JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

Representation, Comparison, and Interpretation of Metabolome Fingerprint Data for Total Composition Analysis and Quality Trait Investigation in Potato Cultivars

MANFRED BECKMANN, DAVID P. ENOT, DAVID P. OVERY, AND JOHN DRAPER*

Institute of Biological Sciences, Edward Llwyd Building, University of Wales, Aberystwyth, Ceredigion SY23 3DA, United Kingdom

Understanding attributes of crop varieties and food raw materials underlying desirable characteristics is a significant challenge. Metabolomics technology based on flow infusion electrospray ionization mass spectrometry (FIE-MS) has been used to investigate the chemical composition of potato cultivars associated with quality traits in harvested tubers. Through the combination of metabolite fingerprinting with random forest data modeling, a subset of metabolome signals explanatory of compositional differences between individual genotypes were ranked for importance. Interpretative analysis of highlighted signals based on ranking behavior, intensity correlations, and mathematical relationships of ion masses correctly predicted metabolites associated with flavor and pigmentation traits in potato tubers. GC-MS profiling was used to further validate proposed compositional differences. The potential for the development of a database strategy for large scale, long-term projects requiring comparison of chemical composition in plant breeding, mutant population analysis in functional genomics experiments, or food raw material analysis is described.

KEYWORDS: MS-based metabolomics; metabolite fingerprint modeling and data interpretation; crop plant trait analysis; food raw material composition database strategy

INTRODUCTION

Despite wide interest in human nutrition we are relatively ignorant of the detailed composition of the food we consume. Total chemical content is a major determinant of many crop plant variety quality characteristics, and food raw materials typically contain many hundreds of different metabolites present at a wide concentration range (*1*−*3*). Thus, not only biotechnology companies and plant breeders with an interest in crop improvement but also food producers and distributors, as well as government advisory agencies, increasingly desire objective and meaningful information relating to the "global" chemical composition of food raw materials.

Metabolomics technology (*1*−*4*) has been proposed for the investigation of crop plant attributes underlying desirable characteristics (*5*−*11*); however, the dimensionality and intrinsic variance of metabolomics data make representation and meaningful comparison challenging (*12*−*16*). Metabolite "fingerprinting" approaches (*1*−*4*) provide a rapid, comprehensive, and nonbiased assessment of the metabolome (*6*−*11*), and flow infusion electrospray ionization mass spectrometry (FIE-MS) has proved to be valuable for the analysis of plant breeding populations and harvested food raw materials (*6*, *8*−*11*). Mass spectrometry fingerprints are directly interpretable through linkage between specific ion signals (mass to charge ratio: *m*/*z*) to candidate metabolite derivatives sharing the same atomic mass (*5*, *6*, *8*−*11*). The high dimensionality of FIE-MS fingerprints (often >1000 *m*/*z* signals) and intrinsic variability in metabolite concentrations demand powerful multivariate methods for meaningful data analyses (*12*−*21*). A range of different statistical and machine-learning techniques give high predictive accuracy (*8*−*16*, *19*−*23*). However, this is often at the expense of either providing multiple solutions (*12*, *24*) or generating complex models that can be opaque to further interpretability (*12*−*15*) and impossible to compare. In the present study, we aimed to approach both the representational and interpretability problems by ranking *m*/*z* signals according to statistical measures related to the behavior of each individual variable in the full data set. The importance score (*17*, *18*) derived from the analysis of an ensemble of decision trees (*19*, *21*, *22*) produced by the random forest (RF) (*17*, *18*) tree classification algorithm can provide such information (*6*). Importantly, this is achieved without requiring an initial reduction of dimensionality so that each variable has an equal chance of being included in the final model, thus allowing direct comparison of all classifiers (*6*).

The present study focused on an analysis of compositional differences in potato cultivars where there was no prior genetic, biochemical, or analytical chemistry data available with which to guide the interpretation of the metabolome data models. A primary objective was to validate a strategy to identify a subset of variables (*m*/*z* signals) that adequately describe significant metabolome differences between several genotypes grown under

* Corresponding author [telephone 44-(0) 1970 622789; 44-(0) 1970 622350; e-mail jhd@aber.ac.uk].

10.1021/jf0701842 CCC: $37.00
Published on Web 04/07/2007

Table 1. Model Statistics in Pairwise Comparison of Potato Cultivars by Random Forest

pairwise comparison	classification accuracy (%)	model margin[a]
Ag_De1	100	0.62
Ag_Gr	97	0.61
Ag_Li	97	0.44
Ag_So	100	0.64
De1_Gr	100	0.66
De1_Li	100	0.42
De1_So	100	0.68
Gr_Li	100	0.59
Gr_So	100	0.74
Li_So	100	0.58

[a] Margin is a measure of model sensitivity.

field conditions. Further interpretation of the highlighted *m*/*z* signals was explored to link metabolome differences to genotype traits. Finally, a strategy for standardized comparison of the global composition of individual crop genotypes or food raw materials is explored.

MATERIALS AND METHODS

Plant Material. The potato tuber material and procedures for sample preparation and extraction have been described previously (*10*). The present study utilized five *Solanum tuberosum* cultivars: Agria (Ag), Désirée (De), Granola (Gr), Linda (Li), and Solara (So). Two closely related populations of Désirée, which were independently propagated (De1 and De2) (*6*, *10*), were included in the sample set.

Sample Preparation and Metabolite Analysis. FIE-MS analysis of the potato tuber extracts was performed using an LCT mass spectrometer (Micromass, Manchester, U.K.) as described previously (*10*). Data were collected in positive and negative ionization modes. Raw data of the whole infusion profiles were exported and mass intensities of each scan electronically binned to 1 amu (between −0.2 and 0.8 amu). The resulting mass spectrum for each analysis was calculated as the mean of 11 scans around the apex of the infusion profile. Mass spectra of all analytical runs per tissue and ionization mode were combined in a single intensity matrix (runs × *m*/*z* ratios). GC-MS analysis of potato tuber extracts was performed using an Agilent 5973 N MSD. A volume of 50 μL of each extract was dried in vacuo (SpeedVac) and derivatized in two steps with 100 μL of 20 mg/mL methoxyamine hydrochloride (Fluka Chemicals from Metlab Supplies Ltd., Sandycroft, U.K.) in dry pyridine (Fluka) at 30 °C for 90 min and subsequently with 100 μL of *N*-methyl-*N*-(trimethylsilyl)trifluoroacetamide (MSTFA) (Macherey-Nagel from Fisher Scientific U.K., Loughborough) at 37 °C for 30 min. GC-MS conditions for potato tuber samples were as follows: split injection (25:1, 250 °C) of 1 μL on a 30 m × 0.25 mm i.d. and 25 μm film DB5-MS column at 1 mL/min (He) flow rate; 85 °C initial oven temperature held for 2 min, then increased at a rate of 30 °C/min to a final temperature of 330 °C and subsequently held at 330 °C for 5 min. The transfer line temperature was set to 300 °C, and data were acquired between 80 and 500 amu at a rate of six spectra/s with a detector voltage of 1400 V. Of the 2304 analyses performed, 48 representative chromatograms were manually deconvoluted. A final list of 91 peaks was used to locate metabolites in all CSV-exported runs (Chemstation, Agilent) using a script in Matlab (V.6.5.1, The Math Works Inc., Matrix House, Cambridge, U.K.). The final data matrix contained the background-subtracted intensity of a characteristic mass ion at the apex for each targeted peak. When possible, peak identity was confirmed by comparison with standards.

Construction of a "Mastermix" Fingerprint Population. One approach to the compositional analysis of different plant genotypes is to compare each to a Mastermix class effectively representative of metabolome variance found in all classes. This technique is known in the microarray community as "pooling", where individual samples are combined either in vitro (*25*) or in silico (*26*) to reduce inherent subject-to-subject variability. In our case, the objective was to form a virtual new potato class wherein the new comparator encapsulated the overall metabolic diversity of the potato cultivar population. Our strategy (Supporting Information) centered on randomly pooling a single FIE-MS fingerprint from each class where the median intensity, at all *m*/*z* of the selected fingerprints, was used to create a virtual comparator sample. To restrict the inevitable reduction in variance, a sample was used only once when the electronic Mastermix was generated.

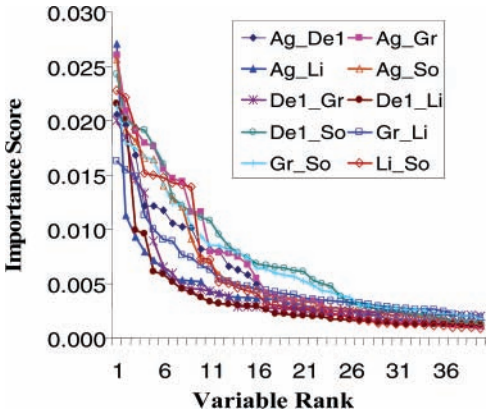
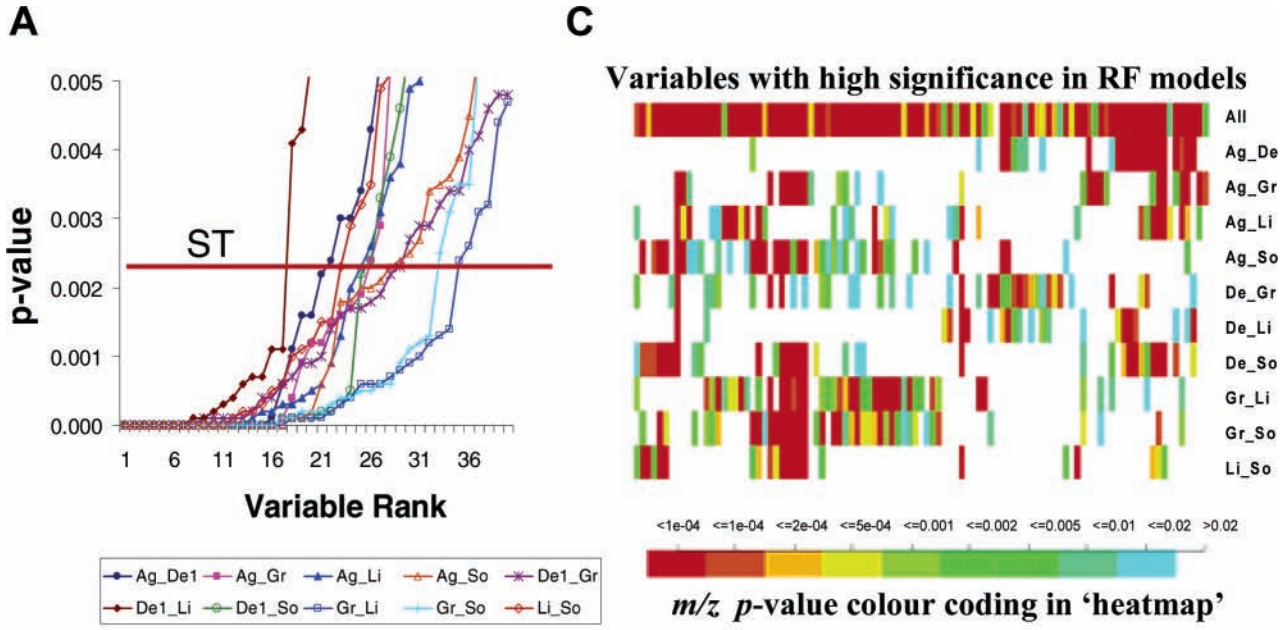


Figure 1. Relationship between importance score and random forest ranking in pairwise comparison of potato cultivars (FIE-MS positive ion data). Variables (*m*/*z*) are shown ranked by random forest importance score for each pairwise classification of the five cultivars (Ag, Agria; De, Désirée; Gr, Granola; Li, Linda; So, Solara).

Data Analysis. GC-MS and FIE-MS raw data were transformed before data analysis as described previously (*6*, *10*). All computation and subsequent visualization were carried out in the R environment (*27*) using packages available from the R website (*randomForest* and *ROCR*). Original datasets were split into a training set (used for model construction, feature selection, and visualization) and an independent test, which was used for assessing model generalizability; thus, 160 samples were used for training and 80 samples for testing in FIE-MS data, and 654 samples were used for training and 300 samples for testing in GC-MS data. Sample classification and the selection of potentially explanatory variables in both FIE-MS and GC-MS data were achieved as demonstrated previously using the RF tree classification algorithm (*6*) (Supporting Information). In the case of FIE-MS data the first stage in the analysis was to define a ranked list of potentially explanatory signals by computing a RF importance score for each *m*/*z* in each classification task. Permutation testing was subsequently used to determine the statistical significance (*p* value) of each importance score (*6*). Significant signals ($p \leq 0.005$) found in at least one of the models were kept for the "heatmap" visualization. Rows (*m*/*z* signals) were reordered according to the dendrogram built by hierarchical clustering using the absolute value of the correlation coefficient between signals as the similarity measure.

FIE-MS Signal Interpretation. A comprehensive list of metabolites present in the potato metabolome was compiled from an exhaustive literature and database search as well as by including additional primary and conserved secondary metabolites extrapolated from The Arabidopsis Information Resource (TAIR) database (*28*) (http://www.arabidopsis.org/tools/aracyc/). This information was used to generate potato-specific entries in ArMec, a metabolite signal identification database developed in Aberystwyth (http://www.armec.org/MetaboliteLibrary/index.html) to interpret FIE-MS data. Using ArMec, a potato ESI prediction spreadsheet was created for both positive and negative ESI modes by calculating masses of potential (de)protonated molecular ions and their associated isotopes, salt adducts of alkali metals, neutral losses, and homogeneous dimer and dimer ion pair adducts.

The initial data analysis by RF produced a list of *m*/*z* signals ranked by importance scores or *p* value for each classification task. Using an importance score of 0.003 as a validated threshold for explanatory power in FIE-MS data (*6*), the lists of the top-ranked *m*/*z* (generally 25−30) were used to make metabolite putative assignments by querying the potato ESI predictions in ArMec and selecting the most likely candidate metabolites based upon the occurrence or nonoccurrence of a corresponding ion mass signal (as expected from the profiling of the chemical



B

All cultivars		Pairwise comparison of cultivars (+ ve ion data)										High ranking correlated positive ions only evident in pair-wise comparisons									
Rank	m/z	Ag_De	Ag_Gr	Ag_Li	Ag_So	De_Gr	De_Li	De_So	Gr_Li	Gr_So	Li_So	Ag_De	Ag_Gr	Ag_Li	Ag_So	De_Gr	De_Li	De_So	Gr_Li	Gr_So	Li_So
1	146	543	146	146	146	280	146	280	124	237	297	543	48	146	146	280	146	280	124	237	297
2	136	120	86	166	324	146	121	120	280	296	448	120	86	166	324	146	121	120	280	296	448
3	121	166	124	324	142	124	136	146	190	324	280	166	124	324	142	124	136	146	190	324	280
4	105	544	120	120	123	121	543	297	201	215	237	544	120	120	123	121	543	297	201	215	237
5	268	182	280	86	280	332	184	142	237	280	296	182	280	86	280	332	184	142	237	280	296
6	119	105	237	513	448	706	393	543	362	298	160	105	237	513	448	706	393	543	362	298	160
7	166	104	298	121	160	646	431	237	297	124	298	104	298	121	160	646	431	237	297	124	298
8	188	188	119	607	297	165	545	296	205	297	142	188	119	607	297	165	545	296	205	297	142
9	87	165	296	205	258	699	544	544	245	281	437	165	296	205	258	699	544	544	245	281	437
10	324	545	297	188	376	690	122	160	215	127	121	545	297	188	376	690	122	160	215	127	121
11	280	136	166	145	237	123	394	298	296	123	258	136	166	145	237	123	394	298	296	123	258
12	182	119	121	498	104	673	189	448	246	258	143	119	121	498	104	673	189	448	246	258	143
13	237	184	617	119	139	659	400	166	221	245	324	184	617	119	139	659	400	166	221	245	324
14	297	126	256	515	617	161	105	136	281	274	188	126	256	515	617	161	105	136	281	274	188
15	165	649	118	189	127	182	104	268	127	259	120	649	118	189	127	182	104	268	127	259	120
16	296	129	84	265	145	637	362	188	192	142	166	129	84	265	145	637	362	188	192	142	166
17	104	123	129	182	125	366	675	182	425	273	124	123	129	182	125	366	675	182	425	273	124
18	220	86	294	514	286	105	301	286	497	264	453	86	294	514	286	105	301	286	497	264	453
19	654	95	147	484	296	675	126	121	145	425	481	95	147	484	296	675	126	121	145	425	481
20	184	675	156	497	260	719	699	126	513	190	123	675	156	497	260	719	699	126	513	190	123
21	123	690	188	190	511	137	704	324	259	201	125	690	188	190	511	137	704	324	259	201	125
22	120	131	110	619	486	674	711	545	324	120	259	131	110	619	486	674	711	545	324	120	259
23	86	687	111	284	298	365	580	125	216	207	469	687	111	284	298	365	580	125	216	207	469
24	448	655	215	256	259	298	799	272	514	362	161	655	215	256	259	298	799	272	514	362	161

Figure 2. Analysis of potato cultivar FIE-MS (positive mode) fingerprints by random forest ranking and correlation analysis of signals. (**A**) Variables (*m*/*z*) are shown ranked by *P* value for each pairwise classification of the five cultivars (Ag, Agria; De, Désirée; Gr, Granola; Li, Linda; So, Solara). A significance threshold (ST) validated in previous studies (*6*) is shown. (**B**) The top 24 *m*/*z* ranked by importance score in different models generated by random forest analysis. The left panel indicates signals selected in a single combined model involving all cultivars, and *m*/*z* considered to be related (following correlation analysis and calculation of mathematical relationships) are color coded. The middle panel shows the *m*/*z* ranked by random forest in pairwise comparisons of potato cultivars; all signals previously selected in the combined model are similarly color coded. The right panel illustrates new sets of potentially related ions found only in pairwise comparisons. (**C**) Correlation analysis of signals (positive ion data) using a subset of significant ($p \leq 0.01$) variables selected by a random forest model comparing all five cultivars. A heatmap representation is shown where each variable is color coded according to significance in the model.

standards). As several overlapping solutions predicting the presence of different metabolites were often possible, the most likely combination of ions putatively identifying a specific metabolite were confirmed by further examination of signal relationships in a correlation analysis using just *m*/*z* with an appropriate *p* value. Further investigation of the highly explanatory signals was explored by MS/MS^n experiments in (i) potato extracts, (ii) pure authentic standards, and (iii) potato extracts spiked with an authentic standard on an LTQ LCMS (Thermo Finnigan, San Jose, CA). When possible, metabolite identification was validated by GC-MS analysis using authentic standards.

RESULTS AND DISCUSSION

Potato Cultivar Classification by FIE-MS Fingerprinting. FIE-MS fingerprints were generated in both ionization modes (*6*, *10*) from tuber extracts of the potato cultivars Agria, Désirée, Granola, Linda, and Solara. The merits of analyzing both the positive ion and negative ion data are twofold. First, it cannot be predicted which ionization mode data set will be the most explanatory because changes evident in the metabolome depend

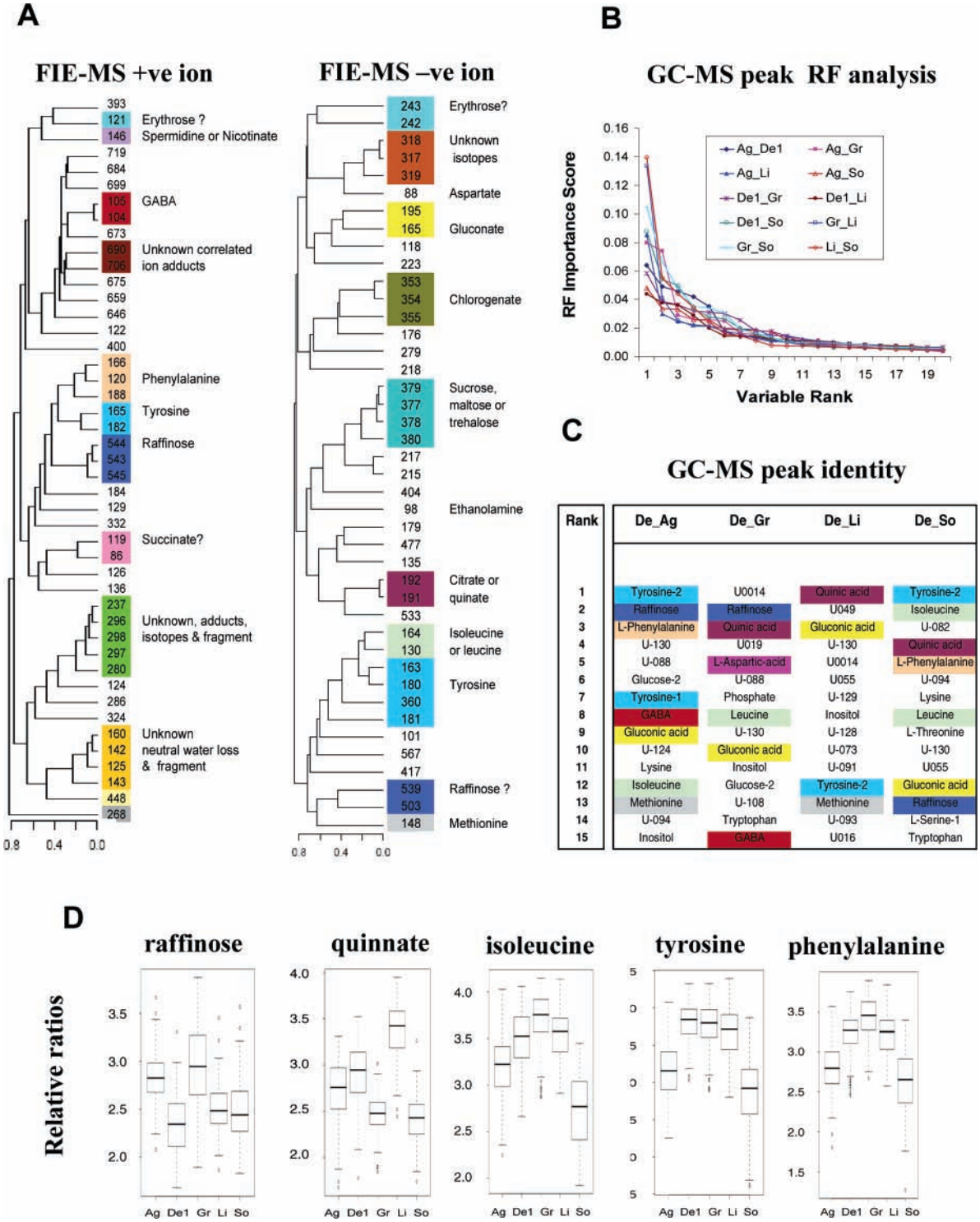


Figure 3. Interpretation of FIE-MS explanatory signals and confirmation by GC-MS profiling. (**A**) Detail of a correlation analysis indicating the putative identity of FIE-MS signals (positive and negative ion data) discriminating Désirée from other cultivars. Groups of correlated signals representing putative metabolites are individually color coded. (**B**) Relationship between importance score and variable rank in random forest analysis of GC-MS data. (**C**) Identity of top 15 explanatory GC-MS peaks selected by random forest analysis data in pairwise comparisons of Désirée with other cultivars. GC-MS peaks identified in data share the same color coding as FIE-MS signals corresponding to the same metabolites. (**D**) Box plots of the concentration (relative ratios) of selected explanatory metabolites in five potato cultivars.

on the actual matrix under investigation; analyzing (or even acquiring) in only one ionization mode will ignore all of those potentially important metabolites that can be ionized only in opposite polarity. Second, explanatory variables in both data sets can be linked to the same parent molecule, and thus both highlight the importance and aid the identification of a named metabolite. Using random forest, an initial pairwise comparison of the positive ion FIE-MS fingerprints (*6*), representing each cultivar, generated models with high classification accuracy and high sensitivity (**Table 1**). These data suggested that surprisingly large differences in composition exist between tubers of individual cultivars. Random forest importance scores in all pairwise comparisons leveled off before variable 30 in the ranked lists, with anywhere between 9 and 24 *m*/*z* signals having

Table 2. Potato Cultivar Quality Traits[a]

trait	Agria	Desiree	Granola	Linda	Solara
after cooking blackening	none–trace	trace–little	trace–little	N/D	none–trace
taste	good–excellent	moderate–good	moderate–good	good	good
French fry suitability	good–very good	moderate–good	poor–moderate	N/D	N/D
frying color	pale	medium	N/D	N/D	N/D

[a] Data summarized from European Cultivated Potato Database (http://www.europotato.org/).

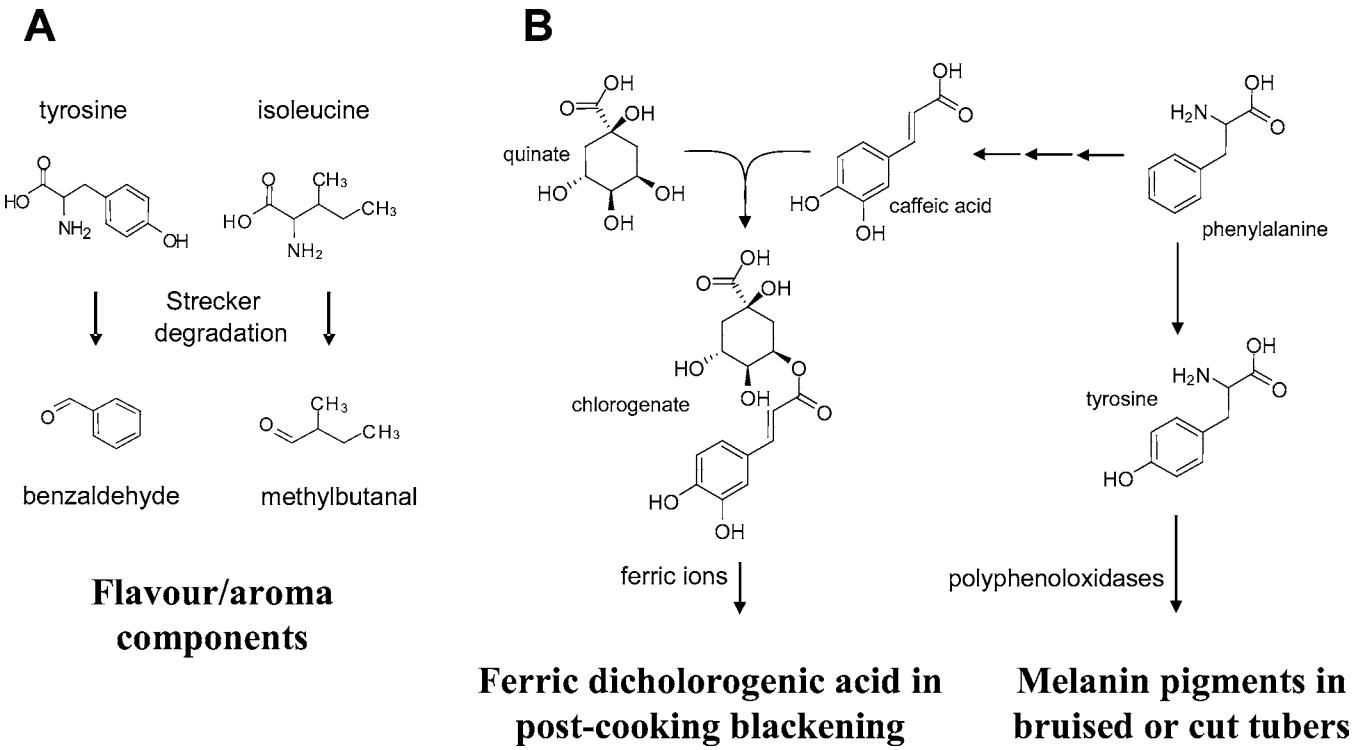


Figure 4. Relationships between explanatory metabolites and potato quality traits: (**A**) Strecker reactions converting tyrosine and isoleucine into flavor and aroma components during potato cooking; (**B**) metabolism leading to undesirable discoloration of potato raw materials during storage, processing, or cooking.

a value greater than the significance threshold (ST) of 0.003 that we have proposed previously (*6*) for FIE-MS pairwise models (**Figure 1**). Permutation testing (*6*) revealed that of 1000 signals, only a small number of variables had significant discriminatory potential (**Figure 2A**). Analysis of negative ion data yielded comparable results (data not shown).

Interpretation of Metabolome Differences Detected by FIE-MS in Potato Cultivars. For each cultivar combination, the behavior of the top 24 variables selected by RF in both ionization modes was examined in detail. The list of the top 24 variables for discrimination of all cultivars in positive ion data (left panel, **Figure 2B**) included many that were potentially related (e.g., 104/105 and 296/297 could represent isotopes). In the pairwise comparisons these putatively linked (color coded) signals clustered and were even more highly ranked (center panel, **Figure 2B**). Each pairwise model contained specific combinations of such groupings and, importantly, further signals potentially linked to the same metabolites were additionally highly ranked, strengthening the possibility that they were derived from the same metabolite. Furthermore, a range of potentially related ions not present in the top 24 signals discriminating all cultivars were also evident in specific subsets of the pairwise comparisons (right panel, **Figure 2B**) and in negative ion data (Supporting Information). A correlation analysis (*6*) performed using variables with p values of <0.02 similarly identified clusters of positive ion signals typical of individual cultivar comparisons (**Figure 2C**) and also in negative ion results (Supporting Information). Using the relationship between Désirée and other cultivars as an example, it is evident (**Figure 3A**) that many such groups of signals potentially represent isotopes, salt adducts, and fragments of known metabolites (Supporting Information). Some metabolites were highlighted in both ionization modes (e.g., raffinose and tyrosine), whereas others were found only in negative or positive ion data (e.g., aspartate, gluconate, leucine/isoleucine, GABA, quinate, and chlorogenate). In all instances the related m/z were found in the top 20−30 variables ranked by RF, suggesting that an importance score threshold of 0.003 was adequate.

GC-MS Profiling Analysis of Potato Cultivars. The potato extracts analyzed by FIE-MS fingerprinting were further subjected to GC-MS metabolite profiling (*10*), and the resulting data matrix of peak relative ratios in each sample was used for RF analysis. Pairwise comparison of the GC-MS cultivar data generated higher importance scores in comparison to the much more highly dimensional FIE-MS data, with values leveling off between ranks 8 and 15 (**Figure 3B**). All models representing pairwise comparisons of GC-MS profiles had a high classification accuracy (>92%), and variables with p values of >0.001 were only evident below rank 17 in all models. When peak identity was known, the highlighted GC-MS variables were matched against the list of explanatory metabolites predicted by high-throughput FIE-MS analysis shown in **Figure 3B**. Using

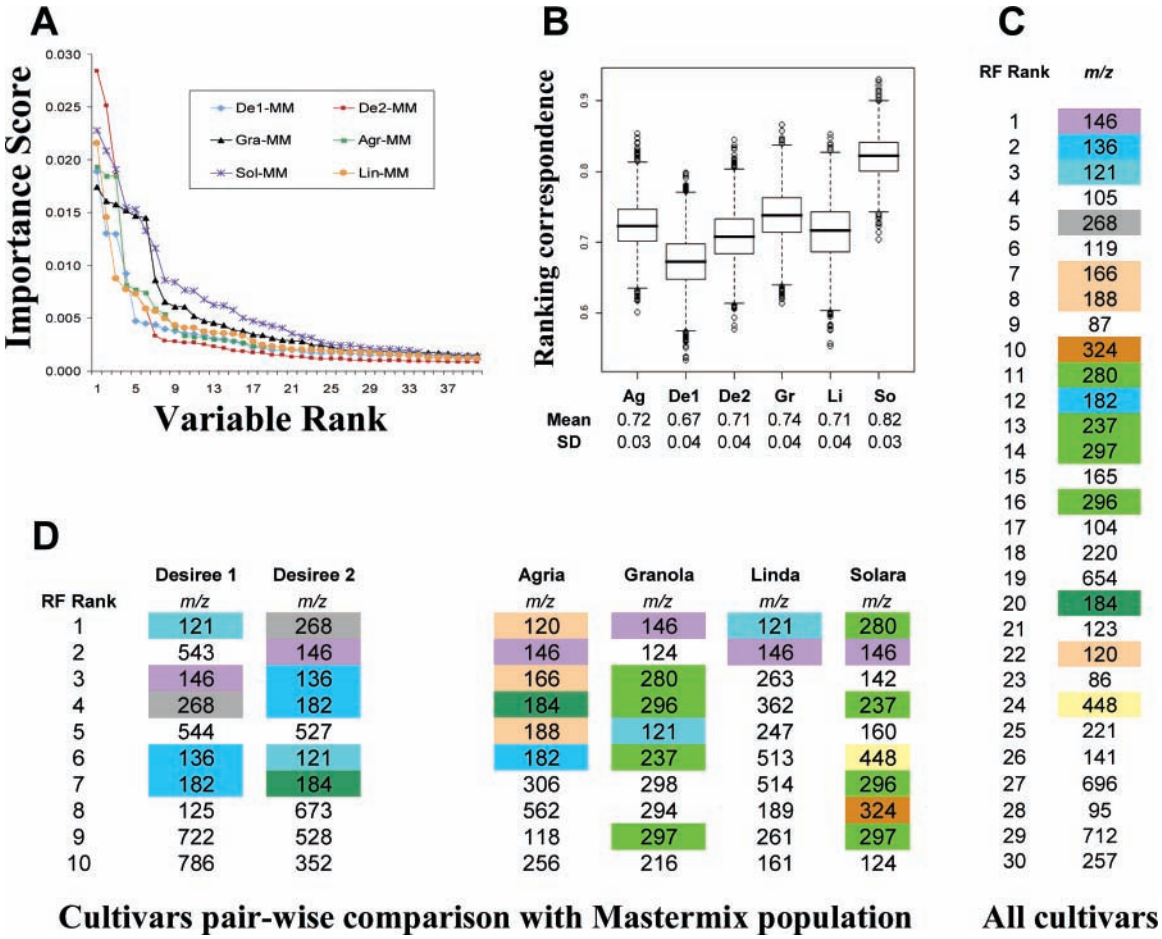


Figure 5. Use of a Mastermix reference population to generate a unique cultivar representation based on explanatory variables ranked by random forest: (**A**) relationship between importance score and variable rank in a random forest analysis of FIE-MS data involving the pairwise comparison of each cultivar with a Mastermix fingerprint population; (**B**) signal ranking correspondence in random forest pairwise comparisons of cultivar FIE-MS fingerprints with 100 different randomly generated Mastermix populations; (**C**) top 30 ranked explanatory variables identified in a random forest model comparing FIE-MS fingerprints of all five potato cultivars; (**D**) top 10 ranked explanatory *m/z* identified in the random forest pairwise comparison of cultivar FIE-MS fingerprints with a Mastermix fingerprint population. The color coding of highlighted signals is presented in **Figure 3A**.

compositional comparisons with Désirée as an example, it can be seen that explanatory GC-MS peaks (color coded) corresponding to tyrosine, raffinose, phenylalanine, GABA, gluconate, isoleucine, leucine, methionine, and aspartate were generally all highly ranked in the same cultivar-specific pattern as in the FIE-MS data (**Figure 3C**). Furthermore, the presence of correlated signals in FIE-MS data representing unidentified metabolites (**Figure 3A**) reflected the presence of a similar proportion of unknown peaks in GC-MS analysis shown in **Figure 3C**.

Relationship between Metabolite Content and Potato Cultivar Characteristics. A good correspondence was evident between high RF ranking and the relative concentrations of the selected metabolites in individual cultivars (**Figure 3D**). For example, tyrosine was present at a considerably higher concentration in a Désirée background than in either Solara or Agria and was the most highly ranked metabolite signal for discrimination between these cultivars. Many of the identified metabolites that contributed significantly to compositional differences between the potato cultivars are linked closely to quality traits in potato tubers (**Table 2**). Isoleucine, leucine, tyrosine, and phenylalanine are all known to be important precursors of flavor and aroma compounds in cooked potatoes (*29*, *30*). For example, isoleucine/leucine and tyrosine are major substrates for Strecker reactions (**Figure 4A**) that produce volatile aldehydes (methylbutanals and benzaldehyde, respectively), which contribute "almond" and "toasted/sweet" aromas to boiled potatoes. Free tyrosine is also a major substrate for polyphenol oxidases (*31*), which are responsible for undesirable melanin biosynthesis in mechanically damaged potatoes (**Figure 4B**). Tyrosine is relatively low in cultivars such as Solara and Agria, which are suitable for slicing and frying (**Table 2**), and high in Désirée and Granola, which are unsuitable (**Figure 3D**). Similarly, chlorogenic acid (and by implication its precursors phenylalanine, quinate, and caffeate) is linked to nonenzymatic reactions associated with postcooking blackening (*32*) and is also a substrate for polyphenol oxidases in cultivars such as Granola and Désirée (**Figure 4B**).

"Mastermix" Strategy for Comparison of Potato Tuber Composition. To provide a more useful framework for compositional comparisons between larger numbers of plant genotypes and to provide possibilities for data integration between laboratories, a single metabolome representation will be required as a comparator. In pilot experiments an electronic Mastermix population of FIE-MS fingerprints was generated (Supporting Information). Importance score ranking by RF revealed that 10−20 *m/z* signals were highly significant to discriminate each cultivar from the Mastermix population (**Figure 5A**). To confirm the robustness of this approach 100 Mastermix populations were generated by random pooling of fingerprints and the pairwise comparison of each cultivar performed using RF. In all instances the explanatory signals identified and their rank order showed

high correspondence in relation to importance score (**Figure 5B**). **Figure 5C** shows in detail the rank order of explanatory signals selected in a RF analysis model comparing all potato cultivars. Representative (color coded as in **Figure 3A**) combinations of these highlighted signals populate the top 10 rank positions in pairwise comparisons of individual cultivars with the Mastermix population (**Figure 5D**). Interestingly, samples representing two independently propagated clones of the cultivar Désirée (De1 and De2) exhibited high correspondence.

Both representation and subsequent comparison of global chemical composition in plant breeding germplasm and food raw materials are difficult tasks. Additionally, important quality assessments by sensory panels rely heavily on subjective data rather than quantitative measures. Against this background there is an urgent need to develop a data structure and a data analysis strategy that will allow robust, high-throughput comparative assessment of metabolite content. In pairwise comparisons between cultivars, we show that highly interpretable models generated by the RF analysis of FIE-MS fingerprints can both accurately classify potato tubers and also explicitly identify significant compositional differences. From a utility perspective, the metabolites we predicted to be responsible for compositional differences between potato cultivars by both FIE-MS and GC-MS were indeed associated with important quality traits. Interestingly, several unknown signals were also highly significant, providing scope for the discovery of novel chemical attributes potentially relating to quality characteristics of individual cultivars. In the future we expect that FIE-MS fingerprinting can be used effectively to generate a meaningful and comprehensive representation of compositional differences within/between complex biological samples. The high-throughput nature and the requirement of only small sample volumes easily allow for the generation of sufficient replicate measurements (*33*) to ensure a rigorous generalizability assessment (*12*−*16*). Unlike many other powerful data mining approaches that strive to produce classifiers comprising largely nonredundant features (*12*−*14*, *23*), the approach we describe using RF allows all variables an equal chance of being both explicitly identified and highly ranked. By comparison of FIE-MS data representing individual genotypes to a common Mastermix fingerprint population, chemical "bar codes" based on importance score ranking of explanatory variables (e.g., top 25−30 *m*/*z* in adequate models) can be generated and might form part of a future database strategy regarding the chemical composition of foodstuffs. For example, RF ranking could be used to assess compositional similarity between batches of food raw materials or form part of a phenotyping strategy to evaluate large genotype populations encountered in either plant breeding or functional genomics experiments. From the perspective of directed plant breeding, a similar approach could be envisaged to link metabolome fingerprints to complex quality traits and to identify genetic sources of significant compositional novelty (*5*, *8*, *9*, *18*, *33*, *34*).

ABBREVIATIONS USED

DT, decision trees; FIE-MS, flow injection electrospray ionization mass spectrometry; NMR, nuclear magnetic resonance; *m*/*z*, mass to charge ratio; RF, random forest.

ACKNOWLEDGMENT

We acknowledge the valuable contributions made to this work by Oliver Fiehn and colleagues (M.P.I., Golm, and now University of California—Davis), who provided the potato samples. We thank Jim Heald and Robert Darby (Biological Sciences, Aberystwyth) for supporting the LCT analysis.

Supporting Information Available: Construction and validation of a Mastermix fingerprint population and supporting Figures 1 (random forest ranking of explanatory FIE-MS fingerprint negative ions in pairwise comparisons of potato cultivars), 2 (correlation analysis of explanatory FIE-MS fingerprint negative ions highlighted by random forest in pairwise comparisons of potato cultivars), and 3 (predicted identity of explanatory FIE-MS fingerprint ions highlighted in random forest analysis of potato cultivars). This material is available free of charge via the Internet at http://pubs.acs.org.

LITERATURE CITED

(1) Fiehn, O. Metabolomics: the link between genotype and phenotype. *Plant Mol. Biol.* **2002**, *48*, 155−171.

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

(3) Bino, R. J.; Hall, R. D.; Fiehn, O.; Kopka, J.; Saito, K.; Draper, J.; Nikolau, B. J.; Mendes, P.; Roessner-Tunali, U.; Beale, M. H.; Trethewey, R. N.; Lange, B. M.; Wurtele, E. S.; Sumner, L. W. Potential of metabolomics as a functional genomics tool. *Trends Plant Sci.* **2004**, *9*, 418−425.

(4) Dunn, W. B.; Bailey, N. J. C.; Johnson, H. E. Measuring the metabolome: current analytical technologies. *Analyst* **2005**, *130*, 606−625.

(5) Keurentjes, J. J. B.; Fu, J.; de Vos, R. C. H.; Lommen, A.; Hall., R. D.; Bino, R.; van der Plas, L. H. W.; Jansen, R. C.; Vreugdenhil, D.; Kornneef, M. The genetics of plant metabolism. *Nat. Genet.* **2006**, *38*, 842−849.

(6) Enot, D. P.; Beckmann, M.; Overy, D. P.; Draper, J. Predicting interpretability of metabolome models based on behavior, putative identity and biological relevance of explanatory signals. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 14865−14870.

(7) Ward, J. L.; Harris, C.; Lewis, J.; Beale, M. H. Assessment of ^{1}H-NMR spectroscopy and multivariate analysis as a technique for metabolite fingerprinting of *Arabidopsis thaliana*. *Phytochemistry* **2003**, *62*, 949−957.

(8) Scholz, M.; Gatzek, S.; Sterling, A.; Fiehn, O.; Selbig, J. Metabolite fingerprinting: detecting biological features by independent component analysis. *Bioinformatics* **2004**, *20*, 1−8.

(9) Aharoni, A.; De Vos, C. H. R.; Verhoeven, H. A.; Maliepaard, C. A.; Kruppa, G.; Bino, R.; Goodenowe, D. B. Nontargeted metabolome analysis by use of Fourier transform ion cyclotron mass spectrometry. *OMICS* **2002**, *6*, 217−234.

(10) Catchpole, G. S.; Beckmann, M.; Enot, D. P.; Mondhe, M.; Zywicki, B.; Taylor, J.; Hardy, N.; Smith, A.; King, R. D.; Kell, D. B.; Fiehn, O.; Draper, J. Hierarchical metabolomics demonstrates substantial compositional similarity between genetically-modified and conventional potato crops. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 14458−14462.

(11) 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*, 137−148.

(12) Kell, D. B.; Darby, R. M; Draper, J. Genomic computing. Explanatory analysis of plant expression profiling data using machine learning. *Plant Physiol.* **2001**, *126*, 943−951.

(13) Somorjai, R. L.; Dolenko, B.; Baumgartner, R. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics* **2003**, *12*, 1484−1491.

(14) Goodacre, R.; Vaidyanathan, S.; Dunn, W. B.; Harrigan, G. G.; Kell, D. B. Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol.* **2004**, *22*, 439−444.

(15) Baumgartner, C.; Bohm, C.; Baumgartner, D.; Mariani, G.; Weinberger, K.; Olgemoller, B.; Liebl, B.; Roscher, A. A. Supervised machine learning techniques for the classification of metabolic disorders in newborns. *Bioinformatics* **2004**, *20*, 2985−2996.
(16) Bijlsma, S.; Bobeldijk, I.; Verheij, E. R.; Ramaker, R.; Kochhar, S.; Macdonald, I. A.; van Ommen, B.; Smilde, A. G. Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. *Anal. Chem*. **2006**, *78*, 567−574.
(17) Breiman, L. Random forests. *Mach. Learn*. **2001**, *45*, 5−32.
(18) Lunetta, K. L; Hayward, L. B.; Segal, J.; Van Eerdewegh, P. Screening large-scale association study data: exploiting interactions using random forest. *BMC Genet*. **2004**, *5*, 32.
(19) Quinlan, J. R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann: San Mateo, CA, 1993.
(20) Taylor, J.; King, R.; Altmann, T.; Fiehn, O. Application of metabolomics to plant genotype discrimination using statistics and machine learning. *Bioinformatics* **2002**, *18*, S241−S248.
(21) Baumgartner, C.; Bohm, C.; Baumgartner, D. Modelling of classification rules on metabolic patterns including machine learning and expert knowledge *J. Biomed. Inf.* **2005**, *38*, 89−98.
(22) Hastie T.; Tibshirani R.; Friedman, J. *The Elements of Statistical Learning*; Springer-Verlag: Berlin, Germany, 2001.
(23) Shawe-Taylor, J.; Cristianini, N. *Kernel Methods for Pattern Analysis*; Cambridge University Press: Cambridge, U.K., 2004.
(24) Rowland, J. J. Model selection methodology in supervised learning with evolutionary computation. *BioSystems* **2003**, *72*, 187−196.
(25) Kendziorski, C.; Irizarry, R. A.; Chen, K. S.; Haag, J. D.; Gould, M. N. On the utility of pooling biological samples in microarray experiments, *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 4252−4257.
(26) Bakay, M.; Chen, Y-W.; Borup, R.; Zhao, P.; Nagaraju K.; Hoffman, E. P. Sources of variability and effect of experimental approach on expression profiling data interpretation. *BMC Bioinformatics* **2002**, *3*, 4.
(27) R environnent, http://www.R-project.org (accessed Nov 2006).
(28) The Arabidopsis Information Resource (TAIR) database, http://www.arabidopsis.org/tools/aracyc/ (accessed Nov 2006).
(29) Martin, F.; Ames, J. M. Formation of Strecker aldehydes and pyrazines in a fried potato model system. *J. Agric. Food Chem.* **2001**, *49*, 3885−3892.
(30) Duckham, S. C.; Dodson, A. T.; Bakker, J.; Ames, J. M. Volatile flavour components of baked potato flesh. A comparison of eleven potato cultivars. *Nahrung* **2001**, *45*, 317−323.
(31) Corsini, D. L.; Pavek, J. J.; Dean, B. Differences in free and protein-bound tyrosine among potato genotypes and the relationship to internal blackspot resistance. *Am. Potato J*. **1992**, *69*, 423−435.
(32) Dao, L.; Friedman, M. Chlorogenic acid content of fresh and processed potatoes determined by ultraviolet spectrophotometry. *J. Agric. Food Chem.* **1992**, *40*, 2152−2156.
(33) Ein-Dor, L.; Zuk, O.; Domany, E. Thousands of samples are needed to generate a robust list for predicting outcome in cancer. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 5923−5928.
(34) Schauer, N.; Semel, Y.; Roessner, U.; Gur, A.; Balbo, I.; Carrari, F.; Pleban, T.; Perez-Melis, A.; Bruedigam, C.; Kopka, J.; Willmitzer, L.; Zamir, D.; Fernie, A. R. Comprehensive metabolic profiling and phenotyping of intersecific introgression lines for tomato improvement. *Nat. Biotechnol*. **2006**, *24*, 447−454.

Received for review January 22, 2007. Revised manuscript received March 5, 2007. Accepted March 8, 2007. The potato mass spectrometry data were partly generated in a research program (G02006) funded by the U.K. Food Standards Agency. M.B. and D.P.E. were supported by the University of Wales, Aberystwyth, and D.P.O. was supported by a Biotechnology and Biological Sciences U.K. ISIS award.

JF0701842