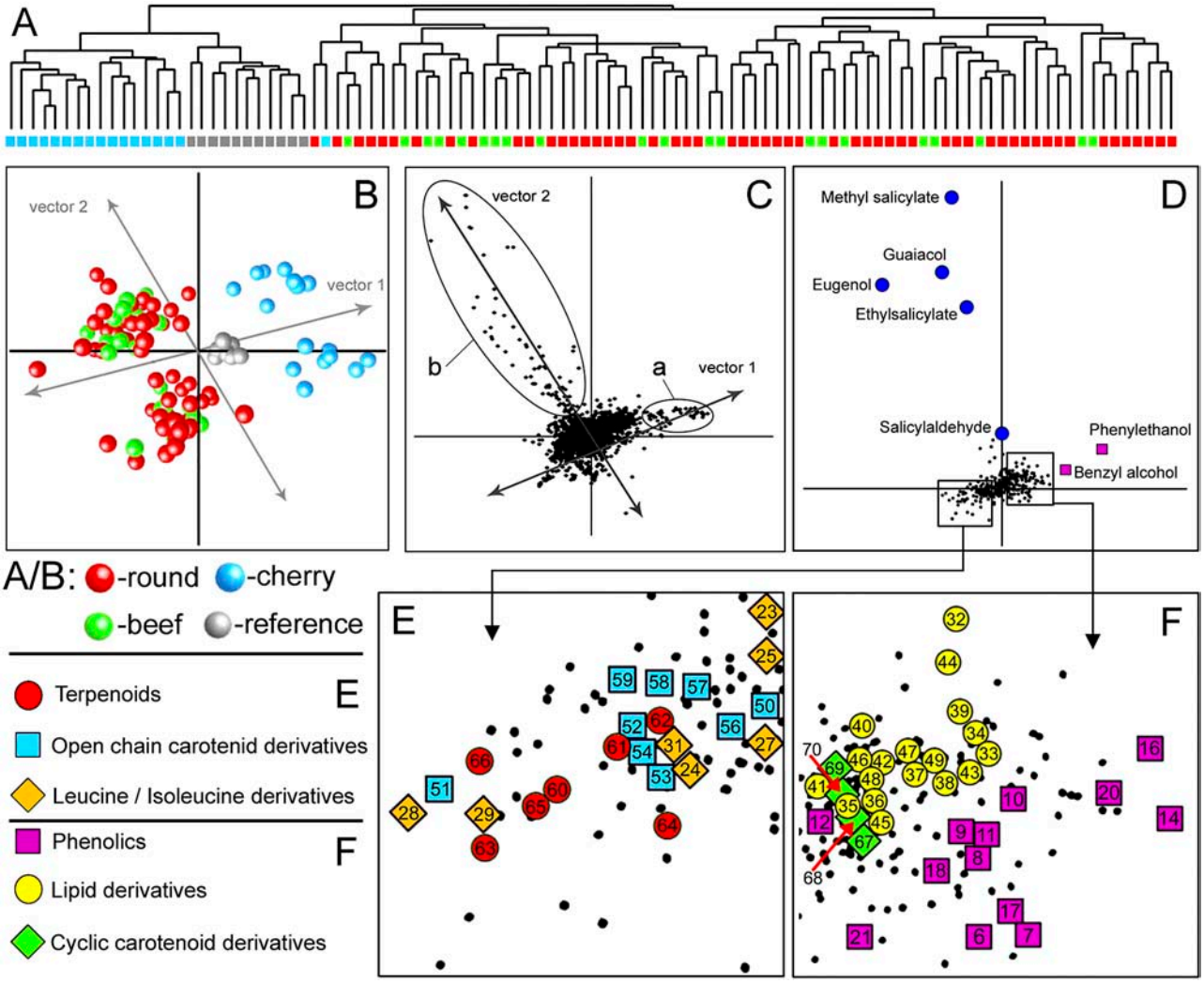

Figure 3. Multivariate analyses of 94 tomato genotypes. A, Hierarchical tree of the 94 tomato genotypes based on intensity patterns of >20,000 individual molecular fragments. B, PCA plot showing two major types of differences between the tomato genotypes: between-type variation, discriminating the cherry tomatoes from round and beef tomatoes along vector 1, and within-type variation, independent of fruit type, along vector 2. C, PCA plot showing the distribution of >20,000 molecular fragments: Those molecular fragments (a) distributed along vector 1 determine the between-type variation, and molecular fragments (b) distributed along vector 2 determine the within-type variation. D, PCA plot showing the distribution of the identified volatile metabolites determining the main differences between the tomato genotypes. E and F, Two enlarged parts of the PCA plot shown in D: Compounds are shown as colored shapes and the numbers refer to the compounds presented in Table II. The smaller black dots represent unknown compounds.

situated within a maximal deviation in retention time of ≤6 s (corresponding to an average peak width at one-half height in the chromatograms we obtained), were considered to belong to the mass spectrum of one and the same metabolite. In total, 322 molecular fragment clusters were obtained. The mass spectra of the 15 key flavor-related tomato volatiles (Baldwin et al., 2000) were in agreement with the mass spectra reconstructed from the molecular fragment clusters at their corresponding retention times, as shown by the example of 2-isobutylthiazole in Figure 4C. This suggests that the 322 molecular fragment clusters each represent the mass spectrum of an individual volatile compound. Overlapping mass spectra of coeluting compounds could also be successfully discriminated from each other using MMSR. Molecular fragments of coeluting compounds were clustered based on the similarity of their patterns throughout the samples and the number of clusters indicated the number of overlapping chemical compounds at a certain retention time (Fig. 4, A and B). In many cases, MMSR allowed extraction of all major fragments of a mass spectrum of a particular coeluting compound (Fig. 4C; Supplemental Data II). In others, it revealed a few compound unique fragments (data not shown). For compound identification, the AMDIS software package, dedicated to chromatogram deconvolution, was used as a bridge to match the compound spectral information derived by

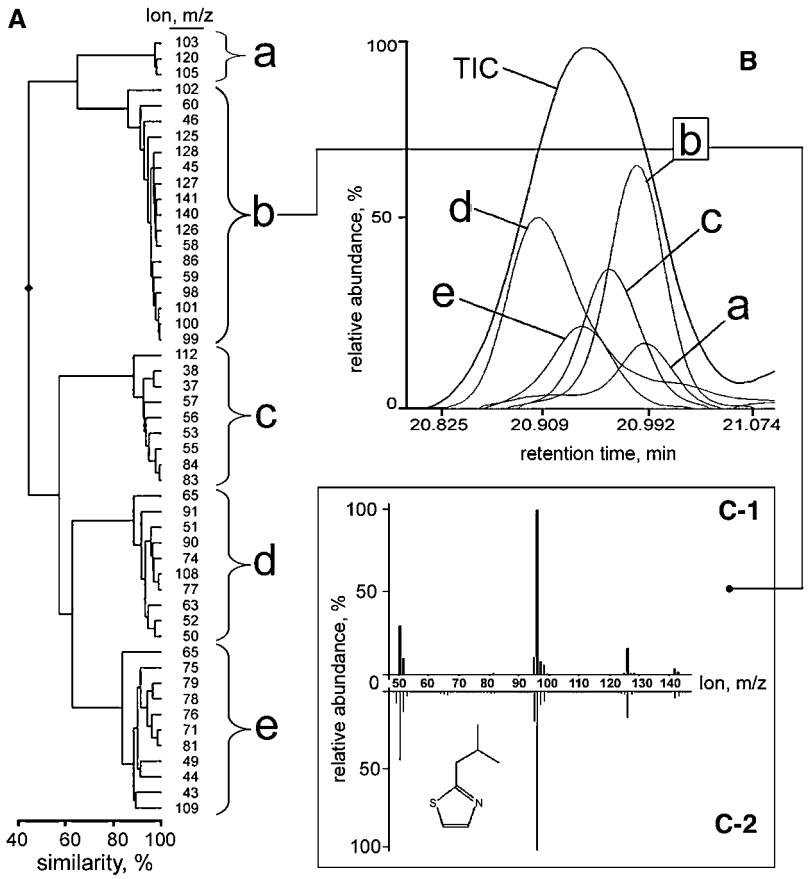

Figure 4. MMSR-driven discrimination of mass spectra. A, Dendrogram showing a clustering of intensity patterns of ions situated in the retention time window 20.8 to 21.07 min into several molecular fragment clusters. B, MMSR indicated the presence of five individual compounds within a visually single total ion count (TIC) peak within the chosen time window. C-1, An experimental mass spectrum, obtained by plotting of the original intensities of the molecular fragments of compound b could be matched to the mass spectrum of the chemical standard analog of 2-isobutylthiazole (C-2), which also has a retention time falling within the chosen window.

MMSR to entries of the National Institute of Standards and Technology (NIST) library of chemical compound mass spectra (as described in "Materials and Methods").

Phe-Derived Volatiles Mostly Explain the Difference in the Composition of the Tomato Fruit Volatile Metabolome

The MMSR and NIST library matching results revealed that the molecular fragments (Fig. 3C), which were most discriminative between the tomato genotypes (Fig. 3B), belonged to two groups of volatile metabolites, derived from the phenolic (depicted in pink) and phenylpropanoid (depicted in blue) pathways (Fig. 3D). Interestingly, both groups originate from the amino acid Phe. Cherry tomatoes could be distinguished from round and beef by a relatively high accumulation of phenolic-derived volatiles (Fig. 3, D and F, vector 1). Two phenolic alcohols, phenylethanol and benzyl alcohol, showed the highest contribution to the cherry versus round/beef contrast. Volatiles derived from the phenylpropanoid pathway, including methyl and ethyl salicylate, guaiacol, eugenol, and salicylaldehyde, were responsible for the division of the genotypes into two groups independent of the tomato fruit type (Fig. 3D, vector 2). Both types of Phe-derived volatiles revealed the largest relative variation across the 94 genotypes (Table I).

Besides phenolic volatiles, cherry tomatoes also contained relatively high levels of lipid derivatives (Fig. 3F) and low levels of terpenoids, open-chain carotenoid derivatives, and Leu/Ile-derived products (Fig. 3E).

Patterns of the 322 Volatiles Are Correlated According to Their Precursor or Metabolic Pathway

The 322 compounds were subjected to HCA using the Pearson correlation coefficient. This revealed the presence of a few major compound clusters, as shown in a correlation matrix (Fig. 5). Compounds situated in the clusters were subjected to a putative identification by matching their mass spectra to the NIST library. Reliable matching results were obtained for 100 of them, of which 70 metabolites had previously been described as tomato fruit volatiles (Petro-Turza, 1987; Table II). The reliability of the identity prediction was assessed through comparison with 46 authentic chemical standards. Each of those standards represented a first hit from the NIST library search (see Supplemental Data I) and together covered a few members of each compound cluster. Of those 46 standards, 43 confirmed the identity of the predicted compound, indicating that the prediction of compound identity was very reliable.

The identification results revealed that each of the compound clusters contained compounds that have a common biochemical precursor or belong to the same

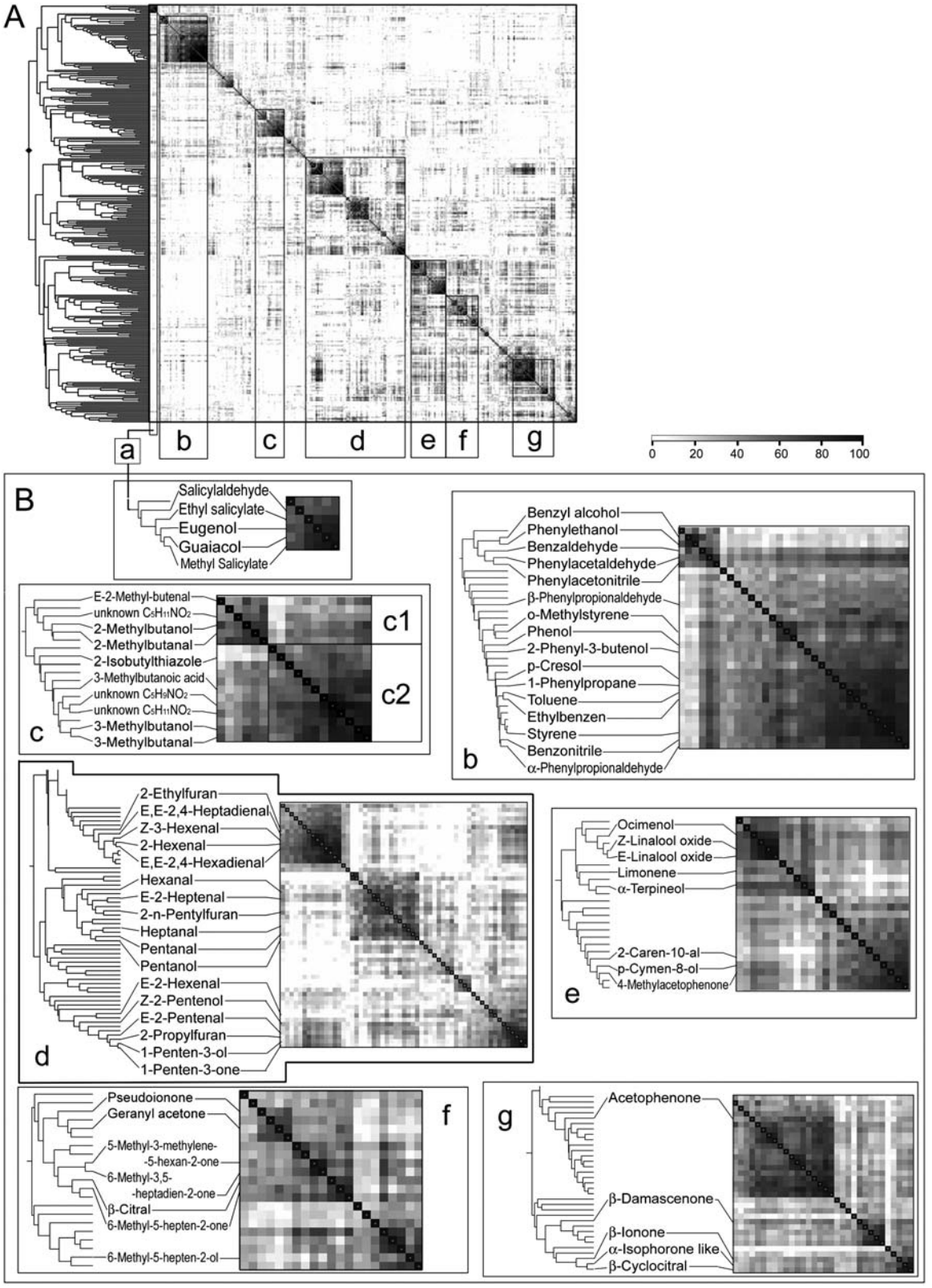

Figure 5. Metabolite-metabolite correlation matrix of the 322 plant-derived compounds. A, The main compound clusters are situated along the diagonal line (groups a–g). Correlations between metabolites are shown in grayscale: the darker the color gray, the higher the percentage of similarity between metabolite expression patterns. B, Detailed dendrogram of each compound cluster with putative compound identity as described in Table II. Compound cluster: a, phenylpropanoid volatiles; b, other phenolic volatiles; c, Leu and Ile derivatives (c1 and c2, respectively); d, lipid derivatives. Isoprenoids: e, terpenoids; f, open-chain carotenoid derivatives; g, cyclic carotenoid derivatives.

Table II. *Putative identity of volatile metabolites present within the clusters obtained using HCA (Fig. 5)*

Metabolites were identified by matching their mass spectra to the NIST library. RT, Retention time; specific ion (*m/z*), mass (*m/z* value) of a compound-specific molecular fragment; identity, putative identity, according to the highest NIST library match; NIST match, matching score (1,000 = 100% identical to the NIST library entry), (+) or (−) after the NIST match value (the NIST match was confirmed [+] or was not confirmed [−] by an authentic chemical standard injection); biochemical group, corresponding cluster in Figure 5.

Comp. No.	Retention Time, min	Specific Ion, *m/z*	Identity	NIST Match	Biochemical Group
1	21.55	122	Salicylaldehyde	838 (+)	a (corresponding to the clusters of Fig. 5)
2	23.06	81	Guaiacol	924 (+)	a
3	26.89	120	Methyl salicylate	960 (+)	a
4	29.32	120	Ethyl salicylate	951	a
5	31.90	164	Eugenol	920 (+)	a
6	10.79	91	Toluene	953	b
7	14.45	91	Ethylbenzene	946 (+)	b
8	15.61	104	Styrene	964 (+)	b
9	18.05	91	1-Phenylpropane	934	b
10	18.34	106	Benzaldehyde	944 (+)	b
11	18.55	94	Phenol	931 (+)	b
12	19.11	118	*p*-Methylstyrene	705	b
13	19.24	103	Benzonitrile	841 (+)	b
14	20.80	117	2-Phenyl-3-buten-ol	782	b
15	20.94	108	Benzyl alcohol	942 (+)	b
16	21.43	120	Phenylacetaldehyde	950 (+)	b
17	22.09	107	*p*-Cresol	951 (+)	b
18	23.63	105	α-Phenylpropionaldehyde	862	b
19	23.95	91	Phenylethanol	944 (+)	b
20	24.82	117	Phenylacetonitrile	856 (+)	b
21	25.66	92	β-Phenylpropionaldehyde	876 (−)	b
22	6.96	44	3-Methylbutanal	837 (+)	c
23	7.21	57	2-Methylbutanal	887 (+)	c
24	9.32	43	3-Methylbutanol	894 (+)	c
25	9.50	57	2-Methylbutanol	922 (+)	c
26	9.77	42	E-2-Methyl-2-butenal	922	c
27	12.77	60	3-Methylbutanoic acid	922 (+)	c
28	15.85	41	3-Methylbutanol nitrite	835 (−)	c
29	16.39	41	Unknown, $C_5H_9NO_2$-like		c
30	16.42	46	Unknown, $C_5H_{11}NO_2$-like		c
31	21.04	99	2-Izobutylthiazole	902 (+)	c
32	7.67	57	1-Penten-3-ol	903	d
33	7.83	55	1-Penten-3-one	882 (+)	d
34	8.18	44	*n*-Pentanal	883 (+)	d
35	8.37	81	2-Ethylfuran	934 (+)	d
36	10.20	55	E-2-Pentenal	926	d
37	10.51	42	1-Pentanol	895 (+)	d
38	10.65	57	Z-2-Penten-1-ol	891	d
39	11.76	80	Z-3-Hexenal	847 (+)	d
40	11.83	72	Hexanal	856 (+)	d
41	13.60	41	2-Hexenal	904	d
42	13.90	98	E-2-Hexenal	944 (+)	d
43	13.95	67	Z-3-Hexenol	899 (+)	d
44	14.36	56	1-Hexenol	932 (+)	d
45	15.75	70	Heptanal	898 (+)	d
46	16.13	81	E,E-2,4-Hexadienal	938 (+)	d
47	17.91	41	E-2-Heptenal	895 (+)	d
48	19.26	81	2-*n*-Pentylfuran	925	d
49	19.44	81	E,E-2,4-Heptadienal	721	d
50	18.99	43	6-Methyl-5-hepten-2-one	919 (+)	f
51	19.17	95	6-Methyl-5-hepten-2-ol	810	f
52	20.07	43	5-Hexen-2-one, 5-methyl-3-methylene	757	f
54	23.41	109	6-Methyl-3,5-heptadien-2-one	916	f
56	28.10	69	β-Citral	906 (+)	f
57	28.98	41	α-Citral	941 (+)	f
58	34.36	43	Geranyl acetone	904 (+)	f

(*Table continues on following page.*)

Table II. (*Continued from previous page.*)

Comp. No.	Retention Time, min	Specific Ion, *m/z*	Identity	NIST Match	Biochemical Group
59	38.06	69	Pseudoionone	711	f
60	20.90	68	Limonene	621 (+)	e
61	22.44	59	Linalool oxide, Z-	876 (+)	e
62	22.99	59	Linalool oxide, E-	799 (+)	e
63	24.91	93	Ocimenol	821	e
64	26.39	43	*p*-Cymen-8-ol	863	e
55	26.51	119	Acetophenone, 4-methyl	913 (+)	e
65	26.69	59	α-Terpineol	889 (+)	e
66	29.92	79	2-Caren-10-al	718	e
67	22.06	82	α-Isophorone	801 (−)	g
53	22.32	105	Acetophenone	928 (+)	g
68	27.80	152	β-Cyclocitral	878 (+)	g
69	32.82	121	β-Damascenone	910	g
70	35.74	177	β-Ionone	851 (+)	g

metabolic pathway (Fig. 5): Phe derivatives (phenolic and phenylpropanoid volatiles), Leu and Ile derivatives, lipid derivatives, and isoprenoid derivatives, consisting of open-chain and cyclic carotenoid breakdown products and terpenes (Buttery and Ling, 1993; Baldwin et al., 2000).

DISCUSSION

High-Throughput Screening of Volatiles

Volatile tomato fruit metabolites have been profiled using headspace SPME-GC-MS, which is a procedure that has been used in the past for many plant matrices including tomato fruits (Song et al., 1998; Deng et al., 2004). SPME is superior to other sampling methods in both speed and robustness (Yang and Peppard, 1994). Only direct thermal desorption exceeds SPME in terms of sensitivity (Pfannkoch and Whitecavage, 2000). In fact, SPME-based methods, as well as other methods based on headspace extraction, are so-called semiquantitative due to the presence of a matrix effect and relatively short linearity of the dynamic range—the drawbacks, which in many cases do not allow an absolute quantification. However, metabolomics, as well as other profiling techniques such as microarray analyses, mostly operate with intensity patterns formed by relative responses, which allow searching for potential differences and performing multivariate comparative analyses. Absolute quantification of the levels of these volatiles will be performed using more sophisticated methods in our future experiments.

For a high-throughput analysis of a large number of biological samples, an automated sequential manipulation of the samples is required. To obtain reliable data in this way, the metabolic composition has to be stable during the entire period of experimentation. This is especially important when analyzing complex native plant materials such as fruit tissue. To develop an automated high-throughput SPME-GC-MS method to screen and profile fruit volatiles of 94 tomato cultivars, the initial focus was placed on 15 volatile metabolites that are of particular importance in determining tomato fruit flavor (Buttery and Ling, 1993; Baldwin et al., 2000). First, we optimized the stability of the metabolites by adding NaOH/EDTA/$CaCl_2$ at the end of the sample preparation procedure. This procedure stabilizes the fruit matrix for at least 12 h, presumably by increasing the pH and exploiting the chelating effect of EDTA to prevent compound oxidation. The method was found to be suitable, reliable, and accurate and enabled the automated measurement of the large numbers of fruit samples required for this investigation.

Full Spectral Alignment Enables Unbiased Comparative Metabolomics

Metabolomics aims to generate a comprehensive overview of the identity and quantity of metabolites in biological materials. The general principle currently used is that all compounds are identified prior to their, often relative, quantification and subsequent comparison throughout the biological samples. When using GC-MS, each chemical compound is classified both on its relative retention time and its mass spectrum. This mass spectrum gives a unique fingerprint of the chemical resulting from its fragmentation on entering the mass spectrometer. However, when using complex plant extracts, despite effective prior chromatographic separation, mass spectra of many compounds inevitably often coelute, thus complicating their discrimination. Consequently, the compound discrimination step (not always unbiased) limits the comprehensiveness of the metabolomic analyses. As an alternative strategy for comparative metabolomic analysis, we propose here a protocol that is based upon an unbiased empirical quantification and search for metabolic differences at the level of molecular fragments (ions) prior to compound identification. This approach avoids the time-consuming need for any prior assignment of chemical information to the molecular structure for hundreds of datasets and thus makes it possible to gain a faster, more unbiased and nontargeted metabolomic overview.

Furthermore, this approach facilitates our desire to home in specifically on those mass peaks that are discriminatory between samples. This approach, however, depends on the initial ability to align the spectral patterns of the tens of thousands of molecular fragments present throughout all the GC-MS datasets to be compared. For this, we used the MetAlign software package to eliminate noise, compensate retention time shifts, and align the mass spectral information. This resulted in a data matrix of about 4,000,000 data points (198 datasets $\times$ 20,000 mass peaks detected). Each row of this data matrix displays the intensities of a unique molecular fragment throughout the 198 GC-MS datasets. However, a number of contaminants resulting from the fiber material can usually be found when using SPME. A multivariate analysis (HCA) of the molecular fragment patterns throughout the GC-MS profiles obtained allowed us to extract the fragments related to the fiber and to remove them from the dataset automatically. The complete mass spectral alignment of metabolic profiles has thus allowed us to perform a reliable, multivariate comparative analysis of the 94 genotypes studied. This analysis revealed both between-fruit-type metabolic differences, discriminating cherry tomatoes from round and beef, as well as within-fruit-type metabolic differences, which were independent of fruit type, and allowed the discrimination of the molecular fragments determining the variation between genotypes. However, to get subsequently biologically relevant information, we have to be able to relate these discriminative molecular fragments to their parent compounds in order to perform a putative identification. To overcome the limitations of metabolite recognition and identification that are due to high metabolome complexity and variability, we developed an approach that allows an automated reconstruction of the mass spectra of individual compounds (MMSR). This approach is based on the fixed ratio of molecular fragment intensities resulting from the fragmentation of a particular molecule. Logically, even if the abundance of a compound varies between samples, the ratios of its molecular fragment intensities derived from the parent molecule should remain the same throughout all the samples. Consequently, when molecular fragments cluster together after being subjected to a multivariate analysis such as HCA and their relative retention time does not exceed a predefined window, it can be concluded that they relate to the same chemical compound.

Using MMSR, we were able to discriminate the full array of chemical compounds present in all datasets using one automated procedure, even in cases of complex overlapping mass spectra (Fig. 4). In comparison, when using AMDIS alone—a software package dedicated to resolve compound overlap cases by means of automated mass spectra deconvolution—for chemical compound discrimination, we were unable to get an equally reliable prediction of the number and chemical identity of overlapping compounds. This is due to their variable mass intensities in the wide range of the different samples (data not shown).

Deconvolution procedures are generally reliable and frequently used to handle individual GC-MS datasets. However, when analyzing hundreds of samples, a limited number of datasets that are assumed to fully represent the compound diversity of the entire sample set analyzed have to be selected for deconvolution visually. The compounds that can be discriminated in these representative datasets are subsequently used for a comparative analysis of the entire sample set. Such procedures, based on a prior mass spectral deconvolution of GC-MS profiles, have been used successfully, and this has allowed the discrimination and identification of many compounds in plant extracts (Taylor et al., 2002). However, in contrast to this conventional procedure, MMSR is not limited and uses all available spectral information, thus allowing discrimination and recognition of all individual compounds based on their variability patterns. This significantly improves comprehensiveness, since even when a particular compound is only abundant in one of the 94 samples it will still be included in the analysis.

In our tomato study, a total of 322 tomato volatile compounds could be discriminated in 198 datasets. This is approximately 80% of all the volatile metabolites (>400 different volatiles) that have so far been reported in tomato fruit (Petro-Turza, 1987). The multivariate analyses (HCA, PCA) revealed that most of the compounds, which could be identified, clustered on the basis of their biochemical nature (Fig. 5) and the entire metabolic organization could be characterized by the existence of a few large compound groups, which unite (e.g. lipid derivatives, phenolic and phenylpropanoid volatiles, isoprenoids, etc.). The main metabolite clusters could subsequently be divided into smaller subclusters. For example, the compounds of cluster c (Fig. 5B) could be clearly divided into two distinct subclusters based on their biochemical precursors, Leu and Ile. Interestingly, volatiles derived from Leu include, besides alcohols (e.g. 3-methylbutanol) and aldehydes (e.g. 3-methylbutanal), a number of nitrogen-containing compounds (yet to be identified) and even a sulfur-containing heterocyclic compound, 2-isobutylthiazole, which are all known to be Leu derived (Buttery and Ling, 1993).

All isoprenoid volatiles can be roughly separated into three subclusters representing terpenoids, open-chain, and cyclic carotenoid derivatives (Fig. 5B, groups e, f, and g, respectively). Interestingly, the terpenoids α- and β-citral appeared in the group of open-chain carotenoid derivatives. This is in line with previous observations that the citral isomers may be derived as a degradation product of lycopene (Cole and Kapur, 1957; Schreier et al., 1977). For several other compounds, such as acetophenone and 4-methylacetophenone, the biosynthetic pathway is still unclear and their clustering with terpenoids and cyclic carotenoid volatiles, respectively, may shed new light on their biochemical origin.

Mathematical analyses of metabolic pathway databases of many organisms have led to the concept of

hierarchical modularity in the organization of metabolic networks. This concept implies that cellular functionality is organized in a set of functional modules, which consequently are organized in a few large modules, which in turn can be grouped into even larger modules (Jeong et al., 2000; Ravasz et al., 2002; Ihmels et al., 2004). The hierarchical modularity of metabolic network organization would allow robust, error-tolerant, and energetically efficient functioning of biological systems. Our experimental results based on an analysis of metabolic expression patterns clearly reflect the features of this concept: structurally related metabolites resulting from different enzymatic or nonenzymatic reactions, but originating from a common metabolic precursor, were clustered into groups and subgroups representing distinct metabolic pathways. This modularity may be due to the existence of a coordinate regulation of these metabolic pathways, e.g. by specific transcription factors activating the expression of the structural genes in a pathway. It may also reflect regulation by the activity of key enzymes, which determines the flux through the downstream pathway or the availability of metabolite precursors at the beginning of a metabolic pathway. Although functional implications of this modular clustering still remain to be elucidated, an existence of such functional modularity can be assumed for the group of phenylpropanoid metabolites and their derivatives, as seen here for tomato, since phenylpropanoid metabolism is known to contribute to plant stress responses (Dixon and Paiva, 1995) and methyl salicylate has been shown as an airborne signaling agent in plant pathogen resistance (Shulaev et al., 1997; Seskar et al., 1998). In this light, it is possible that genotypes with increased levels of these compounds may have been selected through the years for their increased capability to respond to biotic or abiotic stress.

CONCLUSION

The high-resolution, comprehensive, and unbiased strategy for metabolomic data analysis presented here is novel and opens new directions of discovery in the field of metabolomics. Full mass spectral alignment of GC-MS metabolic profiles followed by a universal strategy for chemical compound discrimination has allowed us to perform a high-resolution, unbiased, and fast multivariate comparative analysis of volatile biochemical composition of 94 tomato genotypes (198 complex plant extracts) based on metabolic information derived by the analytical method. The large-scale picture of the volatile part of the tomato fruit metabolome reflects the hierarchical modularity of metabolism organization that is assumed to be common for different levels of a biological system. Further projecting the data into data from other ''omics'' technologies will pave the way for a true systems biology approach to investigating cell networks and more directed gene discovery.

The main goal of this study was to describe this novel efficient approach for unbiased analysis of complex biochemical datasets. A detailed biological interpretation of the data obtained is beyond the scope of this article, but it is anticipated that this will provide much new information on the heterogeneity in biochemical composition within tomato varieties, and this will be the subject of our future investigations.

MATERIALS AND METHODS

Plant Material

Ninety-four tomato (*Lycopersicon esculentum* Mill.) genotypes were obtained from six different tomato seed companies, each with its own breeding program. As such, the cultivars should represent a considerable collection of genetic and therefore phenotypic variation, not just between tomato types (cherry, round, and beef), but also within the individuals of each type. This study was deliberately performed blind. We only received information from the tomato breeders of the companies supplying the material concerning the tomato fruit types and not their genetic background. For classification, breeders generally use a combination of (1) fruit diameter and (2) number of locules in the fruit (fl). For the latter, the criteria were as follows: cherry-type fl = 2; round fl = 3; beef fl = 4 or more. All cultivars were grown in the summer of 2003 under greenhouse conditions at a single location in Wageningen, The Netherlands. Nine plants, randomly distributed over three adjacent greenhouse compartments, were grown for each cultivar. Pink-staged tomato fruits of all plants were picked on two consecutive days. To mimic the conditions from the farm to the fork, fruits were stored for 1 week at 15°C and turned to 20°C at 24 h prior to freezing. During this period, the fruits continued to ripen slowly and, at the moment of sampling, the fruits were fully red ripe, resembling the conditions at the time of consumption. 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 fruit material was immediately frozen in liquid nitrogen, ground in an analytical electric mill, and stored at −80°C before analyses.

Standard Chemicals

Fifteen analytical grade chemicals (all obtained from Sigma) were used as authentic standards to optimize the SPME-GC-MS method for automated sequential analysis of hundreds of samples. These were cis-3-hexenal, β-ionone, hexanal, 1-penten-3-one, 2-methylbutanal, 3-methylbutanal, trans-2-hexenal, 2-izobutylthiazole, trans-2-heptenal, phenylacetaldehyde, 6-methyl-5-hepten-2-one, cis-3-hexenol, 2-phenylethanol, 3-methylbutanol, and methyl salicylate. For metabolite identification, an additional set of standards was used. These include 2-methylbutanol, 3-methylbutanoic acid, 3-methylbutanol nitrite, 1-hexanol, pentanal, 2-ethylfuran, 1-pentanol, heptanal, E,E-2,4-hexadienal, salicylaldehyde, eugenol, guaiacol, ethylbenzene, styrene, benzaldehyde, benzonitrile, benzyl alcohol, phenylacetonitrile, β-phenylpropionaldehyde, phenol, *p*-cresol, acetophenone, 4-methylactophenone, geranylacetone, α-isophorone, β-cyclocitral, α-citral, β-citral, limonene, cis- and trans-linalool oxide, and α-terpineol.

Sample Preparation Procedure and Headspace SPME-GC-MS Analysis

Frozen fruit powder (1 g fresh weight) was weighed in a 5-mL screw-cap vial, closed, and incubated at 30°C for 10 min. An EDTA-NaOH water solution was prepared by adjusting of 100 mM EDTA to a pH of 7.5 with NaOH. Then, 1 mL of the EDTA-NaOH solution was added to the sample to a final EDTA concentration of 50 mM. Solid $CaCl_2$ was then immediately added to give a final concentration of 5 M. The closed vials were then sonicated for 5 min. A 1-mL aliquot of the pulp was transferred into a 10-mL crimp cap vial (Waters), capped, and used for SPME-GC-MS analysis.

Each of the 94 tomato fruit samples was analyzed using two replicated aliquots. In total, 22 freshly prepared samples were measured per day (two series of 11 samples). In addition, reference tomato samples were made by mixing fruit powders from several genotypes of the round, beef, and cherry

fruit phenotypes. This mixture was routinely analyzed every day of experimentation as an external control in order to monitor the stability of the analytical system. The samples were automatically extracted and injected into the GC-MS via a Combi PAL autosampler (CTC Analytics AG). Headspace volatiles were extracted by exposing a 65-μm polydimethylsiloxane-divinylbenzene SPME fiber (Supelco) to the vial headspace for 20 min under continuous agitation and heating at 50°C. The fiber was inserted into a GC 8000 (Fisons Instruments) injection port and volatiles were desorbed for 1 min at 250°C. Chromatography was performed on an HP-5 (50 m $\times$ 0.32 mm $\times$ 1.05 μm) column with helium as carrier gas (37 kPa). The GC interface and MS source temperatures were 260°C and 250°C, respectively. The GC temperature program began at 45°C (2 min), was then raised to 250°C at a rate of 5°C/min, and finally held at 250°C for 5 min. The total run time, including oven cooling, was 60 min. Mass spectra in the 35 to 400 *m/z* range were recorded by an MD800 electron impact MS (Fisons Instruments) at a scanning speed of 2.8 scans/s and an ionization energy of 70 eV. The chromatography and spectral data were evaluated using Xcalibur software (http://www.thermo.com).

Data Analyses: Multivariate Comparative Analysis and MMSR

1. For automated baseline correction, mass spectra extraction, and subsequent spectral data alignment, in total 198 GC-MS datasets were processed simultaneously using the dedicated MetAlign metabolomics software package (http://www.metalign.nl; Fig. 2A).

2. The metabolic profiles aligned were subjected to multivariate analyses: HCA (Pearson correlation coefficient was used) and PCA to search for metabolic differences between the tomato genotypes at the level of molecular fragments (Fig. 2, A–C). The multivariate analyses were performed using the GeneMaths software package (http://www.applied-maths.com). A $\log_2$ transformation was applied to the data prior to the multivariate analyses.

3. MMSR was used to assign the molecular fragments to compounds. For this, the patterns of all molecular fragments were subjected to HCA. Those molecular fragments that revealed a Pearson correlation equal to or more then 0.8 and were situated within a 6-s retention time window (which corresponds to an average peak width at one-half height in the chromatograms we obtained) were considered as belonging to the spectrum of one compound.

4. For compound identification, the following steps were used: (1) for each compound selected for putative identification, the most optimal chromatogram is selected with respect to relative abundance and overlap with other compounds at the specific position; (2) for each selected compound, specific molecular fragments (ions, *m/z*) were selected from the corresponding fragment cluster derived by MMSR; (3) the selected fragments were used as a basis for deconvolution of the chromatographic peak at the corresponding retention time using AMDIS (Stein, 1999). Mass spectral models derived in this way were matched to the NIST mass spectral library (http://www.nist.gov).

ACKNOWLEDGMENTS

The authors are grateful to Syngenta Seeds, Seminis, Enza Zaden, Rijk Zwaan, Nickerson-Zwaan, and De Ruiter Seeds for providing seeds of the 94 tomato cultivars. We would like to thank Mrs. Fien Meijer-Dekens, Mrs. Petra van den Berg, Dr. A.W. van Heusden, and Dr. Pim Lindhout for excellent greenhouse management and plant cultivation, and Dr. Harro Bouwmeester and Mr. Francel Verstappen for helpful discussions and technical support.

Received July 6, 2005; revised September 13, 2005; accepted September 13, 2005; published November 11, 2005.

LITERATURE CITED

Arimura G, Ozawa R, Nishioka T, Boland W, Koch T, Kuhnemann F, Takabayashi J (2002) Herbivore-induced volatiles induce the emission of ethylene in neighboring lima bean plants. Plant J **29:** 87–98

Augusto F, Valente ALP, dos Santos Tada E, Rivellino SR (2000) Screening of Brazilian fruit aromas using solid-phase microextraction-gas chromatography-mass spectrometry. J Chromatogr A **873:** 117–127

Baldwin EA, Goodner K, Plotto A, Protchett K, Einstein M (2004) Effect of volatiles and their concentration on perception of tomato descriptors. J Food Sci **69:** 310–318

Baldwin EA, Scott WJ, Shewmaker CK, Schuch W (2000) Flavor trivia and tomato aroma: biochemistry and possible mechanisms for control of important aroma components. HortScience **35:** 1013–1022

Bezman Y, Mayer F, Takeoka GR, Buttery RG, Ben-Oliel G, Rabinowitch HD, Naim M (2003) Differential effects of tomato (*Lycopersicon esculentum* Mill) matrix on the volatility of important aroma compounds. J Agric Food Chem **51:** 722–726

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

Buttery RG, Ling LC (1993) Volatiles of tomato fruits and plant parts: relationship and biogenesis. *In* R Teranishi, R Buttery, H Sugisawa, eds, Bioactive Volatile Compounds from Plants. ACS Books, Washington, DC, pp 23–24

Cole ER, Kapur NS (1957) The stability of lycopene. I. Degradation of oxygen. II. Oxidation during heating of tomato pulps. J Sci Food Agric **8:** 360–368

Deng C, Zhang X, Zhu W, Qian J (2004) Investigation of tomato plant defense response to tobacco mosaic virus by determination of methyl salicylate with SPME-capillary GC-MS. Chromatographia **59:** 263–268

Desbrosses GG, Kopka J, Udvardi MK (2005) *Lotus japonicus* metabolic profiling. Development of gas chromatography-mass spectrometry resources for the study of plant-microbe interactions. Plant Physiol **137:** 1302–1318

Dicke M, Agrawal AA, Bruin J (2003) Plants talk, but are they deaf? Trends Plant Sci **8:** 403–405

Dixon RA, Paiva NL (1995) Stress-induced phenylpropanoid metabolism. Plant Cell **7:** 1085–1097

Dudareva N, Pichersky E, Gershenzon J (2004) Biochemistry of plant volatiles. Plant Physiol **135:** 1893–1902

Engelberth J, Alborn HT, Schmelz EA, Tumlinson JH (2004) Airborne signals prime plants against insect herbivore attack. Proc Natl Acad Sci USA **101:** 1781–1785

Fiehn O, Kopka J, Dofmann P, Altmann T, Trethewey RN, Willmitzer L (2000a) Metabolite profiling for plant functional genomics. Nat Biotechnol **18:** 1157–1161

Fiehn O, Kopka J, Trethewey RN, Willmitzer L (2000b) Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. Anal Chem **72:** 3573–3580

Fraga CG, Prazen BJ, Synovec RE (2001) Objective data alignment and chemometric analysis of comprehensive two-dimensional separations with run-to-run peak shifting on both dimensions. Anal Chem **73:** 5833–5840

Hall RD, de Vos CHR, Verhoeven HA, Bino RJ (2005) Metabolomics for the assessment of functional diversity and quality traits in plants. *In* G Harrigan, S Vaidyanathan, R Goodacre, eds, Metabolic Profiling. Kluwer Academic Publishers, Dordrecht, The Netherlands (in press)

Holtorf H, Guitton M-C, Reski R (2002) Plant functional genomics. Naturwissenschaften **89:** 235–249

Huhman DV, Sumner LW (2002) Metabolic profiling of saponins in Medicago sativa and Medicago truncatula using HPLC coupled to an electrospray ion-trap mass spectrometer. Phytochemistry **59:** 347–360

Ihmels J, Levy R, Barkai N (2004) Principals of transcriptional control in the metabolic network of Saccharomyces cerevisiae. Nat Biotechnol **22:** 86–92

Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi A-L (2000) The large-scale organization of metabolic networks. Nature **407:** 651–654

Johnson KJ, Wright BW, Jarman KH, Synovec RE (2003) High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis. J Chromatogr A **994:** 141–155

Jonsson P, Gullberg J, Nordström A, Kusano M, Kowalczyk M, Sjöström M, Moritz T (2004) A strategy for identifying differences in large series of metabolomic samples analysed by GC/MS. Anal Chem **76:** 1738–1745

Kopka J, Fernie A, Weckwerth W, Gibon Y, Stitt M (2004) Metabolite profiling in plant biology: platforms and destinations. Genome Biol **5:** 109

Krumbein A, Peters P, Brukner B (2004) Flavour compounds and quantitative descriptive analysis of tomatoes (Lycopersicon esculentum Mill.) of different cultivars in short-term storage. Postharvest Biol Technol **32:** 15–28

Liechti R, Farmer EE (2002) The jasmonate pathway. Science **296:** 1649–1650

Matich AJ, Rowan DD, Banks NH (1996) Solid phase microextraction for quantitative headspace sampling of apple volatiles. Anal Chem **68:** 4114–4118

Nielsen NPV, Carstensen MJ, Smedsgaard J (1998) Alignment of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. J Chromatogr A **805:** 17–35

Ozawa R, Arimura G, Takabayashi J, Shimoda T, Nishioka T (2000) Involvement of jasmonate- and salicylate-related signaling pathways for the production of specific herbivore-induced volatiles in plants. Plant Cell Physiol **41:** 391–398

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

Pfannkoch E, Whitecavage J (2000) Comparison of the sensitivity of static headspace GC, solid phase microextraction, and direct thermal extraction for analysis of volatiles in solid matrices. *In* AppNote 6/2000. Gerstel GmbH & Co. KG, Mulheim an der Ruhr, Germany, http://www.gerstel.com/an-2000-06.pdf (January 6, 2004)

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi A-L (2002) Hierarchical organization of modularity in metabolic networks. Science **297:** 1551–1555

Roessner U, Luedemann A, Brust D, Fiehn O, Linke T, Willmitzer L, Fernie AR (2001) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. Plant Cell **13:** 11–29

Roessner U, Wagner C, Kopka J, Trethewey RN, Willmitzer L (2000) Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. Plant J **23:** 131–142

Roessner-Tunali U, Hegemann B, Lytovchenko A, Carrari F, Bruedigam C, Granot D, Fernie AR (2003) Metabolic profiling of transgenic tomato plants overexpressing hexokinase reveals that the influence of hexose phosphorylation diminishes during fruit development. Plant Physiol **133:** 84–99

Ruiz JJ, Alonso A, García-Martínez S, Valero M, Blasco P, Ruiz-Bevia F (2005) Quantitative analysis of flavour volatiles detects differences among closely related traditional cultivars of tomato. J Sci Food Agric **85:** 54–60

Ryu CM, Farag MA, Hu CH, Reddy MS, Kloepper JW, Pare PW (2004) Bacterial volatiles induce systemic resistance in Arabidopsis. Plant Physiol **134:** 1017–1026

Schreier P, Drawert F, Junker A (1977) The quantitative composition of natural and technologically changed aromas of plants. IV. Enzymic and thermal reaction products formed during the processing of tomatoes. Z Lebensm Unters Forsch **165:** 23–27

Seskar M, Shulaev V, Raskin I (1998) Endogenous methyl salicylate in pathogen-inoculated tobacco plants. Plant Physiol **116:** 387–392

Shulaev V, Silverman P, Raskin I (1997) Airborne signalling by methyl salicylate in plant pathogen resistance. Nature **385:** 718–721

Simkin AJ, Schwartz SH, Auldridge M, Taylor MG, Klee HJ (2004) The tomato carotenoid cleavage dioxygenase 1 genes contribute to the formation of the flavor volatiles beta-ionone, pseudoionone, and geranylacetone. Plant J **40:** 882–892

Song J, Fan L, Beaudry RM (1998) Application of solid phase microextraction and gas chromatography/time-of-flight mass spectrometry for rapid analysis of flavor volatiles in tomato and strawberry fruits. J Agric Food Chem **46:** 3721–3726

Song J, Gardner BD, Holland JF, Beaudry RM (1997) Rapid analysis of volatile flavor compounds in apple fruit using SPME and GC/time-of-flight mass spectrometry. J Agric Food Chem **45:** 1801–1807

Stein SE (1999) An integrated method for spectrum extraction and compound identification from GCMS data. J Am Soc Mass Spectrom **10:** 770–781

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

Tandon KS, Baldwin EA, Scott JW, Shewfelt RL (2003) Linking sensory descriptors to volatile and nonvolatile components of fresh tomato flavor. J Food Sci **68:** 2366–2371

Taylor J, King DR, Altmann T, Fiehn O (2002) Application of metabolomics to plant genotype discrimination using statistics and machine learning. Bioinformatics **18:** 241–248

Tolstikov VV, Fiehn O (2002) Analysis of highly polar compounds of plant origin: combining of hydrophilic interaction chromatography and electrospray ion trap spectrometry. Anal Biochem **301:** 298–307

Verdonk JC, de Vos CHR, Verhoeven HA, Haring MA, van Tunen AJ, Schuurink RC (2003) Regulation of floral scent production in petunia revealed by targeted metabolomics. Phytochemistry **62:** 997–1008

Verhoeven HA, Beuerle T, Schwab W (1997) Solid-phase micro extraction: artefact formation and its avoidance. Chromatographia **46:** 63–66

Weckwerth W, Fiehn O (2002) Can we discover novel pathways using metabolomic analysis? Curr Opin Biotechnol **13:** 156–160

Wiener MC, Sachs JR, Deyanova EG, Yates NA (2004) Differential mass spectrometry: a label-free LC-MS method for finding significant differences in complex peptide and protein mixtures. Anal Chem **76:** 6085–6094

Willse A, Belcher AM, Preti G, Wahl JH, Thresher M, Yang P, Yamazaki K, Beauchamp GK (2005) Identification of major histocompatibility complex-regulated body odorants by statistical analysis of a comparative gas chromatography/mass spectrometry experiment. Anal Chem **77:** 2338–2347

Yang X, Peppard T (1994) Solid-phase microextraction for flavor analysis. J Agric Food Chem **42:** 1925–1930

Yilmaz E, Tandon KS, Scott JW, Baldwin EA, Shwfelt RL (2001) Absence of a clear relationship between lipid pathway enzymes and volatile compounds in fresh tomatoes. J Plant Physiol **158:** 1111–1116