JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

Discriminating between Cultivars and Treatments of Broccoli Using Mass Spectral Fingerprinting and Analysis of Variance−Principal Component Analysis

DEVANAND L. LUTHRIA,*,† LONG-ZE LIN,† REBECCA J. ROBBINS,†,# JOHN W. FINLEY,§,⊥ GARY S. BANUELOS,|| AND JAMES M. HARNLY†

Food Composition and Methods Development Laboratory, Beltsville Human Nutrition Research Center, Agricultural Research Service, U.S. Department of Agriculture, Beltsville, Maryland 20705; Grand Forks Human Nutrition Research Center, Agricultural Research Service, U.S. Department of Agriculture, Grand Forks, North Dakota 58202; and Water Management Research Laboratory, Agricultural Research Service, U.S. Department of Agriculture, Parlier, California 93648

Metabolite fingerprints, obtained with direct injection mass spectrometry (MS) with both positive and negative ionization, were used with analysis of variance−principal components analysis (ANOVA-PCA) to discriminate between cultivars and growing treatments of broccoli. The sample set consisted of two cultivars of broccoli, Majestic and Legacy, the first grown with four different levels of Se and the second grown organically and conventionally with two rates of irrigation. Chemical composition differences in the two cultivars and seven treatments produced patterns that were visually and statistically distinguishable using ANOVA-PCA. PCA loadings allowed identification of the molecular and fragment ions that provided the most significant chemical differences. A standardized profiling method for phenolic compounds showed that important discriminating ions were not phenolic compounds. The elution times of the discriminating ions and previous results suggest that they were common sugars and organic acids. ANOVA calculations of the positive and negative ionization MS fingerprints showed that 33% of the variance came from the cultivar, 59% from the growing treatment, and 8% from analytical uncertainty. Although the positive and negative ionization fingerprints differed significantly, there was no difference in the distribution of variance. High variance of individual masses with cultivars or growing treatment was correlated with high PCA loadings. The ANOVA data suggest that only variables with high variance for analytical uncertainty should be deleted. All other variables represent discriminating masses that allow separation of the samples with respect to cultivar and treatment.

KEYWORDS: Broccoli; *Brassica oleracea*; spectral fingerprinting; analysis of variance; principal component analysis; ANOVA-PCA; direct injection mass spectrometry

INTRODUCTION

Metabolite fingerprinting is a potentially powerful tool for the rapid analysis of foods and nutritional and herbal supplements (*1*−*7*). The ever-changing nature of the U.S. food and supplement market renders classical methods of analysis inadequate for maintaining accurate and timely databases. There is a constant flux of new produce, new cultivars of familiar fruits and vegetables, and new formulations of prepared foods. The supplement market is even more challenging with the availability of many new botanicals, herbal supplements, and variety of formulations with additional nutrients and biologically active compounds. The development of new analytical tools is necessary for the rapid identification of compounds in foods and characterization and classification of foods to keep up with the dynamic market place.

Metabolite fingerprinting is an untargeted method used to identify chemical patterns in organisms without identification or quantification of specific components (*1*−*6*). Fingerprints are acquired without any chromatographic separation by direct analysis of the solid sample or the sample extract. Detection methods include Fourier transform infrared (FT-IR), mass (MS), nuclear magnetic resonance (NMR), and ultraviolet (UV)

* Corresponding author [telephone (301) 504-7247; fax (301) 504-8314; e-mail d.luthria@ars.usda.gov].

† Food Composition and Methods Development Laboratory, U.S. Department of Agriculture.

Present address: Analytical and Applied Sciences Group, Mars Snackfood US, LLC, 800 High St., Hackettstown, NJ 07840.

§ Grand Forks Human Nutrition Research Center, U.S. Department of Agriculture.

⊥ Present address: 150A Domorah Dr., Montgomeryville, PA 18936.

|| Water Management Research Laboratory, U.S. Department of Agriculture.

10.1021/jf801606x
Published on Web 10/09/2008

absorption spectrometry (*4*−*8*). Pattern recognition analysis is necessary to interpret the data and may be unsupervised, for example, principal component analysis (PCA) or hierarchical clustering analysis (HCA), or supervised, for example, discriminate analysis (DA) or partial least-squares (PLS) (*1*−*3*). In the case of plants, fingerprinting with pattern recognition analysis has been used to discriminate between plant genus, species, and genotypes (*5*−*8*).

PCA allows discriminating variables to be identified by their loadings. For MS fingerprints, this means that discriminating ions and, hence, corresponding compounds can be potentially identified (*7*). This is an attractive possibility that would allow metabolite fingerprinting with pattern recognition analysis to be combined with metabolite identification. However, identification of compounds based strictly on direct injection MS is difficult. All molecular and fragment ions are formed simultaneously, and their relationship is undetermined. With chromatographic profiling, the relationship between molecular and fragment ions and UV spectra is well established. Thus, chromatographic profiling provides more information on which to base compound identification and will most likely be required for identification of most compounds.

A new method for the analysis of fingerprints has been developed that combines analysis of variance with principal component analysis (ANOVA-PCA) (*8*−*10*). One of the appeals of PCA, in general, is that it allows visual, as well as statistical, analysis of the data. However, as more variables are incorporated into the experimental design, data evaluation becomes more complicated. With ANOVA-PCA, submatrices are constructed that isolate each experimental variable and simplify PCA (*9*). ANOVA-PCA score plots provide separation based on the first principal component and make visual inspection and statistical analysis very easy. Because prior knowledge of the analytical variables is used to construct the submatrices, ANOVA-PCA is a supervised pattern recognition technique.

The submatrices constructed for ANOVA-PCA also make it possible to compute the relative variance associated with each experimental variable (*8*). Relative variance can be calculated for the whole spectra or for individual masses. In general, the directions of maximum variance correspond to the directions of maximum information (*11*). Masses with high variance, other than that associated with analytical uncertainty, will provide maximum information. Thus, PCA loadings and relative variance will provide similar information toward identifying key masses.

Recently, ANOVA-PCA was used to analyze UV fingerprints (molecular absorption in the ultraviolet region from 200 to 400 nm) of two cultivars of broccoli grown under seven distinctly different growing conditions (treatments) (*8*). These broccoli samples had previously been shown to have significant differences in levels of glucosinolates, phenolic acids, and free amino acids (*12*−*14*). We were able to construct score plots that allowed visual distinction between the cultivars and each pair of the seven treatments (growing conditions). On the basis of the integrated spectra, the relative variance contributions were calculated to be 33% for cultivar, 66% for treatment, and 1% for analytical uncertainty. It was also possible to identify wavelength intervals that had minimal relative analytical uncertainty and contributed strongly to the variance associated with either cultivar or treatment, although the low resolution of the method made it impossible to identify specific compounds. The PCA spectral fingerprinting approach may become an important tool for food-source verification and evaluation of adulteration in nutraceutical, botanical, and dietary supplement formulations.

The study reported here applied ANOVA-PCA to fingerprints obtained from direct injection MS of aqueous methanol extracts of the same broccoli samples previously analyzed (*8*). Loadings from PCA and high treatment variance from ANOVA analyses of the submatrices were used to identify masses of interest. These masses were correlated with peaks obtained using a standardized phenolic profiling method based on liquid chromatography with diode array and mass spectrometric detection (LC-DAD-ESI/MS) developed in our laboratory (*15*). Relative variances for cultivar and sample treatment were computed and compared to results obtained for UV fingerprinting. The relationship between PCA loadings and relative variance for the experimental parameters was examined.

MATERIALS AND METHODS

Plant Materials. Samples were freeze-dried and powdered composites of two varieties of broccoli (*Brasscia oleracea*): Majestic, provided by Dr. John W. Finley (ARS, USDA); and Legacy, provided by Dr. Gary Banuelos (ARS, USDA).

Greenhouse Study. The cv. Majestic broccoli was grown in a greenhouse with four different concentrations of sodium selenate as previously described (*12*). Approximately 2 weeks prior to head formation, 10 mL of four concentrations of sodium selenate (0, 0.17, 0.52, and 5.2 mM) was applied to the developing plants in pots every other day for 8 days. Then 20 mL of sodium selenate solution of each concentration was applied every other day for two additional applications. These treatments with various concentrations of sodium selenate applied resulted in 0.4, 5.7, 98.6, and 879.2 μg/g of selenium (dry weight) in the broccoli florets. In the text, the samples from the four selenium (Se) growing conditions are referred to as 0, 5, 100, and 1000 ppm, respectively.

Field Study. The cv. Legacy broccoli was obtained from field studies from two different 4 ha field sites in central California (Harris Farms, Five Points, CA), one field using conventional farming methods and the other field using organic farming methods on a certified organic field (*13*). Both farms represented typical organic and conventional broccoli production in the Central California Valley Region, where the soil type is classified as Panoche clay loam. Conventionally and organically grown broccoli was planted by direct seed at both sites, and water was initially applied with a sprinkler irrigation system for the first 30 days. After this interval, water was provided by surface drip irrigation (T-tape drip line, T-Systems Int., San Diego, CA) for the remainder of the season until harvest. For conventionally grown broccoli, two irrigation levels were used representing 100 and 80% of the evapotranspiration (Eto) rate reported by the Westlands California Irrigation Management Information System weather station. Organically grown broccoli was produced using a single level of irrigation at 100% Eto rate. In the text, these three growing conditions are referred to as C100, C80, and Org, respectively.

Broccoli plants were harvested at each site, and samples for each growing condition were processed separately. Samples from field crops were collected for at least four growing seasons (*13*). Whole plants were separated into leaf, stems, and florets. Broccoli florets were then freeze-dried and later coarsely ground in food processors and composited. Ground samples were kept below −20 °C. Prior to analysis or extraction, samples were sieved through standard 20 mesh sieves (particle size < 0.850 mm) to obtain uniform homogenized particle size sample.

Chemicals. HPLC-grade MeOH was purchased from Fisher Chemicals (Fair Lawn, NJ). HPLC-grade acetone was purchased from Burdick & Jackson (Muskegon, MI). Deionized water (18.2 MΩ·cm) was obtained in-house using a Nanopure diamond analytical ultrapure water purification system (model D11901, Branstead Internationals, Dubuque, IA). Poly(vinylidene difluoride) (PVDF) syringe filters with pore size $= 0.45$ μm were procured from National Scientific Co. (Duluth, GA).

Extractions. The weighed freeze-dried and powdered broccoli samples were placed in a 16 × 125 mm screw-cap vial with 5 mL of MeOH/H_2O (60:40, % v/v) (*15*). The mixture was sonicated in an ultrasonic bath (Branson 2510, Branson Ultrasonic Corp., Danbury, CT) at 40 °C for 30 min. The mixture was centrifuged (model GT2, West Chester, PA) at low speed (5000 rpm) for 10 min. The supernate was transferred into a separate vial, and the residue was extracted two more times with 2.5 mL of fresh MeOH/H_2O (60:40, % v/v). The volume of the combined extract was adjusted to 10 mL with MeOH/H_2O (60:40, % v/v). All extracts were stored in 2 mL of HPLC vials under nitrogen at −70 °C until analyzed. An appropriate aliquot of an extract was filtered using PVDF syringe filters (pore size = 0. 45 μm) prior to UV and MS analysis. Each of the seven growing conditions (0, 5, 100, and 1000 ppm, C100, C80, and Org) was extracted five times.

Instrumentation. All data were acquired with an Agilent 1100 HPLC (Agilent, Palo Alto, CA) coupled with a diode array detector (DAD) and mass spectrometer detector (MSD, SL mode) (*15*). The MSD (SL) used electrospray ionization (ESI) and was programmable to acquire data in positive and negative ionization (PI and NI) modes at low (100 V) and high (250 V) fragmentation voltages, in rapid sequence. A drying gas flow of 13 L/min, a drying gas temperature of 350 °C, a nebulizer pressure of 50 psi, and capillary voltages of 4000 V for PI and 3500 V for NI were used.

For direct injection, the sample was injected directly into the ionizer with no column at 1 mL/min using an infusion pump. The MSD was programmed to scan masses from *m*/*z* 50 to 2000 in 10 min.

Chromatographic profiles were acquired for the C100 sample using a Waters (Waters Corp., Milford, MA) Symmetry column (C18, 5 mm, 250 × 4.6 mm) with a Sentry guard column (Symmetry 5 mm, 3.9 × 20 mm) at flow rate of 1.0 mL/min. The column oven temperature was set at 25 °C. The mobile phase consisted of a combination of A (0.1% formic acid in water) and B (0.1% formic acid in acetonitrile). The gradient was varied linearly from 10 to 26% B (v/v) in 40 min, to 65% B at 70 min, and finally to 100% B at 71 min and held at 100% B to 75 min.

Data Analysis. All spectral data were converted to the American Standard Code for Information Interchange (ASCII) files and exported for chemometric analysis. Preprocessing of the data matrices was performed using Excel (Microsoft, Inc., Belleview, WA), and PCA was performed using Pirouttte 3.1 (Infometrix, Inc., Bothell, WA) (*8*).

MS fingerprints were one-dimensional spectra, counts versus mass for *m*/*z* 50−2000. Five repeat analyses of the 7 different broccoli extracts provided 35 spectra. This yielded an initial matrix of 35 rows (samples) by 1950 columns (variables, mass ions). The Agilent software listed all counts as a percentage of the highest peak (normalized at 100). The mass with the highest peak count varied between *m*/*z* 91 and 191, for negative ionization, and between *m*/*z* 69 and 83, for positive ionization. The maximum counts varied by approximately 10%.

The first step was to discard all masses for individual fingerprints that did not exceed 1% of the highest count. Second, the masses for the 35 samples were aligned. This was not a trivial task because the instrument software did not list masses if the counts did not exceed a minimum level. Third, individual masses were discarded for all repeats of a sample if four of the five repeats did not register a count; this was primarily for masses with low counts. Fingerprints were discarded in two cases for both the positive and negative spectra because the counts were consistently 40% low. This pattern was observed for both positive and negative ionizations. In addition, no peaks were observed in the negative mode for two of the samples. At this stage, the data matrix for the negative ionization profiles was 31 × 99 and 33 × 167 for positive ionization.

The ANOVA data preprocessing has been described in detail previously (*8*) on the basis of the method of Harrington et al. (*9*). Briefly, the data matrix was transformed by scaling to unit variance for each sample and each mass. Unlike the UV data, the MS data were not transformed to the first derivative.

A grand means matrix was computed and subtracted from the double-scaled matrix to provide a grand means residuals matrix (also called a mean-centered matrix). The cultivar means matrix was computed and subtracted from the grand means residuals matrix to give the cultivar means residuals matrix. The treatment means matrix was computed and subtracted from the cultivar residuals matrix to give the treatment means residuals matrix. The treatment residuals matrix represents the analytical uncertainty matrix. The two matrices tested by PCA were (1) the matrix resulting from the sum of the treatment mean matrix and the analytical uncertainty matrix (i.e., the cultivar residuals matrix) and (2) the matrix resulting from the sum of the cultivar means matrix and the analytical uncertainty matrix.

The variance contribution of the experimental factors was computed by squaring and summing the data for the following: (1) the grand means residuals matrix (total variance), (2) the cultivar means matrix (between cultivar variance), (3) the cultivar means residuals matrix (within cultivar variance), (4) the treatment means matrix (between treatment variance), and (5) the treatment means residuals matrix (within treatment variance). Summing was implemented for each mass to provide variance as a function of mass and across mass and sample to provide variance as a function of the integrated spectra.

Statistical Calculations. The significance of the separation of two compared populations was computed using the Student *t* test. The Student *t* value was computed as the difference between the population means divided by the shared standard deviation:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}} \quad (1)$$

The mean is calculated as

$$\bar{X}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i} \quad (2)$$

The standard deviation is calculated as

$$s_i = \sqrt{\frac{(x_{i1} - \bar{x}_i)^2}{n_i - 1}} \quad (3)$$

The shared standard deviation is computed as

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (4)$$

The probability, *p*, of the two populations being similar was determined from Student *t* tables based on *t* and $n = n_1 + n_2 - 2$.

RESULTS AND DISCUSSION

ANOVA-PCA. The chemical differences of broccoli as a function of cultivar and treatment (growing conditions) were investigated using both negative and positive ionization MS fingerprints obtained from direct injection of the aqueous methanol extracts using a low excitation energy (100 V) to minimize in-source collision-induced dissociation (CID). The fingerprints were one-dimensional arrays with counts as a function of mass as shown in **Figure 1** for the average of the multiple analyses of sample C100.

The MS data were processed as described under Materials and Methods to yield two-dimensional matrices that were 31 samples × 99 masses (between *m*/*z* 100 and 650) for negative ionization and 33 × 167 for positive ionization. No significant counts were observed for masses above *m*/*z* 650. This is consistent with the observation of Shulaev et al. (*5*) that direct injection MS lacked high mass ions due to ion suppression.

The array sizes show that approximately 80% of the masses in the negative ionization fingerprint and 67% in the positive ionization fingerprint were not useful (i.e., low counts and/or poor precision). These data matrices were then preprocessed as described under Materials and Methods, and appropriate submatrices were subjected to PCA.

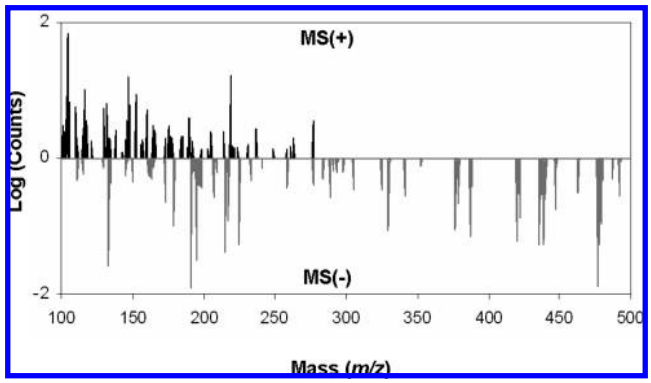

Figure 1. Positive and negative mass spectral fingerprints of Majestic cultivar of broccoli grown with 1000 ppm of selenium.

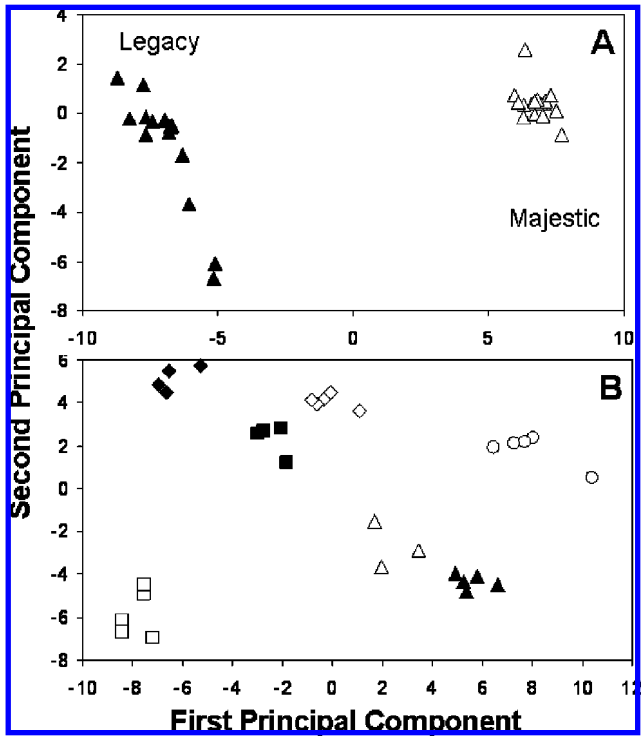

Figure 2. Principal component analysis score plots based on negative ionization fingerprints for (**A**) the comparison of the broccoli cultivars Legacy (▲) and Majestic (△) and (**B**) the comparison of all broccoli grown with different selenium concentrations (■, 0 ppm; △, 5 ppm; ▲, 100 ppm; ◆, 1000 ppm; ◇, C100; ○, C80; □, Org).

The chemical differences between the cultivars of broccoli were initially investigated by PCA of the matrix obtained by combining the cultivar means matrix and the analytical uncertainty matrix. The score plot for the negative ionization fingerprints is shown in **Figure 2A**. The data for the two cultivars were strongly separated on the basis of the first principal component, which accounted for 60% of the total variance. Statistically, the separation was significant at greater than the 99.95% confidence level based on subsamples of the composited samples (**Table 1**). Similar results were observed for the positive ionization fingerprints (**Figure 3A** and **Table 1**).

The chemical differences between treatments were investigated by PCA of the cultivar residual matrix (the combination of the treatment means matrix and the analytical uncertainty matrix). The score plots for inclusion of all seven treatments (**Figures 2B** and **3B**) show that the separation on the *x*-axis is slightly better for the negative fingerprints. In each case, the first principal component is almost sufficient to differentiate

Table 1. Computed Student *t* Values for ANOVA-PCA[a]

	UV			MS(−)			MS(+)		
comparison of treatment	*n*	*t*	*P*=*p*	*n*	*t*	*p*	*n*	*t*	*p*
Legacy vs Majestic	63	105	<0.0005	30	44	<0.0005	33	67	<0.0005
0 vs 5 ppm	15	63	<0.0005	7	13	<0.0005	8	37	<0.0005
100 vs 1000 ppm	20	86	<0.0005	9	32	<0.0005	10	26	<0.0005
C100 vs Org	19	12	<0.0005	10	27	<0.0005	10	29	<0.0005

[a] *n* is the number of subsample measurements, *t* is the computed Student *t* value, and *p* is the probability that the two populations are the same.

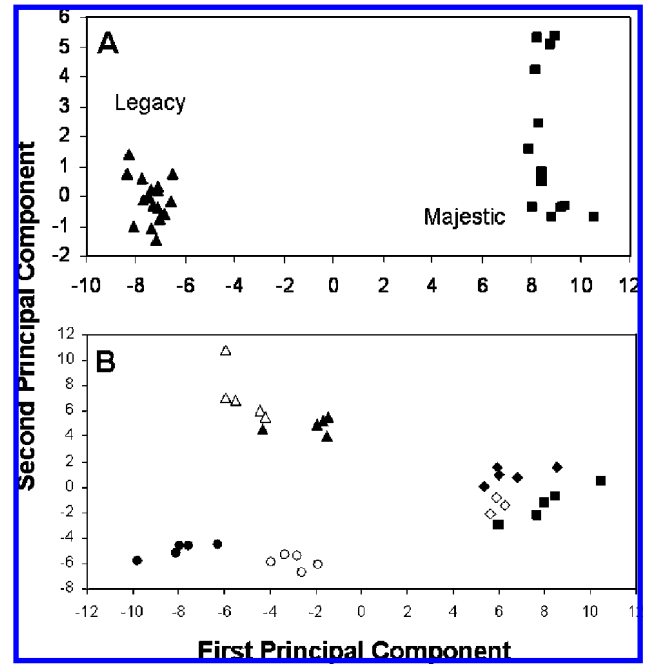

Figure 3. Principal component analysis score plots based on positive ionization fingerprints for (**A**) the comparison of Legacy (▲) and Majestic (■) broccoli cultivars and (**B**) the comparison of all broccoli treatments (▲, 0 ppm; ◇, 5 ppm; ◆, 100 and 1000 ppm; ○, C100; ■, C80, △, Org).

between all treatments. In both cases, however, the second principal component is needed to provide discrimination between treatments.

PCA for pairs of treatments (between cultivars or between individual treatments) provides much clearer visual and statistical discrimination. **Figure 4** and **Table 1** provide score plots and statistical analyses for 3 of the 21 possible pairings of treatments. The results for the other 18 pairings were similar and permitted visual and statistical differentiation. The pairs selected for PCA in **Figure 4** represent three comparisons of treatments within a cultivar (0 vs 5 ppm and 100 vs 1000 ppm for cv. Majestic and C100 vs Org for cv. Legacy).

Despite the dramatic difference between the negative and positive fingerprints (**Figure 1**), PCA of the data yielded very similar results (**Figures 2**−**4** and **Table 1**). The difference in positive and negative ionization of the sample components is also illustrated by the total ion count (TIC) chromatographic profiles shown in **Figure 5B**,**C**.

Although both show intense peaks between 2 and 7 min, there are far fewer peaks at longer retention times with positive ionization. The masses contributing to the peaks between 2 and 7 min also differed considerably, as will be discussed below. Thus, the ions constituting the fingerprints for the two ionization modes were different, but the PCA (**Figures 2**−**4**) and statistical results (**Table 1**) were almost identical.

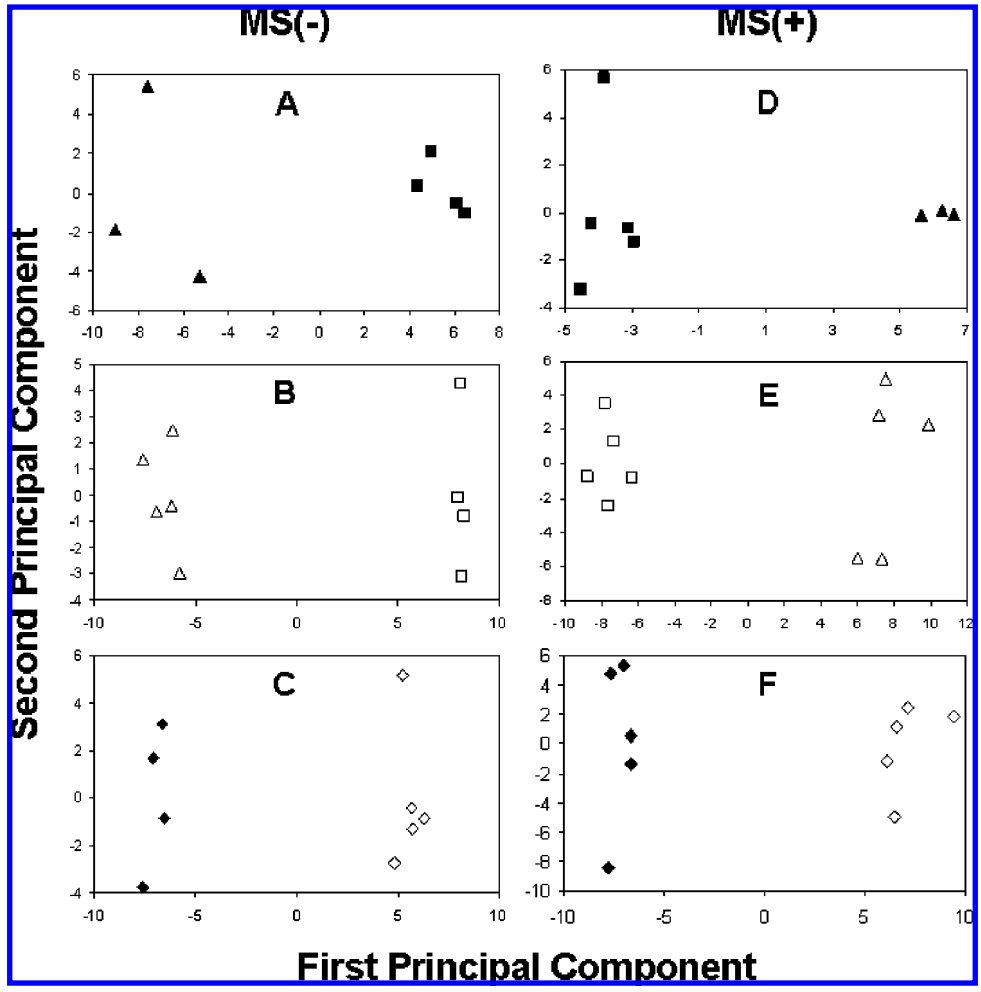

Figure 4. Principal component analysis score plots based on negative (**A**−**C**) and positive (**D**−**F**) ionization for the comparison of treatments: (**A**, **D**) 0 versus 5 ppm; (**B**, **E**) 100 versus 1000 ppm; (**C**, **F**) C100 and Org (■, 0 ppm; ▲, 5 ppm; △, 100 ppm; ■, 1000 ppm; ◆, C100; ◇, Org).

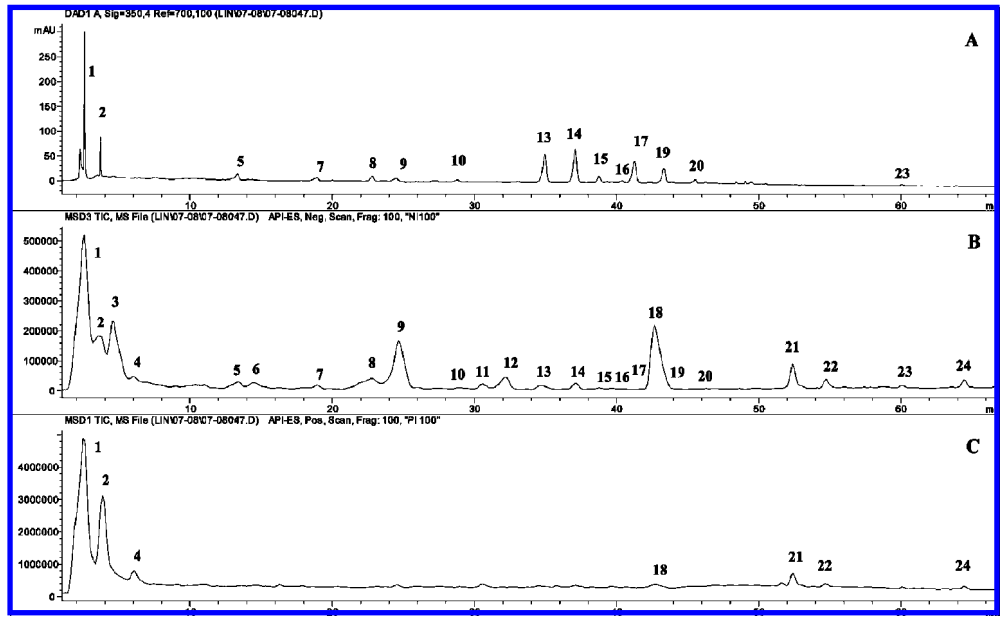

Figure 5. Liquid chromatography−mass spectrometry chromatograms with (**A**) ultraviolet detection (350 nm), (**B**) negative ionization, and (**C**) positive ionization.

PCA Loadings. PCA of the MS fingerprints provides data on the loadings for each ion. Ions with high loadings for the first principal component can be used to identify those compounds that are affected most by cultivar and treatment and contribute the most to the horizontal separations in **Figures 2**−**4**. **Table 2** provides X-axis loading data for PCA of the negative ionization fingerprints shown earlier, that is, comparison of cultivars (**Figure 2A**), all treatments (**Figure 2B**), and three

Table 2. Relative Counts, PCA Loading, and Variance as a Function of Mass—Negative Ionization

		loadings for ANOVA PCA					% of total variance		
	rel			C100 vs	0 vs	100 vs			
m/z	counts	cultivar	treatment	Org	5	1000	cultivar	treatment	analytical
115	2.0		X				56	32	12
131	0.3			X			43	47	10
133	40.9	X	X	X			34	63	3
145	3.5	X	X	X	X		54	44	2
150	2.1				X		8	54	38
161	2.9					X	10	66	24
179	9.0				X		0	89	11
191	76.0				X		3	89	8
195	26.6				X		20	74	6
199	0.5	X		X			50	49	1
207	2.8						1	34	65
209	2.0	X					79	15	6
215	18.4					X	16	32	52
225	15.5				X		6	80	14
233	1.5		X				50	35	15
242	2.5				X		0	84	16
243	0.7		X	X		X	1	89	10
277	3.1	X	X	X			29	68	3
284	1.9			X			49	38	13
293	2.6					X	1	90	8
296	0.9			X			83	6	11
308	1.6	X		X			42	52	6
311	1.0	X		X			74	18	8
325	5.1	X	X	X		X	46	52	2
327	3.6			X	X		37	51	12
328	0.2	X		X			20	79	1
341	2.5		X			X	7	84	10
352	0.1	X			X		10	89	1
363	0.8			X		X	22	70	8
377	6.7		X		X	X	0	87	13
378	1.1					X	3	47	49
379	3.2		X				2	86	12
387	9.1		X	X	X	X	5	93	2
388	1.6		X	X			0	94	6
404	1.3				X		23	72	5
420	5.5	X			X	X	63	36	1
422	3.8	X				X	57	38	5
432	0.2					X	18	73	10
436	11.4	X		X	X	X	39	58	3
439	11.2		X	X	X	X	2	93	4
440	2.2		X		X		6	86	8
446	0.2	X					20	76	4
447	14.7	X					96	3	1
448	3.4	X				X	94	5	1
449	1.5	X					96	2	2
463	1.0					X	32	56	12
465	0.3			X			20	78	2
473	0.3	X					20	78	2
477	51.0		X		X	X	2	94	5
478	12.5		X		X		1	93	6
479	6.4		X		X		0	91	9
487	2.1	X	X			X	65	30	4
488	0.4		X				6	88	6
492	2.1		X		X	X	4	88	8
493	1.1			X			0	80	19
613	1.5					X	0	86	14

Table 3. Relative Counts, PCA Loading, and Variance as a Function of Mass—Positive Ionization

		loadings for ANOVA-PCA					% of total variance		
	rel			C100 vs	0 vs	100 vs			
m/z	counts	cultivar	treatment	Org	5	1000	cultivar	treatment	analytical
105	7.7			X		X	13	83	4
106	0.9		X				1	56	44
110	4.1		X		X	X	0	84	16
111	0.6		X			X	6	85	9
116	9.3		X	X	X	X	2	95	2
117	0.9			X			17	43	40
118	6.3	X	X	X	X	X	53	43	4
131	1.4	X	X				53	44	4
132	12.2	X		X	X	X	77	22	1
133	3.8	X			X		66	31	3
134	2.1				X		37	41	23
136	0.5		X				9	60	31
138	2.3	X		X			60	39	1
146	5.2		X	X		X	21	77	3
147	30.5	X	X	X	X	X	25	74	1
148	4.7	X		X			72	27	1
152	7.1	X		X	X	X	79	20	1
156	2.2	X			X		79	18	3
157	1.8		X	X	X		0	97	3
159	2.0						48	33	19
160	3.0	X	X		X		48	47	5
161	0.8	X			X		16	83	1
164	4.3	X			X	X	26	72	2
166	6.3	X			X		93	6	1
171	1.1		X		X		0	74	25
173	3.4	X		X			71	26	4
174	0.2			X			16	52	32
175	9.4	X			X	X	77	22	1
176	2.7		X	X		X	34	60	6
177	0.8		X		X	X	2	96	2
178	1.7	X			X	X	56	42	1
182	0.4			X			41	29	29
184	2.3	X	X		X	X	66	33	1
185	4.6			X	X	X	9	88	3
186	0.6		X		X	X	12	77	11
187	0.4	X					51	48	1
190	3.6	X					89	10	1
191	1.6		X			X	38	55	7
192	1.5	X	X				79	19	2
198	0.9	X					84	11	5
203	1.6			X			8	42	51
207	1.8			X			39	35	25
208	0.0						4	14	83
219	13.4	X			X		82	16	2
224	0.9	X					87	5	9
231	1.3	X					71	22	7
235	0.8		X			X	7	81	12
237	1.5				X		59	29	12
249	1.6	X	X	X		X	46	52	2
258	0.8					X	16	65	19
261	1.3	X					71	22	8
263	1.6	X					77	17	6
266	1.1		X			X	17	79	4
275	0.7		X			X	0	85	15
279	1.0		X				53	37	10
293	1.3		X	X		X	11	83	6
296	0.6			X			22	37	41
304	0.9		X				0	59	41
306	0.2	X	X			X	15	85	1
307	0.4		X				0	41	59
309	1.3		X		X	X	4	95	1
314	0.2		X			X	22	57	21
336	1.0	X			X		81	15	4
348	0.4			X			56	28	16
365	1.8			X			8	82	10
367	0.5		X	X		X	0	91	9
369	0.3			X			22	49	29
381	16.1	X	X	X	X	X	16	84	0
382	2.8	X	X	X	X	X	22	78	0
383	1.8			X			0	88	12
399	2.2			X	X		0	95	5
400	0.9			X	X	X	22	70	7
404	0.3		X			X	22	65	12
438	2.9	X			X	X	43	56	1
439	0.7				X		39	55	6
474	0.4			X			43	40	17
476	0.3				X		6	66	28
518	0.7			X	X		0	78	22

pairs of treatments (**Figure 4**). Those masses providing the top 20% of the loadings for each PCA are denoted with an "X". Only masses that had at least one significant loading are shown; masses with no "X" are not listed. The exception is *m/z* 207, which will be discussed later. This form of data display was chosen, instead of a loading plot, to make the loading for each ion easier to see.

There is very little overlap between the PCA loadings for the comparison of cultivars (column 3) and all treatments (column 4) (**Tables 2** and **3**). The patterns for PCA loadings for pairs of treatments (columns 5−7) were difficult to characterize (**Tables 2** and **3**). There was little correlation between the pair loadings and either the cultivar or treatment loadings. Loadings for PCA of the other 18 pairs of cultivars are not shown, but they were equally difficult to characterize.

In general, the masses that provided high loadings were highly dependent on the treatments being compared. This observation will be discussed later in more detail.

Data for the loadings of the positive ionization fingerprint comparisons are shown in **Table 3**. In general, the patterns for the loadings were similar to those of the negative fingerprints. There was little overlap of loadings for the PCA of cultivars and all treatments and loadings for the PCA of pairs of treatments. The loadings for the PCA of pairs of treatments were highly dependent on the treatments being compared.

Compound Identification. Identification of compounds based strictly on a single mass is highly problematic. If we assumed that only phenolic compounds were present (because we used a methanol/water extraction optimized for phenolic compounds), then six of the negative ion masses in **Table 2** could be tentatively identified from our previous studies of plant phenolics (*15*). Masses *m/z* 191 and 209 are common fragments of the hydroxycinnamate derivatives. Masses *m/z* 447, 463, 477, and 479 are suggestive of flavonol hexosides: kaempferol, quercetin, isorhamnetin, and myricetin, respectively. However, no masses were observed for the aglycones of the flavonols (*m/z* 285, 310, 315, and 317), and it was not possible to determine if the ions identified as flavonol glycosides were molecular or fragment ions.

A more detailed analysis of the phenolic components in broccoli was made using the chromatographic profiling method developed in our Beltsville laboratory (*15*). This LC-DAD-ESI/MS method provides retention times, UV spectra (200−600 nm), and positive and negative mass spectra at high and low fragmentation voltages and allows clear identification of molecular and fragment ions. The profiling method has been used to analyze more than 360 plant materials and 200 standard compounds. Analysis of an extract of the C100 sample (Legacy broccoli, grown conventionally with 100% water) produced the UV (350 nm) and positive and negative TIC chromatograms shown in **Figure 5** and the identification of compounds shown in **Table 4**.

None of the ions listed in **Tables 2** and **3** are phenolic compounds. The provisionally identified flavonoid glycosides in **Table 4** (quercetin and kaempferol sophorosides and kaempferol diglucoside) do not have molecular or fragment ions that correspond to any of the ions in **Tables 2** and **3**. Extracted ion count (EIC) chromatograms of *m/z* 191, 447, and 477 (**Figure 6**) and *m/z* 209, 463, and 479 (not shown) provide peaks that do not correlate with the retention times of the suggested phenolic compounds or with UV absorption peaks of phenolic compounds (**Figure 5A**). Masses *m/z* 191 (**Figure 6B**) and 209 (not shown) eluted early and were associated with peaks 1 and 2 (**Figure 5A**). Masses *m/z* 447 and 477 (**Figure 6C**,**D**) constituted peaks 18 and 8/9, respectively. Masses *m/z* 463 and 479 (not shown) gave low-intensity peaks that did not register in **Figure 5**. None of these peaks absorbed at 350 nm (**Figure 5A**).

Additionally, the main ions in each peak in **Figure 5B**,**C** were compared to the ions in **Tables 2** and **3**. Approximately 50% and 30% (bolded and italicized) of the masses in **Tables 2** and **3**, respectively, were found in the different chromatographic peaks. Masses *m/z* 447 and 477 (**Figure 6C**,**D**) were the only contributors to peaks 18 and 8/9, respectively. They have not been identified.

The identities of some of the ions in **Tables 2** and **3** are suggested by recent plant metabolomics papers. Goodacre et al. (*16*) analyzed *Pharbitis nil* sap extracts (50% aqueous isopropanol) by direct injection ESI/MS and reported negative ions consistent with organic acids and sugars: fumarate (*m/z* 115), malate (*m/z* 133), 2-oxoglutarate (*m/z* 145), citrate (*m/z* 191), hexoses (*m/z* 179), and sucrose and isomers (*m/z* 341). Dunn et al. (*17*) examined aqueous extracts of tomato fruit by both positive and negative ionization MS and observed many of the same negative ions reported by Goodacre et al. as well as glucoheptonate (*m/z* 225). In addition, they reported positive ions consistent with amino acids: serine (*m/z* 106), proline (*m/z*

Table 4. Peak Assignment for Aqueous Methanol Extracts of Broccoli

peak	t_R (min)	$[M + H]^+/[M - H]^-$ (*m/z*)	aglycones λ_{max} *m/z*	nm	identification
1	2.53	133/191, 195			
2	3.40	295/191, 195			
3	4.62	—/436			
4	6.00	205/173, 195, 203			
5	13.32	—/385			
6	14.06	—/421, 431, 463			
7a	18.92	627/625	303/301	nd[a]	quercetin 3-*O*-sophoroside
7b	18.92	611/609	287/285	nd	kaempferol 7,3-*O*-diglucoside
8	22.81	611/609	287/285	nd	kaempferol 3-*O*-sophoroside
9	24.63	—/447			
10	28.78	—/365			
11	32.03	—/144			
12	32.03	—/447			
13	34.96	—/753	nd	240, 330	1,2-disinapoylgentiobiose
14	37.09	—/723	nd	240, 330	1-sinapooyl-2-feruloylgentiobiose
15	38.77	—/693	nd	240, 330	1,2-diferuloylgentiobiose
16	40.38	—/591		240, 330	
17	41.26	—/959	nd	240, 330	1,2,2′-trisinapooylgentiobiose
18	42.68	—/477			
19	43.33	—/929	nd	240, 330	1,2′-disinapoyl-2-feruloylgentiobiose
20	46.29	—/175, 737			
21	52.38	351/327			
22	54.75	353/211, 329			
23	60.03	—/327			
24	64.44	335/311			

[a] nd, not determined.

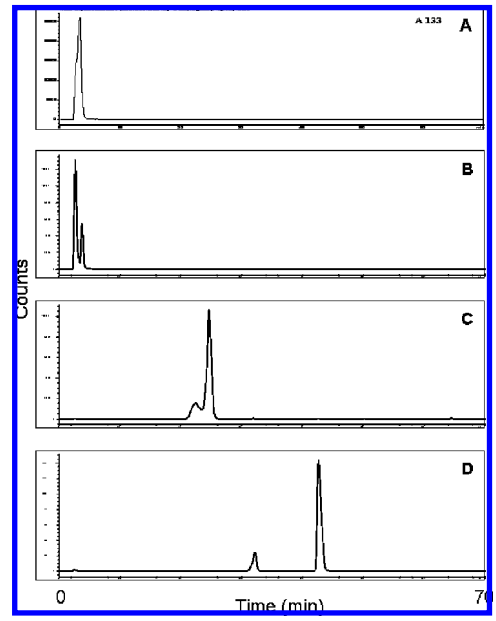

Figure 6. Extracted ion chromatograms with negative ionization for (**A**) *m/z* 133 (possibly malate), (**B**) *m/z* 191 (possibly citrate), (**C**) *m/z* 447, and (**D**) *m/z* 477.

Table 5. Relative (Percent) Variance for Experimental Factors for UV and MS Detection

source of variance	UV	MS(−)	MS(+)
total	100	100	100
cultivar mean	33	32	34
treatmentmean	66	59	59
analytical error	1	9	7

116), valine (*m*/*z* 118), leucine/isoleucine (*m*/*z* 132), aspartic acid (*m*/*z* 134), glutamine (*m*/*z* 147), glutamic acid (*m*/*z* 148), histidine (*m*/*z* 156), phenylalanine (*m*/*z* 166), arginine (*m*/*z* 175), and tyrosine (*m*/*z* 182). These masses contributed some of the highest relative ion counts in **Table 3**. The EICs for malate and citrate are shown in **Figure 6A**,**B**. They eluted early from the column, contributing to peaks 1−4 (**Figure 5**). The early elution times are consistent with the polar natures of the organic and amino acids. The least polar compound, the hexose, had a low intensity and coeluted with peak 9 (**Figure 5B**).

The presence of these compounds in broccoli seems to be reasonable, but positive identification will require further investigation with a column and solvent system appropriate for more polar compounds and standards. Regardless of their identity, the early elution time of these compounds and the role they play in discriminating between cultivars and treatments suggest that fingerprints of an aqueous or more polar extract may be more informative than the MeOH/H_2O (60:40, % v/v) extraction solvent that we have used in the present study. We assumed that phenolic compounds would be environmentally sensitive and strongly reflect growing conditions. However, it is essential to evaluate the role of small polar molecules such as amino acids and simple carbohydrates for classification of foods and other plant-derived products. Certainly, they are present in higher concentrations than the phenolic compounds. By analogy, compounds at the opposite end of the polarity spectrum, the nonpolar lipid fraction, may also contribute valuable information for discriminating between plant genus, species, genotypes, and treatments.

Sources of Variance. The submatrices generated for ANOVA-PCA can also be used to calculate variance data for the fingerprints (*8*). These calculations allow for partitioning of the total variance of the data between the experimental factors (cultivar and treatment) and analytical uncertainty for the integrated spectra or individual variables (masses in this study). Results for the integrated variance calculations for the positive and negative MS fingerprints are shown in **Table 5** along with the previous results for the UV fingerprints. All three show that 33% of the variance was attributed to the cultivars. The analytical uncertainty for MS was about 8% compared to 1% for UV. This higher level of uncertainty associated with MS is not surprising because the injection-to-injection variability is generally acknowledged to be around 10% without an internal standard. As a consequence, the variance contributed by treatment was 59% for the MS fingerprints, 7% lower than calculated for UV.

The agreement of the data in **Table 5** may not at first seem too surprising because the same extract (MeOH/H_2O, 60:40, %v/v) was used for both the UV and MS measurements. However, the molar absorption coefficient of a compound does not necessarily correlate with its ability to be positively or negatively ionized, as shown in **Figure 5**. Thus, it is possible that a specific compound may contribute strongly to one fingerprint and not another. From this perspective, the agreement of the variance between the different methods in **Table 5** is remarkable. This suggests that the change in plant chemistry

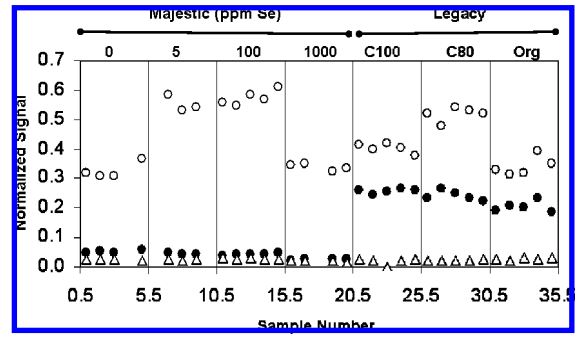

Figure 7. Scaled counts as a function of sample treatment for (○) *m*/*z* 477, (●) *m*/*z* 447, and (△) *m*/*z* 207.

arising from the difference in the cultivars and treatments affects a wide range of metabolites, not just a few isolated compounds.

The relative distribution of variance can also be computed for individual masses. These results are shown in **Tables 2** and **3** (columns 7−9). We show only masses (with the exception of *m*/*z* 207 in **Table 2** and *m*/*z* 208 in **Table 3**) that were in the top 20% of the loadings for one of the PCAs. All of the masses that provided less significant loading were not listed. We observed that the vast majority of the masses in **Tables 2** and **3** have less than 10% of their variance attributed to analytical uncertainty. Thus, high loadings corresponded to high variance associated with either cultivar or treatment, whereas minimal variance (generally less than 15%) was associated with analytical uncertainty.

Close inspection of **Tables 2** and **3** shows that, in general, masses with high loadings for cultivar have a high variance for cultivar. Those with high loadings for treatment have a high variance for treatment. This pattern is well illustrated by *m*/*z* 447, 477, and 207 in **Table 2**. For *m*/*z* 447, high loading was observed for the cultivar comparison, and the variance distribution was 96, 3, and 1%, respectively, for cultivar, treatment, and analytical uncertainty. For *m*/*z* 477, with a high loading for PCAs of the treatments, the variance distribution was 2, 94, and 4%, respectively. Finally, for *m*/*z* 207, which failed to provide significant loading for any of the PCA comparisons, the variance distribution was 1, 34, and 65%, respectively.

There are notable exceptions to the obvious patterns discussed above. As shown in **Table 2**, some masses provided high loading for some comparisons, although the relative variance of the analytical uncertainty fell between 38 and 52% (*m*/*z* 150, 215, and 378). There are also masses that provided high loading for the cultivar comparison despite the variance associated with cultivar being 20% or less (*m*/*z* 328, 352, 446, and 473). Conversely, some masses provide significant loadings for treatment with a very low variance for treatment (*m*/*z* 296). There are also masses with surprisingly high variance for analytical uncertainty that still provided significant loading for one of the comparisons (*m*/*z* 150, 215, and 378).

The previous discussion of the PCA loadings and variance raises interesting questions as to what constitutes a good mass for discriminating between cultivars and treatments. **Figure 7** shows scaled data for three ions: mass *m*/*z* 477 had high loadings for treatments, *m*/*z* 447 had high loadings for cultivars, and *m*/*z* 207 showed no significant loadings for any of the PCAs. The vertical gridlines serve to isolate the five subsamples for each of the seven treatments. The total variance is a function of the vertical distribution of the data. The variance associated with cultivars is determined by the difference between the averages of the two cultivars (i.e., the average of 0, 5, 100, and 1000 ppm vs the average of C100, C80, and Org). The treatment

variance will depend on the spacing between the averages of each treatment, and the variance associated with analytical uncertainty will be determined by the distribution of the data around each treatment average.

We observed that the data for *m*/*z* 447 had almost 100% of its variance associated with cultivar. For *m*/*z* 477, the variance associated with cultivar was much less. The variance associated with analytical uncertainty was a little larger, and variance associated with treatment was much larger. Finally, we observed that the variance for *m*/*z* 207 was associated primarily with the analytical uncertainty. **Figure 7** illustrates how the PCA loading for either cultivar or treatment will correlate with high variance for the same experimental variable. These data reaffirm Wold's statement that the directions of maximum variance correspond to the directions of maximum information (*11*).

Figure 7 demonstrates why it is difficult to select a few marker ions for discrimination. Every ion will have a distinctive pattern. For example, mass *m*/*z* 447 is useful for discriminating between cultivars but not between treatments. Mass *m*/*z* 477 is useful for discriminating between some, but not all, treatments. Mass *m*/*z* 477 allows for discrimination between 0 and 5 ppm and between 100 and 1000 ppm, but not between 0 and 1000 ppm or between 5 and 100 ppm. Plots of other ions show a large variety of patterns. Thus, the ability to discriminate on the basis of experimental variables becomes more reliable as more variables are used.

These data suggest that rather than selecting a few marker ions with which to discriminate between cultivar and treatment, noncontributors should be discarded. That is, it would be better to discard those ions that offer no significant loading or that have variance primarily associated with analytical uncertainty. This can be accomplished by establishing a minimum threshold for loading or a maximum allowable threshold for the relative variance of analytical uncertainty. Alternatively, the selected ions can represent a preset fraction of the variables.

The ions listed in **Tables 2** and **3** represent the top 20% of the loadings. The two lists represent data reductions of factors of 10 and 7 from the total number of ions for negative and positive ionization, respectively. The number of useful masses will be larger if the definition for high loading is extended to the top 30 or 40%. Obviously, this threshold is arbitrary. In this study, the full data set was initially considered. All data below a threshold of 1% relative counts were dropped, and the remaining data were submitted to PCA and the variance calculation. For archival purposes, the most complete data set possible is desirable.

Conclusions. In this study, mass spectral fingerprinting allowed a global comparison of broccoli cultivars and treatments and identified discriminating components. ANOVA-PCA provided clear visual and statistical differentiation of the cultivars and treatments. PCA variable loadings and relative variance were correlated and would appear to be better used to discard those compounds associated primarily with analytical uncertainty rather than isolating a few positive marker compounds. The data illustrate the pitfalls of making compound identifications on the basis of a single ion and point out the need for ion tables for common plant components.

ACKNOWLEDGMENT

We thank Dr. Sudarsan Mukhopadhyay, a postdoctoral fellow from our laboratory, for his assistance.

LITERATURE CITED

(1) Sumner, L. W.; Mendes, P.; Dixon, A. Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry* **2003**, *62*, 817–836.

(2) Goodacre, R.; Luned, R.; Ellis, D. I.; Thorogood, D.; Reader, S. M.; Ougham, H.; King, I. From phenotype to genotype: whole tissue profiling for plant breeding. *Metabolomics* **2007**, *3*, 489–501.

(3) Dixon, R. A.; Gang, D. R.; Charlton, A. J.; Fiehn, O.; Kuiper, H. A.; Reynolds, T. L.; Tjeerdema, R. S.; Jeffery, E. H.; German, J. B.; Ridley, W. P.; Seiber, J. N. Application of metabolomics in agriculture. *J. Agric. Food Chem.* **2006**, *54*, 8984–8994.

(4) Dunn, W. B.; Ellis, D. I. Metabolomics: current analytical platforms and methodologies. *Trends Anal. Chem.* **2005**, *24*, 285–294.

(5) Krishnan, P.; Kruger, N. J.; Ratcliffe, R. G. Metabolite fingerprinting and profiling in plants using NMR. *J. Exp. Biol.* **2005**, *56*, 255–265.

(6) Shulaev, V.; Cortes, D.; Miller, G.; Mittler, R. Metabolomics for plant stress response. *Physiol. Plant.* **2008**, *132*, 199–208.

(7) Goodacre, R.; York, E. V.; Heald, J. K.; Scott, I. M. Chemometric discrimination of unfractionated plant extracts analyzed by electrospray mass spectrometry. *Phytochemistry* **2003**, *62*, 859–863.

(8) Luthria, D. L.; Mukhopadhyay, S.; Robbins, R. J.; Finley, J. W.; Banuelos, G. S.; Harnly, J. M. UV spectral fingerprinting and analysis of variance-principal component analysis: a useful tool for characterizing sources of variance in plant materials. *J. Agric. Food Chem.* **2008**, *56*, 5457–5462.

(9) Harrington, D. B.; Vieira, N. E.; Espinoza, J.; Kien, J. K.; Romero, R.; Yergey, A. L. Analysis of variance−principal component analysis: a soft tool for proteomic discovery. *Anal. Chim. Acta* **2005**, *544*, 118–127.

(10) Harrington, D. B.; Vieira, N. E.; Ping, C.; Espinoza, J.; Kien, J. K.; Romero, R.; Yergey, A. L. Proteomic analysis of amniotic fluids using analysis of variance-principal component analysis and fuzzy rule-building expert systems applied to matrix-assisted laser desorption/ionization mass spectrometry. *Chemom. Intell. Lab. Syst.* **2006**, *82*, 283–293.

(11) Wold, S. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52.

(12) Finley, J. W.; Sigrid-Keck, A.; Robbins, R. J.; Hintze, K. J. Selenium enrichment of broccoli: interactions between selenium and secondary plant compounds. *J. Nutr.* **2005**, *135*, 1236–1238.

(13) Robbins, R. J.; Keck, A. S.; Banuelos, G.; Finley, J. W. Cultivation conditions and selenium fertilizations alter the phenolic profile, glucosinolate and sulforaphane content of broccoli. *J. Med. Food* **2005**, *8*, 204–214.

(14) Lee, J.; Finley, J. W.; Harnly, J. Effect of selenium fertilizer on free amino acid composition of broccoli (*Brassica oleracea* cv. Majestic) determined by gas chromatography with flame ionization and mass selective detection. *J. Agric. Food Chem.* **2005**, *53*, 9105–9111.

(15) Lin, L.-Z.; Harnly, J. M. A screening method for the identification of glycosylated flavonoids and other phenolic compounds using a standard analytical approach for all plant materials. *J. Agric. Food Chem.* **2007**, *55*, 1084–1095.

(16) Goodacre, R.; York, E. V.; Heald, J. K.; Scott, I. M. Chemometric discrimination of unfractionated plant extracts analyzed by electrospray mass spectrometry. *Phytochemistry* **2003**, *62*, 859–863.

(17) Dunn, W. B.; Overy, S.; Quick, W. P. Evaluation of automated electrospray−TOF mass spectrometry for metabolic fingerprinting of the plant metabolome. *Metabolomics* **2005**, *1*, 137–145.

Received for review May 22, 2008. Revised manuscript received July 15, 2008. Accepted July 29, 2008. This research was supported by the Agriculture Research Service of the U.S. Department of Agriculture and the Office of Dietary Supplements at the National Institutes of Health.

JF801606X