Predicting interpretability of metabolome models based on behavior, putative identity, and biological relevance of explanatory signals

David P. Enot, Manfred Beckmann, David Overy, and John Draper*

Institute of Biological Sciences, University of Wales, Aberystwyth SY23 3DA, United Kingdom

Edited by Maarten Koornneef, Wageningen University and Research Centre, Wageningen, The Netherlands, and approved August 2, 2006
(received for review June 20, 2006)

Powerful algorithms are required to deal with the dimensionality of metabolomics data. Although many achieve high classification accuracy, the models they generate have limited value unless it can be demonstrated that they are reproducible and statistically relevant to the biological problem under investigation. Random forest (RF) generates models, without any requirement for dimensionality reduction or feature selection, in which individual variables are ranked for significance and displayed in an explicit manner. In metabolome fingerprinting by mass spectrometry, each metabolite can be represented by signals at several *m*/*z*. Exploiting a prior understanding of expected biochemical differences between sample classes, we aimed to develop meaningful metrics relevant to the significance both of the overall RF model and individual, potentially explanatory, signals. Pair-wise comparison of related plant genotypes with strong phenotypic differences demonstrated that robust models are not only reproducible but also logically structured, highlighting correlated *m*/*z* derived from just a small number of explanatory metabolites reflecting the biological differences between sample classes. RF models were also generated by using groupings of samples known to be increasingly phenotypically similar. Although classification accuracy was often reasonable, we demonstrated reproducibly in both *Arabidopsis* and potato a performance threshold based on margin statistics beyond which such models showed little structure indicative of either generalizibility or further biological interpretability. In a multiclass problem using 25 *Arabidopsis* genotypes, despite the complicating effects of ecotype background and secondary metabolome perturbations common to several mutations, the ranking of metabolome signals by RF provided scope for deeper interpretability.

mass spectral fingerprinting | phenotyping | random forest data analysis

Conceptually, the analysis of high dimensional metabolomics data (1–3) is data-driven and dependent on powerful multivariate modeling techniques (4–10). As in proteomics and transcriptomics research (11, 12), there is an imminent need to develop strategies to verify both the reproducibility and significance of such classification results. In the present study, an assumption is made that the goal of any data modeling experiment is not only to attempt to cluster or discriminate sample classes but also to identify the major "explanatory" metabolome signals important in any model construction. Using well characterized plant genotypes, many with known or predictable biochemical differences, we describe a strategy to validate the robustness and interpretability potential of models generated from high dimensional metabolomics data. Metabolite "fingerprinting" (13–16) provides relatively comprehensive metabolome representations, which is especially important when differences between sample classes are unknown. Approaches using mass spectrometry (such as flow infusion electrospray ionization mass spectrometry, FIE-MS) have the advantage that signals (ion mass to charge ratio, m/z) can be linked to candidate metabolites by virtue of atomic mass (3, 17–20).

Data mining techniques build models (classifiers) describing the relationship between a predictor (genotype or cultivar class label for example) and the metabolome fingerprint (21–31). Supervised methods use a range of different strategies including statistical, neural, and rule-based methods (30). Operationally, these approaches build discriminatory models between predefined classes by using training data and then subsequently test models on previously unseen test data, often derived from the same data batch. However, an adequate test of the robustness of any model is only achieved when the same classifier demonstrates high predictive accuracy on validation data derived from an independent experiment. Supervised data mining algorithms may be categorized into those that produce directly interpretable models that represent the data in an explicit way (e.g., as in a mathematical formula or tree structure) and others that cannot easily be described in terms of the original variables. Although many of the former methods can achieve high predictive accuracy, they commonly generate exceedingly complex classification models that are opaque to further interpretability (4, 6–8). When using supervised methods, great care also has to be taken to avoid production of overoptimistic models using essentially variance that is unrelated to the problem under consideration during model construction. Thus, in practical terms, a key attribute of any model is simplicity, both to hopefully avoid irrelevant background noise and to allow for efficient targeting of just a few potentially explanatory variables for further investigation. Decision tree (DT) methods can be very efficient at selecting variables with explanatory power from data sets with high dimensionality (23, 24) and are particularly useful in metabolomics studies where variables may be associated in a nonlinear fashion (i.e., networks). Although accurate DT models can be produced, the resultant tree may miss out on adequate solutions (multiplicity problem) involving alternative explanatory variables to the ones considered in the final tree. To overcome this problem, we describe the use of random forest (RF), an extension of DT methods based on the generation and comparison of an ensemble of trees (28). RF models cope well with high dimensional data sets and multiclass problems and, more importantly, also provide insight into the structure of the data under study by quantifying the confidence in classification voting and by indicating the importance of each variable for the classification task (28, 31, 32).

When high-throughput data analysis is desired, it is important to be able to validate with confidence that any highlighted

Author contributions: J.D. designed research; D.P.E. and M.B. performed research; D.P.E., M.B., D.O., and J.D. analyzed data; and D.P.E., M.B., D.O., and J.D. wrote the paper.

The authors declare no conflict of interest.

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: FIE-MS, flow infusion electrospray ionization mass spectrometry; DT, decision tree; RF, random forest.

*To whom correspondence should be addressed. E-mail: jhd@aber.ac.uk.

www.pnas.org/cgi/doi/10.1073/pnas.0605152103

Table 1. Properties of RF models comparing sample classes with little relevant biological differences

Model	Block De	De1_De2	Co2_Co0	SST	SST/FFT	Ammonium transporters	BR antisense
Accuracy (tr)	25	65.6	61.1	55.2	45.8	33.3	57.4
Accuracy (te)	56.3	81.3	58.3	68.8	56.3	38.9	61.1
90%	**28.1**	57.8	58.3	38.5	38.5	37	37
95%	31.3	**62.5**	**61.1**	40.6	41.7	**38.9**	40.7
99%	37.5	67.2	69.4	**44.8**	**43.8**	44.4	**46.3**
Margin (tr)	−0.092	0.12	0.04	0.02	−0.03	−0.04	0.03
Margin (te)	−0.004	0.15	0.11	0.06	0.02	−0.01	0.08
90%	**−0.088**	0.03	0.03	−0.05	−0.05	−0.04	−0.04
95%	−0.081	0.05	0.05	−0.04	**−0.04**	**−0.04**	−0.04
99%	−0.061	**0.09**	0.09	**−0.03**	−0.02	−0.02	**−0.03**

tr, training; te, test. Numbers in bold represent significance threshold.

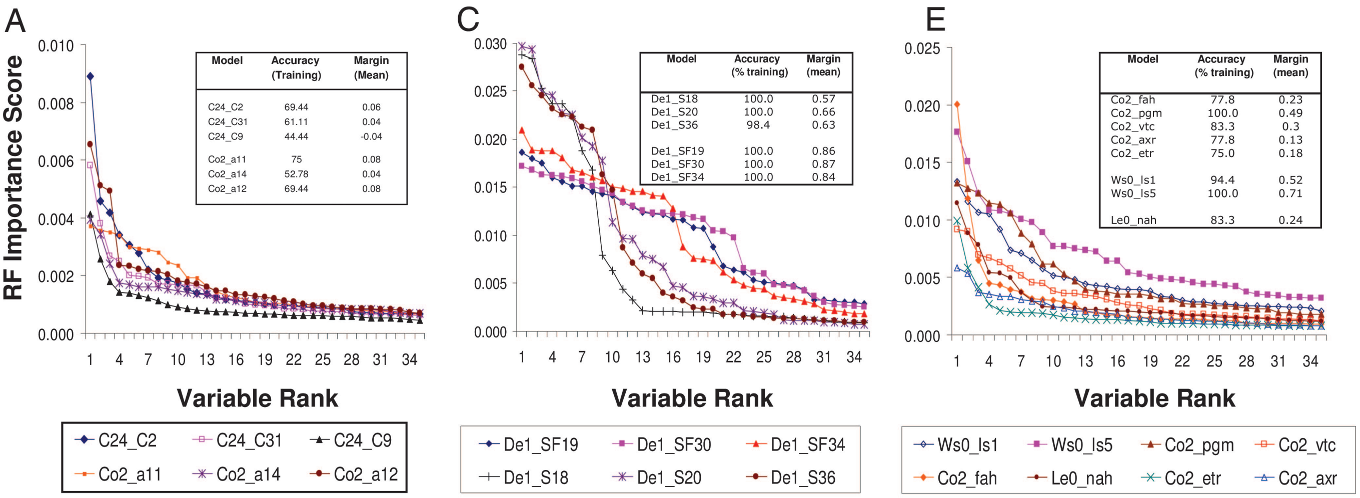
variables in models with apparently good classification accuracy are truly explanatory. In the present study, we describe a strategy both to validate the explanatory potential of RF models and explore approaches to develop significance metrics appropriate for different types of experimental situations. An initial aim was to define a baseline indicative of a significant difference in models involving binary comparisons of sample classes. Building on this information, a key objective was to develop a rationale for the detection of models with potentially sufficient explanatory power to guide deeper investigation of any significant metabolic phenotype; as part of this process, we evaluated a meaningful threshold for variable significance in ranked lists of potentially explanatory metabolome (m/z) signals generated by RF. Finally, based on these model interpretability measures, we discuss a strategy for the future *de novo* assessment of phenotype class membership in larger scale genotype screening experiments.

Results

Determining Metrics for Model Significance and Phenotypic Class Membership. To define a baseline for significance, RF models were developed that attempted to discriminate four near-identical field plots of potato tubers (cultivar Désirée), compare near isogenic lines in two plant species (potato, De1 and De2; and *Arabidopsis*, Co0 and Co2) and classify samples representing three independent examples each of four classes of genetically modified plants. The internal classification accuracy and average margins computed from the training sample, together with the level of significance determined by permutation testing (11), are compared with those calculated for an independent test set in Table 1. Classification accuracies are almost always not significant at the 0.99 quantile, and margins are <0.1, reflecting a very low confidence in class votes, indicating that lines within the same metaclass had very similar metabolomes. Two types of genetically modified *Arabidopsis* lines were selected for metabolite fingerprinting that were either not directly effected in metabolism (ammonium transporter T-DNA insertion lines) or effected in metabolites present at concentrations unlikely to be detectable in fingerprinting experiments [brassinosteroid hormone (BR) antisense lines]. Three independent lines of each metaclass were compared with the progenitor ecotypes by RF (Fig. 1*A*). Only three genotypes (C2 and a11 and a12) had significant accuracies (threshold = 69.4% at $P = 0.01$), but model margins all fell below 0.1 (significance threshold = 0.09 at $P = 0.01$). The top 20 variables ranked by importance score in each RF model are shown in Fig. 1*B*. In classifiers with good generalizability, it is expected that the same explanatory variables should be highly ranked in all models of the same metaclass and each should be accompanied by other metabolome signals representing isotopes, adducts, or neutral losses of the same metabolite. Very few common features were found between the two weakest models of each metaclasses (m/z boxed in Co2_a12, Co2_a14, C24_C9 and C24_C31), whereas isotope pairs are evident in the two strongest models (Co2_a11 and C24_C2).

Determining Significance Thresholds for Explanatory Variables. Potato and *Arabidopsis* lines (see Table 2, which is published as supporting information on the PNAS web site) were selected that had been shown to exhibit detectable changes in metabolism when compared with a progenitor genotype. Two transgenic potato metaclasses (three independent representatives of each in a Désirée background) had been genetically engineered to synthesize fructans of different degrees of polymerization by expression of novel enzyme activity (SST and SST/FFT genotypes; ref. 33). Five *Arabidopsis* lines in a Co2 background had been mutated in genes coding for specific enzymes in important metabolic pathways (*fah*, *pgm*, *vtc*) or genes involved in hormone signaling (*axr* and *etr*). Two spontaneous lesion mutants in a Ws0 background (*ls1* and *ls5*) and a further genotype expressing a transgene coding for salicylate hydroxylase (*nah*) had strong defense-related phenotypes.

RF discriminated all of the transgenic potato lines with a near perfect classification accuracy, and in each instance the model margin exceeded 0.5 (Fig. 1*C Inset*). We have demonstrated (19) that these transgenic potato lines contain no detectable metabolome differences except those signals associated with novel fructans, which are shaded in the ranked lists of explanatory signals shown in Fig. 1*D* (for identity of fructan m/z signals and a confirmatory correlation analysis, see Table 3 and Fig. 4, which are published as supporting information on the PNAS web site). Thus, one logical explanatory "significance" threshold would be the point at which signals not associated with fructans start to enter the list of variables ranked by RF analysis. This point is reached around rank 15–20 in the SST lines and at approximately rank 30 in the SST/FFT lines (Fig. 1*D*). In models of both genotype classes, this threshold occurs at an importance score of ≈ 0.003 (see Fig. 1*C*). Classification accuracies in the *Arabidopsis* binary comparisons approached or were higher than 80%, and model margins (with the exception of Co2_axr and Co2_etr) were above 0.2 (Fig. 1*E Inset*). In six of the *Arabidopsis* lines, an importance score >0.003 was reached at variable rankings from 10 to 30, but in Co2_axr and Co2_etr, the importance scores were generally much lower (Fig. 1*E*). In the three well studied *Arabidopsis* Co2 lines mutated in genes coding for key enzymes in specific metabolic pathways (*fah*, *pgm* and *vtc*), almost 80% of the top 20 electrospray ionization (ESI) variables were predicted to be either salt adducts or isotopes of a small number (6–8) of metabolites in both ionization modes (ESI $^{+}$ m/z are shaded in Fig. 1*F*; see Table 4, which is published as supporting informa-



Model	Accuracy (Training)	Margin (Mean)
C24_C2	69.44	0.06
C24_C31	61.11	0.04
C24_C9	44.44	-0.04
Co2_a11	75	0.08
Co2_a14	52.78	0.04
Co2_a12	69.44	0.08

Model	Accuracy (% training)	Margin (mean)
De1_S18	100.0	0.57
De1_S20	100.0	0.66
De1_S36	98.4	0.63
De1_SF19	100.0	0.86
De1_SF30	100.0	0.87
De1_SF34	100.0	0.84

Model	Accuracy (% training)	Margin (mean)
Co2_fah	77.8	0.23
Co2_pgm	100.0	0.49
Co2_vtc	83.3	0.3
Co2_axr	77.8	0.13
Co2_etr	75.0	0.18
Ws0_ls1	94.4	0.52
Ws0_ls5	100.0	0.71
Le0_nah	83.3	0.24

B

RF Rank	Co2_a11	Co2_a12	Co2_a14	C24_C2	C24_C31	C24_C9
	m/z	*m/z*	*m/z*	*m/z*	*m/z*	*m/z*
1	959	336	793	658	348	207
2	852	503	182	724	168	428
3	850	782	290	657	207	260
4	835	873	669	681	533	466
5	246	516	663	667	112	533
6	960	890	642	484	114	611
7	382	240	940	662	182	936
8	937	752	450	666	963	767
9	381	128	447	510	617	298
10	140	789	329	255	736	307
11	800	578	613	219	335	974
12	685	447	941	714	524	508
13	463	760	666	696	153	265
14	504	502	914	116	937	529
15	797	163	308	551	301	193
16	760	780	414	988	683	595
17	765	826	195	233	751	172
18	849	781	939	588	846	114
19	383	344	597	443	566	673
20	686	726	605	114	522	137
	NH4 transporters			BR antisense		

D

RF	SST Lines			SST/FFT Lines		
Rank	De1_S18	De1_S20	De1_S36	De1_SF19	De1_SF30	De1_SF34
	m/z	*m/z*	*m/z*	*m/z*	*m/z*	*m/z*
1	544	689	706	543	851	705
2	543	543	543	677	507	868
3	706	705	690	706	677	689
4	705	706	705	867	543	677
5	689	690	707	596	705	434
6	545	544	544	689	434	706
7	690	707	689	527	867	344
8	707	545	545	544	868	543
9	527	434	691	705	596	544
10	528	527	434	545	528	515
11	524	691	867	434	515	851
12	675	524	524	869	545	867
13	691	528	527	515	706	869
14	387	867	708	868	527	545
15	546	708	272	528	344	596
16	263	345	353	777	777	690
17	360	263	85	344	689	507
18	403	851	528	690	707	588
19	708	546	344	851	544	852
20	434	85	851	507	869	707
21	359	675	345	546	690	527
22	521	272	69	852	426	777
23	124	542	474	426	345	528
24	72	121	675	758	588	546
25	409	344	255	678	336	255
26	79	81	327	588	417	516
27	142	75	443	336	758	435
28	495	353	73	523	696	426
29	469	239	72	516	546	696
30	851	529	137	435	255	336
31	837	255	82	345	516	272
32	868	501	837	255	521	72
33	655	387	394	707	678	73
34	364	506	160	272	72	495
35	69	65	75	417	852	77
36	272	359	529	607	67	858
37	474	563	263	533	272	533
38	529	485	692	858	82	82
39	605	66	68	514	334	708
40	649	487	256	573	69	758

F

RF	Metabolic mutants			Signalling mutants	
Rank	Co2_fah	Co2_pgm	Co2_vtc	Co2_axr	Co2_etr
	m/z	*m/z*	*m/z*	*m/z*	*m/z*
1	207	266	843	311	381
2	883	398	261	381	382
3	960	365	685	382	383
4	382	397	382	789	805
5	381	778	818	503	959
6	685	399	797	797	800
7	867	428	819	798	550
8	800	175	381	803	797
9	959	203	179	393	290
10	708	245	447	336	501
11	617	187	329	687	666
12	417	478	383	959	769
13	803	130	980	234	424
14	802	804	841	585	771
15	797	336	803	793	883
16	383	367	873	800	269
17	848	268	960	953	749
18	981	282	773	128	961
19	985	261	247	804	733
20	862	335	138	954	807
21	503	173	852	785	806
22	896	298	347	962	337
23	778	131	391	685	168
24	961	150	299	817	548
25	912	229	491	383	411
26	379	169	301	819	802
27	424	381	799	961	184
28	619	666	817	204	685
29	472	168	937	181	869
30	416	234	842	791	686
31	769	382	127	788	778
32	397	152	821	449	993
33	511	779	829	406	257
34	223	503	787	768	229
35	832	184	460	967	530
36	733	657	830	121	128
37	259	183	438	869	804
38	559	462	804	385	683
39	861	174	953	400	131
40	915	324	820	786	594

Fig. 1. Determining the characteristics of robust RF models. (*A*) Variable importance score versus ranking in weak RF models comparing *Arabidopsis* ammonia transporter mutant lines and brassinosteroid synthesis antisense lines to progenitor ecotypes. (*B*) Ordered list of top ranking signals from data depicted in *A*; correlated variables (e.g., isotopes) are color-coded and variables shared between models in the same metaclass are boxed. Variable importance score versus ranking in stronger RF models comparing pair wise with progenitor genotypes in potato transgenic lines (*C*) and *Arabidopsis* mutants (*E*). (*D* and *F*) Top ranking signals (descending order) from a selection of models depicted in *A* and *E; m/z* representing correlated variables (e.g., isotopes, salt adducts, and common fragments) are shaded in both lists.

tion on the PNAS web site, for an explanation of signal relationships in both +ve and −ve ion data). As in the potato transgenic lines, many of the top ranking ($P = < 0.01$) signals in the binary comparison of *Arabidopsis* lines were highly correlated, suggesting that the proposed isotopes and adducts were indeed likely to be derived from the same metabolite (Fig. 5, which is published as supporting information on the PNAS web site). The lowest ranking signals putatively associated with the biochemical lesions in all three lines (*pgm*; rank 31 +ve ion; *vtc* rank 24 +ve ion and *fah*, rank 20 −ve ion) are located where the importance scores level off at values between 0.002 and 0.003. Permutation testing was applied to determine the significance of the variable importance score in each RF model. In data sets from both plant species, it can be seen that the *P* value of individual variables start to rise at a different position in the RF ranking, anywhere from rank 2 in Co2_*axr* to position 26 in De1_SF19 (Fig. 2 *A* and *C*). A threshold for variable significance in both potato and *Arabidopsis* RF models was reached at a *P* value between 0.0025 and 0.01 where a decrease in margin was evident when *m*/*z* with larger *P* values were included in the modeling process (Fig. 2 *B* and *D*).

Visualization and Interpretation of Phenotypic Relationships in Larger Scale Experiment. The average margins of all possible pair-wise RF models were projected into a 2D space by non linear mapping to illustrate the relationships between 25 *Arabidopsis* lines (Fig. 3*A*). Each *Arabidopsis* ecotype is generally well separated and genotypes with weak metabolome differences cluster close to their progenitor ecotype (e.g., Co2 with a2, a11, a12, and a14 and C24 with C2, C9, and C31). Similarly, genotypes with increasingly stronger phenotypes are found at increasing distances from the progenitor ecotype (e.g., Co2 with *fah*, *vtc*, and *pgm* and Le0

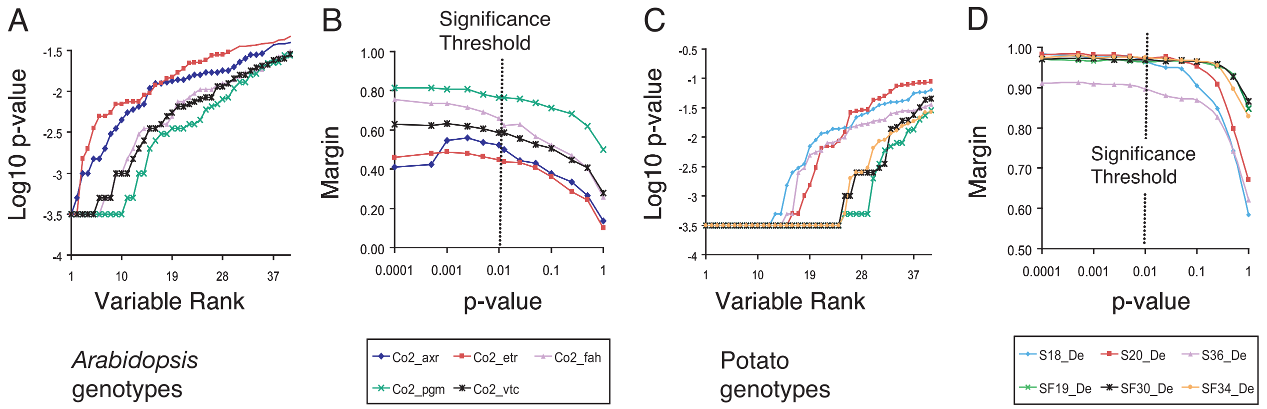


Fig. 2. Relationship between margin and variable significance in metabolome fingerprint models. Overall model *P* value (log 10) when including increasing numbers of top ranking variables in RF Analysis of *Arabidopsis* (*A*) and potato (*C*) genotypes. Overall model margins when including variables with increasing *P* value in RF analysis of *Arabidopsis* (*B*) and potato (*D*) genotypes. A suggested significance threshold is indicated.

with *uvr*, *eds*, and *nah*). The metabolome differences with the progenitor ecotype Ws0 in the case of *ls1* and *ls5* are so great that there is no apparent connectivity. The *Arabidopsis* genotypes were chosen to include six lines (*uvr*, *eds*, *vtc*, *ls1*, *ls5*, and *nah*) with described defense- or stress-related phenotypes. In a pairwise comparison to each other all of the defense-related mutants in Le0 and Ws0 backgrounds had much smaller margins than when compared with the progenitor cultivar (data not shown). The *uvr*, *eds*, and *nah* genotypes shared many top-ranking signals (Fig. 3*B*), whereas phenotypically unrelated mutants such as *pgm* and *fah* had little in common. Interestingly, despite being in a different ecotype background *vtc* also shared many explanatory features with the other defense mutants. The RF models comparing lesion mimic mutants to Ws0 had very large margins, but they were also substantially different from each other (Fig. 3*C*), and in both cases, 25–30 top ranking variables had importance scores >0.003 (see Fig. 1*E*). *Ls1* and *ls5* had a large number of top ranked variables in common (color coded in Fig. 3*C*); however, each line equally had a large proportion of highly significant ($P < 0.001$) correlated signals that had no explanatory power in the other line (Fig. 3*D*).

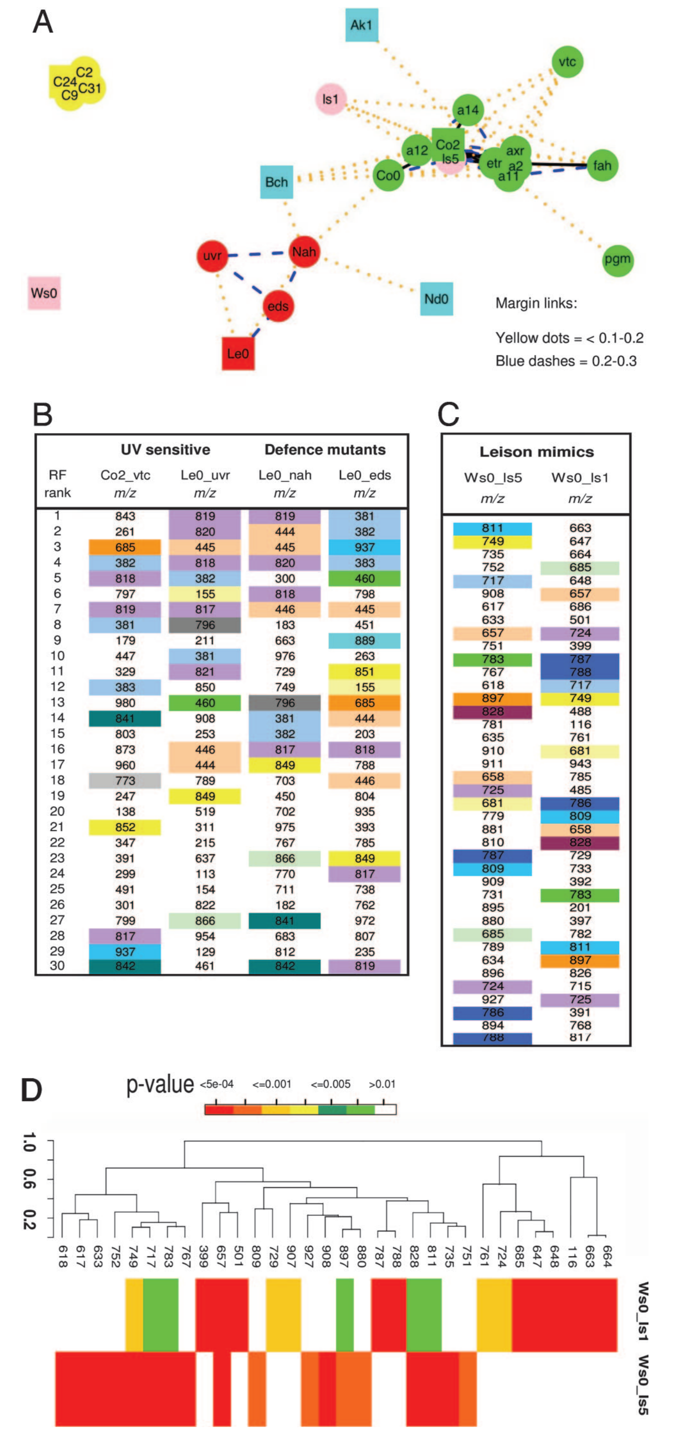
Discussion

Choice of Data Mining Technique. There are a range of strategies available to analyze metabolomics data (4–10). In preliminary work (see Table 5, which is published as supporting information on the PNAS web site), we showed that classification accuracies achieved by using RF were equivalent to those obtained by using three common supervised learning algorithms. Feature selection is of primary importance from an interpretation perspective. One problem associated with "naïve" modeling of metabolome fingerprint data are the multiplicity of possible good solutions. It is rarely the case that one unique signal (or combination of very few uncorrelated variables) will adequately describe the property under study. Indeed, previous studies have highlighted the importance of including all variables in the final model to identify "silent phenotypes" or unexpected metabolic pathways (19, 20, 34–36). Several powerful data mining approaches combine feature selection and classification in one analytical run; for example, genetic algorithms or genetic programming evaluate feature subsets by using accuracy estimates provided by a machine learning algorithm (4, 7). These "wrapper" techniques produce parsimonious models using very few variables that are generally dominated by the stronger attributes. Effectively, this means that correlated (essentially redundant) variables are selected only rarely, consequently missing out on potentially informative solutions. We suggest that RF has additional utility because the aim is not to determine the smallest feature set but to identify a complete set of statistically significant explanatory variables.

Assessing Model Robustness. Validation of a classifier demands not only a consistent predictive power but also that the variables selected for high explanatory potential should be the same in replicate experiments. Thus to achieve an adequate assessment of generalizibility, it is valuable to use algorithms, such as RF, which produce directly interpretable models that represent the data in an explicit way. We suggest that a more stringent representation of class boundary complexity than classification accuracy alone is essential for assessing model quality and further interpretability potential. By definition, the sample margin encompasses a measure of confidence in votes for the right class. In contrast to margin-based classifiers (e.g., support vector machines) or discriminant techniques (such as linear discriminant analysis or partial least squares discriminant analysis), RF does not explicitly maximize the margin, thus making this measure valuable because it is both unbiased and related directly to the generalization error.

In high-throughput metabolite fingerprinting, deciding which models to consider for deeper analysis of signals is usually problem-specific due to constraints associated with sample size (in relation to data variance and dimensionality characteristics) and the lack of prior knowledge about expected margin distributions. Permutation-based tests have been used to provide such information (11). We describe an alternative approach to evaluate any new experimental system based on a combined examination of statistical significance and biological information content to validate robustness and interpretability potential. Thus, using plants of known genotype and predicable biochemical phenotypes, we have explored both margin characteristics and variable behavior in FIE-MS fingerprint models that we expect to be either poor, or possibly adequate or robust in terms of generalizibility. These observations suggest that FIE-MS fingerprinting in combination with RF analysis will be valuable as a prescreen to detect lines with novel metabolic phenotypes in large populations. In model systems, such as *Arabidopsis*, targeted, even quantitative, profiling approaches using high mass resolution instruments (3, 17, 20) will become more routine as metabolite identity in the sample matrix is better understood, allowing signals relating to specific molecules to be predesignated.

Selecting Significance Thresholds and Evaluating Model Interpretability Potential. Importance score ranking combined with significance testing of m/z signals provides an excellent metric con-



	UV sensitive		Defence mutants	
RF rank	Co2_vtc *m/z*	Le0_uvr *m/z*	Le0_nah *m/z*	Le0_eds *m/z*
1	843	819	819	381
2	261	820	444	382
3	685	445	445	937
4	382	818	820	383
5	818	382	300	460
6	797	155	818	798
7	819	817	446	445
8	381	796	183	451
9	179	211	663	889
10	447	381	976	263
11	329	821	729	851
12	383	850	749	155
13	980	460	796	685
14	841	908	381	444
15	803	253	382	203
16	873	446	817	818
17	960	444	849	788
18	773	789	703	446
19	247	849	450	804
20	138	519	702	935
21	852	311	975	393
22	347	215	767	785
23	391	637	866	849
24	299	113	770	817
25	491	154	711	738
26	301	822	182	762
27	799	866	841	972
28	817	954	683	807
29	937	129	812	235
30	842	461	842	819

Leison mimics	
Ws0_ls5 *m/z*	Ws0_ls1 *m/z*
811	663
749	647
735	664
752	685
717	648
908	657
617	686
633	501
657	724
751	399
783	787
767	788
618	717
897	749
828	488
781	116
635	761
910	681
911	943
658	785
725	485
681	786
779	809
881	658
810	828
787	729
809	733
909	392
731	783
895	201
880	397
685	782
789	811
634	897
896	826
724	715
927	725
786	391
894	768
788	817

Fig. 3. Metabolome modeling with larger multiple class problems. (*A*) Two-dimensional mapping of 25 *Arabidopsis* lines using Sammon nonlinear mapping. Control ecotypes are colored blue, and progenitor ecotypes of mutant lines are presented as squares. The ecotype background of mutant lines is depicted by color: red, LeO; yellow, C24; pink, Ws0; green, Columbia. The lines linking phenotypically related genotypes represent margins in pair-wise comparisons and are color coded as follows: black solid line, <0.1; yellow dotted line, 0.1–0.2; blue dashed line, 0.2–0.3. Margins >0.3 have been omitted for the representation. (*B*) Top ranking signals in common (color coded) between RF models representing pair-wise comparisons between selected defense related and UV sensitive genotypes and their progenitor ecotypes. (*C*) RF models comparing lesion mimic mutants (*ls1* and *ls5*) with the progenitor genotype (Ws0) indicating the presence of many common signals (color coded). (*D*) A correlation analysis of variables contributing significantly ($P = < 0.005$) to models discriminating mutant lines *ls1* and *ls5* from the progenitor ecotype Ws0.

tributing to a rapid assessment of model robustness and interpretability potential. A standardized significance cutoff based on an arbitrary RF rank was not appropriate, because significance testing revealed that *P* values for individual variables started to rise at different rank positions in individual models. By generating a series of models using only variables with *P* values below predetermined thresholds, it was shown that margins began to drop significantly when variables with *P* values >0.001 were used. Prior knowledge of biochemical difference between genotypes, particularly in the novel fructan-producing transgenic potato lines, allowed us to confirm that signals unrelated to the transgenic phenotype began to populate models if variables with a *P* value >0.0025 were used. In most instances, for pair-wise comparison of genotypes, this *P* value threshold ($P \leq 0.001$ to $P \leq 0.0025$) correlated with an importance score of ≈0.003, below which any variables were unlikely to have any significant explanatory power. Robust models with adequate margins (>0.2) derived from comparison of progenitor genotypes with mutants (or transgenic lines) with strong metabolic phenotypes had a large proportion (>65%) of correlated *m/z* signals (e.g., potential isotopes, salt adducts, and neutral losses representing the same predicted metabolite) in the top 30 ranked variables. Mutants with pleiotrophic effects lacking a distinct metabolic phenotype, such as auxin (*axr-1*) or ethylene (*etr-1*) hormone signaling defective lines, exhibited much lower levels (<35%) of correlated variables. In situations where the metabolic differences between genotypes are more discrete, centering on signals derived from just a small number (2–5) of metabolites, there is clearly a much greater potential for further interpretability. Typically in such models, the top ranking signals are highly correlated and importance scores drop rapidly to the significance threshold.

Interpretation of Phenotypic Relationships in Larger Mutant Genotype Populations. In large, multiple-class experiments, the meaningful representation of high dimensional multivariate models is problematic, particularly if the objective is to assign any phenotypic relatedness based on separation "distances" that encapsulate the diversity of genotypic differences. Few studies have tackled this problem, and our rather small experiment (considering the number of *Arabidopsis* mutants available) already illustrates the richness of the information derived from metabolome fingerprints based on mass spectrometry. A traditional approach would be to compare each genotype to a so called progenitor line (e.g., Desiree or Columbia in the present example) and relate phenotypic differences to a representative example of a given species. However, this strategy runs the risk of missing crucial or novel relationships between specific lines. For example, in the present study, we demonstrate that effectively "unlinked" genotypes (e.g., *vtc* and *eds*) can, in fact, share many highly ranked explanatory variables; in this case, the ecotype background dominates the modeling process. A further factor impinging on effective phenotyping is the problem of secondary effects of mutations on the metabolome that mask the true explanatory differences between classes; this is demonstrated in the present study by the fact that the defense-related mutants had a large subset of highly ranked variables that were probably associated with their general light (UV) sensitivity. Similarly, the two lesion mimic mutants had a large number of signals in common related to secondary effects after the induction of cell death. However, in both of these situations, by examining the list of top-ranking signals, it is possible to identify real differences between such genotypes.

In conclusion, we suggest that direct interpretability, and a nonbiased capacity to deal with multiple adequate solutions, are just as important as achieving a high classification accuracy in any metabolome modeling procedure using high dimensional data. By representing and ranking all potentially explanatory variables in an explicit way, RF models provide an increased opportunity for assessment of model generalizibility and deeper phenotypic investigation. With more standardized metabolome fingerprinting procedures in the future, we suggest that ranked lists of top explanatory variables (perhaps 20–30 with an adequate model margin) may provide robust, directly comparable representations of sample composition in situations requiring high throughput classification tasks.

Materials and Methods

Plant Material, Sample Preparation, and Metabolite Analysis. Experimental transgenic potato genotypes engineered to synthesize fructans (33) were derived from the cultivar Désirée and have been described (19). A somaclonal variant (De2) generated via tissue culture provided a near-isogenic line of the commercial Désirée cultivar (De1). Procedures for sample preparation and extraction have been described (19). Information on the *Arabidopsis* genotypes selected for this study and additional details of metabolite analysis are presented in *Supporting Text*, which is published as supporting information on the PNAS web site. A minimum of 30 biological replicates were used to develop FIE-MS fingerprints in both ionisation modes.

Data Modeling. Sample classification, selection, and ranking of potentially explanatory variables in FIE-MS data were achieved by using an implementation of RF as described in *Supporting Text*. Training/test set partitioning was carried out on the basis of independent analytical batches with 18 and 12 plant replicates selected to form the training and test set, respectively. For each RF model, classification accuracies and average margins were computed from the ‘‘out of bag’’ training samples. One thousand trees were generated in each modeling experiment by using the overall fingerprint if not stated otherwise. The importance score for each m/z for each classification task to define a ranked list of potentially explanatory signals was computed according to Breiman (28). The levels of significance were determined by a permutation test (11, 37) under the null hypothesis that the importance score is not relevant to the classification task. The P value is defined as the fraction of times an importance score in the class-permuted data are greater or equal to the score in the unpermuted data. Two thousand permutations were performed. Average margins (38) of the training samples were used as input for the Sammon nonlinear mapping algorithm (39) using the library ‘‘MASS’’ in the R environment (http://www.r-project.org).

FIE-MS Signal Interpretation. The initial data analysis by RF produced a list of m/z signals ranked by importance scores or P value for each classification task. The lists of top ranked m/z (generally top 40) were examined for groups of potentially related signals that could represent either the (de)protonated ion (e.g., $[M+H]+ = M + 1$), salt adducts (both single and double charged e.g., $[M+Na]^+ = M + 23$, $[M+K]^+ = M + 39$ or $[M+Na+K]^{2+} = (M + 23 + 39)/2)$, common neutral losses (e.g., $[M+H\text{-}H_2O]^+ = M\text{-}17$ and $[M+H\text{-}HCOOH]^+ = M\text{-}45)$, the homogeneous dimer ion [e.g., $[2M+H]^+ = 2(M + 1)$], and dimer ion pair adducts [e.g., $[2M+Na]^+ = 2(M + 23)$] as well as isotopes ($M + 2$ or $M + 3$ amu) of a single metabolite. Because several overlapping solutions predicting the presence of different metabolites were often possible, the most likely combination of ions putatively identifying a specific metabolite was confirmed by further examining signal relationships in a correlation analysis using just m/z with an appropriate low P value.

We thank Oliver Fiehn and colleagues (Max Planck Institute, Golm, Germany, and now University of California, Davis, CA) who provided the potato samples and seed of many of the *Arabidopsis* lines, and Jim Heald and Robert Darby (Institute of Biological Sciences, University of Wales, Aberystwyth, U.K.) for supporting the LCT analysis. M.B. and D.O. were supported by the University of Wales, Aberystwyth. D.O. was funded by an International Scientific Interchange Scheme award from the Biotechnology and Biological Sciences Research Council (United Kingdom).

1. Dunn WB, Bailey NJC, Johnson HE (2005) *Analyst* 130:606–625.
2. Bino RJ, Hall RD, Fiehn O, Kopka J, Saito K, Draper J, Nikolau BJ, Mendes P, Roessner-Tunali U, Beale MH, *et al.*(2004) *Trends Plant Sci* 9:418–425.
3. Dunn WB, Overy S, Quick WP (2005) *Metabolomics* 1:137–148.
4. Kell DB, Darby RM, Draper J (2001) *Plant Physiol* 126:943–951.
5. Somorjai RL, Dolenko B, Baumgartner R (2003) *Bioinformatics* 12:1484–1491.
6. Baumgartener C, Bohm C, Baumgartner D, Mariani G, Weinberger K, Olgemoller B, Liebl B, Roscher AA (2004) *Bioinformatics* 20:2985–2996.
7. Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB (2004) *Trends Biotechnol* 22:439–444.
8. Bijlsma S, Bobeldijk I, Verheij ER, Ramaker R, Kochhar S, Macdonald IA, van Ommen B, Smilde AG (2006) *Anal Chem* 78:567–574.
9. Lee K, Hwang D, Yokoyama T, Stephanopoulos G, Stephanopoulos GN, Yarmush ML (2004) *Bioinformatics* 20:959–969.
10. Wu B, Abbott T, Fishman D, McMurray W, Mor G, Stone K, Ward D, Williams K, Zhao H (2004) *Bioinformatics* 19:1636–1643.
11. Lyons-Weiler J, Pelikan R, Zeh HJ, Whitcomb DC, Malehorn DE, Bigbee WL, Hauskrecht M (2005) *Cancer Informatics* 1:53–77.
12. Ein-Dor L, Zuk O, Domany E (2006) *Proc Natl Acad Sci USA* 103:5923–5928.
13. Fiehn O (2002) *Plant Mol Biol* 48:155–171.
14. Mouille G, Robin S, Lecomte M, Pagant S, Hofte H (2003) *Plant J* 35:393–404.
15. Ward JL, Harris C, Lewis J, Beale MH (2003) *Phytochemistry* 62:949–957.
16. Allen J, Davey HM, Broadhurst D, Heald JK, Rowland JJ, Oliver SG, Kell DB (2003) *Nat Biotechnol* 21:692–696.
17. Aharoni A, De Vos CHR, Verhoeven HA, Maliepaard CA, Kruppa G, Bino R, Goodenowe DB (2002) *OMICS* 6:217–234.
18. Scholz M, Gatzek S, Sterling A, Fiehn O, Selbig J (2004) *Bioinformatics* 20:1–8.
19. Catchpole GS, Beckmann M, Enot DP, Mondhe M, Zywicki B, Taylor J, Hardy N, Smith A, King RD, Kell DB, *et al.* (2005) *Proc Natl Acad Sci USA* 102:14458–14462.
20. Keurentjes JJB, Fu J, de Vos RCH, Lommen A, Hall RD, Bino R, van der Plas LHW, Jansen RC, Vreugdenhil D, Kornneef M (2006) *Nat Genet* 38:842-849.
21. Manley BFJ (1994) *Multivariate Statistical Methods*: *A Primer* (Chapman and Hall, London).
22. Dietterich TG (1998) *Neural Comput* 10:1895–1923.
23. Baumgartner C, Bohm C, Baumgartner D (2005) *J Biomed Informat* 38:89–98.
24. Hastie T, Tibshirani R, Friedman J (2001) *The Elements of Statistical Learning* (Springer, Berlin).
25. Shawe-Taylor J, Cristianini N (2004) *Kernel Methods for Pattern Analysis* (Cambridge Univ Press, Cambridge, UK).
26. Vapnik VN (1998) *Statistical Learning Theory* (Wiley, New York)
27. Quinlan JR (1993) *C4.5*: *Programs for Machine Learning* (Morgan Kaufmann, San Mateo, CA).
28. Breiman L (2001) *Machine Learning* 45:5–32.
29. Jonsson P, Gullberg J, Nordstrom A, Kusano M, Kowalczyk M, Sjostrom M, Moritz T (2004) *Anal Chem* 76:1738–1745.
30. Weiss SH, Kulikowski CA (1991) *Computer Systems that Learn* (Morgan Kaufmann, San Mateo, CA).
31. Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P (2004) *BMC Genetics* 5:32–45.
32. Diaz-Uriarte R, Alvarez de Andres S (2006) *BMC Bioinformatics* 7:3.
33. Hellwege EM, Czapla S, Jahnke A, Willmitzer L, Heyer AG (2000) *Proc Natl Acad Sci USA* 97:8699–8704.
34. Roessner U, Luedemann A, Brust D, Fiehn O, Linke T, Willmitzer L, Fernie AR (2001) *Plant Cell* 13:11–29.
35. Fiehn O, Kopka J, Altmann T, Trethewey R, Willmitzer L (2000) *Nat Biotechnol* 18:1157–1161.
36. Weckwerth W, Loureiro ME, Wenzel K, Fiehn O (2004) *Proc Natl Acad Sci USA* 101:7809–7814.
37. Good P (2000) *Permutation Tests*: *A Practical Guide to Resampling Methods for Testing Hypotheses* (Springer, Berlin)
38. Schapire R, Freund Y, Bartlett P, Lee W (1998) *Ann Stat* 26:1651–1686.
39. Sammon JW (1969) *IEEE Trans Comput C* 18:401–409.