JOURNAL OF NATURAL PRODUCTS

Review

pubs.acs.org/jnp

Metabolomics for Phytochemical Discovery: Development of Statistical Approaches Using a Cranberry Model System

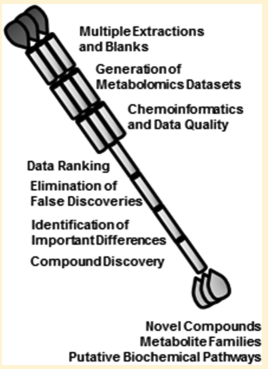
Christina E. Turi,[†] Jamie Finley,[‡] Paul R. Shipley,[†] Susan J. Murch,[†] and Paula N. Brown*[,‡]

[†]Department of Chemistry, University of British Columbia, 3247 University Way, Kelowna, British Columbia, Canada, V1V 1V7

[‡]Natural Health Products and Food Research Group, British Columbia Institute of Technology, 4355 Mathissi Place, Burnaby, British Columbia, Canada, V5G 3H2

Supporting Information

ABSTRACT: Metabolomics is the qualitative and quantitative analysis of all of the small molecules in a biological sample at a specific time and influence. Technologies for metabolomics analysis have developed rapidly as new analytical tools for chemical separations, mass spectrometry, and NMR spectroscopy have emerged. Plants have one of the largest metabolomes, and it is estimated that the average plant leaf can contain upward of 30 000 phytochemicals. In the past decade, over 1200 papers on plant metabolomics have been published. A standard metabolomics data set contains vast amounts of information and can either investigate or generate hypotheses. The key factors in using plant metabolomics data most effectively are the experimental design, authentic standard availability, extract standardization, and statistical analysis. Using cranberry (*Vaccinium macrocarpon*) as a model system, this review will discuss and demonstrate strategies and tools for analysis and interpretation of metabolomics data sets including eliminating false discoveries and determining significance, metabolite clustering, and logical algorithms for discovery of new metabolites and pathways. Together these metabolomics tools represent an entirely new pipeline for phytochemical discovery.



INTRODUCTION

The term "metabolomics" was mentioned for the first time in 1998[1] and has been defined most recently as the "qualitative and quantitative analysis of all small metabolites in a biological sample at a specific time and influence".[2] Developments over the 15-year period subsequently have included significant advancements in analytical technologies, statistical tools, and metabolite databases.[3−5] A Web of Science search on March 10, 2014, using the terms "metabolomics" or "metabonomics" found 8414 publications, and further analysis has revealed that the number of manuscripts published each year has increased steadily since 2002 (R^2 = 0.9931) (Figure 1a). The field of metabolomics is inherently multidisciplinary, and although most studies are based in biological enquiry, successful research teams require expertise in diverse fields, including bioinformatics, analytical chemistry, instrumental sciences, and metabolic biology.[6]

Plants have one of the largest metabolomes, and it is estimated that the average plant leaf contains upward of 30 000 phytochemicals.[7] This is likely attributed to the fact that plants must possess a phytochemical arsenal in order to withstand abiotic and biotic stressors present in the surrounding environment.[8,9] More than 200 000 different secondary metabolites have been isolated, purified, and identified from plants.[10,11] Assuming 300 000 to 350 000 plant species exist[12] on earth and that each contains 4.7 unique metabolites,[13] only a small fraction of the estimated 1.4 to 1.5 million novel chemical structures belonging to the plant kingdom have been found. In the past decade, >1200 articles have been published describing plant metabolomes (Figure 1b), investigating basic questions in plant metabolic processes,[14,15] plant responses to the environment,[16,17] synergy between compounds,[18] and biological activity[2,19] as well as applied studies to authenticate commercial crops through fingerprinting.[20,21]

Metabolomics data sets can investigate hypotheses or generate hypotheses depending on the experimental design and analysis.[22] A typical plant metabolomics analysis generated by a separation technology and mass spectrometry generates a data set of more than 500 000 observations (Figure 2). This wealth of data is both a blessing and a curse since the validity of a given study depends on how the data are managed and used. This review will describe the use of statistical tools for metabolomics data sets that detect and eliminate false discoveries, differentiate between individuals, discover relationships between samples, identify the most important phytochemical signals, generate metabolite families, and putatively identify new compounds and pathways to direct future investigations. These tools provide the backbone for a new compound-discovery pipeline based on metabolomics (Figure 3). There are three different approaches to metabolomics that support the compound discovery efforts: targeted metabolomics, untargeted metabolomics, and a hybrid technique of targeted−untargeted metabolomics.

Targeted Metabolomics. Targeted metabolomics is a term used for the processes of identification and quantification of the known phytochemistry in a biological sample. Targeted approaches use standards and/or spectroscopic data libraries,

Received: August 24, 2014
Published: March 9, 2015

DOI: 10.1021/np500667z
J. Nat. Prod. 2015, 78, 953−966

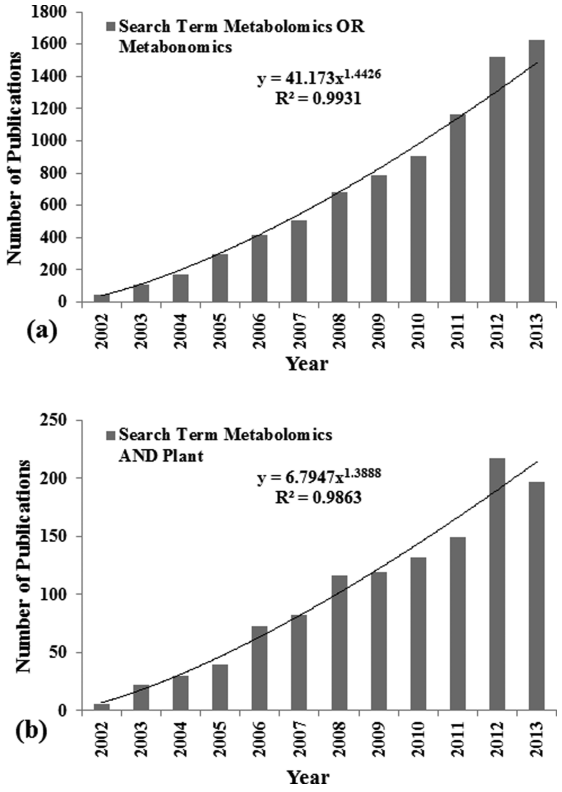


Figure 1. Rate of growth of publications reporting metabolomics and metabonomics studies over the past decade.

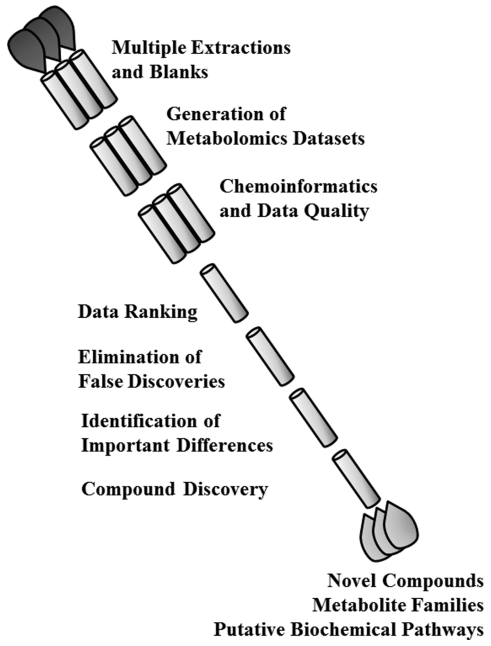


Figure 3. Metabolomics-guided compound discovery pipeline.

often in conjunction with other "omics" technologies (e.g., transcriptomics, genomics, and proteomics), to investigate hypothesis-driven research projects in order to answer a specific research question or questions.[23,24] Targeted approaches closely resemble traditional analytical methods that aim to quantify specific metabolites by optimizing extraction, separa-

Ion Intensity

29747 lines of data

Retention Time	Mass (m/z)	Cranberry fruits	Cranberry fruits	Cranberry fruits	Product 1 (capsules)	Product 1 (capsules)	Product 1 (capsules)	Product 2 (caplets)	Product 2 (caplets)	Product 2 (caplets)	Cranberry fruits	Cranberry fruits	Cranberry fruits
1.4579	100.0747	0	0	0	0	0	0	0	0	0	0	0	0.0148
8.2873	100.0752	0	0	0	0	0.1559	0	0	0	0	0	0	0
7.2062	100.0754	0.1125	0.0356	0.0407	0	0	0.1545	0	0	0	0	0.109	0.1035
2.0723	100.0756	0.0909	0.1449	0.0909	0	0	0	0.2647	0.1634	0.3205	0.2486	0.1281	0.0226
8.4672	100.0757	0	0	0	0.2598	0	0	0	0.1309	0.1278	0	0	0
7.8093	100.0757	0.1152	0	0	0	0	0	0	0.0505	0.2051	0	0	0
6.3382	100.0758	0	0.0615	0	0	0	0	0	0	0.3788	0	0.1117	0
9.1185	100.0758	0	0.0829	0	0	0	0	0	0	0	0	0	0
11.2907	161.0471	0.0462	0	0	0	0	0	0	0	0.0449	0	0	0
11.7713	161.0555	0	0	0.0291	0	0	0	0	0	0	0	0.0351	0.0323
3.4328	161.0631	0.0911	0.0453	0.0598	0.0579	0	0.0643	0.0589	0.0639	0.0678	0.0536	0.0363	0.0653
1.4537	161.0699	0	0	0.0305	0	0	0	0	0	0	0	0.0265	0.0203
7.3071	161.0727	0	0	0	0	0	0	0	0	0	0.0373	0	0
2.7688	161.0732	0	0.0599	0.0674	0	0.0184	0	0.0426	0	0.0148	0.0316	0.0682	0.0366
3.8549	161.0748	0	0.0395	0	0.0399	0.0414	0	0	0	0	0.0296	0	0
11.3764	161.0765	0.047	0.0282	0.0322	0	0.0547	0	0	0	0.0258	0	0.0316	0.0296
14.1015	161.0765	0	0.039	0	0	0	0	0	0	0	0	0	0
13.6362	213.1133	0	0	0	0	0	0	0	0	0	0	0.0381	0
14.4998	213.1138	0.0723	0	0	0	0	0	0	0	0	0	0	0.0399
14.7892	213.1138	0	0	0.162	0.1013	0.0785	0	0.079	0.1097	0	0	0	0
11.1704	213.1167	0	0	0	0.2414	0.0439	0	0	0	0.1804	0	0	0
12.6198	213.1173	0	0	0	0	0	0	0	0.1782	0.1232	0.0548	0	0
12.9116	213.1191	0	0	0.0877	0	0	0	0	0	0	0	0	0
14.2775	213.1203	0	0	0	0	0	0	0	0	0	0	0	0.1044
13.2287	213.1206	0	0.0533	0	0	0.0681	0.0931	0	0.1081	0.1259	0.0864	0	0
11.8914	213.1222	0	0	0	0	0.028	0	0	0	0	0	0	0
6.0577	213.1485	0.1901	0.488	0.2984	0.535	0.3562	0.3225	0.4672	0.5988	1.2925	1.0552	0.5779	1.0415

Figure 2. Screenshot of a typical metabolomics data set generated by reversed-phase separation on ultraperformance liquid chromatography (elution time) and mass-to-charge (m/z) detection by time-of-flight mass spectrometry. Typical analyses generate data sets with 35 000 to 50 000 lines per sample. Computing capacity determines the limit of samples that can be analyzed in a single data cassette. Blank injections, solvent blanks, and extraction blanks provide signals to eliminate background noise and contamination.

DOI: 10.1021/np500667z
J. Nat. Prod. 2015, 78, 953−966

tion, and detection methods for a specified class or classes of compound.

Untargeted Metabolomics. In contrast, nontargeted analysis is the process of generating data sets representing the whole chemical spectrum of a biological sample or samples without bias.[25,26] Untargeted metabolomics approaches are hypothesis generating and can uncover patterns or relationships within the whole data set regardless of whether the identity of a specific analyte is known.[26−28] These nontargeted approaches have great potential in chemometric fingerprinting, compound and pathway discovery, and hypothesis generation, but standardization is difficult and the potential for false discovery is high.

Targeted−Untargeted Metabolomics. More recently, a hybrid of analytical approaches has enabled studies that combine technologies and advantages of targeted and untargeted approaches. Targeted−untargeted approaches can use leads from spectroscopy databases or other "omics" technologies to partially target specific metabolites. Untargeted data sets are mined for known compounds from different species or tissues. The discovery of new metabolites and putative pathways is possible beginning with known chemical leads and finding relationships between compounds in the data sets.[29−31] The resulting data can be used for generating novel hypotheses to target specific characteristic unknowns, to fingerprint samples or extracts, to discover new phytochemicals, to elucidate pathways, to associate chemometrics with physiological responses, and many other applications.[13,14,21,32]

CRANBERRY MODEL SYSTEM

This review will provide perspectives on the use and usefulness of diverse approaches from data generation and gathering to interpretation and discovery of novel compounds using a model system based on cranberry (*Vaccinium macrocarpon* Aiton) (Ericaceae). Cranberry ranks among the top-10-selling dietary supplements in the U.S. market and is used for maintenance of urinary tract health.[33,34] The model data set consisting of cranberry fruits and two cranberry products extracted using identical techniques in three different solvents is used to herein demonstrate these principles. The model database was not designed to address a specific hypothesis but rather to generate a data set suitable for developing data analysis approaches.

Model Samples. Cranberry fruits were obtained from commercial production (Ocean Spray Canada Ltd., Richmond, BC, Canada), frozen and maintained at −20 °C, then freeze-dried (Thermo Scientific Super Modulyo freeze-dryer; Fischer Scientific), and ground to a <60 mesh (250 μm) powder. Cranberry fruit preparations were purchased from two commercial vendors in Burnaby, BC, Canada (Table 1).

Solvent Extraction. The freeze-dried, ground samples were weighed (0.259 ± 0.025 g) into 15.0 mL conical centrifuge tubes (n = 9). Each test material type was extracted in triplicate with 10 mL of the solvent (methanol, 70% aqueous ethanol, or water) by vortexing for 1 min (Thermolyne Maxi Mix One vortex mixer; Fisher Scientific), sonicating for 20 min (Branson model 3510R-MTH ultrasonic cleaner; VWR), and centrifuging at 5000 rpm (4500g) for 5 min (Eppendorf tabletop centrifuge 15804R, VWR). The supernatant was collected and the solvent evaporated to dryness under nitrogen. The residue was then redissolved in methanol−water (80:20), vortexed, and filtered through a 0.45 μm PVDF syringe filter to HPLC vials for analysis.

Table 1. Model Cranberry Samples for Statistical Tools[a]

sample	manufacturer/source	product description label dosage/serving	ingredients
cranberry fruits	Ocean Spray Richmond, BC, Canada Ltd.	2007 harvest, lower mainland of British Columbia	whole cranberry fruits (*V. macrocarpon*), freeze-dried, ground to 60 mesh (<250 μm)
product 1 (capsules)	Swiss Natural Sources, Montreal QC, Canada	Cran-Max cranberry 34:1 fruit extract; cranberry juice powder 20:1 fruit extract; recommended dose: one 500 mg capsule daily	Cran-Max cranberry (*V. macrocarpon*) 34:1 fruit extract (250 mg/capsule); cranberry juice powder (*V. macrocarpon*) 20:1 fruit extract (250 mg/capsule)
product 2 (caplets)	Natural Factors, Coquitlam, BC, Canada	cranberry concentrate 36:1 (*V. macrocarpon* fruit juice); recommended dose: one or two 500 mg softgels daily	500 mg CranRich cranberry concentrate (*V. macrocarpon*, fruit juice) 36:1; softgel (gelatin, glycerin, purified water), rice starch, silica, magnesium stearate

[a]Samples were specifically selected to represent a diversity of data sets.

DOI: 10.1021/np500667z
J. Nat. Prod. 2015, 78, 953−966

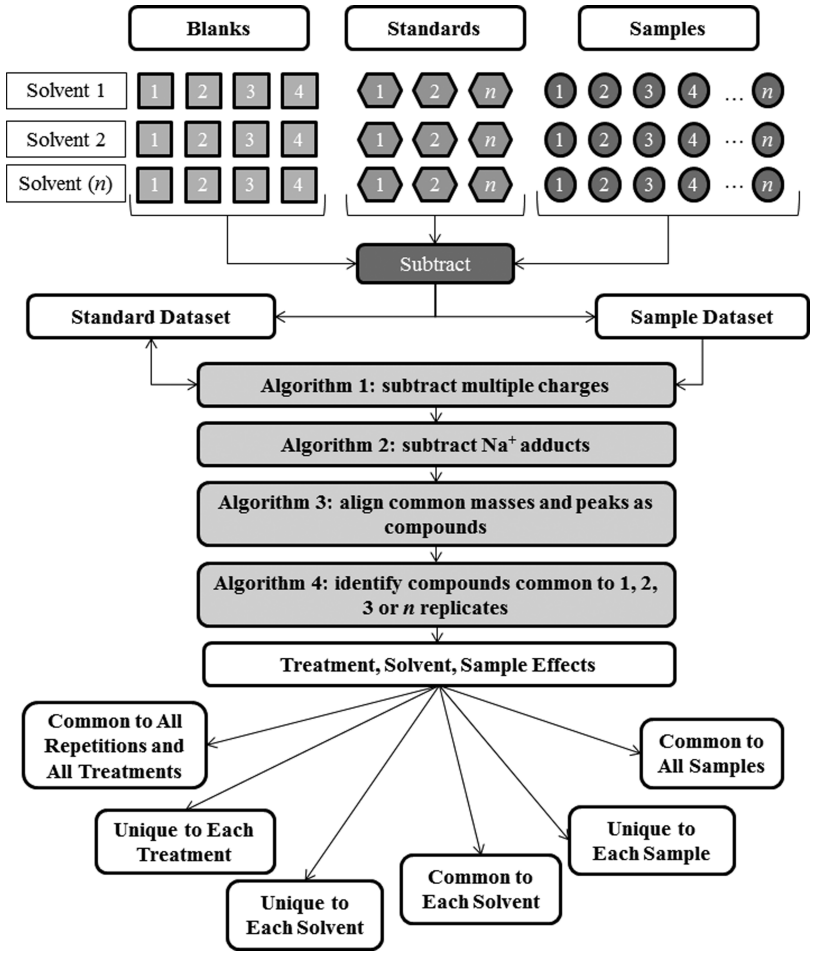


Figure 4. Experimental design for data generation and manipulation in subtractive metabolomics.

Ultrafast Liquid Chromatography with Time-of-Flight Mass Spectrometry. Experiments were performed with an ACQUITY Series Ultra-Performance liquid chromatography system (Waters, Inc., Mississauga, ON, Canada) coupled with a Micromass LCT Premier Series ToF-MS (Waters) and controlled with a MassLynx V4.1 data analysis system (Waters). Chromatographic separation was achieved with a Waters BEH Acquity C_{18} column, 2.1 × 150 mm, 1.7 μm, with the following mobile phase conditions: 0.1% aqueous formic acid−acetonitrile (0.0−10.0 min, 95:5−5:95 v/v, 10.0−15.0 min, 5:95 v/v, 15.0−20.0 min, 5:95−95:5 v/v, 20.0−25.0 min, 95:5 v/v). A 25 min run time was used with a flow rate of 0.25 mL/min and column temperature of 30 °C. The autosampler was at 4 °C with an injection volume of 5 μL. A Waters 1525 HPLC binary solvent manager provided a steady flow of 2 ng/mL leucine enkephalin at 10 μL/min.

Data Acquisition and Export. Data were exported from the mass spectrometer via MarkerLynx (Waters) into an ASCII text file with sample identifiers in colums (objects) and retention time, m/z ratio, and abundance as rows (variables) (Figure 2). The metabolomic data for each product, with blanks removed, were assessed without scaling or further data transformation. Excel (Microsoft), R (R Foundation, GNU), and Solo (Eigenvector Research, Inc.) were used for transposing data when needed, univariate statistics, principal component analysis (PCA), ANOVA simultaneous component analysis (ASCA), and metabolite set enrichment analysis (MSEA), as described in detail below.

METABOLOMICS DATA ANALYSIS

The overall objective of this review is to examine the effectiveness and challenges of the different metabolomics approaches using a model data set. The data set was generated with a whole plant sample and two different types of plant preparations in order to represent diverse sample matrices. It was expected that the extraction efficiency, types of compounds extracted, and overall metabolome would vary significantly. Further, the preparations contain various fillers and manufacturing agents that could complicate the analysis. The model data set was designed to demonstrate the potential of metabolomics for understanding the full spectrum of chemistry in prepared natural products.

Chemometrics/Chemoinformatics. The treatment of metabolomic data by multivariate data analysis to develop models for meaningful interpretations is known collectively as "chemometrics". Chemoinformatics is the specific use of models to generate information about the data set. This may incorporate classification accuracy, model sensitivity, and derivations of eigenvalues from multivariate models, all of which provide tools for the assessment of the metabolome data under study.[35] Several chemometric approaches were used with the model data set to identify unique compounds and to rank importance of these compounds.

Metabolite Counts and Subtractive Metabolomics. The first challenge of metabolomics data set characterization is to capture all of the information generated in the analysis and to

DOI: 10.1021/np500667z
J. Nat. Prod. 2015, 78, 953−966

Table 2. Subtractive Metabolomics Analysis of the Model Dataset

all extracts				
		cranberry	product 1	product 2
A1	common to replicates for all solvent	7080	6311	6157
methanol extracts				
		cranberry	product 1	product 2
M1	average number of compounds in each methanol extract	9826 ± 473	8593 ± 831	8508 ± 816
M2	total number of compounds extracted by methanol	16 475	16 005	15, 148
M3	total number of compounds common to all methanol extracts	7635		
M4	percent of common chemistry	46	48	50
M5	number of compounds only extracted in methanol	2570	3627	3222
M6	number of compounds unique to product type	3025	2028	2034
M7	percent abundance of unique compounds (%TIC)	2.9	5	3.7
70% aqueous ethanol extracts				
		cranberry	product 1	product 2
Et1	average number of compounds in each 70% ethanol extract	7445 ± 3446	7019 ± 453	6864 ± 610
Et2	total number of compounds extracted by 70% ethanol	14 944	13 072	13 384
Et3	total number of compounds common to all 70% ethanol extracts	5832		
Et4	percent of common chemistry	39	45	44
Et5	number of compounds only extracted in ethanol	2035	1974	2420
Et6	number of compounds unique to product type	3636	2220	2604
Et7	percent abundance of unique compounds (%TIC)	5.9	3.7	15.2
water extracts				
		cranberry	product 1	product 2
W1	average number of compounds in each water extract	7959 ± 630	7061 ± 254	6704 ± 296
W2	total number of compounds extracted by water	14 710	13 403	12 746
W3	total number of compounds common to all water extracts	5664		
W4	percent of common chemistry	39	42	44
W5	number of compounds only extracted in water	2513	2444	2231
W6	number of compounds unique to product type	3213	2194	2698
W7	percent abundance of unique compounds (%TIC)	4.2	4.1	4.6

develop some parameters for assessment of the quality of the data. A simple series of algorithms and functions was designed in Excel to remove signals present in blanks and to determine whether an individual signal is present in some or all replicates of a sample. The algorithms can also be used to identify compounds that are unique to a sample or extract and are termed "subtractive metabolomics"[20,21] (Figure 4). Evaluation of the metabolomics fingerprint of the model data set allowed for identification of several interesting patterns within the data set (Table 2). About 16 000 distinct compound signals were found in the metabolomic profiles of the cranberry fruits and cranberry products, with approximately 7000 compounds found in all replicates of the fruits regardless of solvent used. About 39−50% of the phytochemistry described by the metabolomic fingerprinting was consistent between the cranberry fruits and the two commercial products, giving an indication of the number of plant-based compounds that are conserved in the manufacturing process (Table 2, lines M4, Et4, W4). In the cranberry fruits and products, 2570, 3627, and 3222 compounds were found to be unique to the methanol extracts and not in the other solvents (Table 2, line M5) and 3025 of the compounds identified in the methanol extracts of cranberry fruits were not found in the methanol extracts of either of the products (Table 2, line M6). Similarly, chemical fingerprinting of ethanol extracts of cranberry fruits identified 3636 compounds that did not appear in the fingerprints of the products (Table 2, line Et6), and the water extracts were found to contain 3213 compounds not found in either product (Table 2, line W6). Together, these data help to identify phytochemical constituents of cranberry fruits that are retained or lost in the processing and can be used for improved product quality. Further, the identification of a few, high-concentration individual compounds that are unique to the cranberry fruits or their products provides potential biomarkers for product quality and efficacy (Table 2, lines M7, Et7, and W7).

Data Distribution. Chemoinformatics analysis further included an assessment of the data quality by checking the distribution of results. Quantile−quantile (or Q−Q) plots were generated for mean-centered experimental data sets from the stats package available in R using the qqnorm algorithm, which plots calculated quantiles for observed data versus theoretical quantiles for normal distribution of the data.[36] The abundances of compound signals in the experimental data sets [mass to charge (m/z) ratios of 100.0747 to 999.7201] were examined, and Q−Q plots were generated to compare the mean-centered experimental data distribution to a theoretical normal distribution of data for individual experimental conditions, i.e, by product and solvent (Figure 5). Also, comparative Q−Q plots were generated to directly compare the probability distribution between mean-centered sets of products within the same solvent system (data not shown). This was done using the qqplot algorithm in R, which plots calculated quantiles for each data set being compared. According to Figure 5, for analysis of the model data set, the distribution of experimental data for the cranberry fruits in each of the three extraction solvents was compared directly to the theoretical distribution, and the experimental distribution did not follow a normal distribution pattern. The nonparametric distribution of the data for each

DOI: 10.1021/np500667z
J. Nat. Prod. 2015, 78, 953−966

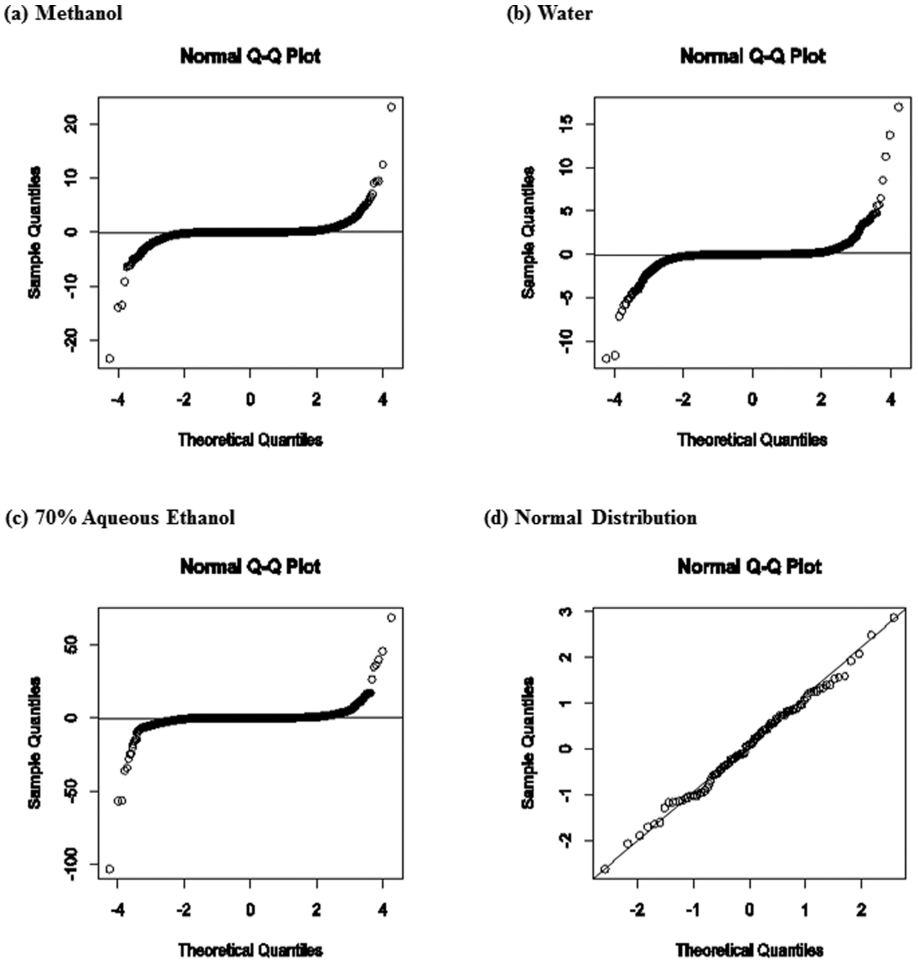


Figure 5. Q–Q plots, observed data quantiles (open circles) versus theoretical quantiles for normal distribution. The solid line represents a theoretical matched distribution between normal data and the observed data. Cranberry fruit product extracted in (a) methanol, (b) water, and (c) 70% ethanol. (d) Example of randomly generated normal data.

Table 3. Statistically Significant Compounds Determined by the Kruskal–Wallis Test and Total Area under the Curve Method

		methanol		70% aqueous ethanol		water	
Kruskal–Wallis p value	total AUC	number of metabolites	m/z value	number of metabolites	m/z value	number of metabolites	m/z value
0.0211	0.8333					1	863.3893
0.0221	0.8333	688		254		488	
0.0234	1.0000					1	100.0756
0.0241	1.0000	49		20		100	
0.0265	1.0000	38		17		71	
0.0273	1.0000	57		19		120	
0.0329	0.9444	2					
0.0336	0.8889					1	437.1010
0.0340	0.9444	235		97		130	
0.0347	0.8889	516		198		376	
0.0349	0.9630	102		6		37	
0.0379	0.9630	43		14		29	
0.0390	0.9630	56		16		61	
0.0394	0.9444	1	859.2513				
0.0439	0.9444	1	667.3990				
0.0447	0.9259	1	765.5578	1	512.3705	1	367.2658
0.0458	0.8704			2		1	409.1991
0.0459	0.9259	163		57		104	
0.0484	0.9259	1	772.3100	1	463.0239		
0.0496	0.9259	34		14		35	
	total	1987		716		1556	

DOI: 10.1021/np500667z
J. Nat. Prod. 2015, 78, 953–966

experimental condition was found to be unique. All statistical operations for evaluating data variance and significance were selected for suitability of use with nonparametric data sets.

ELIMINATING FALSE DISCOVERIES

False discoveries are analytical responses that do not correspond to genuine compounds in the original tissue or extract. In metabolomics analyses, false discovery rates (FDR) include artifacts of the extraction, contamination from solvents, glassware, or instruments, and so on. In statistical analysis, FDR can refer to incorrect assignments. Avoiding false discoveries when interpreting metabolomic data usually requires that more than one approach or model be assessed.[37]

Univariate Analyses. Three different approaches were taken to determine the potential significance of data within the entire data set and subsets of data grouped by extraction solvent. These three approaches are the calculated Kruksal–Wallis p value[38] and the significance analysis of microarrays (SAM) statistic[39] as well as the area under the ROC (receiver operating characteristic) curve (AUC).[40,41]

Kruskal–Wallis p Value. The Kruskal–Wallis test is a variance test used in place of ANOVA when one cannot assume a normal distribution of the data.[42] The use of the Kruskal–Wallis nonparametric method to discriminate the cranberry fruits from the two finished cranberry products provides identification information for the important metabolites in the data sets. Each metabolite was assigned a p value calculated with the Kruskal test algorithm in R. The Kruskal–Wallis test indicates if the differences between conditions are sufficiently large so that they are unlikely to have occurred by chance with an associated significance p value per metabolite of <0.05. To illustrate typical results, the m/z values for compounds with p values less than 0.05 are presented in Table 3. For all product types extracted with methanol (Table 3), 688 compounds were found to have a p value of 0.0221, which represented the case where the compound was detected in only one of the three product types. Next, three sets of p values were observed (0.0241, 0.0265, 0.0273) and could be associated with m/z values of 144 compounds. Some observations can be made from these three sets of p values. For example, at a p value of 0.0273, there were 57 compounds, of which 35 were found in all replicates of all products at abundance levels distinguishable per product type. This p value of 0.0273 merits a further comment: more compounds of true significance (i.e., within a p value of <0.05) do exist in the comparison of all three product types existing in all replicates, but the Kruskal–Wallis test missed compounds and these are not reported. The remaining 22 compounds also existed in all products, but were found to be undetectable in a single replicate of one of the products. For the p value of 0.0265, 38 compounds were found to be in all product types, but undetected in two replicates for one of the products. Compounds found to be present in two of the three product types, of which there were 49, had a p value of 0.0241. In the case where the compound was found only in the finished products, as opposed to the cranberry fruits, it is possible the compounds either are a processing artifact or are concentrated during the manufacturing process and are actually present in the cranberry fruits but below the detection limit. Whether focusing on differentiation of products and cranberry fruits or finding those metabolites found in products and their cranberry fruit origins, sorting the compounds detected by calculating the associated p value provides a way to prioritize characterization. Looking at the compounds retained in the finished products from the cranberry fruits, there were only 35 compounds found in common when extracted by methanol. Considering the effects of the other two solvents under study, five and 78 compounds were found in common in both products when extracted by 70% ethanol and water, respectively (Table 3).

Significance Analysis of Microarrays Statistic. Modifications of t-statistics techniques exist in the literature including the significance analysis of microarrays statistic[39] and are meant to deal with limitations of original univariate techniques. Additionally, the SAM statistic provides a degree of estimation in falsely discovering metabolites of significance (FDR based on permutations). The SAM statistic and associated FDR were calculated for the model data set using the sam algorithm from the siggenes package in R for subsets of data grouped by the extraction solvent category. Due to the limited number of replicates in the model data set (n = 3), permutations were used to set the expected $d(i)$ as the null level of abundance, which allows for the comparison of the observed $d(i)$ based on actual data and the expected $d(i)$ (Table 4). To consider metabolites differing significantly in terms of abundance, an artificially selected threshold (Δ) was applied to flag metabolites beyond this threshold boundary. The associated FDR was determined by starting with the smallest $d(i)$ among the metabolites with largest abundance and the least negative $d(i)$ among the metabolites with the smallest abundance and counting the number of metabolites that exceeded the horizontal cutoffs. The count is averaged from all permutations and then divided by the total number of metabolites called significant (Table 4). As the threshold Δ decreases, the number of significant metabolites increases (closer to the value of $d(i)_{\text{actual}} = d(i)_{\text{expected}}$), but this is at the cost of increasing FDR. For every comparison of the three product types by a common extraction solvent, a list of significant m/z values is given with an associated FDR (Table 4). Only the m/z values identified as significant by the SAM statistic (FDR 10%) extracted by methanol were examined.

Using the Kruskal–Wallis and Significance Analysis of Microarrays Statistic to Eliminate False Discovery. Analyzing the cranberry model system with the SAM statistic is an effective tool for elimination of false discoveries (Table 4). Of the 688 compounds identified by the Kruskal–Wallis and AUC methods (at p value: 0.0221 and total AUC: 0.8333, extracted with methanol), only five were identified by the SAM statistic as significant and only known to exist in cranberry fruits and not in the other two product types. Next, only two compounds present in all test materials were identified by the SAM statistic. One compound (954.6111) came from a group of 57 compounds (p value: 0.0273 and total AUC: 1.0000) identified by the Kruskal–Wallis and AUC methods, a considerable reduction (Table 3). The other compound came from a group of 163 compounds at a p value of 0.0459 and total AUC of 0.9259. Since the SAM statistic originally was designed to minimize the identification of significance difference of minor size in the comparison of a small variance gene taken from different experimental conditions, further interpretation of compounds being present in all test materials can be conducted.[43] For example, this t-statistic modification can be used to select compounds existing in all products but at a much higher amount in one of those products, as this is based on high specificity of a compound in a product but not in other products. This is seen by the exclusion of significant compounds present in all three products identified by the Kruskal–Walls and AUC methods to only two compounds by

DOI: 10.1021/np500667z
J. Nat. Prod. 2015, 78, 953–966

Table 4. Significant Compounds Identified by SAM Supplemented with the Kruskal–Wallis Test and Total AUC Data

description of occurrence	m/z value	$d(i)$	p value	total AUC
methanol extracts: calculated false discovery rate = 10%				
observed only in cranberry fruits	845.7227	1772	0.0221	0.8333
	779.435	1697	0.0221	0.8333
	343.0365	2357	0.0221	0.8333
	261.1252	3824	0.0221	0.8333
	222.1465	6082	0.0221	0.8333
observed in all test materials	954.6111	2131	0.0273	1.0000
	247.1691	3755	0.0459	0.9259
70% aqueous ethanol extracts: calculated false discovery rate = 15%				
observed only in cranberry fruits	318.3207	328	0.0221	0.8333
observed only in product 1	885.5636	326	0.0221	0.8333
	881.1926	638	0.0221	0.8333
	841.2387	303	0.0221	0.8333
observed only in product 2	966.5259	275	0.0221	0.8333
	864.771	351	0.0221	0.8333
	449.8076	353	0.0221	0.8333
	428.2075	271	0.0221	0.8333
	420.268	1786	0.0221	0.8333
	266.1803	432	0.0221	0.8333
observed in all test materials	841.7087	1307	0.0273	1.0000
	387.1789	2730	0.0549	0.9074
	326.3762	274	0.0605	0.8519
	261.2195	1101	0.0459	0.9259
observed in product 1 and product 2	417.2127	400	0.0347	0.8889
	219.2105	311	0.0347	0.8889
observed in cranberry fruits and product 2	475.3222	308	0.0347	0.8889
water extracts: calculated false discovery rate = 7%				
observed only in cranberry fruits	739.5641	1795	0.0221	0.8333
	613.3321	7122	0.0221	0.8333
	200.187	3797	0.0221	0.8333
	194.1179	2066	0.0221	0.8333
observed only in product 1	549.1552	3807	0.0221	0.8333
observed only in product 2	955.2333	1895	0.0221	0.8333
	329.9232	4716	0.0221	0.8333
observed in all test materials	276.1068	4076	0.0552	0.8889
observed in product 1 and product 2	559.1321	4232	0.0347	0.8889

the SAM statistic. Accordingly, these two compounds (m/z = 954.6111 and 247.1691) showed significant differences of major abundance in the three products under study, with small variance seen per product. Analogously, the five compounds that are observed only in cranberry fruits (p value: 0.0221 and total AUC: 0.8333) and identified as significant by the SAM statistic are those with major size differences with small variance when comparing the three products under study. Those compounds with a p value of 0.0221 and total AUC of 0.8333 do not represent the case of major size differences with small variance and thus are not identified by the SAM statistic.

Receiver Operating Characteristic. ROC is an example of the use of predictive modeling to assess metabolomics data by application of a diagnostic statistic that evaluates the optimization and performance of models.[40] By evaluating the statistical significance of a model data set, it is possible to generate a meaningful interpretation of the data. In particular, ROC is useful for detecting small differences between groups in complex data sets as a comparison between the true positive rate (TPR) and the false positive rate (FPR). For the analysis of the model samples, the TPR was equivalent to the sensitivity of the method and the FPR is equal to 1-specificity. For the model cranberry data set, a binary comparison was used to classify products within a solvent category, and ROC values were generated for each model and plotted as sensitivity (TPR) vs 1-specificity (FPR). The ROC curve provides a spectrum of performance assessments for the data set that is interpreted by measuring the AUC. At an AUC value of 1.0, metabolite classes are correctly discriminated, but at low values, it is not possible to discriminate between metabolites. For the model data set, the ROC values were calculated in R and plotted to determine the total AUC values.

Area under the Curve of the Receiver Operator Characteristic. The importance of compounds in discriminating cranberry fruits from the two-finished product forms was evaluated by comparing the total AUC of the ROC for the model data sets grouped by extraction solvent. The m/z values for compounds with total AUC $\geq$ 0.8333, a reflection of distribution across the product types and replicates within a product, are presented in Table 3. When comparing data from a given extraction solvent from the cranberry fruits and two finished products, the total AUC of one reflects two discovered distribution scenarios: either the compound exists in all product types with up to two absent replicates in a product type or the compound exists in two product types but with variation in abundance. For the other total AUC values represented, the compound is present in all or up to two product types with the AUC value being dependent on the absence of replicates within product types and/or similarity in abundance levels between product types. For analysis of the model cranberry data set, the AUC value of each metabolite was obtained from the colAUC algorithm within the caTools package in R.[44] The mean of the three calculated binary AUC values per metabolite was obtained and designated as "total AUC" within each solvent category.[41] Values for which the total AUC is greater than or equal to 0.8333 were identified as significant.

Using the Kruskal–Wallis and Other Statistical Parameters. When comparing the results of the statistical analyses of the model metabolomics data set, the findings were consistent with the interpretation of the Kruskal–Wallis p value analysis. For example, for the samples extracted with methanol (Table 3), 688 compounds were found to have a total AUC of 0.8333, and a total AUC value of 1.0 was found associated with 144 compounds (with three sets of p values observed). Those observations reflect possible scenarios depending on the differentiation in abundance, providing a more meaningful picture to the interpretation of the Kruskal–Wallis p values. For example, having a total AUC value of 1.0 and being present in all product types did not necessarily mean such compounds are equally detected in all product types. Rather, these compounds vary when being detected, in that their AUC value would differ when observing the associated ROC curve at high sensitivity values. Further exploration of partial AUC values (at a selected high sensitivity value; 0.95 for example) of such compounds would indicate the ones that are rated as compounds being detected, while those found to be poorly detected would have lower partial AUC values (and occur with a higher rate of false negatives).

Multivariate Analyses of Metabolomic Data. Metabolites in a biological sample are interconnected in pathways and families and therefore are not truly independent variables.

DOI: 10.1021/np500667z
J. Nat. Prod. 2015, 78, 953–966

Multivariate analyses are applied to identify those metabolites that are most important in differentiating samples when the variance of several factors is considered.

Principal Component Analysis. By far, the most common statistical analysis of metabolomics data is principal component analysis.[45,46] For the model data set, PCA was applied as a method to visualize variance for the multivariate data set by creating covariance matrices with transformation into a coordinate system (Figure 6). In all cases, autoscaling was

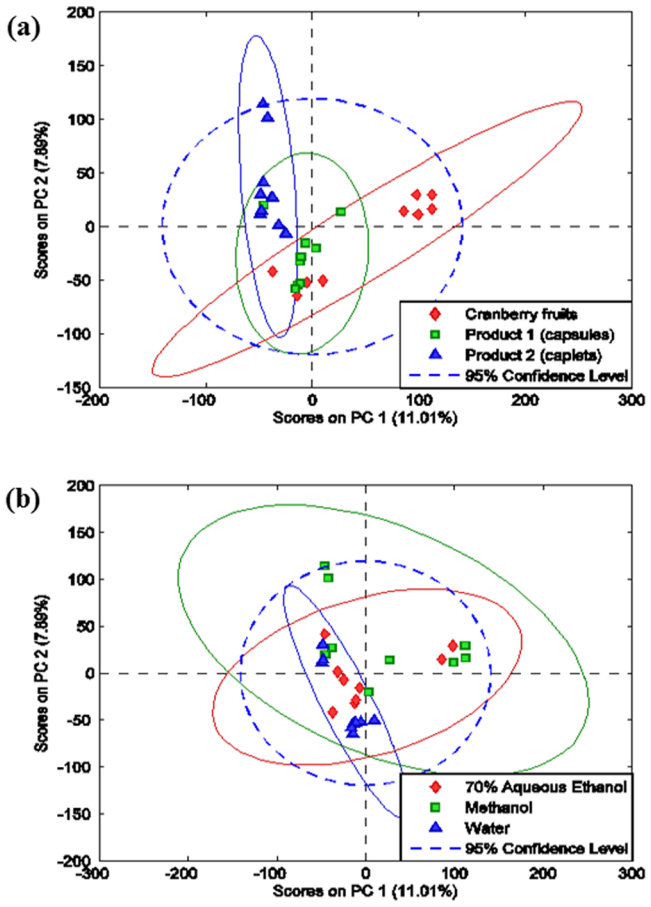


Figure 6. PCA score plots generated for LC-TOF-MS profiles of extracted freeze-dried cranberry fruits and for product 1 and product 2 in methanol, 70% aqueous ethanol, and water, grouped by (a) product type and (b) extraction solvent.

selected for preprocessing before applying the PCA algorithm. The resulting PCA score plot has principal components on each axis that represent a reduction of the number of data dimensions to those preserving most of the variance seen in the measured data set (i.e., the samples). The score plot of the first and second principal component from the PCA model allows visualization of differentiation in the data for all solvents by the product type (Figure 6a) and for all products by the extraction solvent (Figure 6b). Despite the differences noted in the metabolite counts for each solvent, overall the data set is more clearly differentiated within the 95% confidence boundaries by product type than by the varied extraction solvents. It is interesting to note that while almost all replicate extractions of product 1, regardless of solvent, fall within the 95% confidence boundary around the cranberry fruit extracts, there is significantly less overlap between the cranberry fruits and product 2 (Figure 6a).

The associated loadings plot reflects the influence of the variables in the PCA model generated. To accurately interpret loadings plots as generated by PCA algorithms, the method of Yamamoto et al. was used to determine individual significant metabolites. Briefly, the p value of each compound existing in the data set was determined by the "pca_scaled" function in the "mseapca" package within the R environment.[47] In a comparison of the sample types as extracted by methanol—cranberry fruits, product 1, and product 2—and observing that the cranberry fruits sample replicates was positioned on one side of the first principal component (PC1) axis in the scores plot and the processed products on the other side of the same axis, only the PC1 loadings plot data were used. Using the PC1 loadings plot data, the Yamamoto et al. method indicated that 4028 compounds were significant at a p value of <0.05.

On comparing the results of this multivariate analysis with the compounds identified as significant by the AUC and Kruskal–Wallis p value tests (for example, total AUC of 0.8333 and p value of 0.0221), 582 compounds were identified with a Yamamoto et al. p value of <0.05 (84.5% of 688 total). If the AUC value is set to 1.0 and a series of p values of 0.0241, 0.0265, and 0.0273 is applied, altogether 113 compounds were identified as significant (78.5% of 144 total). The remaining compounds above a total AUC value of 0.8889 and below a p value of 0.0496 reflect the same pattern of compounds found to be significant by the comparisons of the two approaches. On examining the SAM statistic results of the methanol-extracted samples, all SAM significant compounds were associated with a p value of <0.05 as calculated by the Yamamoto et al. method. This reduction of significant compounds comes from the association of these compounds with the data illustrated on the PC1 axis only, arising from use of the PCA algorithm to establish a correlation between the position of the products (in scores) and the compounds (in loadings).[47]

Analysis of Variance and Principal Component Analysis. The main goal of performing an analysis of variance–principal component analysis (ANOVA-PCA) (similarly, ANOVA–simultaneous component analysis) is to visualize and cluster the data set while taking into account any preknown factors, possible interactions, and noise.[48−51] For example, along with the designated factors to be studied in an experiment, the measurement of variation due to biological, technical, and analytical sources is also possible. This approach allows for the understanding of the studied factors and whether those factors influence the end result of the data set. For the cranberry fruits, product 1, and product 2 data set, for example, the postharvesting procedures would be interesting to study if they were known. This comparison of cranberry fruits and their products (e.g., capsules and caplets) would be useful to understand the effects of processing on the cranberry metabolome.

To exemplify this method, ASCA was utilized to examine the factors known in the cranberry model system, the product type (cranberry fruits, product 1, and product 2) and the solvent type used for extraction (methanol, water, and 70% ethanol). As Figure 7 shows, the two factors examined were found to be significant (using a null hypothesis of no experimental effect) to a p value of <0.001 as generated in the ASCA tool used in Solo, utilizing 1000 permutations.[52] The product factor clearly shows that the two processed products (product 1 and product 2) are different from the cranberry fruits, consistent with what was seen in PCA itself. In contrast, the solvent factor is less obvious as being significant, with the 70% ethanol replicates clearly

DOI: 10.1021/np500667z
J. Nat. Prod. 2015, 78, 953–966

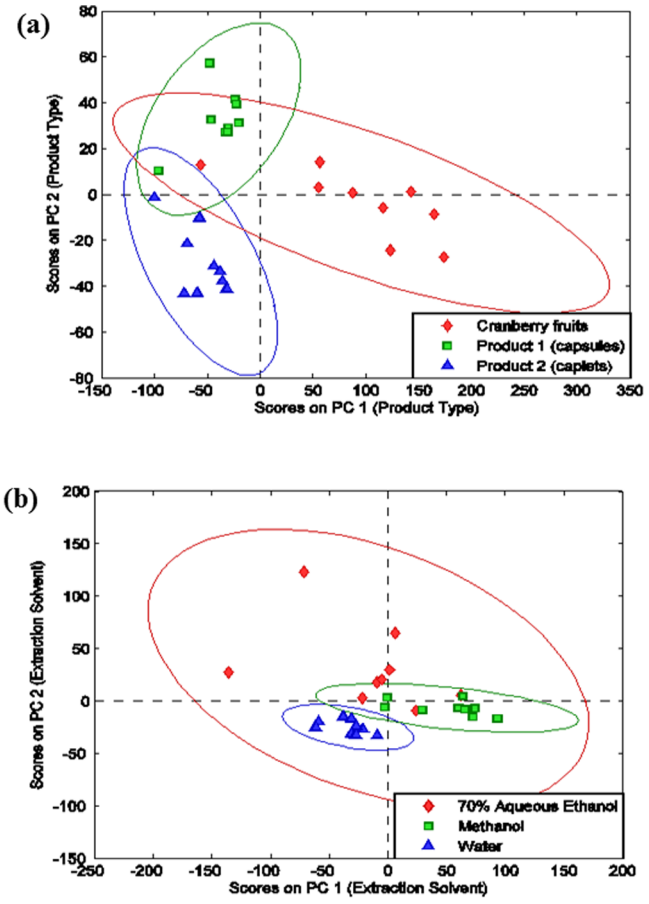


Figure 7. ASCA score plots generated for LC-TOF-MS profiles of extracted freeze-dried cranberry fruits and for product 1 and product 2 in methanol, 70% aqueous ethanol, and water, grouped by (a) product type and (b) extraction solvent.

more dispersed than those extracted with methanol and water. However, the methanol and water replicates were found to be more distinguishable, indicating differences between the two solvents. In terms of the factors contributing to the sum of squares (i.e., total variation) of the cranberry model system, the product factor contributed 20.43%, while the solvent factor contributed only 7.27%, with the overall mean at 51.50% and residuals at 20.81%. These data indicated that the comparison of samples was more important than the extractions, but the extraction solvent may be a complicating factor in the analysis.

Multivariate Curve Resolution. Curve resolution methods (i.e., multivariate curve resolution, MCR) allow the data set under study to be resolved into two defined matrices, typically concentration (the C matrix, i.e., chromatograph profile) and spectra (the S matrix, i.e., the mass spectrometric profile) matrices.[53,54] These two matrices have both physical and chemical meaning, as opposed to other multivariate methods (such as the use of scores and loadings, having no chemical meaning). The major limitation to the use of MCR as a deconvolution method depends on the complexity of the data set under study. The complexity of such a data set under study can be determined by the total rank (i.e., from eigenvalues vs principal component plots). The total rank has to be low enough for the deconvolution method to resolve successfully pure spectra known to exist in the data set. They may be resolved using a hierarchical approach, commonly named "hierarchical MCR", where the chromatograms are split up into windows, with each window exhibiting a total rank low enough to resolve the pure spectra that are desired to be measured. If the pure spectra can be resolved per window segment, the associated concentration profile data can then be further analyzed using supervised methods to obtain loadings plot data, which can then be used for identifying significant compounds using other techniques.

Supervised Analysis Using Classification Methods Based on Regression. Regression methods are utilized to measure data sets (as generated by ToFMS or NMR spectroscopy) to obtain a vector (or matrices) of predictions desired. This is done typically due to an inability to measure directly the desired property due to cost or difficulty. Classification methods based on regression exist, with partial least squares discriminant analysis (PLSDA) and soft independent modeling by class analogy (SIMCA) most commonly utilized, which are typically named "supervised" methods.

From the literature, seemingly important variables identified by supervised methods (PLSDA, SIMCA) as well as non-supervised methods (PCA) are often found not to be influential when the variable (compound) is isolated and the data remodeled.[37,55] For supervised methods, the major concern is the issue of overfitting the data set under study. One solution to this concern is the use of cross validation, where a subset of the data set is modeled and RMSECV data are obtained to evaluate the calibration set. Last, a subset of the data set not included in the calibration set should be tested to obtain RMSEP values to properly evaluate if the established model does indeed predict correctly. One limitation to using cross validation is the use of sufficient samples (and their replicates). For the cranberry data set consisting of three products as described in earlier sections of this review, this has not been done due to the number of replicates existing (n = 3 per product), representing only a limited sample size. In brief, supervised methods are useful to obtain a set of compounds found to be strongly associated with the property desired, as long as precautions (cross validation) and statistical tools are utilized when interpreting loadings plots of supervised methods.

USING METABOLOMICS FOR COMPOUND DISCOVERY

One of the greatest potential applications of metabolomics research is the capacity for the discovery of novel phytochemicals and pathways. Using comparative metabolomics between active and nonactive extracts, species within a genus, treatment groups, or individuals within a species, it is possible to discover new compounds. New compound discovery takes advantage of the richness of the data set for predictive deconvolution and putative identification. Follow-up studies with standard phytochemical procedures and bioassays are used to determine the structure, configuration, and activity of novel molecules.

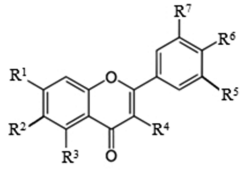
Synthetic Biotransformation. Secondary metabolites in plant tissues and preparations are not truly independent of each other but are largely the products of enzymatic reactions within cells. As such, there are a finite number of biochemical reactions that can occur, and most can be captured with simple algorithms to search for the products of enzymatic reactions.[14,15,56]

Synthetic biotransformation is the process of determining the products of enzymatic reactions by addition of functionalities such as amine, carboxyl, hydroxy, methyl, methoxy, or epoxy

DOI: 10.1021/np500667z
J. Nat. Prod. 2015, 78, 953−966

groups, as well as sugars, to the monoisotopic mass of a specific metabolite.[14,56] Predicted m/z values are then identified within the data set to generate interrelated metabolite families, putative biochemical pathways, or novel hypotheses for plant responses to external stimuli, the environment, or other factors.[15,57] This approach can also be used to identify unknown compounds responsible for the biological activity of extracts.[14] As an example of this approach, synthetic biotransformations were performed for flavone and anthocyanin compounds in the model data set. There are several precautions for the data set to make this possible. First, the mass spectrometer must be stable and calibrated regularly. In the case of this model example, the lockmass with leucine encephalin was determined and corrected for mass drift eight times per second. Second, the m/z values must be accurate to at least the third decimal place. In preliminary studies (data not shown) the time-of-flight mass spectrometer was stable to ±0.0003. For compound discovery, the calculated mass for each biotransformation was queried (±0.02 Da) within the raw data set using methods previously described.[14,15] Logical algorithms (= IF(cellid >0.01,1,0)) were applied in order to identify compounds that were detected in all three samples and possessed counts greater than 0.01. Compounds meeting the above criteria were selected for further analysis.

For example, starting with the monoisotopic masses for flavone (monoisotopic mass = 215.0133, Figure 8) and

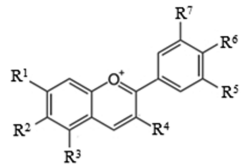


number of compounds detected as flavones	R^1	R^2	R^3	R^4	R^5	R^6	R^7
11	OH	OH	OH	OH	OH	H	H
11	OH	OH	H	H	H	H	O-glucose
11	OH	OH	OH	H	H	H	O-rhamnose
9	OH	OH	OH	H	H	H	O-glucose
9	OH	OH	OH	OH	H	H	O-rhamnose
8	OH	OH	OH	OH	H	H	H
8	OH	OH	OH	OH	OH	OH	H
5	OH	OH	OH	H	H	H	H
5	OH	OH	OH	H	H	H	O-arabinose

Figure 8. Most common biotransformations of flavones in the model cranberry data.

anthocyanin parent structures (monoisotopic mass = 200.0262, Figure 9) the application of the synthetic biotransformation algorithm identified the following transformations: +OH (monoisotopic mass = 17.0027), +H (monoisotopic mass = 1.0078), +O−CH_3 (monoisotopic mass = 31.0184), +O-glucose (monoisotopic mass = 179.0555), +O-rhamnose (monoisotopic mass = 163.0606), and O-arabinose (monoisotopic mass = 149.0445) (Figure 8). Similar processes were used to generate predicted metabolites of anthocyanin (Figure 9), and these metabolite signal masses could be further transformed with predicted enzymatic reactions, resulting in 158 hypothetical derivatives in cranberry.

Overall, a greater number of anthocyanin derivatives was detected compared to flavone biotransformation for cranberry fruits, product 1, and product 2 treatments (Figures 8 and 9).



number of compounds detected as anthocyanins	R^1	R^2	R^3	R^4	R^5	R^6	R^7
11	OH	OH	OH	H	H	H	O-glucose
11	OH	OH	OH	OH	H	H	O-rhamnose
11	OH	OH	OH	OH	OH	OH	H
9	OH	OH	OH	OH	H	H	O-glucose
9	OH	OH	OH	OH	OH	H	O-rhamnose
8	OH	OH	OH	OH	OH	H	H
8	OH	OH	OH	OH	OH	OH	OH
8	OCH_3	OCH_3	H	H	H	H	H
8	OCH_3	OCH_3	OCH_3	OCH_3	OCH_3	OCH_3	OCH_3
8	OCH_3	OCH_3	OCH_3	OCH_3	OCH_3	OCH_3	OCH_3
6	H	H	H	H	H	H	H
5	OH	OH	OH	OH	H	H	H
5	OCH_3	OCH_3	OH	OH	OH	OH	O-rhamnose
5	CH_3	CH_3	CH_3	OH	OH	OH	O-arabinose
5	H	H	H	H	H	H	O-rhamnose
5	OH	OH	OH	OH	H	H	O-arabinose

Figure 9. Most common biotransformations of anthocyanins in the model cranberry data.

The total number of anthocyanin and flavone derivatives detected in each sample varied depending on the extraction solvent (Table 5). The largest numbers of anthocyanin and flavone derivatives were found in methanol extracts, with significantly fewer metabolites found in the 70% ethanol extraction (Table 5). The most common type of biotransformation detected was the OH/H/sugar modification. The overall composition of anthocyanin and flavone derivatives detected in cranberry fruits, product 1, and product 2 was also affected by the extraction solvent (Table 5).

Putative Identification. Several databases are currently available to assist with compound annotation such as The Dictionary of Natural Products, KNApSAcK,[13] Plant Metabolic Network (http://plantcyc.org), PubChem (http://pubchem.ncbi.nlm.nih.gov/), and Chemspider.[25] Fiehn et al. and Fukushima have comprehensively reviewed the application of metabolite databases for plant metabolomics; thus this topic will not be detailed here.[25,58,59] Although metabolite databases represent a resourceful tool, most are incomplete for the majority of plant species and are not universally compatible with all analytical platforms.[60]

For the cranberry model data set, the KNApSAcK database was used to determine the identity of the predicted metabolites identified by the synthetic biotransformations. Many of the observed 233 anthocyanin and 193 flavone biotransformations identified, 56% and 11%, respectively, could not be identified using KNApSAcK (Tables S1 and S2, Supporting Information). Delphinidin and glycosylated derivatives of aurantinidin and pelargonidin were the anthocyanins most commonly identified (Table S1, Supporting Information). Among the flavone biotransformation metabolites, pentahydroxyflavones and tri- and tetrahydroxylated glycosides were most commonly detected (Table S2, Supporting Information).

Metabolite Set Enrichment Analysis. Metabolite set enrichment analysis is a technique using a supervised generalized linear model to associate the property (i.e., the samples under study) to the predefined set of variables

DOI: 10.1021/np500667z
J. Nat. Prod. 2015, 78, 953−966

Table 5. Metabolomic Summary of Synthetic Biotransformations in Cranberry Fruits and Cranberry Products

	anthocyanin derivatives			flavone derivatives		
methanol	cranberry fruits	product 1	product 2	cranberry fruits	product 1	product 2
total number of biotransformations detected	323	296	294	289	272	272
average number of biotransformations detected	222 ± 2	192 ± 15	193 ± 10	185 ± 3	164 ± 17	170 ± 11
total number of biotransformations common to all treatments	63	47				
total number of biotransformations common to all replicates	145	106	111	106	73	86
total number of unique biotransformations	59	13	27	45	9	24
70% aqueous ethanol	cranberry fruits	product 1	product 2	cranberry fruits	product 1	product 2
total number of biotransformations detected	284	232	250	261	216	221
average number of biotransformations detected	153 ± 55	148 ± 9	141 ± 15	137 ± 51	131 ± 9	117 ± 13
total number of biotransformations common to all treatments	14	13				
total number of biotransformations common to all replicates	19	78	68	14	64	48
total number of unique biotransformations	3	27	19	0	28	13
water	cranberry fruits	product 1	product 2	cranberry fruits	product 1	product 2
total number of biotransformations detected	278	245	230	266	220	209
average number of biotransformations detected	170 ± 10	154 ± 2	148 ± 2	158 ± 7	129 ± 2	131 ± 3
total number of biotransformations common to all treatments	45	33				
total number of biotransformations common to all replicates	84	82	91	72	62	75
total number of unique biotransformations	21	7	28	26	9	29

(compounds).[61,62] These sets of variables (compounds) can be either known biochemical pathways or defined functional groups (i.e., those of phenolic groups, flavones, anthocyanins, etc.). MSEA can determine if the defined set of compounds is associated with the samples under study in some way. Typically, two approaches are used: competitive and self-contained testing. Generally, self-contained testing is recommended, as this rests on the assumption that samples are independent and compounds are not.[62]

The gt function in the globaltest package was utilized to determine the p value for the comparison of the products using all compounds.[62] Then, from the synthetic biotransformation, the list of calculated monoisotopic masses for anthocyanins was utilized to mine the data set for the presence of anthocyanins. With the class of anthocyanins found in the data set, this class was designated in the gt function as a subset to determine the p value for the comparison of the anthocyanin class between the two manufactured products and the cranberry fruits. Additionally, the p value of each compound in the anthocyanin class was determined using the covariates function in the globaltest package in the R environment.

For example, as 23 541 metabolites were found by extraction by methanol, the comparison of all compounds in the three data sets resulted in a p value of 3.49×10^{-4}, an indication of the difference between the sample types being significant. Next, using the phenolic class of anthocyanin derivatives in the model data set, this class could be detected within the cranberry set, as a total of 25 compounds from a list of 57 theoretically possible anthocyanin structures. With MSEA, the anthocyanin class was compared between the manufactured products and the cranberry fruits, which indicated a p value of 2.07×10^{-2}.[61,63] To examine this significant difference further, the individual compounds in this class were examined, with 13 of 25 compounds identified with a p value of <0.05, representing significant differences in abundance between the three sample types. In comparison to the univariate statistics collected, none of the 13 compounds were associated with a total AUC of >0.8333 and none could be associated with a Kruskal–Wallis p value of <0.05. Also, none were found to be significant as determined by the SAM technique. Only three of the 13 compounds (m/z 404.1055, 480.1169, 508.1479) were found to be significant by the Yamamoto et al. method of selecting compounds in the loadings plot of PCA.

This emphasis on the identification of the metabolic classes of interest allowed the possibility of reducing false discoveries as opposed to using the other statistical tests, as only the Yamamoto et al. method was able to partially identify compounds in the anthocyanin class, while the total AUC, Kruskal–Wallis, and SAM procedures did not. As members of the anthocyanin class offer potential benefits for human health, examining the differences within this class between cranberry products allows a direct measurement of the inherent biological variation apparent among cranberry products.[64] Associating the anthocyanin class (and the level present) with clinical outcomes is possible[61] and allows a quantified approach to responses seen when consuming cranberry products.

In addition to the comparison of the cranberry products, MSEA can be utilized to examine the optimization of the extraction method and/or extraction steps with selected solvents of the products. For example, the MSEA tool was used to evaluate the differences of compounds extracted in cranberry fruit by the three solvents under study. Using Goeman's globaltest,[61] a p value of 0.0127 was calculated in comparing the solvent replicates (n = 3 each of water, methanol, 70% ethanol) of the cranberry fruit product, indicating significant differences between solvents. This is in contrast to the observed overlap of the solvents (Figure 6b) per solvent type with PCA, indicating a strong commonality between solvents. However, with ASCA (Figure 7b), the solvent factor was found to be significant, consistent with the MSEA technique.

Furthermore, examining metabolic class(es) within the cranberry fruit product metabolome can be conducted between solvents, for example by looking at the anthocyanin class and observing which solvent is appropriate for this class. In the present investigation, a p value of 0.411 was calculated, indicating no significant difference in the extraction of the anthocyanin class; thus, any of the three solvents would be suitable. Of course, examining additional phenolic metabolic classes may lead to different conclusions in terms of the solvent

DOI: 10.1021/np500667z
J. Nat. Prod. 2015, 78, 953–966

selected (for example, when anthocyanins and flavones are to be extracted at the same time).

Since no difference was seen between solvents when examining the anthocyanin class, other tools were also considered as ways to select a solvent for further analysis and workup. The abundance data of the compounds identified in the anthocyanin class were examined, showing only nine of 54 compounds had abundance values seen in all three replicates when extracting with water or methanol, and none of the compounds extracted by 70% ethanol were seen in all three replicates. Next, the results from Goeman's globaltest algorithm indicated that 24 of 54 anthocyanin compounds are associated with water (i.e., at least one replicate is seen in this solvent), 20 of 54 anthocyanin compounds were associated with methanol, but none of the 54 were associated with 70% ethanol. Three compounds were found associated with a p value of <0.05 for the solvent water (significantly different compared to the other two solvents), and one compound was found associated with the solvent methanol (with a p value of <0.05). No compound was found that was associated with a p value of <0.05 when extracted with 70% ethanol. From these observations, it became clear that the 70% ethanol extraction of the cranberry fruits exhibits poor reproducibility and should not be utilized, and rather the data sets extracted by methanol and water should be examined further.

CONCLUSIONS

Practical methods for understanding complex biological samples such as standardized natural product extracts and botanical dietary supplements will greatly improve their quality and functionality. Metabolomics tools offer new opportunities for the discovery of novel bioactive compounds and for the establishment of standard compound profiles to ensure product quality. Other "omic" approaches may have less application for plant products and their manufacture. Recently, a genomics screening procedure was used to identify compounds that were interpreted as "contamination and substitution" in natural health products,[65] but such studies may not fully account for the phytochemical profiles of plant preparations, and the genomics may be unrelated to the final product.[66] A model data set of cranberry and cranberry extracts was generated for evaluation of statistical approaches. In 2008, a Cochrane review on the effectiveness of cranberry products in preventing urinary tract infections (UTIs) found that although cranberry juice decreased the number of symptomatic UTIs, the optimum dosage and method of administration (e.g., juice, tablets, capsules) were not clear.[67] With the wide variety of products in the marketplace and limited amount of product information provided in published clinical studies,[68] it is important that tools are developed to better characterize these complex products, ensuring their quality and efficacy. For the purpose of the present review, metabolomic profiles were generated by UFLC-TOF-MS for cranberry fruits and two cranberry products to exemplify how such studies can be performed. The statistical tools and modeling used to assess these products employed a combination of approaches to identifying compounds of significance and allowed for the development of a more robust approach to interpret and utilize metabolomic profiling as a quality control tool and for identifying compounds potentially important to biological efficacy. The use of chemometric and statistical approaches for determining data quality and to prioritize data is essential for developing, validating, and standardizing methods. Univariate statistical tools such as the Kruksal–Wallis p value and the ROC/AUC parameters provide information about the distribution of the data and the confidence that the analyst can have in the detection of individual signals. The SAM statistic ranks significance based on the abundance of metabolites across a data set with consideration given to the deviation within the replicates. Once confidence is established in the data, a PCA model or similar multivariate technique can be used to find patterns within the large data sets. Finding significant compounds by PCA modeling increases the confidence that individual metabolites are important, further supplemented by the use of the ASCA tool. Together, these procedures can establish the starting point to understand how the metabolomes of each sample are related. Once important compounds are identified in the data set, synthetic biotransformation and database searches can be implemented to discover new metabolites within predefined functional classes and previously undiscovered biochemical pathways for future investigations. Applying MSEA allows the quantification of differences (and similarities) of functional classes or pathways in metabolomes. Together, these approaches create new methods and opportunities for compound discovery, quality assurance, and the generation of novel hypotheses that may lead to new understandings of the phytochemistry of plants.

ASSOCIATED CONTENT

Supporting Information

This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Tel: 604-412-7484. Fax: 604-436-0286. E-mail: paula_brown@bcit.ca.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Oliver, S.; Winson, M.; Kell, D.; Baganz, F. *Trends Biotechnol.* **1998**, *16*, 373–378.
(2) Yuliana, N. D.; Khatib, A.; Choi, Y. H.; Verpoorte, R. *Phytother. Res.* **2011**, *25*, 157–169.
(3) Wilcoxen, K. M.; Uehara, T.; Myint, K. T.; Sato, Y.; Oda, Y. *Expert Opin. Drug Discovery* **2010**, *5*, 249–263.
(4) Wang, L.; Chen, C. *AAPS J.* **2013**, *15*, 941–950.
(5) Gad, H. A.; El-Ahmady, S. H.; Abou-Shoer, M. I.; Al-Azizi, M. M. *Phytochem. Anal.* **2013**, *24*, 1–24.
(6) Lv, H. *Mass Spectrom. Rev.* **2013**, *32*, 118–128.
(7) Verpoorte, R.; Choi, Y. H.; Kim, H. K. *Phytochem. Anal.* **2010**, *21*, 2–3.
(8) Wink, M. *Phytochemistry* **2003**, *64*, 3–19.
(9) Wink, M., Ed. *Biochemistry of Plant Secondary Metabolism*, 2nd ed.; Annual Plant Reviews; Wiley-Blackwell: Oxford, UK, 2010; Vol. *40*, p 464.
(10) Dixon, R.; Strack, D. *Phytochemistry* **2003**, *62*, 815–816.
(11) Hartmann, T. *Phytochemistry* **2007**, *22–24*, 2831–2846.
(12) Miller, J. S. *Econ. Bot.* **2011**, *65*, 396–407.
(13) Afendi, F. M.; Okada, T.; Yamazaki, M.; Hirai-Morita, A.; Nakamura, Y.; Nakamura, K.; Ikeda, S.; Takahashi, H.; Altaf-ul-Amin, M.; Darusman, L. K.; Saito, K.; Kanaya, S. *Plant Cell Physiol.* **2012**, *53*, 1–12.
(14) Turi, C. E.; Murch, S. J. *Planta Med.* **2013**, *79*, 1370–1379.
(15) Murch, S. J.; Rupasinghe, H. P.; Goodenowe, D.; Saxena, P. K. *Plant Cell Rep.* **2004**, *23*, 419–425.

DOI: 10.1021/np500667z
J. Nat. Prod. 2015, 78, 953–966

(16) Arbona, V.; Manzi, M.; Ollas, C.; Gomez-Cadenas, A. *Int. J. Mol. Sci.* **2013**, *14*, 4885−4911.
(17) Kral'ova, K.; Jampilek, J.; Ostrovsky, I. *Ecol. Chem. Eng. Sci.* **2012**, *19*, 133−161.
(18) Ulrich-Merzenich, G.; Panek, D.; Zeitler, H.; Wagner, H.; Vetter, H. *Phytomedicine* **2009**, *16*, 495−508.
(19) Yuliana, N. D.; Jahangir, M.; Verpoorte, R.; Choi, Y. H. *Phytochemistry* **2013**, *12*, 293−304.
(20) Brown, P. N.; Murch, S. J.; Shipley, P. R. *J. Agric. Food Chem.* **2012**, *60*, 261−271.
(21) Brown, P. N.; Turi, C. E.; Shipley, P. R.; Murch, S. J. *Planta Med.* **2012**, *78*, 630−640.
(22) Goodacre, R.; Vaidyanathan, S.; Dunn, W.; Harrigan, G.; Kell, D. *Trends Biotechnol.* **2004**, *22*, 245−252.
(23) Davis, M. C.; Fiehn, O.; Durnford, D. G. *Plant, Cell Environ.* **2013**, *36*, 1391−1405.
(24) Rochfort, S. J.; Trenerry, V. C.; Imsic, M.; Panozzo, J.; Jones, R. *Phytochemistry* **2008**, *69*, 1671−1679.
(25) Hall, R. D., Ed. *Biology of Plant Metabolomics*; Annual Plant Reviews; Wiley-Blackwell Publishing: Oxford, UK, 2011; Vol. *43*, p 448.
(26) Lankadurai, B. P.; Nagato, E. G.; Simpson, M. J. *Environ. Rev.* **2013**, *21*, 180−205.
(27) Patti, G. J.; Tautenhahn, R.; Siuzdak, G. *Nat. Protoc.* **2012**, *7*, 508−516.
(28) Wishart, D. S. *Briefings Bioinf.* **2007**, *8*, 279−293.
(29) Gao, W.; Sun, H.; Xiao, H.; Cui, G.; Hillwig, M. L.; Jackson, A.; Wang, X.; Shen, Y.; Zhao, N.; Zhang, L.; Wang, X.; Peters, R. J.; Huang, L. *BMC Genomics* **2014**, *15*, 73−73.
(30) Heuberger, A. L.; Broeckling, C. D.; Kirkpatrick, K. R.; Prenni, J. E. *Plant Biotechnol. J.* **2014**, *12*, 147−160.
(31) Wahyuni, Y.; Stahl-Hermes, V.; Ballester, A.; de Vos, R. C. H.; Voorrips, R. E.; Maharijaya, A.; Molthoff, J.; Zamora, M. V.; Sudarmonowati, E.; Arisi, A.; Bino, R.; Bovy, A. *Mol. Breed.* **2014**, *33*, 503−518.
(32) Jandric, Z.; Roberts, D.; Rathor, M. N.; Abrahim, A.; Islam, M.; Cannavan, A. *Food Chem.* **2014**, *148*, 7−17.
(33) Klein, M. A. In *Encyclopedia of Dietary Supplements*, 2nd ed.; Coates, P. M., Betz, J. M., Blackman, M. R., Gragg, G. M., Levine, M., Moss, J., White, J. D., Eds.; Informa Healthcare: London, UK, 2010; pp 193−201.
(34) Upton, R., Ed. In *Cranberry Fruit Vaccinium macrocarpon Aiton Standards of Analysis, Quality Control, and Therapuetics*; American Herbal Pharmacopoeia and Therapeutic Compendium; American Herbal Pharmacopoeia: Santa Cruz, CA, 2002; p 28.
(35) Enot, D. P.; Draper, J. *Metabolomics* **2007**, *3*, 349−355.
(36) R Core Team. In *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, 2014.
(37) Broadhurst, D. I.; Kell, D. B. *Metabolomics* **2006**, *2*, 171−196.
(38) Kruskal, W. H.; Wallis, W. A. *J. Am. Stat. Assoc.* **1952**, *47*, 583−621.
(39) Tusher, V.; Tibshirani, R.; Chu, G. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 5116−5121.
(40) Fawcett, T. *Pattern Recog. Lett.* **2006**, *27*, 861−874.
(41) Hand, D.; Till, R. *Mach. Learning* **2001**, *45*, 171−186.
(42) Khan, A.; Rayner, G. D. *J. Appl. Math. Decis. Sci.* **2003**, *7*, 187−206.
(43) Jeffery, I. B.; Higgins, D. G.; Culhane, A. C. *BMC Bioinf.* **2006**, *7*, 359.
(44) Tuszynski, J. In *caTools: Tools: Moving Window Statistics, GIF, Base64, ROC AUC, etc.*; R Foundation for Statistical Computing: Vienna, 2010.
(45) Jolliffe, I. T. In *Prinicipal Component Analysis*, 2nd ed.; Springer Series in Statistics; Springer: New York, 2002; p 488.
(46) Bro, R.; Smilde, A. K. *Anal. Methods* **2014**, *6*, 2812−2831.
(47) Yamamoto, H.; Fujimori, T.; Sato, H.; Ishikawa, G.; Kami, K.; Ohashi, Y. *BMC Bioinf.* **2014**, *15*, 51.
(48) Harrington, P. B.; Vieira, N. E.; Espinoza, J.; Nien, J. K.; Romero, R.; Yergey, A. L. *Anal. Chim. Acta* **2005**, *544*, 118−127.
(49) Smilde, A. K.; Jansen, J. J.; Hoefsloot, H. C.; Lamers, R. J.; van der Greef, J.; Timmerman, M. E. *Bioinformatics* **2005**, *21*, 3043−3048.
(50) Sarembaud, J.; Pinto, R.; Rutledge, D. N.; Feinberg, M. *Anal. Chim. Acta* **2007**, *603*, 147−154.
(51) Luthria, D. L.; Lin, L. Z.; Robbins, R. J.; Finley, J. W.; Banuelos, G. S.; Harnly, J. M. *J. Agric. Food Chem.* **2008**, *56*, 9819−9827.
(52) Zwanenburg, G.; Hoefsloot, H. C. J.; Westerhuis, J. A.; Jansen, J. J.; Smilde, A. K. *J. Chemometr.* **2011**, *25*, 561−567.
(53) Jonsson, P.; Johansson, A. I.; Gullberg, J.; Trygg, J.; A, J.; Grung, B.; Marklund, S.; Sjostrom, M.; Antti, H.; Mortiz, T. *Anal. Chem.* **2005**, *77*, 5635−5642.
(54) Juan, A.; Jaumot, J.; Tauler, R. *Anal. Methods* **2014**, *6*, 4964−4976.
(55) Westerhuis, J. A.; Hoefsloot, H. C. J.; Smit, S.; Vis, D. J.; Smilde, A. K.; van Velzen, E. J. J.; van Duijnhoven, J. P. M.; van Dorsten, F. A. *Metabolomics* **2008**, *4*, 81−89.
(56) Li, L.; Li, R.; Zhou, J.; Zuniga, A.; Stanislaus, A. E.; Wu, Y.; Huan, T.; Zheng, J.; Shi, Y.; Wishart, D. S.; Lin, G. *Anal. Chem.* **2013**, *85*, 3401−3408.
(57) Kai, K.; Takahashi, H.; Saga, H.; Ogawa, T.; Kanaya, S.; Ohta, D. *Plant Biotechnol.* **2011**, *28*, 379−385.
(58) Fiehn, O.; Barupal, D. K.; Kind, T. *J. Biol. Chem.* **2011**, *286*, 23637−23643.
(59) Fukushima, A.; Kusano, M. *Front. Plant Sci.* **2013**, *4*, 1−11.
(60) Dunn, W. B.; Erban, A.; Weber, R. J. M.; Creek, D. J.; Brown, M.; Breitling, R.; Hankemeier, T.; Goodacre, R.; Neumann, S.; Kopka, J.; Viant, M. R. *Metabolomics* **2013**, *9*, S44−S66.
(61) Goeman, J. J.; van de Geer, S. A.; de Kort, F.; van Houwelingen, H. C. *Bioinformatics* **2004**, *20*, 93−99.
(62) Goeman, J. J.; Buhlmann, P. *Bioinformatics* **2007**, *23*, 980−987.
(63) Hendrickx, D. M.; Hoefsloot, H. C.; Hendriks, M. M.; Canelas, A. B.; Smilde, A. K. *Anal. Chim. Acta* **2012**, *719*, 8−15.
(64) Brown, P. N.; Shipley, P. R. *J. AOAC Int.* **2011**, *94*, 459−466.
(65) Newmaster, S. G.; Grguric, M.; Shanmughanandhan, D.; Ramalingam, S.; Ragupathy, S. *BMC Med.* **2013**, *11*, 1−13.
(66) Gafner, S.; Blumenthal, M.; Harbaugh-Reyaud, D.; Foster, S.; Techen, N. *HerbalEGram* **2013**, *10*, number 11.
(67) Jepson, R. G.; Williams, G.; Craig, J. C. *Cochrane Database Syst. Rev.* **2012**, *10*, CD001321.
(68) Gagnier, J. J.; DeMelo, J.; Boon, H.; Rochon, P.; Bombardier, C. *Am. J. Med.* **2006**, *119*, 800.e1−800.e11.

DOI: 10.1021/np500667z
J. Nat. Prod. 2015, 78, 953−966