nature biotechnology

High-throughput classification of yeast mutants for functional genomics using metabolic footprinting

Jess Allen[1], Hazel M Davey[1], David Broadhurst[1], Jim K Heald[1], Jem J Rowland[2], Stephen G Oliver[3] & Douglas B Kell[1]

Many technologies have been developed to help explain the function of genes discovered by systematic genome sequencing. At present, transcriptome and proteome studies dominate large-scale functional analysis strategies. Yet the metabolome, because it is 'downstream', should show greater effects of genetic or physiological changes and thus should be much closer to the phenotype of the organism. We earlier presented a functional analysis strategy that used metabolic fingerprinting to reveal the phenotype of silent mutations of yeast genes[1]. However, this is difficult to scale up for high-throughput screening. Here we present an alternative that has the required throughput (2 min per sample). This 'metabolic footprinting' approach recognizes the significance of 'overflow metabolism' in appropriate media. Measuring intracellular metabolites is time-consuming and subject to technical difficulties caused by the rapid turnover of intracellular metabolites and the need to quench metabolism and separate metabolites from the extracellular space. We therefore focused instead on direct, noninvasive, mass spectrometric monitoring of extracellular metabolites in spent culture medium. Metabolic footprinting can distinguish between different physiological states of wild-type yeast and between yeast single-gene deletion mutants even from related areas of metabolism. By using appropriate clustering and machine learning techniques, the latter based on genetic programming[2–8], we show that metabolic footprinting is an effective method to classify 'unknown' mutants by genetic defect.

After optimization of electrospray ionization–mass spectrometry (ESI-MS) parameters for the analysis of yeast minimal medium plus a metabolite cocktail (see **Supplementary Table 1** online), we initially compared metabolic footprints from different growth phases of yeast batch culture. Visual inspection of the resulting mass spectra revealed that the metabolic footprint was characteristic of each specific phase of culture growth (**Fig. 1a**). Footprints taken early, during lag and early exponential phase, were comparatively simple, with clear peaks for many of the exogenously supplied amino acids visible against the background of the basal medium. The most marked changes occurred across the transition from exponential to stationary phase, when the spectra became increasingly complex with the appearance of numerous small peaks. This implies that the cells are secreting or excreting metabolites into the culture medium. At the dilutions we used, and in these media, the normalized spectra were qualitatively independent of the exact dilution over a broad range. In particular, no new spectral peaks appeared as samples were diluted (see **Supplementary Fig. 1a,b online**). In addition to visual analysis of the resulting mass spectra, principal components analysis (PCA) was conducted on the preprocessed mass spectral data (see Methods). As a useful aid to multivariate data visualization, PCA seeks to rotate the data points into a new coordinate system, such that the majority of the variance in the data is accounted for in the directions of a subset of these rotated axes. Hence, plotting the points in this new coordinate system makes it easier to visualize the significant effects within the data. In the resulting PCA scores plot (**Fig. 1b**), data points are separated along a curve across the first two principal components (which account for some 98% of the variance) in a pattern that relates to the time of sampling (growth phase).

Using direct-injection MS alone, we could not unambiguously identify unknown peaks in the metabolic footprint. However, it is important to note that metabolic footprinting was not devised as a metabolite profiling strategy[9], but instead relies on the development of rules to describe trends in the data that involve only a small number of the variables (masses)[6]. This type of pattern recognition approach[10] then allows one to confine identification by tandem MS to those substances contributing to the rules, thus avoiding the necessity for time-consuming identification of all metabolites in first-round analyses.

We next studied the robustness of the method with respect to variances in conditions that might be anticipated. These included variations in batches of medium, different inoculum levels and inoculation from different starter cultures, and the use of filter sterilization as opposed to centrifugation for separating cells from medium before MS analysis. In addition, we used strains in which the *kan*MX deletion cassette had been inserted into the *HO* mating locus—a reportedly phenotypically neutral site[11,12]. We found that, in all cases, equivalent data points clustered together in a robust and reproducible manner; representative PCA plots are shown (**Fig. 2**).

[1]Institute of Biological Sciences, Cledwyn Building, University of Wales, Aberystwyth, Aberystwyth SY23 3DD, UK. [2]Department of Computer Science, University of Wales, Aberystwyth, Aberystwyth SY23 3DB, UK. [3]School of Biological Sciences, University of Manchester, 2.205 Stopford Building, Oxford Rd., Manchester M13 9PT, UK. [4]Present address (from 1 September 2003): Department of Chemistry, UMIST, Sackville St., P.O. Box 88, Manchester M60 1QD, UK. Correspondence should be addressed to D.B.K. (dbk@umist.ac.uk).

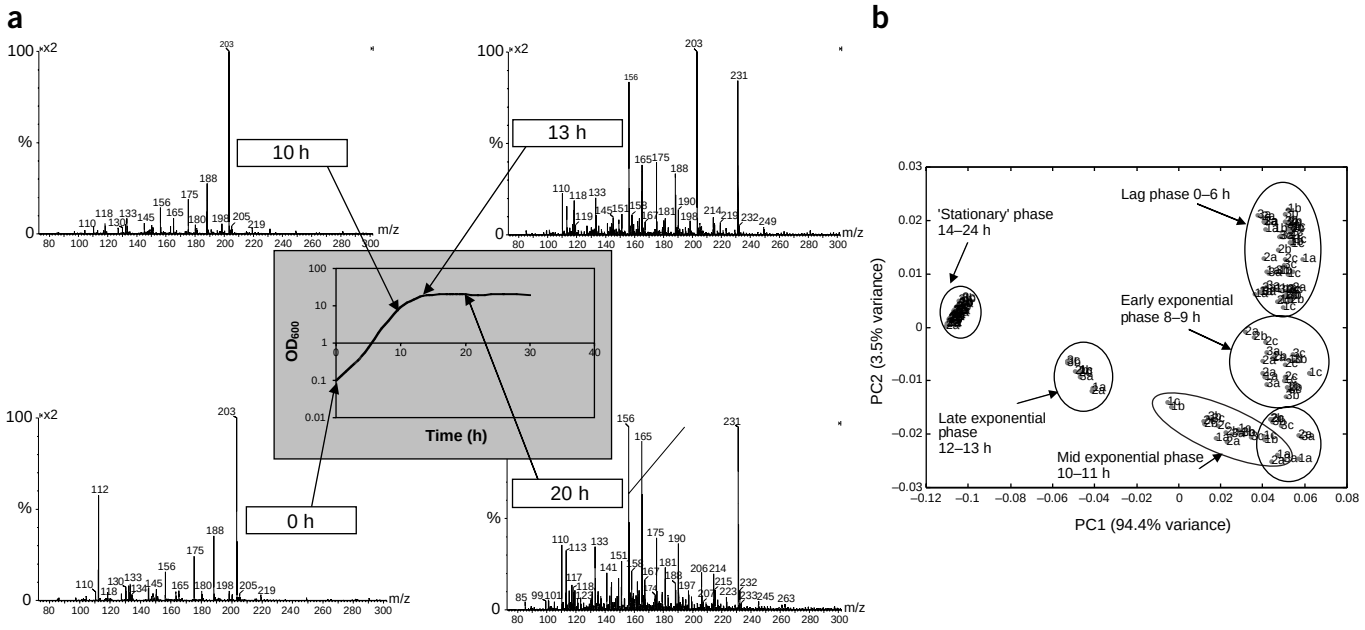

Figure 1 Metabolic footprinting of *Saccharomyces cerevisiae*. Cultures were grown on minimal medium supplemented with a metabolite cocktail (**Supplementary Table 1** online). Samples were removed from batch culture throughout the fermentation and prepared for mass spectral analysis (positive ionization, only *m/z* 65–300 displayed) as described in the text. TOF, time of flight. (**a**) Representative spectra. (**b**) Principal components analysis (PCA) of the data in **a**.

Yeast single-gene deletants with defects in amino acid metabolism (see **Supplementary Table 2** online) generated by the EUROFAN project (that is, carrying the *kan*MX resistance cassette in place of the target open reading frame) were compared. Wild-type yeast grown on footprinting medium in batch culture exhausts available glucose between 12 and 14 h after inoculation, at which time the cell density, as measured by optical density, is sufficiently great to inhibit growth by respiration of the ethanol formed earlier, and optical density subsequently remains constant for many hours (**Fig. 1**). Consequently, 24 h was chosen as both a biologically and technically convenient time for taking samples. Microtiter-plate yeast cultures of deletant strains were set up, and 24-h samples were processed and analyzed (see Methods). PCA conducted on the data gave good groupings of biological (culture) replicates (data not shown) but was not otherwise very informative. Consequently, we applied discriminant function analysis (DFA), a supervised technique that allows groups in the data to be defined[1]. Groups were assigned simply on the basis of (biological) replicate number, so that the analysis was not biased by preconceived notions of how the footprint data should group according to our knowledge of the respective metabolic defect. By defining the groups in this way, one is essentially informing the model that each strain is different and thereby encouraging it to preserve those differences. Consequently, when strains do cluster together, this demonstrates the presence of a real underlying biological relationship. In addition, cross-validation of the DFA model was done (see **Supplementary Fig. 2** online). Each strain forms a distinct cluster, with the exception of *ho*Δ, which clusters together with the wild type. This supports earlier conclusions[11] from competition experiments in chemostat culture that the *HO* mating-type locus was a selectively neutral site for deletion cassette marker insertion and that the *ho*::*kan*MX strain would therefore prove a suitable reference for functional analysis of EUROFAN deletants.

The EUROFAN systematic gene deletion project generated *MATa* and *MATα* haploid as well as homozygous and heterozygous diploid single-gene deletants for the purposes of large-scale functional analysis. Clearly, for essential genes, only heterozygous diploids are viable, so it was important to establish whether mutants could be distinguished on the basis of their metabolic footprints regardless of their mating type. Footprints of *MATa* and *MATα* haploids and heterozygous diploid strains of three different gene deletion mutants (strains 1, 3 and 5; **Supplementary Table 2**) plus wild type were compared. Data were processed and analyzed as above. Replicates of the same gene deletion (that is, all three mating types) were assigned to the same DFA class. Mutants in genes from even nominally closely related areas of metabolism (in this case, amino acid biosynthesis) may be distinguished on the basis of their footprints, regardless of their mating type (see **Supplementary Fig. 3** online).

A common strategy in functional genomics is to carry out expression profiling on a series of strains of known genotype and use the patterns to determine the 'closeness' to these expression profiles of genes of 'unknown' function. Such a strategy lies behind the FANCY (Functional ANalysis by Co-response in Yeast) method used previously for metabolic fingerprinting[1], and when calibrated in this way is known as 'supervised learning'[10,13,14]. To demonstrate the potential application of metabolic footprinting as a 'guilt-by-association'[15] functional genomics strategy, we compared 24-h footprints from microtiter-plate cultures of 19 different deletion mutants with defects in a broad range of metabolic categories. Strain 18 (which carries a deletion in *ROX1*, a heme-dependent transcriptional repressor of hypoxic genes) was very slow growing (see **Supplementary Fig. 4** online) and had not reached stationary phase by the time of sampling; it was consequently omitted from further analyses. Two sets of very similar pairs of enzyme deletants were included in the strain set, including the *pfk26Δ* and *pfk27Δ* strains studied earlier[1], so that we

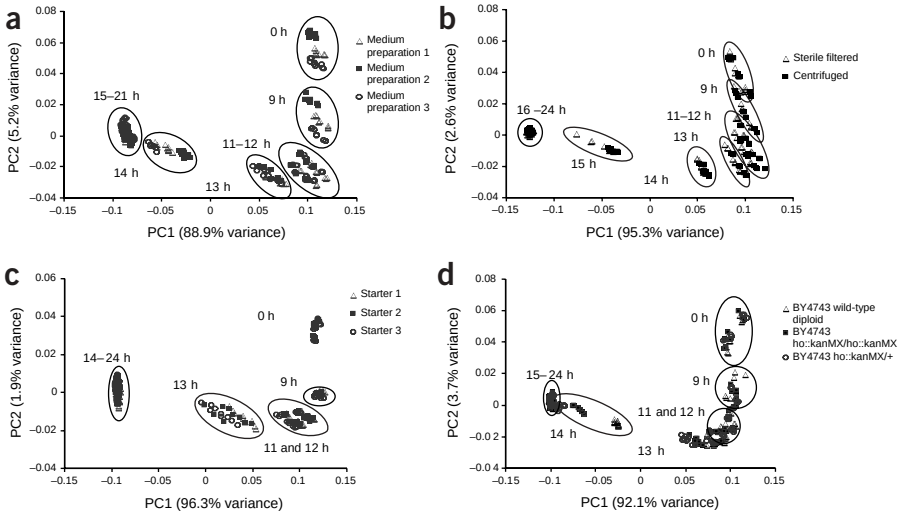

Figure 2 PCA plots of metabolic footprint data to illustrate the robustness of the method with respect to variations that may be expected. (**a**–**d**) Experiments were done as described in the legend to **Figure 1** with the following variations: different preparations of footprinting medium (**a**), different sample preparation methods (centrifugation versus filter sterilization; **b**), inoculation from different starter cultures (**c**) and the comparison of true wild type with a phenotypically 'wild-type' reference strain carrying the *kan*MX deletion cassette in a 'neutral' site (mating type switching locus, *HO*; **d**).

could determine whether simple DFA-based clustering of metabolic footprints would serve to identify the 'unknowns' correctly. Footprint data from strains harboring the *nit3* and *pfk27* deletions clustered closely together with strains carrying deletions in the related genes *nit2* and *pfk26* (**Fig. 3a,b**). The cluster analysis relies on using just the most significant discriminant functions for its display in this way, and this conceals some of the complexity. To this end, we also carried out an agglomerative hierarchical cluster analysis (HCA; ref. 1) on the basis of all the discriminant functions used in the model (**Fig. 3c**). This shows clearly that the *nit* and *pfk* mutants do indeed group most closely with each other, and that in terms of the overall variance in the data in this figure these genotypes were well distributed, with the *nit* knockouts located fairly near the center of the dendrogram and the *pfk* knockouts near the edge in this multidimensional space (**Fig. 3c**). The dendrograms are based on data from all the samples and, as would be expected, were consequently little affected by removal of individual genotypes from the analysis (data not shown). In addition, the HCA shows that *cki1*Δ and *faa3*Δ knockouts group most closely with the wild-type hoΔ. Note that they are indistinguishable from it in the two-dimensional DFA plot of **Figure 3b**. This is reasonable, because there is almost certainly redundancy here—that is, other related genes with the same activities are present to provide equivalent functionality. In addition to *CKI1* (YLR133W, specifying choline kinase), there is an ethanolamine kinase gene *EKI1* (YDR147W) whose product shows great similarity to choline kinase as determined from the MIPS database (http://mips.gsf.de/proj/yeast/CYGD/db/index.html). The product of *FAA3* (YIL009W) (acyl CoA synthase) is 100-fold lower in activity than is the product of *FAA1* (YOR317W). *FAA2* (YER015W) and *FAA4* (YMR246W) are also closely related and probably arose from gene duplication of *FAA1*; indeed the Faa4 protein has 78% identity to Faa3p and also has this acyl CoA synthase activity (http://mips.gsf.de/proj/yeast/CYGD/db/index.html). This provides a useful control experiment.

Although the DFA strategy was highly successful, it was not very informative as to which masses were important in the discrimination of the mutant classes. Rule-based methods (for example, refs. 14,16,17) are much more descriptive but can be less accurate[18]. We therefore used a variant of genetic programming (gmax-bio; Aber Genomic Computing) to evolve a rule that could be used to give a simple explanation of what best explained the differences between the classes of interest. In the case of nitrilase, trained as in **Figure 3**, the rule "IF normalized *m/z*_201 > 0.00126 THEN mutant = nitrilase$^-$" correctly identified all examples in training, validation and test sets with no false positives.

The continuing need for methods to analyze complex biological systems with high throughput and high information content has led many researchers to measure expression profiles at the level of the metabolome or 'metabolic fingerprint'. We have here demonstrated that the metabolic footprint—the quota of low-molecular-weight metabolites present in the extracellular medium—is rich in biochemical information, can be determined rapidly by direct-injection MS, differs reproducibly between strains and mutant types, and is thus a suitable method for the purposes of functional genomics. Specific advantages of footprinting over fingerprinting include its speed, ease of performance, the ability to automate it and the opportunity—by identifying the metabolites involved—to establish the biochemical basis for a defect. It is obvious that other spectrometric techniques might be used as easily as MS, and that the formulation of the culture medium might be varied to steer the analysis toward particular functional domains. Finally, because gene knockouts can be discriminated reproducibly, it is reasonable that the same strategy might be used in assays of the mode of action of drugs[19] or in the assessment of cytotoxicity[20].

METHODS

Yeast strains. All method development was carried out with the wild-type diploid strain BY4743 (*MAT**a**/MATα ura3Δ0/ura3Δ0 leu2Δ0/leu2Δ0 his3Δ1/his3Δ1 +/lys2Δ0 met15Δ0/+*). Later comparisons with reference strains carrying the *kan*MX resistance cassette (conferring resistance to geneticin (G418)) were done with both homozygous and heterozygous BY4743 *ho::kan*MX. Experiments with EUROFAN mutants were done with BY4741 *MAT**a*** haploid (*ura3Δ0 leu2Δ0 his3Δ1 met15Δ0*), in which single genes of interest had been replaced with the *kan*MX deletion cassette. Yeast cells were grown in synthetic defined minimal medium (0.67% Yeast Nitrogen Base without Amino Acids (Difco) (**Supplementary Table 1** online), 2% dextrose) supplemented with a metabolite cocktail of amino acids, bases and organic acids, all at a final concentration of 1 mM.

Flask culture. Batch cultures of 50 ml were inoculated with 500-μl washed inoculum from 5-ml starter cultures and grown in 250-ml flasks in a rotary shaker at 30 °C, 200 rpm. Typically, cells were counter in a hemocytometer after washing to ensure that cultures were inoculated to an initial concentration of approximately 5×10^5 cells ml^{-1} (optical density at 600 nm (OD_{600}) ≈ 0.1).

Microtiter plate culture. For larger-scale comparison of metabolism mutants, microtiter plate batch cultures were used. 100-well Honeycomb II plates (Labsystems) were filled with 100 μl per well (ten replicate wells per strain) of yeast cells diluted in fresh medium to a concentration of 5×10^5 cells ml^{-1} from washed 5-ml starter cultures. Plates were incubated at 30 °C

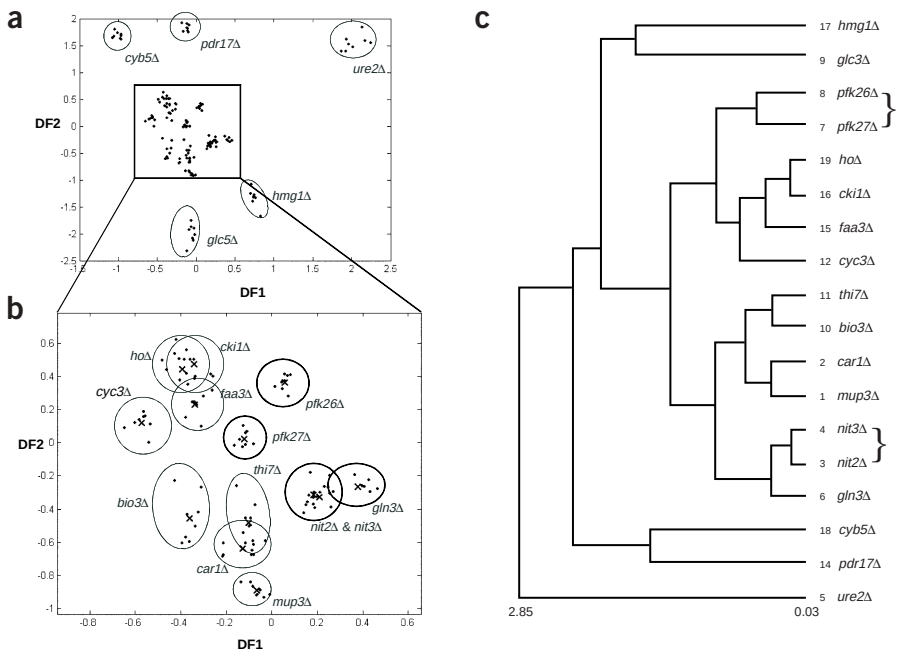

Figure 3 Metabolic footprinting may be used to classify strains on the basis of the deletion they carry. An experiment was set up in which 24-h microtiter plate footprints of 19 different deletant strains with a broad range of metabolic defects were compared. (**a**,**b**) Footprint data were used to train a DFA model (20 PCs, 99.6% of the variance). Footprint data from strains harboring the *nit3* and *pfk27* deletions clustered closely together with strains carrying deletions in the respective isoenzymes *nit2* and *pfk26*. Box in **a** indicates region enlarged in **b**. DF, discriminant function. (**c**) Hierarchical cluster analysis of the data using all 18 DFs. The scale represents the Euclidean distance in DF space.

and shaken continuously at 'medium' intensity in an automatic plate reader (Bioscreen C; Labsystems) which measured OD_{600} every 20 min for the duration of the experiment.

Sampling procedure. For flask cultures, 1-ml samples were removed at appropriate points during culture growth and either centrifuged (8,000*g*, 10 minutes) or filter-sterilized (through a prerinsed 0.2-μm Acrodisc filter, Gelman) to remove cells. The supernatant or filtrate was stored in 200-μl aliquots at –40 °C until later preparation for mass spectral analysis. For microtiter-plate cultures, 90 μl was removed at appropriate points during culture growth from each of three to eight adjacent replicate wells per strain and spun down; 85 μl of supernatant was removed and stored in microtiter plates at –40 °C until later analysis.

Mass spectrometry and sample preparation. Before mass spectral analysis, samples were diluted 10-fold in 30:70 (vol/vol) methanol/water. Formic acid was added to a final concentration of 0.1% (27 mM) to aid in the ionization of amino acids. The samples were degassed and large particles were removed by microcentrifugation (8,000*g*, 3–5 min). Mass spectra were obtained using a Micromass (Manchester, UK) LCT electrospray ionization time-of-flight mass spectrometer. Collection of mass spectra was automated by linking the MS to a Waters Alliance 2690 liquid chromatography (LC) system in which the HPLC column had been replaced with three m of 0.13 mm internal diameter polyetheretherketone (PEEK) tubing[21]. Diluted, degassed sample was dispensed in 200-μl aliquots into polypropylene inserts within HPLC vials (Waters). The vials were placed in the autosampling carousels of the LC system, arranged in order of time of sampling and replicate number. Using the MassLynx software (Micromass), the system was set up as if for an automated LC-MS run: 20 μl of sample was automatically injected into the sample loop and carried through the LC system in 30:70 (vol/vol) methanol with 0.1% formic acid at a flow rate of 500 μl/min. The flow was split between LC and MS such that flow into the MS did not exceed 50 μl/min. Spectra were collected in positive ion mode every second (0.9-s scan time, 0.1-s interscan delay) for 2 min per sample from m/z 65 to 1,000. A single 2-min run was performed for each sample consecutively, with a 1-min intersample wash period, then this cycle was repeated twice more to obtain three machine replicates per sample. The ESI conditions were as follows: capillary voltage 3,000 V, source temperature 80 °C, desolvation temperature 120 °C, RF lens 100 V, sample cone voltage 30 V and extraction cone voltage 10 V. Sample cone voltage, which determines the degree of fragmentation of the analyte, was chosen such that the molecular ions of a set of amino acids standards (histidine, leucine, lysine, methionine, tryptophan and uracil each at 100 μM) gave a clear peak at the expected m/z without excessive fragmentation.

MS data preprocessing. The mass spectrometric methods described above produce vast amounts of potentially useful data. Automated direct-injection MS, since it uses the LC system for automation, produces a spectrochromatogram (an array of the mass spectra versus time) for each sample analyzed that can typically hold 10^6 values (depending upon the MS range and sampling rates). In their native form, such data are extremely difficult to interpret. To convert the data into information of chemical or biological interest, some sort of multivariate statistical analysis must be used. To simplify any subsequent statistical analysis, two simple preprocessing algorithms were applied to the direct-injection LC-MS spectrochromatograms. Each LC-MS array was reduced into a single 'aggregate' MS vector by summing the ion counts of a given *m/z* ratio over the total scan cycle. Then each MS vector was 'binned' to unit *m/z* ratio (that is, ion counts of fractional *m/z* ratios were added to the nearest integer *m/z*). Thus, after this initial data reduction, an MS spectrochromatogram with *m/z* range 65 to 1,000 will be reduced to a single vector having 935 values.

Multivariate data analysis. Before any multivariate analysis is carried out, each mass spectral vector is normalized to the total ion count for that vector, so that different spectra can be compared quantitatively. Once a set of *N* spectra (with mass range *p*) is concatenated into a single matrix (*N* objects × *p* variables), each column of the data set can be optionally normalized to unit variance. This is done to eliminate bias, in subsequent analysis, toward any column that contains either large absolute values or large variances[22]. Notably, however, normalization (to unit variance) can sometimes be more detrimental than helpful. If there are a large number of redundant variables in the data, the noise affecting such variables is amplified to the same importance as relevant variables, and this can easily cloud any underlying statistical trends. Consequently, in this instance scaling was not done before statistical analysis. To reduce the dimensionality of the mass spectral data, PCA[23] was used in the first stage of analysis (as described in ref. 1). PCA involves projecting the original **X**-matrix (*N* objects; *p* variables) onto a *d*-dimensional subspace using a projection (or 'loading') matrix, thus creating object coordinates (a 'scores' matrix) in a new coordinate system. This is achieved by the method known as singular value decomposition (SVD) of **X**:

$$\mathbf{X}_{N\times p} = \mathbf{U}_{N\times d}\mathbf{\Lambda}_{d\times d}\mathbf{L}^{T}_{p\times d} = \mathbf{T}_{N\times d}\mathbf{L}^{T}_{p\times d}$$

where **U** is the unweighted (normalized) score matrix and **T** is the weighted (or biased) score matrix. **L** is the loading matrix where the columns of **L** are known as eigenvectors or loading principal components. **Λ** is a diagonal matrix (that is, all of the off-diagonal elements are equal to 0) containing the square roots of the first *d* eigenvalues of the covariance matrix ($\mathbf{X}^{T}\mathbf{X}$), where $d < N$ and $d < p$.

The principal components can be considered as a basis set used to project the original data matrix, **X**, onto the scores, **T**. In other words, the new coordinates are linear combinations of the original variables. For example, the elements of the first principal component can be represented as:

$$t_{11} = x_{11}l_{11} + x_{12}l_{21} + \ldots + x_{1p}l_{p1}$$
$$t_{21} = x_{21}l_{11} + x_{22}l_{21} + \ldots + x_{2p}l_{p1}$$
$$\vdots$$
$$t_{n1} = x_{n1}l_{11} + x_{12}l_{n1} + \ldots + x_{np}l_{p1}$$

The influence of each of the original variables on the new principal components (that is, the contents of the loading matrix) is determined on the basis of the maximum variance criterion. The first principal component is considered to lie in the direction describing maximum variance in the original data. Each subsequent principal component lies in an orthogonal direction of maximum variance that has not been considered by the former components. The number of principal components computed for a given data set is up to the analyst. However, usually as many principal components are calculated as are needed to explain a preset percentage of the total variance in the original data (the number of principal components is always less than or equal to the number of original variables).

The second stage of the data analysis involves using DFA to separate the samples further into groups of replicates using the principal components calculated in stage 1 as the source data[24,25]. In contrast to PCA, DFA is a supervised method that allows groups in the data to be defined. All DFA classes were defined only according to replicates of the same strain, such that the analysis was unbiased by prior knowledge of how deletant strains might group according to their underlying metabolic defect.

Finally, the Euclidean distance between group centers in DFA space was used to construct a similarity measure, which was transformed into a dendrogram using an agglomerative HCA[26,27].

Machine learning. We used the genetic programming software gmax-bio (Aber Genomic Computing). This takes data in the form described above, together with knowledge of the target class it is desired to learn, and evolves rules which effect the necessary nonlinear mapping[6,28]. We used 50% of the examples in the training set as an internal cross-validation set.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

This work was supported by a grant from the Biotechnology and Biological Sciences Research Council, UK, to D.B.K. and S.G.O., and by a grant from the Wellcome Trust to S.G.O. J.A. was the recipient of a BBSRC CASE studentship with Bayer CropScience (formerly Aventis CropScience). We thank John Pillmoor, Steve Dunn and Jane Dancer for their careful supervision, Bharat Rash and Nicola Burton of the Manchester laboratory for technical assistance and Roy Goodacre (Aberystwyth/UMIST) for useful discussions.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Biotechnology* website for details).

Received 6 November 2002; accepted 28 February 2003
Published online 12 May 2003; doi:10.1038/nbt823

1. Raamsdonk, L.M. *et al.* A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat. Biotechnol.* **19**, 45–50 (2001).
2. Cramer, N.L. A representation for the adaptive generation of simple sequential programs. in *Proceedings of the First International Conference on Genetic Algorithms and their Applications* (ed. Grefenstette, J.J.) 183–187 (Lawrence Erlbaum, Mahwah, New Jersey, 1985).
3. Koza, J.R. Genetic Programming: On the Programming of Computers by Means of Natural Selection (MIT Press, Cambridge, Massachusetts, 1992).
4. Banzhaf, W., Nordin, P., Keller, R.E. & Francone, F.D. *Genetic Programming: An Introduction* (Morgan Kaufmann, San Francisco, 1998).
5. Langdon, W.B. Genetic Programming and Data Structures: Genetic Programming + Data Structures = Automatic Programming! (Kluwer, Boston, 1998).
6. Kell, D.B., Darby, R.M. & Draper, J. Genomic computing: explanatory analysis of plant expression profiling data using machine learning. *Plant Physiol.* **126**, 943–951 (2001).
7. Kell, D.B. Genotype:phenotype mapping: genes as computer programs. *Trends Genet.* **18**, 555–559 (2002).
8. Langdon, W.B. & Poli, R. *Foundations of Genetic Programming* (Springer, Berlin, 2002).
9. Fiehn, O. Metabolomics: the link between genotypes and phenotypes. *Plant Mol. Biol.* **48**, 155–171 (2002).
10. Kell, D.B. & King, R.D. On the optimization of classes for the assignment of unidentified reading frames in functional genomics programmes: the need for machine learning. *Trends Biotechnol.* **18**, 93–98 (2000).
11. Baganz, F., Hayes, A., Marren, D., Gardner, D.C.J. & Oliver, S.G. Suitability of replacement markers for functional analysis studies in *Saccharomyces cerevisiae*. *Yeast* **13**, 1563–1573 (1997).
12. Oliver, S.G., Winson, M.K., Kell, D.B. & Baganz, F. Systematic functional analysis of the yeast genome. *Trends Biotechnol.* **16**, 373–378 (1998).
13. Duda, R.O., Hart, P.E. & Stork, D.E. *Pattern Classification*, edn. 2 (John Wiley, London, 2001).
14. Hastie, T., Tibshirani, R. & Friedman, J. The Elements of Statistical Learning: Data Mining, Inference and Prediction (Springer, Berlin, 2001).
15. Oliver, S.G. Proteomics: guilt-by-association goes global. *Nature* **403**, 601–603 (2000).
16. Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. *Classification and Regression Trees* (Wadsworth International, Belmont, California, 1984).
17. Quinlan, J.R. *C4.5: Programs for Machine Learning* (Morgan Kaufmann, San Mateo, California, 1993).
18. Alsberg, B.K., Goodacre, R., Rowland, J.J. & Kell, D.B. Classification of pyrolysis mass spectra by fuzzy multivariate rule induction—comparison with regression, K-nearest neighbour, neural and decision-tree methods. *Anal. Chim. Acta* **348**, 389–407 (1997).
19. Aranibar, N., Singh, B.K., Stockton, G.W. & Ott, K.-H. Automated mode-of-action detection by metabolic profiling. *Biochem. Biophys. Res. Commun.* **286**, 150–155 (2001).
20. Griffin, J.L. *et al.* Metabolic profiling of genetic disorders: a multitissue H-1 nuclear magnetic resonance spectroscopic and pattern recognition study into dystrophic tissue. *Anal. Biochem.* **293**, 16–21 (2001).
21. Goodacre, R., Vaidyanathan, S., Bianchi, G. & Kell, D.B. Metabolic profiling using direct infusion electrospray ionisation mass spectrometry for the characterisation of olive oils. *Analyst* **127**, 1457–1462 (2002).
22. Martens, H. & Næs, T. *Multivariate Calibration* (John Wiley, Chichester, UK, 1989).
23. Jolliffe, I.T. *Principal Component Analysis* (Springer, New York, USA, 1986).
24. MacFie, H.J.H., Gutteridge, C.S. & Norris, J.R. Use of canonical variates in differentiation of bacteria by pyrolysis gas-liquid chromatography. *J. Gen. Microbiol.* **104**, 67–74 (1978).
25. Windig, W., Haverkamp, J. & Kistemaker, P.G. Interpretation of sets of pyrolysis mass spectra by discriminant analysis and graphical rotation. *Anal. Chem.* **55**, 81–88 (1983).
26. Manly, B.F.J. *Multivariate Statistical Methods: A Primer* (Chapman and Hall, London, UK, 1994).
27. Goodacre, R. *et al.* Rapid identification of urinary tract infection bacteria using hyperspectral, whole organism fingerprinting and artificial neural networks. *Microbiology* **144**, 1157–1170 (1998).
28. Kell, D.B. Defence against the flood: a solution to the data mining and predictive modelling challenges of today. *Bioinformat. World* **1**, 16–18 (http://www.abergc.com/biwpp16-18_as_publ.pdf, 2002).