Enhancement of Plant Metabolite Fingerprinting by Machine Learning[1][W]

Ian M. Scott*, Cornelia P. Vermeer, Maria Liakata, Delia I. Corol, Jane L. Ward, Wanchang Lin, Helen E. Johnson, Lynne Whitehead, Baldeep Kular, John M. Baker, Sean Walsh, Anuja Dave, Tony R. Larson, Ian A. Graham, Trevor L. Wang, Ross D. King, John Draper, and Michael H. Beale

Institute of Biological, Environmental, and Rural Sciences, Aberystwyth University, Aberystwyth SY23 3DA, United Kingdom (I.M.S., C.P.V., W.L., H.E.J., J.D.); Department of Computer Science, Aberystwyth University, Aberystwyth SY23 3DB, United Kingdom (M.L., R.D.K.); National Centre for Plant and Microbial Metabolomics, Rothamsted Research, Harpenden AL5 2JQ, United Kingdom (D.I.C., J.L.W., J.M.B., M.H.B.); Centre for Novel Agricultural Products, University of York, York YO10 5YW, United Kingdom (L.W., A.D., T.R.L., I.A.G.); and John Innes Centre, Norwich NR4 7UH, United Kingdom (B.K., S.W., T.L.W.)

Metabolite fingerprinting of Arabidopsis (*Arabidopsis thaliana*) mutants with known or predicted metabolic lesions was performed by ^{1}H-nuclear magnetic resonance, Fourier transform infrared, and flow injection electrospray-mass spectrometry. Fingerprinting enabled processing of five times more plants than conventional chromatographic profiling and was competitive for discriminating mutants, other than those affected in only low-abundance metabolites. Despite their rapidity and complexity, fingerprints yielded metabolomic insights (e.g. that effects of single lesions were usually not confined to individual pathways). Among fingerprint techniques, ^{1}H-nuclear magnetic resonance discriminated the most mutant phenotypes from the wild type and Fourier transform infrared discriminated the fewest. To maximize information from fingerprints, data analysis was crucial. One-third of distinctive phenotypes might have been overlooked had data models been confined to principal component analysis score plots. Among several methods tested, machine learning (ML) algorithms, namely support vector machine or random forest (RF) classifiers, were unsurpassed for phenotype discrimination. Support vector machines were often the best performing classifiers, but RFs yielded some particularly informative measures. First, RFs estimated margins between mutant phenotypes, whose relations could then be visualized by Sammon mapping or hierarchical clustering. Second, RFs provided importance scores for the features within fingerprints that discriminated mutants. These scores correlated with analysis of variance *F* values (as did Kruskal-Wallis tests, true- and false-positive measures, mutual information, and the Relief feature selection algorithm). ML classifiers, as models trained on one data set to predict another, were ideal for focused metabolomic queries, such as the distinctiveness and consistency of mutant phenotypes. Accessible software for use of ML in plant physiology is highlighted.

Functional genomics extends Johannsen's 1909 concept of the phenotype as the total of an organism's expressed characters (Nachtomy et al., 2007). In practice, nonetheless, phenotype data inevitably remain partial, contextual, and subject to acquisition methods. These complexities are very evident in metabolomics, whose phenotyping opportunities, while exciting, are fraught with technical and conceptual challenges. Some of these issues are explored in this paper, along with practical guidance on relevant strategic approaches that have emerged in recent years.

One challenge in metabolomics is the extent to which its ambition to define global metabolic phenotypes is constrained by lack of technologies encompassing metabolite diversity (Hall, 2006). This has caused more proliferation of methods than in other functional genomics. We examine how the apparent metabolic phenotype depends on the analytical frame of reference, an obvious theoretical issue but infrequently evaluated in practice.

The progenitor of metabolomics, metabolite "profiling" by gas chromatography-mass spectrometry (GC-MS), is unsurpassed for identifying multiple metabolites (Lisec et al., 2006), and its phenotyping applications increase. Examples are association of tomato (*Solanum lycopersicum*) quantitative trait loci for yield and metabolism (Schauer et al., 2006), growth rate prediction from metabolite composition (Meyer et al., 2007), and starch mutant classification (Messerli et al., 2007). These GC-MS studies quantified 43 to 181 known or unidentified compounds.

While progress in metabolic phenotyping is impressive, its acceleration is desirable. The AraCyc database, for example, is contextualizing Arabidopsis (*Arabidop-*

[1] This work was supported by the United Kingdom Biotechnology and Biological Sciences Research Council (BBSRC). SM line production was supported by a BBSRC award to M.W. Bevan and J.D.G. Jones (John Innes Centre and the Sainsbury Laboratory).

* Corresponding author; e-mail ias@aber.ac.uk.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Ian M. Scott (ias@aber.ac.uk).

[W] The online version of this article contains Web-only data.

www.plantphysiol.org/cgi/doi/10.1104/pp.109.150524

sis thaliana) genes in metabolic pathways but still has many "pathway holes," or functional annotations lacking evidence (Zhang et al., 2005). Metabolomic knowledge deficits may expand as genomes are sequenced.

For such reasons, optimization of time and resources is a theme in metabolomics (Hall, 2006). Time-consuming chromatography in "hyphenated" technologies such as GC-MS can impede progress, while chromatographic deconvolution (Lisec et al., 2006) is an extra interpretive complexity. An alternative that prioritizes speed over metabolite identification is metabolite (or metabolic) "fingerprinting," whose minimalist versions have little sample preparation and no chromatography (Hall, 2006). We compared three primary spectroscopies in analyses of unfractionated plant extracts. First, NMR is less sensitive than GC-MS (Lisec et al., 2006) but is a dominant biomedical technique with proven utility for plants (Ward et al., 2007). Then, an MS approach was tested in the form of flow injection electrospray (FIE)-MS (Beckmann et al., 2008), where bioextracts are infused without liquid chromatography (LC). Finally, the vibrational spectroscopy Fourier transform infrared (FTIR) gives the option of analyzing whole-tissue preparations (Gidman et al., 2006), although with less molecular information. These technologies were compared for phenotyping in parallel analyses of replicate Arabidopsis plants grown simultaneously in a single environment. Metabolic distinctiveness from the Columbia (Col-0) wild type, and its consistency, were gauged in known or predicted mutants with little morphological impact.

A key question was whether loss of information or precision in fingerprinting diminished phenotype discrimination relative to conventional, targeted profiling, which monitors metabolite groups by selective methods (Hall, 2006). Therefore, we also produced LC and GC profiles of more than 100 amino acids, isoprenoids, fatty acids, and acyl-CoAs. This number of metabolites was typical for profiling studies (see above) but 1 order of magnitude less than the fingerprints (901–1,852 spectral values) with which they were compared.

This numerical comparison illustrates a primary challenge in functional genomics: how to deal with so-called high-dimensional data spaces, where hundreds of variables define samples (Clarke et al., 2008). Data modeling is a complex, evolving field, often unfamiliar to plant biologists. Therefore, we present ways in which it might be employed in metabolic phenotyping and compare methodologies. We minimize technical description but provide a glossary in Supplemental Table S1.

Effectively, the standard data-modeling tool in plant metabolomics is principal component analysis (PCA). The rationale behind a PC is that it captures a global pattern in the data by weighting variables (here, metabolome signals) with high covariance. PCA is conceptually suited to systems biology, where such global patterns may reflect some coordinated cellular network (Janes and Yaffe, 2006). It should be cautioned, however, that some data analysts regard PCs as artificial mathematical entities and too often uncritically "reified" as biologically meaningful (Mahoney and Drineas, 2009).

If a PC encapsulates differences between sample classes, these may separate by their coordinates in PC space, or "scores." Two-dimensional (2D) scatterplots of PC scores are consequently a vivid and ubiquitous form of data exploration and were a benchmark in our evaluations. If data variance is dominated by factors irrelevant to one's biological hypothesis, on the other hand, PCA is limited and more may be gained from "supervised" methods using knowledge of sample classes (Tarca et al., 2007). One of the latter is partial least squares-discriminant analysis (PLS-DA), which like PCA produces a series of multivariate components. Whereas PCA encapsulates only data variance in the PCs, however, PLS-DA also seeks to discriminate classes (Janes and Yaffe, 2006).

Reputedly yet more powerful supervised tools are machine learning (ML) classifiers, a generic term for algorithms that "learn" from class-labeled training data to predict classes among test data (Tarca et al., 2007). We evaluated two approaches ranking among the most important ML developments in recent years (Friedman, 2006). One was the support vector machine (SVM), highly regarded for its roots in statistical learning theory and often unsurpassed performance (Ben-Hur et al., 2008). The other came from the decision tree category of ML (Tarca et al., 2007) in the form of random forest (RF), whose capabilities for metabolomics have been examined by Enot et al. (2006).

Importantly, we used widely available software, mostly open source (Table I), so the data-modeling strategies illustrated in this paper could be widely adopted. Software accessibility has been recognized for several years as a limiting factor for metabolomics. Sweetlove et al. (2003), referring to ML, remarked: "Although such methods are extremely powerful, they are not readily approachable.... Software interfaces will be required that allow the user to input and define their data while the algorithms are applied automatically and the results displayed in a readily interpretable format. The display of multidimensional data brings its own challenges." In the interim, ML has been underexploited in plant metabolomics, given its theoretical fitness for high-dimensional data.

This situation may change, however, with the proliferation of software packages such as those in Table I. Among the newest is the MetaboAnalyst Web server (Xia et al., 2009), whose online data analysis tools include those just outlined. The open-source software we used came from two major projects of the data-mining community, Weka and R (Table I). Weka has a well-developed graphical user interface (GUI) by which a renowned suite of ML algorithms can be accessed in basic form. R has greater functionality and processing speed but primarily requires command-

Table I. *Open-source classification and feature selection software for high-dimensional biodata*

Resource (URL)	Reference	Functionality
Weka project		
Weka (www.cs.waikato.ac.nz/ml/weka)	Frank et al. (2004)	Extensive ML suite with Java GUI
BioWeka (www.bioweka.org)	Gewehr et al. (2007)	Bioinformatics extensions for Weka
R project		
Bioconductor project (www. bioconductor.org):	Gentleman et al. (2004)	Growing assemblage of biodata tools in R
(e.g. MLInterfaces, CMA)	Tarca et al. (2007), Slawski et al. (2008)	Command line R packages with advanced ML
FIEmspro (users.aber.ac.uk/jhd)	Enot et al. (2008)	FIE-MS-oriented command line R package
Metabonomic (cran.r-project.org)	Izquierdo-Garcia et al. (2009)	R GUI with ML for proprietary NMR data
Online metabolomics data analysis		
MetaboAnalyst (www.metaboanalyst.ca)	Xia et al. (2009)	RF, SVM, PLS-DA, PCA, HCA online
MetaGeneAlyse (metagenealyse. mpimp-golm.mpg.de)	Daub et al. (2003)	PCA, ICA, HCA online
MeltDB (meltdb.cebitec.uni-bielefeld.de)	Neuweger et al. (2008)	PCA, ICA, HCA online
MetNet (metnet.vrac.iastate.edu)	Wurtele et al. (2003)	Arabidopsis functional genomics software
Further resources		
Automics (code.google.com/p/automics)	Wang et al. (2009)	C++ GUI for proprietary NMR data
MetaFIND (mlg.ucd.ie/metafind)	Bryan et al. (2008)	Java GUI with feature analysis tools
Additional ML, feature selection, metabolomics, and bioinformatics software	Arita (2004), Stajich and Lapp (2006), Saeys et al. (2007)	Reviews listing over 60 relevant resources

line interaction with modular "packages" created by the statistical computing community.

We sought to benchmark the power and, if possible, the mode of operation of newer data-modeling tools relative to established ones. We examined how the scope and strategies of hypothesis interrogation can be extended by ML. The distinctiveness and consistency of mutant phenotypes, as defined by different metabolomic techniques, were thereby explored.

RESULTS AND DISCUSSION

Phenotyping Strategies: Unsupervised versus Supervised

Two generic strategies were used for phenotyping metabolome data of Arabidopsis wild type (Col-0) and mutants. One, unsupervised phenotyping (Fig. 1A), involved PCA in SIMCA-P, a commercial software widely used in metabolomics (Xia et al., 2009). Unsupervised methods such as PCA may reveal natural groupings of data objects, without reference to predefined classes (Tarca et al., 2007).

An initial decision in PCA is whether data should be scaled for analysis. By default in SIMCA-P, each variable is scaled by dividing by its SD. Scaling Arabidopsis NMR and FIE-MS fingerprints in this way reduced the influence of larger spectral peaks, and this made PCA less interpretable. When interpreting a single PC, consideration should be given to the proportion of overall data variance it encapsulates. Successive PCs are ranked by this criterion, PC[1] representing the most prominent pattern in the data, and so on. Scaling NMR or FIE-MS fingerprints resulted in more PCs, each representing less of the data variance. We saw this as increasing the danger of reification (i.e. imputing unmerited biological meaning to PCs; Mahoney and Drineas, 2009). Therefore, we analyzed these fingerprints without scaling, although we did use the standard practice of mean centering each variable (by subtracting its average in the data).

These PCA routines were compared with phenotyping by supervised classifiers (Fig. 1B). These algorithms are "supervised" by being provided with the desired outcomes for training examples (Tarca et al., 2007). In our context, their task was to find mathematical transformations ("models") to consistently associate metabolome data patterns with specified phenotype "classes." We trained classifiers to recognize mutant versus Col-0 samples and validated the resultant models by blind testing for prediction of a different set of mutant and Col-0 samples. A central point here is that the supervised-training and blind-validation routine is ideal for phenotyping. The classifier is trained with multivariate examples of biological characters and directed to find a model relating these to hypothetical phenotypes. The resultant classifier models are phenotype models, and the blind validation is a pattern-recognition scenario to test the phenotype hypothesis. Consequently, while ML algorithms are internally complex, their conceptual employment as classifiers for phenotyping should be second nature to an experimental biologist.

The flow chart illustrates two validation designs we used (Fig. 1B). In 9-fold cross-validation (9×CV), we trained classifiers on eight of nine of the plant replicates from one or more experiments and validated on the held-out one of nine replicates. This procedure was repeated for all nine subsets in turn to get overall classification accuracies. In an alternative, physiolog-

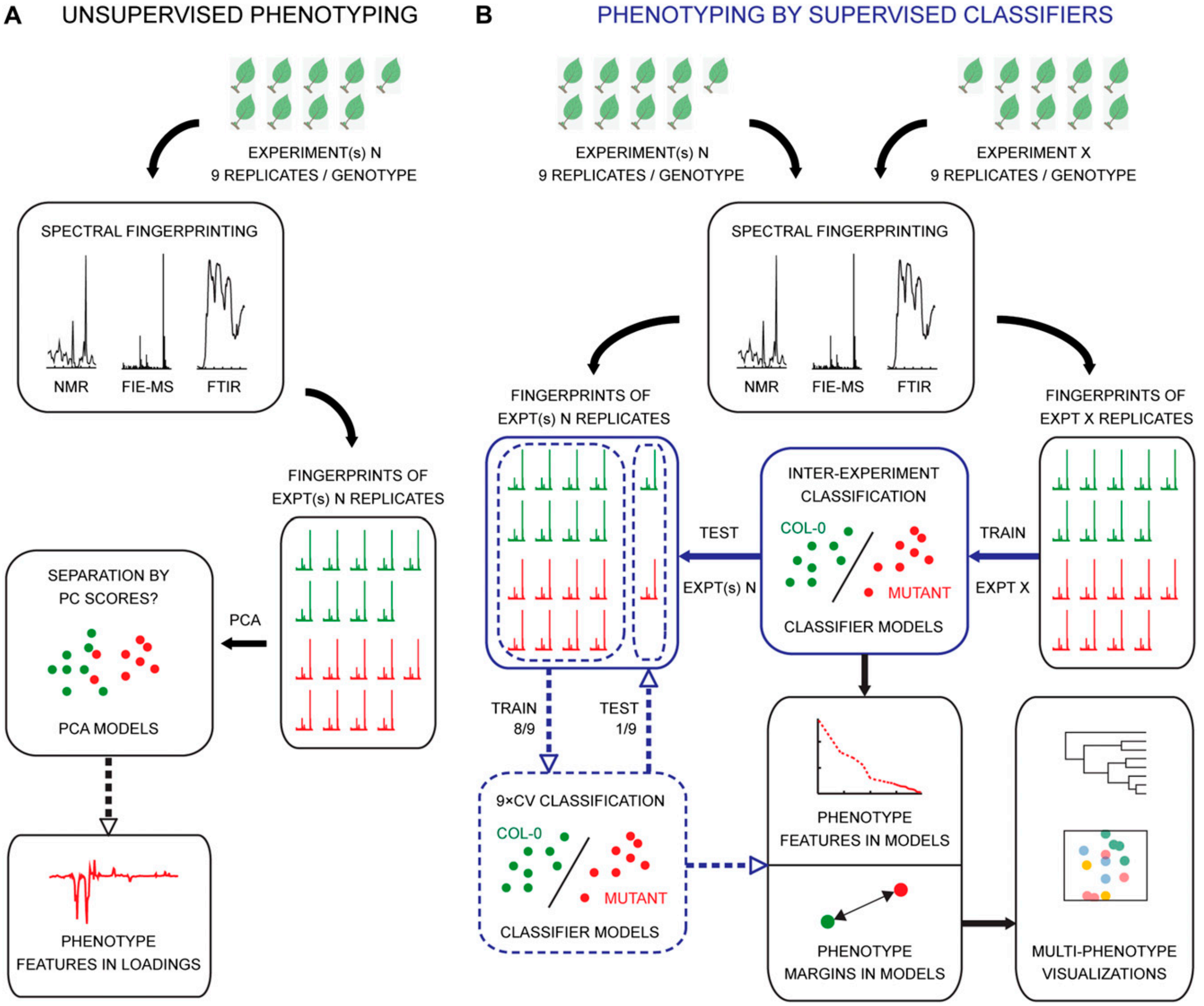

Figure 1. Flow charts for unsupervised phenotyping (A) versus phenotyping by supervised classifiers (B). Blue dashed lines and boxes depict 9×CV, and blue solid lines and boxes depict classification using different experiments for training and testing.

ically more stringent "interexperiment classification," we used a test set that did not comprise plant replicates of those in the training data but rather plants grown in a different experiment. Of these two validation designs, CV is more common due to the paucity of biological data (Tarca et al., 2007).

Intuitive front ends for basic ML are today available in open-source software (Table I). Both validation designs (Fig. 1B) were readily implemented via the Weka GUI (Frank et al., 2004). The main familiarization needed to use Weka is in preparation of input files in a specific format called ARFF, a specimen of which is shown in Supplemental Protocol S1. For supervised classification with PLS-DA, we used SIMCA-P.

It should be stressed that appropriate training/validation relations are crucial for supervised classifiers (Broadhurst and Kell, 2006). The power of the ML classifiers was such that predictions were meaningless unless different samples were used for training and testing. Classifiers were always 100% accurate (n = 114) in predicting genotype classes of the same samples on which they were trained. Even for training data with randomized class labels, accuracies were 99.9%. When these random data classifiers were blind tested on different samples, predictions were, of course, no better than chance. This was an extreme demonstration of "overfitting," whereby fortuitous data noise is used to classify training samples, so that the resultant overoptimistic model is destined to perform poorly when validated on new samples lacking the noise "pattern" (Broadhurst and Kell, 2006).

We next describe how these basic strategies (Fig. 1) were used in different aspects of phenotyping. (1) Discrimination. Could the mutant be distinguished from Col-0? (2) Distance. How different was the mutant from Col-0 and other phenotypes? (3) Consistency. Did the mutant have a characteristic phenotype in different experiments? (4) Features. What metabolomic characteristics of the mutant were distinctive?

Discrimination: Fingerprinting and Modeling Methods Benchmarked

We first evaluated methods for simply discriminating Col-0 from each of 19 mutants that had experimental evidence for primary lesions in metabolism (but without major growth abnormalities). These lesions were in starch (*adg1*, *adg2*, *pgm*, *sex1*), lipid (*ats1*/*act1*, *fad2*, *fad3*, *fad4*, *fad5*, *fad7*, *fae1*, *tag1*), amino acid (*trp1*, *trp5*/*asa1*, *val1*), and ascorbate (*vtc1*, *vtc2*, *vtc3*, *vtc4*) metabolism. (Mutants in this paper are detailed in Supplemental Table S2.) We tallied the numbers of these mutants discriminated by the alternative strategies (Fig. 1) of unsupervised PCA or supervised classifiers (Table II).

2D scatterplots of PC scores showed differential separation from Col-0 of fingerprint data for the various mutants (Fig. 2). To use these ubiquitous plots as a benchmark to compare classifiers, we adopted two metrics. To quantify a simple visual interpretation, we counted the partitioning of samples across the zero axis of the best PC for phenotype separation in pairwise analyses of Col-0 with each mutant. When the significance of these distributions was assessed as the binomial probabilities of their occurrence by random chance, up to 12 mutants (using NMR) were discriminated at $P < 0.01$ (Table II). Mann-Whitney comparisons of scores on the same PC discriminated slightly more (Table II).

In 79% of these cases, optimal separation of Col-0 and mutant occurred on PC[1], which represented a mean 63% (SD 13%) of variance. This confirmed that fingerprint variation was phenotype dominated but also that some variance was neglected if only one PC was used. Therefore, we explored whether incorporation of additional PCs enhanced the phenotyping process. This necessitated a means of identifying the number of PCs to retain for further investigation, an issue for which alternative solutions have been proposed over the years. For this purpose, the SIMCA-P software includes a CV routine (Eastment and Krzanowski, 1982) in which data elements are predicted by PCA models from which they were excluded. The CV is performed using models composed of varying numbers of the PCs. PCs are informative ("significant" in the manufacturer's terminology) if their inclusion in a model reduces predictive error in CV; otherwise, they may be noise. We used this SIMCA-P procedure to identify which PCs to use in phenotyping.

Table II. *Phenotyping performances of data-mining and fingerprinting methods*

NMR, FIE-MS, and FTIR fingerprints of 19 mutants (see text) were each compared with Col-0 by several methods, and the numbers of mutants significantly discriminated in these binary comparisons are shown. PCA scores of the classes (Col-0, mutant) were compared several ways. Score plots counted data on each side of the zero axis of the PC that best partitioned classes. Univariate tests were Mann-Whitney analyses of single PCs on which classes were most significantly different. Multivariate tests compared Mahalanobis distances between multivariate means of class scores on all PCs found informative by CV. Supervised classifiers were tested for Col-0 or mutant class prediction in 9×CV.

Variable	No. of Mutants Discriminated ($P < 0.01$)			
	NMR	FIE-MS	FTIR	Cumulative
Col-0 and mutant PCA scores compared				
Score plots	12	10	7	29
Univariate tests	14	10	7	31
Multivariate tests	14	12	11	37
Supervised binary (Col-0, mutant) classifiers				
PLS-DA	13	12	8	33
RF	15	13	9	37
SVM	17	15	11	43

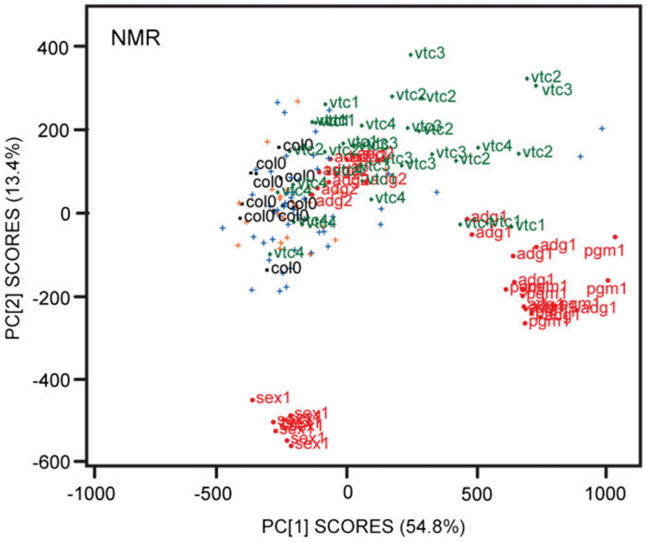

Figure 2. PCA of NMR fingerprints of Col-0 and mutants (*adg1*, *adg2*, *pgm*, *sex1*, *ats1*, *fad2*, *fad3*, *fad7*, *fae1*, *trp1*, *val1*, *vtc1*, *vtc2*, *vtc3*, *vtc4*). Amino acid and lipid mutant labels are omitted for clarity. Variance encompassed by PCs is shown on axes. Metabolic phenotypes are color coded by lesion: starch (red), lipid (blue), amino acid (yellow), ascorbate (green).

The mean numbers of informative PCs found by the SIMCA-P CV were 3.7 (for NMR data), 4.2 (FIE-MS), and 6.8 (FTIR). Phenotyping was then performed by multivariate comparison of Col-0 and mutant scores on all these PCs using a nonparametric significance test in the freeware package PAST (Hammer et al., 2001). This procedure did extend the numbers of mutants discriminated with FIE-MS or FTIR data (Table II) and indeed was competitive with supervised classifiers (Table II). We would again, however, caution about reification (Mahoney and Drineas, 2009): is it biologically plausible that PCA is finding up to seven genuine patterns in the cellular systems analyzed?

Supervised classifiers were evaluated for discrimination of each mutant from Col-0 in 9×CV. (Validation strategies were employed here only for supervised classifiers, as they are not standard practice for PCA.) As depicted in Figure 1B, 9×CV involved predicting the full set of fingerprints from each spectroscopy, over nine rounds of model building on eight of nine of the replicates and testing on the held-out one of nine. The overall accuracies of these predictions were assessed as the binomial probabilities of their chance occurrence. Table II shows the numbers of mutants (out of 19) for which these predictions were significant ($P < 0.01$). PLS-DA did better than analyses of single PCs, but the best classifiers were SVMs. Overall, ML (RF and SVM) matched all other methods. In fact, analysis confined to PC score plots might have overlooked one-third of mutants discriminated by SVM (Table II).

Classifier results are commonly displayed in a "confusion matrix," whose rows are the true classes and columns are those predicted (Tarca et al., 2007). This format is used in Supplemental Figure S1 to show full details of the classifications of each mutant relative to Col-0, for each supervised classifier, and our PC score plot metric, applied to each fingerprinting technique.

A summary comparison of the three spectroscopies, using SVM as the best-performing classifier (Fig. 3), shows that NMR discriminated most mutants, despite using solvent favoring polar metabolites. For FIE-MS, we used solvents for both lipids and polar metabolites, obtaining 100% accuracy for the lipid-related mutants *ats1*, *fad2*, and *fad5*. FTIR of dried whole shoots discriminated *vtc*, amino acid, *fae1*, and strong starch mutants, but fewer than NMR or FIE-MS. Eight mutants were discriminated by all spectroscopies, and two lipid mutants (*fad4*, *tag1*) by none (Fig. 3).

A common device to enhance supervised classification of high-dimensional data is a "filter" algorithm that preselects promising features to reduce dimensionality and hence the danger of overfitting (Tarca et al., 2007). We found, however, that reducing fingerprint dimensionality to 100 to 150 features preselected by Relief (a well-known algorithm in data mining) had no consistent effects on SVM or RF performances. This was not unexpected, as RF in particular was conceived to be resistant to overfitting (Enot et al., 2008).

Discrimination: Fingerprints versus Chromatographic Profiles

The strategic case for fingerprinting was evident from the fact that we analyzed a total of 2,268 plants by FIE-MS, while only 441 were concurrently processed by chromatographic profiling. Therefore, we were interested to know how much extra discrimination was achievable by the more conventional approaches. Using replicate plants grown alongside those taken for fingerprinting, we obtained amino acid, fatty acid, acyl-CoA, and isoprenoid chromatographic profiles for 15 amino acid (*trp1*, *trp5*, *val1*), starch (*adg1*, *adg2*, *pgm*, *sex1*), lipid (*ats1*, *fad2*, *fad3*, *fad4*, *fae1*, *tag1*), and ascorbate (*vtc1*, *vtc3*) mutants. The metabolite data and SVM classification rates are shown in Supplemental Table S3.

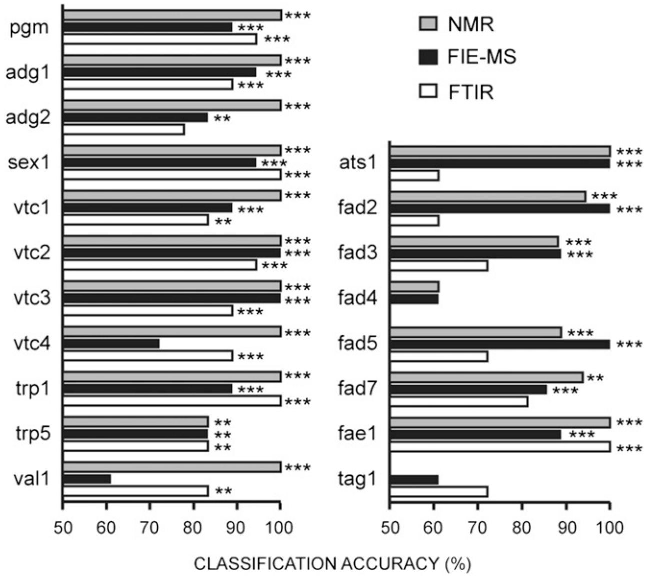

Figure 3. Comparison of NMR, FIE-MS, and FTIR fingerprints for discrimination of mutants from Col-0 by SVMs. Asterisks indicate binomial significance of predictive accuracies (*** $P < 0.001$, ** $P < 0.01$).

SVM classifiers using amino or fatty acid profiles discriminated 13 mutants from Col-0 ($P < 0.01$) compared with 12 or 13 using FIE-MS or NMR fingerprints. All but *fad4* and *tag1* were discriminated by amino acids, all but *trp5* and *vtc1* by fatty acids (Supplemental Table S3). Most notably, the two mutants (*fad4* and *tag1*) for which all fingerprints failed were discriminated by fatty acids. Sensitivity to minor metabolites explained this superiority of fatty acid profiles. Mean quantities in Col-0 of individual fatty acids that were different in mutants (by ANOVA, $P < 0.001$) were 20 to 66 μmol g^{-1} dry weight for *ats1*, *fad2*, *fad3*, or *fae1* but only 0.12 or 2.86 μmol g^{-1} for *tag1* or *fad4*.

SVMs using acyl-CoA profiles discriminated *fad2* and *fad3* perfectly, but only five mutants in total at $P < 0.01$. Using isoprenoids, eight mutants were discriminated, from each metabolic category (Supplemental Table S3). With all four profiles concatenated, SVM discriminated all 15 mutants. PC score plots of the various profiles separated only one to eight mutants.

Distance: Visualizing Relations of Multiple Phenotypes

Qualitative discrimination from the wild type is basic to phenotyping a mutant, but metrics for their mutual distance, and distances from other phenotypes, are needed to visualize relations of multiple phenotypes. Such relations are conventionally seen by ordination in PC score plots. PCA of NMR fingerprints for Col-0 and 15 mutants (Fig. 2), for example, confirmed the proximity of starch-deficient *adg1* and *pgm* mutants and their separation from the starch-accumulating *sex1*. The PCA showed notably high dispersion of replicates of *fad7* and the ascorbate-deficient *vtc* mutants, whose PC[1] scores had SD values 5.6- to 10.6-fold greater than *sex1*.

In contrast to PCA, distance metrics from ML classifiers have yet to find much use in plant metabolomics. The fact that discrimination was better with ML than PCA (Table II), however, suggested that ML-based ordination of multiple phenotypes, as proposed in Figure 1, could be interesting. In fact, the RF concept includes an internal mechanism for estimating interclass distances. RF classifications are based on "voting" by a large number of trees, each built using different subsets of the original variables, in an intricate procedure to avoid overfitting high-dimensional data (Enot et al., 2008). The difference between correct and incorrect predictive votes from all trees in the model is the "margin" between classes.

We used the margins between phenotypes in RF models for multiphenotype ordinations, using a routine proposed by Enot et al. (2006). RF classifiers were built for all pairwise comparisons of the 19 mutants and Col-0. The resultant matrix of RF margins was projected onto a 2D map, in which interphenotype distances were kept as true as possible to the margins, using an algorithm called Sammon mapping (Fig. 4). This routine is not currently available within an integrated software but is provided in Supplemental Protocol S2 as R code with quick-start instructions.

Sammon mapping minimizes distortion using an error function called "stress," which tends to preserve small separations. Stress values for RF margin maps of each fingerprint type were within the acceptable limit of 0.1 (Fig. 4). Correlations between RF margins and map distances for Col-0 versus the mutants were high (r = 0.95 ± 0.01), although the FIE-MS map (Fig. 4B) somewhat understated the distance from Col-0 of *adg1* and *pgm* (Supplemental Fig. S2).

RF Sammon maps of NMR data showed starch mutants far from Col-0, with *adg1* and *pgm* together but separated from *sex1* (Fig. 4A), as in the PCA (Fig. 2). The *vtc* mutants were also far from Col-0, while lipid mutants were generally closer (Fig. 4A). For FIE-MS, certain lipid and *vtc* mutants were farthest from Col-0 (Fig. 4B). For FTIR, strong starch phenotypes were most distant (Fig. 4C).

We looked for other methods to corroborate the RF Sammon map concept. An alternative to 2D plots for visualizing relations is hierarchical cluster analysis (HCA), where entities are linked stepwise by relative proximity. HCA using RF margins as distance metrics produced clusters (Fig. 5A) transcribable to the Sammon map (Fig. 4B). Comparable clusters were also found by an established HCA method (Enot et al., 2008) using PC-linear discriminant analysis (LDA), another supervised method, which maximizes class separation by linear combinations of PCs (Fig. 5B).

Consistency: Phenotypes in Different Experiments

A biologically important aspect of a phenotype is its consistency. To test this, classifiers, trained on one data set to recognize a pattern in another are conceptually ideal. We examined mutant phenotype consistency across different experiments, encompassing plants grown in the same chamber at different times or in two chambers with different light levels. FIE-MS was used for its sensitivity, versatility, and high throughput.

First, we consider a strong metabolic phenotype grown in several experiments: the plastid glycerolipid mutant *ats1* (Xu et al., 2006), earlier called *act1*. In PCA of 122 Col-0 and *ats1* FIE-MS fingerprints from seven experiments, most variance (58%) was in phenotype separation on PC[1]. The different experiments separated only on various lesser PCs. Thus, by PCA criteria, interexperiment variation in this strong phenotype was minor.

ML classifiers, on the other hand, easily recognized the experiments whence data derived. RFs trained on the 122 Col-0 and *ats1* fingerprints (labeled by experiment, not genotype) identified all seven experiments with 99.2% accuracy in 9×CV. In pairwise comparisons of experiments, moreover, RF margins were very high, irrespective of the phenotypes. Among 111 RF experiment classifiers, mean margins were 0.88 ± 0.10 (SD; perfect discrimination would be 1.0). We explored whether the interexperiment variation was innate to the plant material or due to postharvest analytical processing. When samples from two growth experiments were processed in random order, margins were reduced by about one-third, but the experiments were still discriminated in 9×CV.

These incidental findings on the ability of ML to distinguish sample origins were noteworthy, but our primary classifier hypothesis was that a mutant phenotype should be recognized in plants from different experiments. As alternative RF classifiers could be trained on the same FIE-MS data with either "experiment" or "genotype" class labels, we obtained both interexperiment and interphenotype margins. (Such analyses can be performed with the R code and instructions in Supplemental Protocol S2.) We evaluated mutants across different experiments by the ratio of (1) the "phenotype margin" from Col-0 to (2) the mean "interexperiment margin." These ratios were almost always less than 1.0 (Fig. 6, x axis).

Our other phenotype consistency measure was discrimination by RFs presented with fingerprints from more than one experiment. (We used Weka for this.) Fine resolution of phenotype consistencies for 35 mutants in two to seven experiments was obtained by plots of classification accuracies against the phenotype-experiment margins ratio (Fig. 6). These two measures had high, but nonlinear, correlation. High classification accuracies had a disproportionate margin range (Fig. 6), confirming that RF margins represent class boundaries better than classification

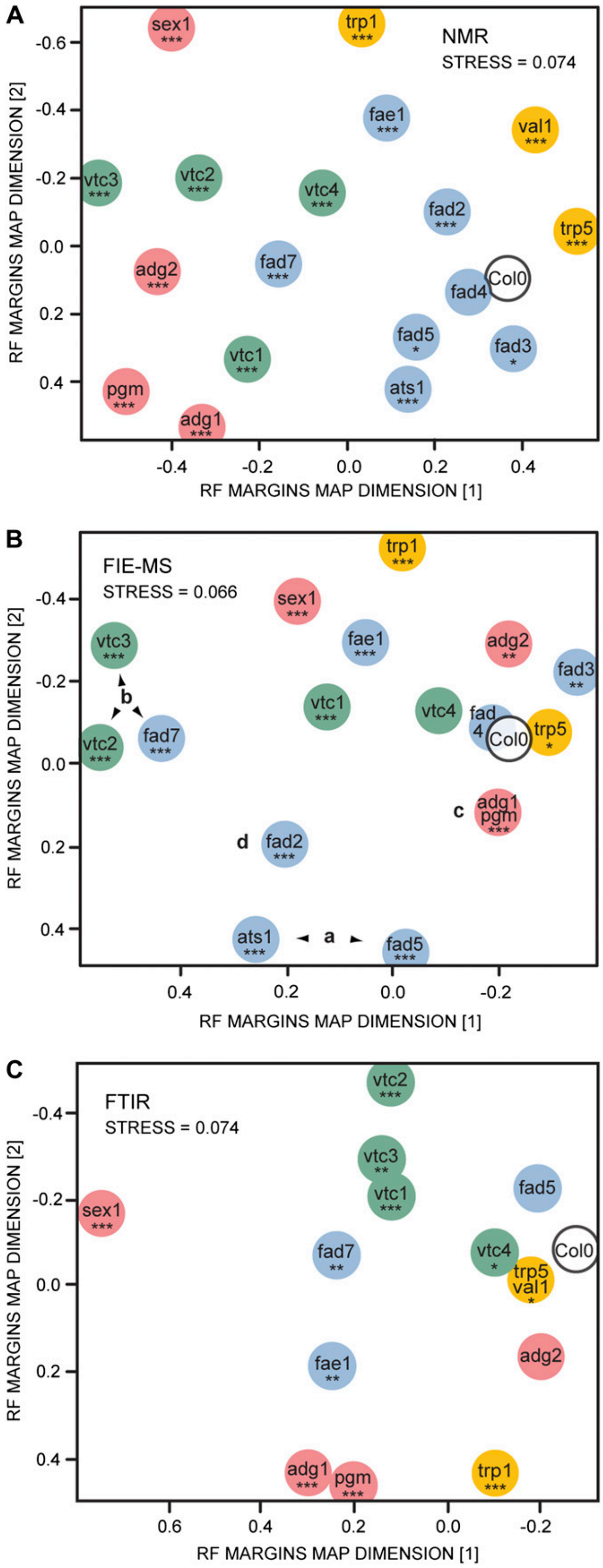

Figure 4. 2D ordinations of metabolic phenotypes as Sammon maps of RF margins between all pairs of Col-0 and mutant fingerprints. Metabolic phenotypes are color coded by lesion: starch (red), lipid (blue), amino acid (yellow), ascorbate (green). A, NMR fingerprints. Not separated from Col-0: *tag1*. B, FIE-MS fingerprints. Not separated: *tag1*, *val1*. Arrowheads and letters refer to clusters in Figure 5. C, FTIR fingerprints. Not separated: *fad2*, *fad3*, *fad4*, *tag1*. Maps approximate RF margins with stress error (for glossary, see Supplemental Table S1) on plots. For correlations between map distances and RF margins, see Supplemental Figure S2. Asterisks indicate significance of 9×CV discrimination of each mutant from Col-0 by RF in Weka (*** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$).

accuracies alone (Enot et al., 2006). However, statistical significance was more easily defined for classification accuracies.

In applying classifiers to phenotypes from more than one experiment, two different validations were used, following the schemes in Figure 1. Figure 6 plots accuracies for "pooled-experiment" classifiers, where fingerprints from all relevant experiments were used in 9×CV. Significant accuracy in this validation meant that the classifier could identify a phenotype for the mutant in more than one experiment. In the 9×CV process, the classifier was trained on replicates from all experiments and then tested on other replicates (Fig. 1). In addition, "single-experiment" RF classifiers were built on one experiment and then tested on others. This was a more stringent test of the phenotype, because the training set did not include any replicates from the test experiment(s). Asterisks in Figure 6 highlight mutants whose single-experiment classifiers were all significant.

An impressive case was *ats1*, identified by classifiers trained on any of seven experiments, in a mean 96.3% of 104 or more tests on the six held-out ones. Other generalizable phenotypes were *fad2*, *fad5*, *adg1*, *pgm*, and the sinapate ester mutant *sng1*. On the other hand, these procedures revealed another type of phenotype, for mutants such as *uge4* and *vtc3* (affected in development or stress physiology), for which pooled-experiment models had significant accuracies but single-experiment ones were poorly generalizable (Fig. 6).

The versatility of FIE-MS allowed us to explore how the apparent phenotype depended on extraction solvents. Alternative FIE-MS fingerprints were produced with the aqueous extraction used for NMR (Fig. 6, white symbols). This improved generalizability for some mutants (e.g. *fae1*) but not others (e.g. *uge4*).

These routines were extended to reverse-genetics screens of putative metabolism genes from the d*Spm* single-copy (SM) collection of insertion mutants (Tissier et al., 1999). RFs found few SM phenotypes rivaling well-known ones from classical phenotype screens, but SM21150 (At1g07720), a putative ketoacyl-CoA synthase, gave fairly strong, generalizable RF models (Fig. 6). Its distinctness was probably not due only to ketoacyl-CoA synthase products, as too many (8%) of its FIE-MS peaks differed from Col-0 (ANOVA, $P <$

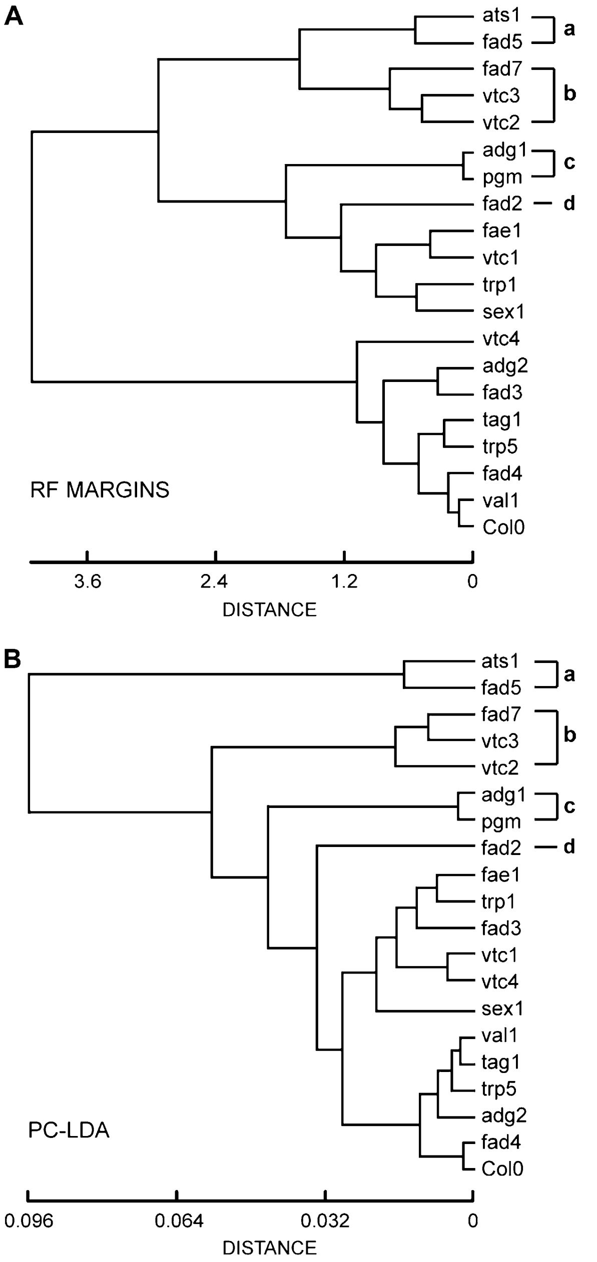

Figure 5. Metabolomic relations of Col-0 and 19 mutants by hierarchical cluster analysis. Phenotypes were compared using FIE-MS fingerprints in clustering by RF margins (A) and 13 discriminant functions from PC-LDA (B).

0.05), and the best discrimination came from aqueous extracts. At1g07720 is highly up-regulated in elongating epidermis (Suh et al., 2005), so abnormal waxes may have affected physiology. This phenotype was confirmed by profiling: ML classifiers discriminated it using amino acid, fatty acid, acyl-CoA, or isoprenoid profiles.

Another ketoacyl-CoA synthase mutant was the SM19881 insertion in *FAE1*. It had a weaker phenotype than the classical *fae1* mutant (mean RF margin from Col-0 of SM19881 was 0.33 of *fae1*), but its aqueous fingerprints were discriminated by pooled-experiment classifiers (Fig. 6). Boyes et al. (2001) found *fae1* altered in development, but this mutant is enigmatic, as vegetative expression has not been detected (Suh et al., 2005).

Another quite distinctive case (Fig. 6) was SM15231 (At4g31970), of the CYP82 cytochrome P450 family, which has stress-responsive members of uncertain functions (Nelson et al., 2004). Many peaks (19.9%) in its aqueous fingerprints differed from Col-0 (ANOVA, $P < 0.05$), perhaps due to a stress-related phenotype.

We also used NMR on five UDP-Glc 4-epimerase (*UGE*) mutants affected in wall synthesis, whose metabolomes have been little studied. ML discriminated *uge4*/*rhd1* ($P < 0.001$), which has defective root hairs, and *uge2* less strongly ($P < 0.05$), but it failed with *uge1*, *uge3*, and *uge5*. This gradation in metabolic phenotypes reflected the *UGE* genes' influences on growth and wall Gal content (Rösti et al., 2007).

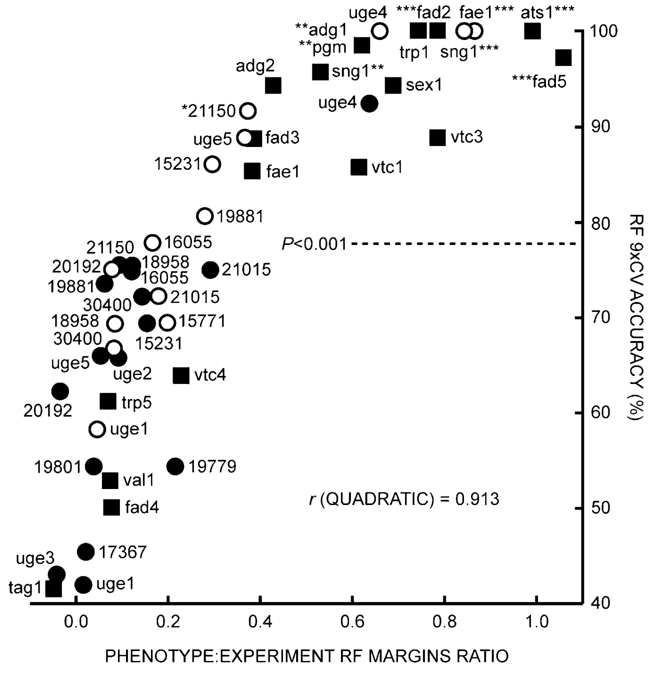

Figure 6. Interexperiment generalizability of RF models of 35 mutant phenotypes. 9×CV accuracies of discrimination from Col-0 by RFs, trained on FIE-MS data pooled from two to seven experiments, are plotted against ratios of between-phenotype margins to mean between-experiment margins. Squares indicate that data encompass different growth conditions. Extractions are shown in white (aqueous) and black (propan-2-ol:methanol:water) symbols. The dotted line shows the 9×CV significance threshold ($P < 0.001$). Correlation (*r*) of accuracies with quadratic function of margins ratios is shown. Asterisks indicate significance for classifiers trained on one experiment and tested on others (*** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$). Five-digit labels are SM mutant identifiers (Supplemental Table S2).

Features: Ranking Distinctive Metabolome Components

Describing rather than demonstrating a phenotype requires insights into its features. While fingerprinting is for speed rather than information richness and precision, its metabolomic insights may still be valuable. Moreover, acceptance of ML in biology would be helped by demonstration of how it uses data features.

Feature utilization by RFs is accessible as "importance scores," obtained from misclassification rates when each variable is randomized. This enabled us to compare RF models with well-known statistical methods like ANOVA and other data-mining measures (more information on which is found in Supplemental Table S1). We used the R package FIEmspro (Enot et al., 2008) for most of these measures, including the RF importance scores.

As a mining exercise for data features, we explored phenotype clusters (a, b, c, d) identifiable in the RF Sammon map, HCA, and PC-LDA (Figs. 4B and 5). These were interesting, as among predictable clusters was a less obvious association of *fad7* with *vtc2* and *vtc3* (cluster b).

Although these clusters emerged in multivariate analyses, we evaluated univariate as well as multivariate measures of their distinctive metabolomic features. The main univariate tool for high-dimensional data is ANOVA, although its parametric assumptions are often ignored (Jafari and Azuaje, 2006). Therefore, we took *F* values (in the Welch test ANOVA, which does not assume equal variances) as a benchmark for feature selection. Very high rank correlations with ANOVA were obtained using the nonparametric Kruskal-Wallis test and the area under the (receiver operating characteristic) curve (AUC) measure of true- and false-positive rates (Table III). Mutual information, a univariate information theory measure, also correlated highly with ANOVA (Table III).

We further compared ANOVA with three multivariate methods, which should be appropriate for interdependent metabolomic features. RF importance scores and Relief correlated highly with ANOVA, although somewhat less than the univariate measures (Table III). Thus, the functioning of the RF algorithm could be validated in relation to traditional statistics. Similarities between *F* values and RF importance scores for spectral peaks were evident, as were the distinctive features of each phenotype cluster (Supplemental Fig. S3).

PCs loadings gave different feature rankings than other methods (Table III). This reflected the distinct operation of PCA. Single PCs generally encompass only part of the data variance. PCA also differs from the other methods in weighting variables for contribution to data variance. This means that (without data scaling) abundant metabolites can be more important than minor ones even if the latter show greater proportional differences.

These alternative perspectives are both useful to understand metabolomes. For interpretation of NMR fingerprints, in fact, we found PCA particularly amenable. PC[1] loadings clearly reflected NMR spectral peaks (Supplemental Fig. S4) and corresponded closely to difference spectra between phenotypes, confirming their biological relevance.

Phenotypes: Metabolomic Insights

One vision for fingerprinting is in "hierarchical metabolomics" (Catchpole et al., 2005), where rapid preliminary identification of distinct phenotypes could guide authoritative profiling. We found that the metabolomic inclusiveness of fingerprints also gave some striking global perspectives. One was the broad consequences of single primary lesions in metabolism. While alternative methods detected metabolic phenotypes differently, a given method was sometimes more informative than expected for a particular class of mutant.

Another perspective was the proportions of metabolites differing in phenotypes. Figure 7 is a global ranking of *F* values of FIE-MS features for the phenotype clusters from Figure 5. We show *F* values, from among several correlated measures (Table III), for their wide acceptance (Jafari and Azuaje, 2006). *F* values for most clusters were skewed to a few high values, less than 10% of peaks being different (at $P < 0.05$) in intensity to Col-0 (Fig. 7).

Cluster b (*fad7*, *vtc2*, *vtc3*) FIE-MS fingerprints had atypically large numbers (24%) of significantly differ-

Table III. *Feature selection tools for data mining in plant metabolomics*

Comparison of selection methods for features of FIE-MS fingerprints that discriminated Col-0 from mutant clusters in Figure 5. Correlations are shown between *F* values, from univariate ANOVAs of each spectral variable, and diverse measures, including loadings on the main PC (representing 56%–72% of data variance) separating each cluster from Col-0.

Mutant Cluster	Rank Correlation (r_s) with ANOVA					
	Univariate Measures			Multivariate Measures		
	Kruskal-Wallis	AUC	Mutual Information	RF	Relief	PC Loadings
a	0.92	0.92	0.80	0.78	0.66	0.61
b	0.97	0.97	0.92	0.87	0.83	0.35
c	0.98	0.98	0.98	0.95	0.93	0.67
d	0.95	0.95	0.87	0.82	0.93	0.09
Mean ± SD	0.96 ± 0.03	0.96 ± 0.03	0.89 ± 0.08	0.85 ± 0.07	0.84 ± 0.13	0.43 ± 0.27

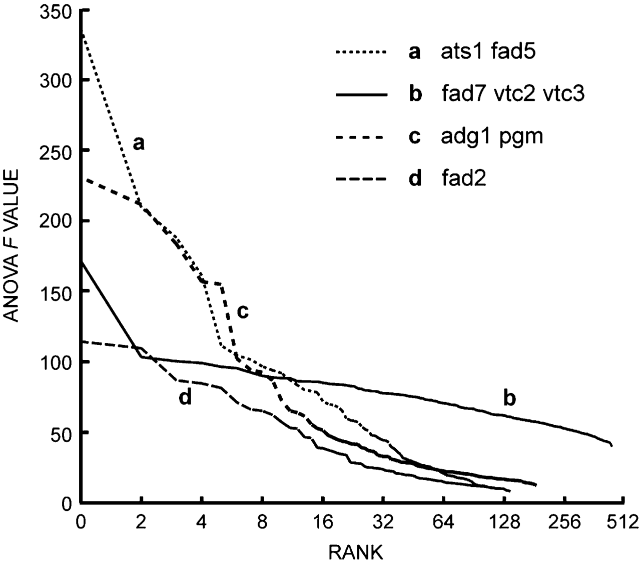

Figure 7. Data mining of FIE-MS fingerprints to compare metabolic phenotypes. Ranked *F* values ($P < 0.05$) are from ANOVAs of features (normalized intensity at each *m/z*) of Col-0 with mutant clusters in Figure 5.

ent peaks (Fig. 7). These mutants are all implicated in stress responses (Kachroo et al., 2005; Conklin et al., 2006), so their extensively modified metabolism may have reflected aberrant physiology. Possibly consistent with altered cell signaling were Suc and Pro loadings in *vtc* mutant NMR fingerprints (Supplemental Fig. S4). These metabolites are influenced by abscisic acid (Verslues and Bray, 2006), which is up-regulated in *vtc1* (Pastori et al., 2003). Amino acid profiles confirmed that Pro was elevated 4-fold in *vtc3* (Supplemental Table S3).

For cluster a (*ats1* and *fad5*), the highest *F* values were for plastid glycerolipids, such as the monogalactosyldiacylglycerols (MGDGs) 18:3-18:3-MGDG $[M+K]^+$ (F = 83.5) and 18:3-16:3-MGDG $[M+K]^+$ (F = 38.8). Galactolipids with 16-carbon fatty acids form in plastids, those with only 18-carbon fatty acids form partly in the endoplasmic reticulum (Mekhedov et al., 2000). Mutants affected in these compartments had distinctive galactolipid spectra (Supplemental Fig. S5). The main positive ion in Col-0 spectra was 18:3-16:3-MGDG (Supplemental Fig. S5). In spectra of *ats1*, where plastid glycerolipid biosynthesis is disrupted at the first reaction, 18:3-18:3-MGDG was the main positive ion. The *fad5* mutant, deficient in plastid desaturation of MGDG 16:0, was distinguished by an 18:3-16:0-MGDG peak. In *fad2*, defective in microsomal 18:1 desaturation, 18:3-18:3-MGDG was diminished (Supplemental Fig. S5).

Galactolipids contributed up to 21% of the FIE-MS positive-ion current, explaining the utility of these fingerprints for lipid mutants. Therefore, the fact that FIE-MS also discriminated so many nonlipid mutants (Fig. 3), as did fatty acid profiles, indicated the pervasiveness of processes where lipids function (Beisson et al., 2003). The same applied to the converse situation, where NMR discriminated mutants such as *ats1* and *fad2*, despite using solvents unsuited to metabolites directly affected by these lesions. The spectra confirmed that NMR was detecting effects on the polar metabolome.

For cluster c (*adg1* and *pgm*), the top *F* values (Fig. 7) in FIE-MS were for Suc ions, such as $[M-H]^-$ (F = 231). Sugars also dominated NMR PC loadings for these starch-deficient mutants and the starch-excess *sex1* mutant (Supplemental Fig. S4). In the diurnal light period (when we sampled), both phenotypes have high soluble sugars due to lack of conversion to starch (Caspar et al., 1991; Gibon et al., 2006). FTIR fingerprints of unfractionated shoots were particularly distinct for these mutants, with differences from Col-0 at 1,200 to 800 cm^{-1} (Supplemental Fig. S6). This region included vibrational frequencies of polysaccharide ring and glycosidic bonds, interpretable as starch and wall absorbances. PC loadings, or ANOVA *F* values, from comparisons of *sex1* and *adg1* (or *pgm*) FTIR fingerprints with Col-0 resembled pure starch spectra.

Malate, which has pivotal metabolic roles and is an important form of fixed carbon in Arabidopsis (Fahnenstich et al., 2007), was prominent in the NMR PC loadings of diverse mutants, particularly *val1* and *uge4* (Supplemental Fig. S4). Val, which accumulated in *val1* (Supplemental Fig. S4; Supplemental Table S3), is made from pyruvate, for which malate is a precursor (Fahnenstich et al., 2007). Malate is also important in nitrogen assimilation (Stitt et al., 2002), which could be affected in *val1*, and also in *uge4*, in view of the latter's altered root morphology.

Extensive phenotypic variation in amino acids was also seen in NMR PC loadings (Supplemental Fig. S4) as well as in amino acid profiles (Supplemental Table S3). The metabolic centrality of amino acids (Stitt et al., 2002) was confirmed by the ability of amino acid profiles to discriminate not only amino acid (*trp1*, *trp5*, *val1*) but starch (*adg1*, *adg2*, *pgm*, *sex1*), lipid (*ats1*, *fad2*, *fad3*, *fae1*), and *vtc* (*vtc1*, *vtc3*) mutants (Supplemental Table S3).

Evaluation: Phenotyping by Supervised ML Classifiers

This benchmarking study identified several merits of supervised ML classifiers for phenotyping. Their power of discrimination exceeded standard PCA, suggesting that studies confined to PCA may miss opportunities. The familiarity and vividness of PCA mean that it is perceived as simple, but this is deceptive unless the primary PCs capture most data variation. For example, only 26% of the correlations between 81 variables are explained by the 2D PC score plots in an impressive recent mutant-screening paper (Lu et al., 2008). Conversely, supervised classifiers, despite reputed complexity, proved straightforward for interrogation of hypothetical phenotypes. Examples were our routines of graded stringency for phenotype consistency in different experiments (Fig. 1).

Open-source ML algorithms were more powerful than another major supervised approach, PLS-DA. Supervised classification is developing, and there are further methodologies. Bayesian modeling was recently used in phenotyping starch mutants, for example, although the programs are not released (Messerli et al., 2007; Davison, 2008). We have emphasized public software, which is likely to be the future trend in the interests of transparency, reproducibility, and methodological efficiency (Gentleman et al., 2004).

ML classifiers are sometimes seen as "black boxes" (Davison, 2008), but importance scores for data features in RF models were accessible (in R, but not in Weka, at the time of this study). Indeed, RF has been advocated for feature ranking (Enot et al., 2008). Importance scores confirmed that feature utilization by RFs was consistent with other measures, such as ANOVA. This differed from loadings of PCA (on unscaled data), where features were weighted by the proportion of overall data variance they represented. In consequence, RF and PCA can provide complementary perspectives, particularly for comparing multiple phenotypes. To be attractive for this, RF needs a visualization format to rival PC score plots, and the 2D mapping of RF margins has potential.

CONCLUSION

The potential of metabolite fingerprinting will be enhanced by new data-modeling methods for high-dimensional data (including, but not limited to, ML). We have tried to show that ML methods are further suited to the biological concept of phenotyping and do not need to be technically inaccessible black boxes. We believe the open-source resources emerging for metabolomics will vindicate this vision in the not-too-distant future.

MATERIALS AND METHODS

Plant Material

Supplemental Table S2 details all mutants of Arabidopsis (*Arabidopsis thaliana*) Col-0. SM single-copy transposon-insertion lines were selected via a TIGR.5 genome annotation in Excel, with insertions hyperlinked to the ATIDB database (Pan et al., 2003). ATIDB entries were matched to MAPMAN for Gene Ontology Consortium categorization (Thimm et al., 2004). SM lines used were homozygosity tested (Tissier et al., 1999).

Plants were grown in 7-cm pots of Levington M2 compost with Intercept insecticide (Scotts) in nine random blocks in one environment of 23°C/18°C, 16-/8-h day/night photoperiods of 250 to 270 μmol m^{-2} s^{-1}, and 70% relative humidity. Where mentioned, other conditions were 23°C, 16/8 h of 100 to 150 μmol m^{-2} s^{-1}, and 60% relative humidity. Aerial tissues from stage 6.00 plants (Boyes et al., 2001) were harvested into liquid N_2 in mid light period, freeze dried, and powdered. Replicate plants from each block were allocated to each analytical method. Shipment and laboratory processing entailed a few days at ambient temperature; otherwise, storage was at -80°C. Metadata were recorded in software compliant with the ArMet plant metabolomics data model (Jenkins et al., 2005).

NMR Fingerprinting

Samples (15 mg) were extracted (50°C, 10 min) in 1 mL of 80:20 ^2H_2O: $C^2H_3O^2H$ with 0.05% (w/v) [$^2H_4$]TSP (for sodium trimethylsilylpropionate). After cooling and centrifugation, supernatants (850 μL) were reheated (90°C, 2 min), refrigerated 45 min, and recentrifuged. ^{1}H-NMR spectra of supernatants (750 μL) were acquired at 300 K on a Bruker Biospin Avance at 600 MHz with a 5-mm inverse probe. A water-suppression pulse sequence with 5-s relaxation delay was used. Spectra were acquired in 128 scans of width 7,310 Hz and Fourier transformed with an exponential window (0.5-Hz line broadening). Chemical shifts were referenced to d_4-TSP (δ0.0). Spectra were binned to 0.01 ppm, and intensities were scaled to d_4-TSP (δ0.05 to -0.05). Signals removed were residual water (δ4.865–4.775), d_4-methanol (δ3.335–3.285), d_4-TSP, and fumarate (δ6.525–6.515), which showed particular diurnal fluctuation. Analytical replicates (three) were averaged. Final fingerprints had 901 bins ($\geq$669 nonzero).

FIE-MS Fingerprinting

For speed, we used one-tube extractions with no solvent partitions. Five milligrams was extracted in 0.5 mL of propan-2-ol (4°C, 1 h), and 0.5 mL of 80:20 methanol:water was added for another 1 h at 4°C. (Including propan-2-ol yielded 29% more ions of above-average intensity at mass-to-charge ratio [m/z] > 750 than methanol:water alone.) RF margins were superior ($P < 0.05$, Wilcoxon tests) or similar to chloroform:methanol:water (1:2.5:1) extracts. Where stated, 20:80 methanol:water was used as for NMR.

Samples, diluted 1:1 with 60:40 methanol:water (salts were avoided to aid throughput), were loaded in the autosampler of a Waters Alliance 2690 LC system delivering this solvent at 0.5 mL min^{-1} (with no LC column), and 10 μL was introduced by split flow of 75 μL min^{-1} to the Z-spray source of a Waters Micromass LCT MS device. Source and desolvation temperatures were 120°C and 250°C; capillary was at 3 kV; sample and extraction cones were at 30 and 5 V; nebulizer and desolvation N_2 gas flows were 70 and 470 L h^{-1}. Spectra (m/z 65–990) scanned in 1-s cycles for 2 min were binned to unit m/z and normalized to total ion current infused. Concatenated positive- and negative-ion spectra had 1,852 variables (all nonzero).

Tandem MS was done on a Waters Micromass nanospray Q-Tof apparatus with 0.8-kV capillary voltage, cone voltages as above, and argon collision gas at 3.1×10^{-5} mbar.

FTIR Fingerprinting

Samples (5 mg) were mixed with 200 μL of water, and 5 μL was slurry loaded on duplicate 400-well aluminum plates. These were dried (50°C, 45 min) and loaded on a motorized stage of a reflectance thin-layer chromatography accessory of an IFS28 FTIR spectrometer with an MCT detector (Bruker Optics). Absorbance spectra were recorded over 4,000 to 600 cm^{-1} at 4 cm^{-1} resolution, and 256 were averaged per sample. Averaged duplicate plate spectra (1,764 variables, all nonzero) were normalized to zero mean and unit SD.

Chromatographic Profiling

Twenty-six amino acids were measured, with norleucine internal standard, on a Thermo LCQ Classic LC-MS device (Thermo Scientific). Samples (2 mg) extracted in 700 μL of 80:20 ethanol:water (4°C, 30 min) were analyzed as isobutyl chloroformate derivatives (Husek, 1998) on a 100-mm (3 mm i.d.) porous graphitic column (5 μm Hypercarb; Thermo Scientific) at 0.4 mL min^{-1} with a 15-min gradient of 100% solvent A (10 mM ammonium trifluoroacetate, 10 mM trifluoroacetic acid in 50:50 ethanol:water) to 100% B (10 mM trifluoroacetic acid in tetrahydrofuran). Amino acids were measured by positive-ion atmospheric pressure chemical ionization-tandem MS, with capillary at 4 V and 150°C, vaporizer at 550°C, and discharge current of 6 μA.

Nineteen acyl-CoAs were measured in 5-mg samples, as etheno derivatives by LC fluorescence, and 38 fatty acids were measured by GC after methylation of lipids from this protocol (Larson and Graham, 2001).

Twenty-nine isoprenoids (carotenes, xanthophylls, tocopherols, ubiquinones, chlorophylls) were analyzed by modifications of Fraser et al. (2000), with α-tocopherol internal standard. Samples (5 mg) on ice were extracted in 200 μL of methanol (5 min), 200 μL of 50 mM Tris-HCl (pH 7.5) was added (10 min), followed by 800 μL of chloroform (10 min). Dried chloroform layers were analyzed on a 250-mm (4.6 mm i.d.) C30 column (5 μm YMC30; YMC), at 1 mL min^{-1}, with 0.2% formic acid/1 mM ammonium formate in methanol or methyl tertiary-butyl ether (solvent B). Solvent B program was as follows: 0%, 6 min; 15%, 5 min; to 90% in 30 min; 90%, 5 min. Isoprenoids were quantified by positive-ion atmospheric pressure chemical ionization-MS, with capillary at 15 V and 150°C, vaporizer at 500°C, and discharge current of 5 μA.

PLS-DA and PCA

PLS-DA and PCA were done in SIMCA-P version 11.0 (Umetrics) on mean-centered unscaled data. PLS-DA models were built for eight of nine replicates, labeled by genotype class (Col-0 and a mutant), and tested in class prediction of the held-out one of nine (in which the two classes were equally represented). Test data were excluded from all model-building stages (Westerhuis et al., 2008). This process, repeated for all nine subsets in turn to get overall classification accuracies, is termed 9×CV. *P* values were binomial probabilities of classifications by chance.

As a comparable metric for PCA, we used partition of replicates across the zero axis of the PC best separating Col-0 and mutant data. Scores on single PCs were also compared by Mann-Whitney tests. Furthermore, Mahalanobis distances (Tarca et al., 2007) between multivariate means of scores for all PCs found to be informative by CV (Eastment and Krzanowski, 1982) were compared using a permutation test in PAST version 1.66 (Hammer et al., 2001). This nonparametric test was used as multivariate normality (Mardia test), and equivalent covariances (Box's *M* test) were often not confirmed. For LDA of PC scores, the R package FIEmspro was used (Enot et al., 2008). HCA (by Ward's method) of discriminant functions from PC-LDA was done in PAST.

ML Classifiers

ML classifications used Weka version 3.4.5 (Frank et al., 2004). A specimen input file in the ARFF format for Weka is shown in Supplemental Protocol S1. SMO (sequential minimal optimization) classifiers with linear kernels (and default parameters) were used for the SVM results shown. We also tested polynomial (exponents 2, 3, 4) and radial basis function (with parameters optimized for each spectroscopy) kernels, but these were poorer or not significantly different from linear kernels, which are often best for high-dimensional data with few samples (Ben-Hur et al., 2008). Weka RF classifers had 2,000 trees and nodes split on $\sqrt{m}$ of the *m* variables. Accuracies were from 9×CV or a defined test set. *P* values were binomial probabilities.

RF margins and Sammon maps were obtained in the R randomForest and MASS packages: R code is shown in Supplemental Protocol S2. HCA (by Ward's method) of RF margins was done in PAST. RF margins were estimated on complete data sets (i.e. without CV or test sets).

Feature Selection

Loadings of the PC that best separated classes were obtained in SIMCA-P. Feature selection used the Weka ReliefF attribute evaluator (Frank et al., 2004) and FIEmspro for all other measures (Enot et al., 2008). RF importance scores were derived for 100 10-fold replicated bootstrap data sets to maximize consistency of the heuristic algorithm. Mutual information was obtained with Shannon entropies estimated using binned kernel densities. Univariate measures were ANOVA (Welch *F* test, with no assumption of equal variances), AUC of receiver operating characteristic curves, and Kruskal-Wallis tests. Position *P* values (Zhang et al., 2006) were found in 100 10-fold replicated bootstrap sets; these proved more conservative than conventional parametric estimates.

Other Statistics

Correlation, Kruskal-Wallis, Mann-Whitney, Wilcoxon, Mardia, and Box's *M* tests on small data sets used PAST.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure S1. Confusion matrices of actual and predicted phenotype classes, based on NMR, FIE-MS, or FTIR fingerprints.

Supplemental Figure S2. Correlations of Sammon map distances with true RF margins.

Supplemental Figure S3. Data mining by ANOVA and RF for distinctive features of mutant metabolic phenotypes.

Supplemental Figure S4. Metabolomic information from PCA of NMR fingerprints.

Supplemental Figure S5. Major spectral peaks attributable to plastid glycerolipids in FIE-MS fingerprints.

Supplemental Figure S6. FTIR fingerprints of homogenized whole shoots.

Supplemental Table S1. Glossary of data analysis techniques used.

Supplemental Table S2. Details of mutants analyzed.

Supplemental Table S3. Mean contents of metabolites by targeted profiling.

Supplemental Protocol S1. Weka ARFF file specimen, containing Col-0 and *ats1* FIE-MS data from seven experiments.

Supplemental Protocol S2. R code for RF margins between multivariate classes (1) and projection as a 2D Sammon map (2).

Supplemental Data S1. NMR fingerprint data used in Figure 4A.

Supplemental Data S2. FIE-MS fingerprint data used in Figure 4B.

Supplemental Data S3. FTIR fingerprint data used in Figure 4C.

ACKNOWLEDGMENTS

For technical assistance or advice, we thank C. Atkinson, M. Beckmann, P. Bull, S. Cang, D. Enot, R. Goodacre, N. Hardy, J. Heald, L. Hill, M. Hills, H. Jenkins, D. Kell, C. Lu, J. McCance, D. Overy, and C. Sumner. We also thank all who supplied seeds (Supplemental Table S2).

Received November 5, 2009; accepted June 19, 2010; published June 21, 2010.

LITERATURE CITED

Arita M (2004) Computational resources for metabolomics. Brief Funct Genomics Proteomics **3:** 84–93

Beckmann M, Parker D, Enot DP, Duval E, Draper J (2008) High-throughput, nontargeted metabolite fingerprinting using nominal mass flow injection electrospray mass spectrometry. Nat Protoc **3:** 486–504

Beisson F, Koo AJK, Ruuska S, Schwender J, Pollard M, Thelen JJ, Paddock T, Salas JJ, Savage L, Milcamps A, et al (2003) Arabidopsis genes involved in acyl lipid metabolism: a 2003 census of the candidates, a study of the distribution of expressed sequence tags in organs, and a Web-based database. Plant Physiol **132:** 681–697

Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G (2008) Support vector machines and kernels for computational biology. PLoS Comput Biol **4:** e1000173

Boyes DC, Zayed AM, Ascenzi R, McCaskill AJ, Hoffman NE, Davis KR, Gorlach J (2001) Growth stage-based phenotypic analysis of *Arabidopsis*: a model for high throughput functional genomics in plants. Plant Cell **13:** 1499–1510

Broadhurst DI, Kell DB (2006) Statistical strategies for avoiding false discoveries in metabolomics and related experiments. Metabolomics **2:** 171–196

Bryan K, Brennan L, Cunningham P (2008) MetaFIND: a feature analysis tool for metabolomics data. BMC Bioinformatics **9:** 470

Caspar T, Lin TP, Kakefuda G, Benbow L, Preiss J, Somerville C (1991) Mutants of Arabidopsis with altered regulation of starch degradation. Plant Physiol **95:** 1181–1188

Catchpole GS, Beckmann M, Enot DP, Mondhe M, Zywicki B, Taylor J, Hardy N, Smith A, King RD, Kell DB, et al (2005) Hierarchical metabolomics demonstrates substantial compositional similarity between genetically modified and conventional potato crops. Proc Natl Acad Sci USA **102:** 14458–14462

Clarke R, Ressom HW, Wang A, Xuan J, Liu MC, Gehan EA, Wang Y (2008) The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nat Rev Cancer **8:** 37–49

Conklin PL, Gatzek S, Wheeler GL, Dowdle J, Raymond MJ, Rolinski S, Isupov M, Littlechild JA, Smirnoff N (2006) *Arabidopsis thaliana VTC4* encodes L-galactose-1-P phosphatase, a plant ascorbic acid biosynthetic enzyme. J Biol Chem **281:** 15662–15670

Daub CO, Kloska S, Selbig J (2003) MetaGeneAlyse: analysis of integrated transcriptional and metabolite data. Bioinformatics **19:** 2332–2333

Davison A (2008) Some challenges for statistics. Stat Methods Appl **17:** 167–181

Eastment HT, Krzanowski WJ (1982) Cross-validatory choice of the number of components from a principal component analysis. Technometrics **24:** 73–77

Enot DP, Beckmann M, Overy D, Draper J (2006) Predicting interpretability of metabolome models based on behavior, putative identity, and biological relevance of explanatory signals. Proc Natl Acad Sci USA **103:** 14865–14870

Enot DP, Lin W, Beckmann M, Parker D, Overy DP, Draper J (2008) Preprocessing, classification modeling and feature selection using flow injection electrospray mass spectrometry metabolite fingerprint data. Nat Protoc **3:** 446–470

Fahnenstich H, Saigo M, Niessen M, Zanor MI, Andreo CS, Fernie AR, Drincovich MF, Flügge UI, Maurino VG (2007) Alteration of organic acid metabolism in Arabidopsis overexpressing the maize C_4 NADP-malic enzyme causes accelerated senescence during extended darkness. Plant Physiol **145:** 640–652

Frank E, Hall M, Trigg L, Holmes G, Witten IH (2004) Data mining in bioinformatics using Weka. Bioinformatics **20:** 2479–2481

Fraser PD, Pinto MES, Holloway DE, Bramley PM (2000) Application of high-performance liquid chromatography with photodiode array detection to the metabolic profiling of plant isoprenoids. Plant J **24:** 551–558

Friedman JH (2006) Recent advances in predictive (machine) learning. J Classif **23:** 175–197

Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol **5:** R80

Gewehr JE, Szugat M, Zimmer R (2007) BioWeka: extending the Weka framework for bioinformatics. Bioinformatics **23:** 651–653

Gibon Y, Usadel B, Blaesing O, Kamlage B, Hoehne M, Trethewey R, Stitt M (2006) Integration of metabolite with transcript and enzyme activity profiling during diurnal cycles in *Arabidopsis* rosettes. Genome Biol **7:** R76

Gidman EA, Stevens CJ, Goodacre R, Broadhurst D, Emmett B, Gwynn-Jones D (2006) Using metabolic fingerprinting of plants for evaluating nitrogen deposition impacts on the landscape level. Glob Change Biol **12:** 1460–1465

Hall RD (2006) Plant metabolomics: from holistic hope, to hype, to hot topic. New Phytol **169:** 453–468

Hammer Ø, Harper DAT, Ryan PD (2001) PAST: paleontological statistics software package for education and data analysis. Palaeontol Electron **4:** 1.4A

Husek P (1998) Chloroformates in gas chromatography as general purpose derivatizing agents. J Chromatogr B Analyt Technol Biomed Life Sci **717:** 57–91

Izquierdo-Garcia J, Rodríguez I, Kyriazis A, Villa P, Barreiro P, Desco M, Ruiz-Cabello J (2009) A novel R-package graphic user interface for the analysis of metabonomic profiles. BMC Bioinformatics **10:** 363

Jafari P, Azuaje F (2006) An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. BMC Med Inform Decis Mak **6:** 27

Janes KA, Yaffe MB (2006) Data-driven modelling of signal-transduction networks. Nat Rev Mol Cell Biol **7:** 820–828

Jenkins H, Johnson H, Kular B, Wang T, Hardy N (2005) Toward supportive data collection tools for plant metabolomics. Plant Physiol **138:** 67–77

Kachroo P, Venugopal SC, Navarre DA, Lapchyk L, Kachroo A (2005) Role of salicylic acid and fatty acid desaturation pathways in *ssi2*-mediated signaling. Plant Physiol **139:** 1717–1735

Larson TR, Graham IA (2001) A novel technique for the sensitive quantification of acyl CoA esters from plant tissues. Plant J **25:** 115–125

Lisec J, Schauer N, Kopka J, Willmitzer L, Fernie AR (2006) Gas chromatography mass spectrometry-based metabolite profiling in plants. Nat Protoc **1:** 387–396

Lu Y, Savage LJ, Ajjawi I, Imre KM, Yoder DW, Benning C, DellaPenna D, Ohlrogge JB, Osteryoung KW, Weber AP, et al (2008) New connections across pathways and cellular processes: industrialized mutant screening reveals novel associations between diverse phenotypes in Arabidopsis. Plant Physiol **146:** 1482–1500

Mahoney MW, Drineas P (2009) CUR matrix decompositions for improved data analysis. Proc Natl Acad Sci USA **106:** 697–702

Mekhedov S, de Ilarduya OM, Ohlrogge J (2000) Toward a functional catalog of the plant genome: a survey of genes for lipid biosynthesis. Plant Physiol **122:** 389–402

Messerli G, Partovi Nia V, Trevisan M, Kolbe A, Schauer N, Geigenberger P, Chen J, Davison AC, Fernie AR, Zeeman SC (2007) Rapid classification of phenotypic mutants of Arabidopsis via metabolite fingerprinting. Plant Physiol **143:** 1484–1492

Meyer RC, Steinfath M, Lisec J, Becher M, Witucka-Wall H, Torjek O, Fiehn O, Eckardt A, Willmitzer L, Selbig J, et al (2007) The metabolic signature related to high plant growth rate in *Arabidopsis thaliana*. Proc Natl Acad Sci USA **104:** 4759–4764

Nachtomy O, Shavit A, Yakhini Z (2007) Gene expression and the concept of the phenotype. Stud Hist Phil Biol Biomed Sci **38:** 238–254

Nelson DR, Schuler MA, Paquette SM, Werck-Reichhart D, Bak S (2004) Comparative genomics of rice and Arabidopsis: analysis of 727 cytochrome P450 genes and pseudogenes from a monocot and a dicot. Plant Physiol **135:** 756–772

Neuweger H, Albaum SP, Dondrup M, Persicke M, Watt T, Niehaus K, Stoye J, Goesmann A (2008) MeltDB: a software platform for the analysis and integration of metabolomics experiment data. Bioinformatics **24:** 2726–2732

Pan X, Liu H, Clarke J, Jones J, Bevan M, Stein L (2003) ATIDB: *Arabidopsis thaliana* insertion database. Nucleic Acids Res **31:** 1245–1251

Pastori GM, Kiddle G, Antoniw J, Bernard S, Veljovic-Jovanovic S, Verrier PJ, Noctor G, Foyer CH (2003) Leaf vitamin C contents modulate plant defense transcripts and regulate genes that control development through hormone signaling. Plant Cell **15:** 939–951

Rösti J, Barton CJ, Albrecht S, Dupree P, Pauly M, Findlay K, Roberts K, Seifert GJ (2007) UDP-glucose 4-epimerase isoforms *UGE2* and *UGE4* cooperate in providing UDP-galactose for cell wall biosynthesis and growth of *Arabidopsis thaliana*. Plant Cell **19:** 1565–1579

Saeys Y, Inza I, Larrañaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics **23:** 2507–2517

Schauer N, Semel Y, Roessner U, Gur A, Balbo I, Carrari F, Pleban T, Perez-Melis A, Bruedigam C, Kopka J, et al (2006) Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. Nat Biotechnol **24:** 447–454

Slawski M, Daumer M, Boulesteix AL (2008) CMA: a comprehensive Bioconductor package for supervised classification with high dimensional data. BMC Bioinformatics **9:** 439

Stajich JE, Lapp H (2006) Open source tools and toolkits for bioinformatics: significance, and where are we? Brief Bioinform **7:** 287–296

Stitt M, Muller C, Matt P, Gibon Y, Carillo P, Morcuende R, Scheible WR, Krapp A (2002) Steps towards an integrated view of nitrogen metabolism. J Exp Bot **53:** 959–970

Suh MC, Samuels AL, Jetter R, Kunst L, Pollard M, Ohlrogge J, Beisson F (2005) Cuticular lipid composition, surface structure, and gene expression in Arabidopsis stem epidermis. Plant Physiol **139:** 1649–1665

Sweetlove LJ, Last RL, Fernie AR (2003) Predictive metabolic engineering: a goal for systems biology. Plant Physiol **132:** 420–425

Tarca AL, Carey VJ, Chen X, Romero R, Drăghici S (2007) Machine learning and its applications to biology. PLoS Comput Biol **3:** e116

Thimm O, Blasing O, Gibon Y, Nagel A, Meyer S, Kruger P, Selbig J, Muller LA, Rhee SY, Stitt M (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J **37:** 914–939

Tissier AF, Marillonnet S, Klimyuk V, Patel K, Torres MA, Murphy G, Jones JDG (1999) Multiple independent defective *Suppressor-mutator* transposon insertions in *Arabidopsis*: a tool for functional genomics. Plant Cell **11:** 1841–1852

Verslues PE, Bray EA (2006) Role of abscisic acid (ABA) and *Arabidopsis thaliana* ABA-insensitive loci in low water potential-induced ABA and proline accumulation. J Exp Bot **57:** 201–212

Wang T, Shao K, Chu Q, Ren Y, Mu Y, Qu L, He J, Jin C, Xia B (2009) Automics: an integrated platform for NMR-based metabonomics spectral processing and data analysis. BMC Bioinformatics **10:** 83

Ward JL, Baker JM, Beale MH (2007) Recent applications of NMR spectroscopy in plant metabolomics. FEBS J **274:** 1126–1131

Westerhuis JA, Hoefsloot HCJ, Smit S, Vis DJ, Smilde AK, van Velzen EJJ, van Duijnhoven JPM, van Dorsten FA (2008) Assessment of PLSDA cross validation. Metabolomics **4:** 81–89

Wurtele ES, Li J, Diao LX, Zhang HL, Foster CM, Fatland B, Dickerson U, Brown A, Cox Z, Cook D, et al (2003) MetNet: software to build and model the biogenetic lattice of *Arabidopsis*. Comp Funct Genomics **4:** 239–245

Xia JG, Psychogios N, Young N, Wishart DS (2009) MetaboAnalyst: a Web server for metabolomic data analysis and interpretation. Nucleic Acids Res **37:** W652–W660

Xu C, Yu B, Cornish AJ, Froehlich JE, Benning C (2006) Phosphatidylglycerol biosynthesis in chloroplasts of Arabidopsis mutants deficient in acyl-ACP glycerol-3-phosphate acyltransferase. Plant J **47:** 296–309

Zhang C, Lu X, Zhang X (2006) Significance of gene ranking for classification of microarray samples. IEEE/ACM Trans Comput Biol Bioinformatics **3:** 312–320

Zhang P, Foerster H, Tissier CP, Mueller L, Paley S, Karp PD, Rhee SY (2005) MetaCyc and AraCyc: metabolic pathway databases for plant research. Plant Physiol **138:** 27–37