# PERSPECTIVES

# Metabolite profiling: from diagnostics to systems biology

*Alisdair R. Fernie, Richard N. Trethewey, Arno J. Krotzky and Lothar Willmitzer*

Abstract | The concept of metabolite profiling has been around for several decades, but only recent technical innovations have allowed metabolite profiling to be carried out on a large scale — with respect to both the number of metabolites measured and the number of experiments carried out. As a result, the power of metabolite profiling as a technology platform for diagnostics, and the research areas of gene-function analysis and systems biology, is now beginning to be fully realized.

Metabolite measurements have been carried out for decades because of the fundamental regulatory importance of metabolites as components of biochemical pathways, the importance of certain metabolites in the human diet and their use as diagnostic markers for a wide range of biological conditions, including disease and response to chemical treatment. Historically, the measurement of metabolites has mostly been achieved by spectrophotometric assays that can detect single metabolites, or by simple chromatographic separation of mixtures of low complexity. Over the past decade, however, methods that offer both high accuracy and sensitivity for the measurement of highly complex mixtures of compounds have been established.

Reflecting differences in metabolite coverage, accuracy and instrumentation, several terms have been adopted to describe these methods — ranging from metabolic fingerprinting and metabolite profiling, to metabolomics and metabonomics. For the

purposes of this article, we prefer the term metabolite profiling, because we contend that it has the widest applicability. We will define the currently used metabolite-profiling technologies and detail the main differences between these technologies. We will also highlight what, in our opinion, are the main technical problems posed in metabolite analyses, and discuss current protocols for data extraction and evaluation. Finally, we will describe some key biological questions that metabolite profiling has helped address so far in the areas of diagnostics, gene-function analysis and systems biology.

**The technical status quo**
The technologies that are used for metabolite profiling, and their respective advantages and limitations, are similar across the breadth of the biological sciences and have been covered extensively in several recent reviews (for example, see REFS 1–3). Therefore, we discuss only the main differences between the technologies and the challenges that lie ahead. Fundamentally, there are two different types of metabolic-profiling approaches: mass-spectrometry and NMR methodologies (BOXES 1,2).

*Technologies compared.* Gas-chromatography–mass-spectrometry (GC–MS), gas-chromatography–time-of-flight–mass-spectrometry (GC–TOF–MS) and liquid-chromatography–mass-spectrometry (LC–MS) are currently the principal mass-spectrometry methods for metabolite analysis. GC–MS technologies allow the

identification and robust quantification of a few hundred metabolites within a single extract[4–6]. The main advantages of this instrumentation stem from the fact that it has long been used for metabolite profiling. Therefore, there are stable protocols for machine set-up and maintenance, sample preparation and analysis, and chromatogram evaluation and interpretation. Recently, GC–TOF–MS technology has emerged, which offers several advantages over QUADRUPOLE TECHNOLOGY (see Glossary): notably, fast scan times, which give rise to either improved deconvolution or reduced run times for complex mixtures, and a higher mass accuracy.

Compared with the gas-chromatography technologies mentioned above, LC–MS offers several distinct advantages, but this newer technology has to be applied with considerable caution. Whereas gas-chromatography-based approaches can only be used with volatile compounds or compounds that can be rendered volatile by derivitization, LC–MS can be adapted to a wider array of molecules, including a range of secondary metabolites such as alkaloids, flavonoids, glucosinolates, isoprenes, oxylipins, phenylpropanoids, pigments and saponins[7–9]. In addition, these technologies are also beginning to find routine applications in drug detection[10,11]. That said, LC–MS in its most common form uses ELECTROSPRAY IONIZATION, which is prone to ION SUPPRESSION when carried out on complex mixtures, and therefore it is imperative that methods are subjected to rigorous validation before quantitative data can be interpreted.

Two recent mass-spectrometry approaches are Fourier-transform-ion-cyclotron-resonance–mass-spectrometry (FTICR–MS) and capillary-electrophoresis–mass-spectrometry (CE–MS). The first of these relies solely on very high-resolution mass analysis, which, potentially, allows the measurement of the empirical formula for thousands of metabolites. However, it is currently limited in two respects. The lack of chromatographic separation renders it incapable of discriminating between isomers, which, given the functional

Box 1 | **Metabolite profiling by mass spectrometry**
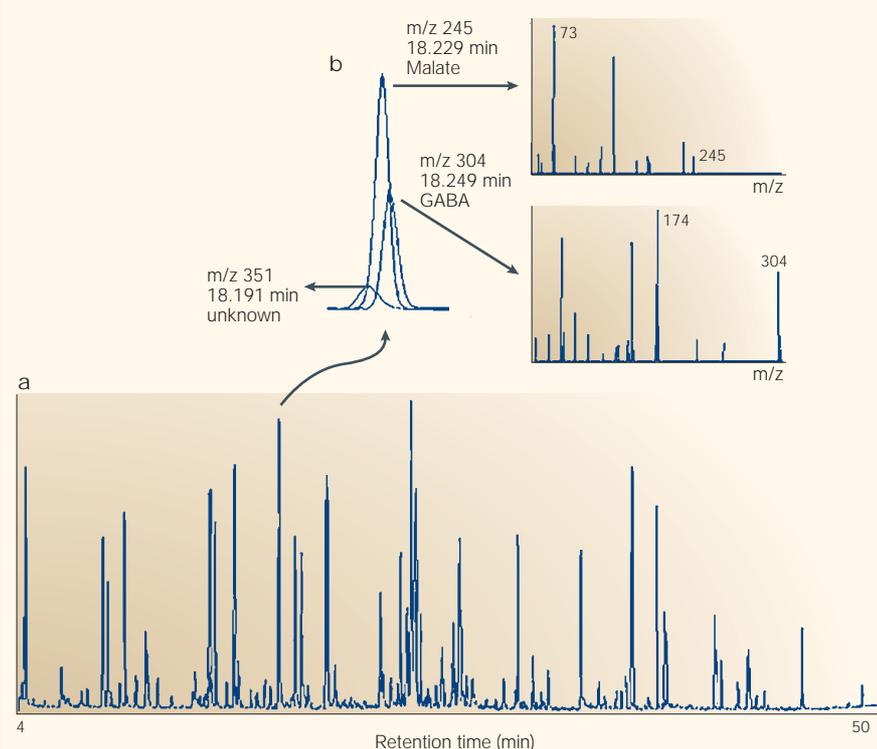


Mass-spectrometry-based metabolite profiling intrinsically generates large and complex datasets. A typical gas-chromatography–mass-spectrometry (GC–MS) profile of a biological sample contains mass-intensity data from 300–500 analytes in a file of around 20 megabytes. So, it has been a considerable challenge to develop software and algorithms that allow the extraction and quantification of many individual component analytes. Good software has long been widely available for finding anticipated analytes within a chromatographic run and for providing a good integration of the signal. However, there has been recent interest in the application of deconvolution algorithms that provide an electronic separation of component peaks without prior knowledge of these components. The most widely used algorithm in this respect is AMDIS ('automated mass spectral deconvolution and identification system'; REF. 48, see online links box), and recent studies have looked specifically at deconvolution strategies in the context of metabolite-profiling applications[49]. The figure shows a total ion chromatogram of the polar phase of an *Arabidopsis thaliana* extract (panel **a**), and the deconvoluted spectra of a single peak of this chromatogram reveals three metabolite signals that can be clearly discriminated (panel **b**). As metabolite profiling normally requires the assessment of sample and control data from multiple chromatograms, it is also necessary to develop software that ensures that the correct peak is identified and integrated in each chromatogram[50]. In addition, high-throughput operations require software that ensures that the data generated by different machines are comparable. Although steps are being taken to tackle these issues (see REF. 51), progress will probably be made through a wider application of chemometric methods[52]. GABA, γ-aminobutyric acid; m/z, ratio of mass to charge. Figure modified, with permission, from *Nature Biotechnology* REF. 5 © (2000) Macmillan Magazines Ltd.

importance of isomers in biology, is an important hindrance. In addition, although this technology is widely discussed, there is, at present, only a single documented report, which failed to provide a robust validation of the methodology[12]. More data are available for CE–MS, which is a highly sensitive methodology that can detect low-abundance metabolites and that provides good analyte separation. CE–MS methods have been rigorously

validated and shown to provide a rich source of metabolic data on over 1,000 metabolites from *Bacillus subtilis* extracts[13].

The main alternative to mass-spectrometry-based approaches for metabolic profiling is provided by NMR approaches. These methods offer the advantage that, because they have been used for years, they have been subjected to intensive validation, and CHEMOMETRICS that are associated with NMR

instrumentation are consequently also well developed. Furthermore, despite limitations in sensitivity and, as a result, severe restrictions in metabolic coverage, the use of NMR is clearly advantageous for certain biological questions (see BOX 2).

*Challenges ahead.* At present, even the combination of a wide range of analytical tools allows us to see only a portion of the total metabolite complement of the cell (FIG. 1). This is largely due to the diversity of cellular metabolites. *Escherichia coli* are estimated to contain ~750 metabolites[14], whereas the total number of metabolites that are present in the plant kingdom has been estimated to be of the order of hundreds of thousands[15]. Current estimates vary greatly, but it seems likely that a typical eukaryotic organism contains between 4,000 and 20,000 metabolites.

However, it is not merely a problem of numbers — the diversity itself yields great complexity because, unlike RNA or proteins, which are composites of highly similar small molecules, the physical and chemical properties of metabolites are highly divergent. This divergence means that there is no single extraction process that does not incur substantial loss to some of the cellular metabolites, let alone a single analytical platform that can measure all metabolites. So, one main problem for the analysis of metabolites is the lack of a totally comprehensive approach. A second problem that needs to be addressed is the high proportion of unknown analytes that is measured in metabolite profiling. Typically, in current GC–MS-based metabolite profiling, a chemical structure has been unambiguously assigned in only 20–30% of the analytes detected[16]. The solutions to both problems will not be rapid and will require successive iterative improvement. Sample pre-processing, such as solid-phase extraction (for example, see REF. 17), and the full integration of a wider range of chemistry-specific measurement platforms could improve metabolite coverage.

In the longer term, new technologies might ameliorate the problem of limited metabolite coverage. However, the level of complexity and other factors, such as the dynamic measurement range that is required, make the development of a single analytical platform that will provide a comprehensive analysis of the cellular metabolome a formidable challenge. The problem of chemical (and biochemical) identification of detected analytes also represents an arduous task. Again, one approach to tackle this problem would be improved instrumentation and protocols. An obvious example of such an

approach would be the coupling of chromatography with FTICR–MS. In addition, improvements in multidimensional mass spectrometry and the combination of NMR with mass-spectrometry technologies are likely to substantially increase metabolite coverage. A second approach is extensive fraction collection and empirical structure determination by NMR. Such work will need to be supported by the establishment of standard operating procedures and database resources that are available to the academic community.
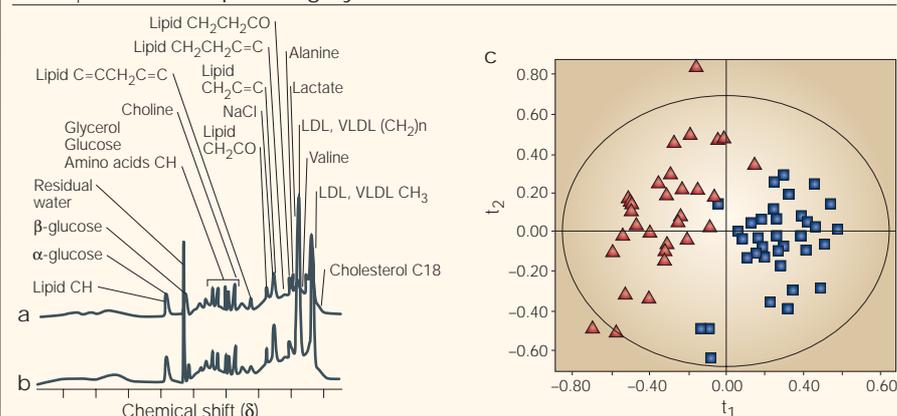
## Quantitative analyses

Although measuring the presence or absence of a metabolite in a biological sample is in some cases sufficient — for example, in drug detection or in the SUBSTANTIAL EQUIVALENCE TESTING of foodstuffs[17,18] — quantitative analysis is often necessary. As is common practice in transcript and protein profiling, metabolite profiles are often expressed as ratio data compared with a control sample. However, there are also many applications where absolute quantification is required. This is important to allow comparison between different sample types; for example, those that originate from different tissues within an organism. In addition, absolute quantification is important for understanding metabolic networks, as it is necessary for the calculation of atomic balances or for using kinetic properties of enzymes to develop predictive models.

Irrespective of whether data are represented as ratios or as absolute values, there are significant chemical, biological and technical issues associated with achieving a validated quantitative dataset. Great care has to be taken in experimental design and sampling. The chemical and technical methods that are used need to be fully appraised for the intrinsic variability of the system[16,17] and, ideally, the limits of detection and limits of quantification of each metabolite should be known. This process can be facilitated by the use of standard reference substances, and would ideally involve a full complement of isotopically labelled metabolites. Because of the interdisciplinary nature of these problems, they tend to be underestimated. It will be a challenge for this emerging technology to establish standard operating procedures to avoid the publication of flawed datasets.

Once quantitative datasets have been obtained, there is a wide range of data-analysis strategies that can be pursued with metabolite profiles[4,16,19–22]. The fundamental approach is, of course, to compare the level of a metabolite between an experimental and a control sample, and to use standard statistics to assess the significance of any differences. These approaches can then be extended by one dimension to look at the correlative behaviour of individual metabolites across different samples[4], and such correlative behaviour can be built into networks[23]. However, much recent interest has been focused on grouping approaches for whole profiles, and both SUPERVISED and UNSUPERVISED methods have been used[4,5,22,23]. Working with such large data sets, a direct visualization of the results onto pathways is highly advantageous to support the interpretation of the data[6–8] (see online links box).

## Applications of metabolite profiling

Improvements in metabolite-profiling technologies have already been effective in addressing biological problems. Until recently, the main applications have been in the area of diagnostics, but now the first applications of metabolite profiling to functional genomics and systems biology have been reported.

*Diagnostics.* Metabolic profiles have been widely used, in conjunction with statistical tools, for diagnosis. One of the first examples of the use of metabolite profiling in diagnostics was the estimation of modes of action of various herbicides[19]. The authors carried out gas-chromatography profiling of barley seeds that had been treated with sublethal doses of various herbicides, and compared the profiles obtained with those of untreated plants. This approach could discriminate between known acetyl CoA carboxylase (ACC) and acetolactate synthase (ALS) inhibitors on the basis of a visual comparison of metabolite profiles of plants treated with these herbicides. This indicates that the mode of action of bioregulators might be diagnosed by the analysis of reflective patterns of metabolite composition. Recently, a more sophisticated analysis of the mode of action has been carried out by Ott and colleagues, who used a combination of
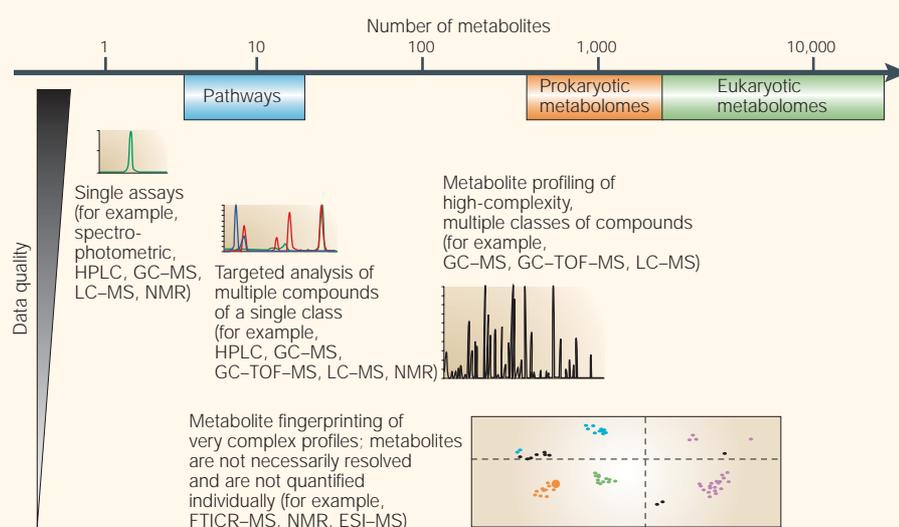
---

### Box 2 | Metabolite profiling by NMR



Despite recent improvements in instrumentation, NMR technologies mostly have a lower sensitivity compared with mass-spectrometry methodologies. However, computational analyses and the chemometric software associated with NMR are highly developed. Furthermore, NMR is non-invasive, which can be a particularly important advantage in certain situations. Although recent methodologies that couple NMR to liquid chromatography and solid-phase extraction columns to improve separation and sensitivity have been reported[52,53], it is unlikely that NMR will ever attain the sensitivity of mass spectrometry. NMR is, however, important in the unequivocal determination of metabolite structures[54] and, perhaps owing to its non-invasive nature, is more commonly used in mammalian systems than mass-spectrometry technologies.

Perhaps the most powerful example so far of the use of NMR for metabolic profiling is provided by the work of the Nicholson group, who acquired and evaluated proton NMR ($^1$H-NMR) spectra of human serum[21]. They were able to distinguish individuals with coronary heart disease from healthy individuals, and even to assess the severity of the disease. The 600-MHz $^1$H-NMR spectra of a healthy patient (panel **a**) and a patient with severe artheriosclerosis (panel **b**) are shown. Partial least squares discriminant analysis of the obtained spectra shows the difference between healthy (triangle) and diseased (square) individuals (panel **c**). This technique is therefore capable of rapid, non-invasive diagnosis of coronary heart disease, and could be used in population screening or in the testing of ameliorative drug treatments. It is currently undergoing clinical evaluation. $t_1$ and $t_2$ represent the combined variance observed across many metabolites (with $t_1$ representing the most variable metabolites). These values are used to discriminate between biological samples using analyses that are fundamentally similar to principal component analyses (see Glossary). LDL, low-density lipoprotein; VLDL, very-low-density lipoprotein. Figure modified, with permission, from *Nature Medicine* REF. 21 © (2002) Macmillan Magazines Ltd.

---

**Figure 1 | The trade-off between metabolic coverage and the quality of metabolic analysis.** The most sensitive and precise analyses are typically those for single metabolites. Targeted methods for metabolite analysis provide high-quality data on a single class of compounds using dedicated and optimized methods. Broad metabolite profiling provides data for a wide range of chemical classes, but the methods represent a compromise and do not provide the same data quality for all of the metabolites covered. Metabolite-fingerprinting approaches also offer the widest coverage, but profiles, and not metabolite levels, are compared by these approaches, and the methodologies can be subject to artefacts that undermine the quantitative assessment of the data. None of the methodologies alone are sufficient to approach full coverage of a complete metabolome. Data quality encompasses overall sensitivity, accuracy, precision of quantification and metabolite identification rates. ESI–MS, electrospray-ionization–mass-spectrometry; FTICR–MS, Fourier-transform-ion-cyclotron-resonance–mass-spectrometry; GC–MS, gas-chromatography–mass-spectrometry; GC–TOF–MS, gas-chromatography–time-of-flight–mass-spectrometry; HPLC, high-performance liquid chromatography; LC–MS, liquid-chromatography–mass-spectrometry.

NMR-based metabolic profiling and bioinformatics to classify herbicides with unknown modes of action[24]. The statistical analysis of proton NMR (^{1}H-NMR) spectra has also been effectively used for non-invasive diagnosis of coronary disease[23] (BOX 2).

The use of two statistical tools, HIERARCHICAL CLUSTER ANALYSIS (HCA) and PRINCIPAL COMPONENT ANALYSIS (PCA; both reviewed in REF. 25), in the diagnostic interpretation of metabolite profiling data sets is now widespread. In the plant field, they have been used for the analysis of four different *Arabidopsis thaliana* genotypes[5] and potato genotypes[4]. Analysis of the *dgd1* and *sdd1-1* mutants of *A. thaliana*, and their respective wild-type ECOTYPES, showed that there were greater differences between the wild-type ecotypes than those observed between the mutants and their corresponding wild-type controls[5]. These examples, however, represent a much greater range of metabolite-profiling-based diagnostic approaches. Other examples include NMR studies of the comparative biochemistry of three wild rodents[26], Fourier-transform infrared spectroscopy (FT–IRS) and direct-injection-electrospray-mass-spectrometry analysis of metabolites that are secreted into the growth media of bacterial[27] or yeast mutants[20]. All these studies

were carried out to assess whether genetically divergent individuals could be distinguished from one another on the basis of their metabolite profiles. Another example is the use of mass spectrometry to screen human blood samples for inborn errors of metabolism[28]. Taken together, metabolite profiling clearly has important applications in the diagnostic characterization and classification of different genetic or environmental conditions.

***Annotation of gene function.*** Metabolite profiling provides direct functional information on metabolic phenotypes and indirect functional information on a range of phenotypes that are determined by small molecules; for example, stress tolerance or disease manifestations. Given this, metabolite profiling has potential as a tool for functional genomics[5,20]. The power of *in vitro* biochemical screening for the assignment of an activity to a cDNA has been shown in a ground-breaking study by Martzen *et al.*[29]. These authors created 6,080 yeast strains, each harbouring a different transgene that encoded a fusion protein, and following expression and purification they subjected each protein to a range of biochemical assays for classification purposes.

More challenging is the use of metabolite profiling to explore gene function *in vivo*[30]. It has recently been shown that the use of co-response analysis in yeast (that is, the quantification of the change of several metabolite concentrations relative to the concentration change of one selected metabolite) can reveal the site of action of a gene. In addition, the comparison of metabolite concentrations in known mutants with those in control strains can allow functional gene assignment[31]. It should be noted here, however, that the analysis of the metabolites is carried out on extracts that represent snap-shots of the metabolic status of the organism under investigation. Elegant studies using proteins that were engineered to fluoresce at levels that are proportional to the local concentrations of a target metabolite[32] show that *in vivo* quantification of metabolites is possible. However, such studies are time consuming and laborious. A more practical candidate technology might be magnetic resonance imaging (MRI). However, this is currently also restricted in scope, with only a handful of measurable metabolites. In any case, the multiparallel approach of profiling methods provides an immediate insight into the behaviour of the whole metabolic network after modulation of a particular gene function. This provides exciting opportunities for defining gene function at the level of metabolic networks and the overall phenotype in the context of a particular organism.

Gain-of-function analysis is emerging as a particularly powerful approach for functional genomics. For example, the analysis of a gene of known function, a member of the threonine aldolase family, that was introduced into *A. thaliana* both confirmed the expected function and revealed new effects on the metabolic network, including the upregulation of the methionine pathway (~2–4-fold increases in homoserine and methionine levels) and the downregulation of the isoleucine pathway (with isoleucine decreasing to 15% of wild-type levels) (R.N.T. and A.J.K, unpublished results). Furthermore, solely on the basis of the metabolic composition, functions can be associated with so far unannotated open reading frames (R.N.T. and A.J.K, unpublished results). In many cases, the effects of overexpression or mutation of a gene can be quite pleiotropic: for example, in the case of the *dgd1* mutant mentioned above[5], there were significant changes in half of the metabolites analysed. Such complex responses involve interactions among metabolic components and interactions between the metabolic network and the mechanisms of gene and protein expression. However, our

knowledge of these interactions is limited at present. It is, however, important that these network behaviours are well understood if the application of functional-genomics information in metabolic-engineering approaches is to be successful (for detailed reviews, see REFS 33,34).

One of the key aspects of metabolite profiling is that it is a technology that can be used in a high-throughput operation. Of all the genomics technologies, metabolite profiling offers one of the best combinations of practical performance and cost per sample. The expression of almost every gene from the yeast and *E. coli* genomes in *A. thaliana*, and subsequent metabolite profiling with GC–MS and liquid-chromatography–tandem-mass-spectrometry (LC–MS/MS) has recently been achieved (FIG. 2). This approach is deliberately non-biased, with respect to both the choice of gene and the metabolites measured, because of the two objectives; that is, to explore gene function at the level of protein activity, and to explore the consequences of introducing a new protein into the metabolic network. Similar genomic approaches such as the RNA-interference-based systematic functional analysis of the *Caenorhabditis elegans* genome[35] will facilitate such large-scale metabolic studies in other species.

The non-biased approach to discovering gene function can be productive. Experience shows that a statistically significant change in the steady-state level of any given metabolite will be triggered by overexpression of 0.1–1.0% of the genes in a genome (conclusions from work represented in FIG. 2; R.N.T. and A.J.K, unpublished results). In some cases, the genes will influence flux directly in a pathway (the theory of metabolic-flux-control analysis implies that small changes in enzyme activity can often lead to large changes in metabolite concentration[36]). In other cases, they might trigger a host of regulatory changes that alter the atomic partitioning or the activity of metabolic networks.

The approach of focusing on individual genes can also be extended to exploring the phenotypic relevance of genome regions. Recently, GC–MS profiling of breeding populations of tomato, wherein genomic segments from the wild tomato species *Lycopersicon pennellii* have been introgressed into the elite cultivated species *L. esculentum*, has been initiated[37]. Once complete, this profiling should allow the identification of several environmentally stable quantitative trait loci and, ultimately, through the study of progressively smaller recombinant introgressions and, finally, map-based cloning, will lead to the identification of genes that
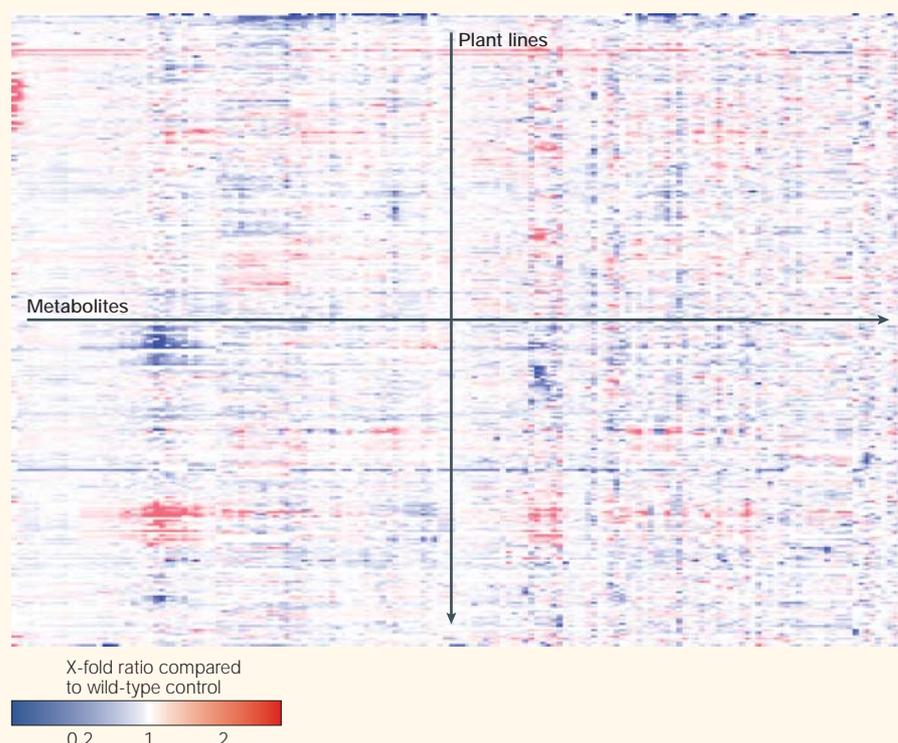


Figure 2 | **Overexpression and metabolite profiling at the transgenomic level.** An example of a heat map of the metabolite profiles of the leaves of around 19,000 mature plants including plant lines that each overexpress essentially every gene of the yeast genome (R.N.T. and A.J.K., unpublished results). Most of this map is white, which reflects the fact that overexpression does not result in a change in metabolite content compared with control plants in most cases. Regions of red or blue indicate that the metabolite content is either increased or decreased, respectively, following overexpression. The colour scale is nonlinear and the maximum increases and decreases detected are around 100-fold. A total of 158 metabolites that have derived from both gas-chromatography–mass-spectrometry (GC–MS) and liquid-chromatography–tandem-mass-spectrometry (LC–MS/MS) analyses are shown; the chemical identity is known for around 60% of them. The chemical classes covered include amino acids, organic acids, sugars, sugar alcohols, vitamins and pigments. Although the individual metabolite columns can be visually distinguished, the pixel resolution of the image is not sufficient to distinguish the rows (which represent the plants and plant lines). The software that is used to generate the image uses smoothing algorithms to circumvent this limitation. Such datasets provide a rich resource for the identification of novel gene–function relationships and provide a foundation for systems-biology approaches.

regulate metabolite content in a species of nutritional significance.

*Systems biology.* Both theoretical and experimental disciplines have seen the emergence of systems-based approaches to biology in the past few years[38–40]. Although historically, the term 'systems biology' was applied exclusively to mathematical-modelling strategies (for example, see REF. 40), it is now more widely used, particularly in genomics. Systems-biology studies are typified by a shift from the more traditional reductionist approach towards more holistic approaches, with experimental strategies aimed at understanding interactions across multiple molecular entities. Whereas most of such approaches have focused on transcript and/or protein levels[41,42], they are also beginning to include metabolite- and pharmaceutical-based approaches[43–45].

In addition to the incorporation of metabolite profiling in systems-biology initiatives, the measurement of many metabolites in parallel gives insights into the complex regulatory circuits that underpin metabolism. Initial systems-based approaches that used data from metabolite profiling involved comprehensive correlation analyses between all metabolites that were profiled in potato tubers. These studies showed that most metabolites had little correlation, although some were tightly co-regulated and others were nonlinearly related, perhaps indicating that the metabolites involved were linked by an enzyme that is subject to strong metabolic regulation[4].

Since this initial study, more sophisticated metabolic-network analyses have been undertaken by plotting networks in which the metabolites are represented by nodes that are interconnected by lines that represent the correlative behaviour between the metabolites[44].

**CHEMOMETRIC**
The application of statistical and computer methods to data analysis in chemistry and related scientific fields.

**ECOTYPE**
The smallest taxonomic subdivision of an ecospecies, which consists of populations that have adapted to a particular set of environmental conditions.

**ELECTROSPRAY IONIZATION**
To analyse compounds effectively by mass spectrometry, they must be ionized in the gas phase. Electrospray is the most widely used atmospheric-pressure ionization technique for the sensitive, comprehensive analysis of polar and ionic compounds. Using electrospray, a strong electric field is applied to the liquid sample stream, which is then nebulized and desolvated with the assistance of a high-temperature gas flow to produce gas-phase ions.

**HIERARCHICAL CLUSTER ANALYSIS**
An agglomerative statistical method that finds clusters of observations within a data set. It allows the grouping of individuals on the basis of the similarity in their properties.

**ION SUPPRESSION**
The common term that is given to a range of phenomena that can occur during the ionization of complex mixtures. An important component of this is the competition between co-eluting compounds for ionization energy, which can lead to varying degrees of ionization of any individual compounds.

**PRINCIPAL COMPONENT ANALYSIS**
A statistical tool in which an orthogonal coordinate system, with axes that are ordered in terms of the amount of variance in a dataset, is produced. This allows the separation of individuals on the basis of differences in their properties and can also be used to evaluate the properties that contribute the most to these separations.

**QUADRUPOLE TECHNOLOGY**
A quadrupole mass filter consists of four parallel metal rods. Two opposite rods have a DC voltage and the other two have an AC voltage. The applied voltages affect the trajectory of ions that travel down the flight path that is centred between the four rods, such that only ions of a certain mass-to-charge ratio pass through the quadrupole filter and all other ions are thrown out of their original path. A mass spectrum is obtained by monitoring the ions that pass through the quadrupole filter as the voltages on the rods are varied.

**SUBSTANTIAL EQUIVALENCE TESTING**
The concept of substantial equivalence embodies the idea that organisms that are used as foods or as food sources can serve as a basis for comparison when assessing the safety of human consumption of a food or food component that has been modified or is new.

**SUPERVISED GROUPING APPROACH**
A method that requires training with known data sets in which the types of groups expected are predefined before being applied to experimental data.

**UNSUPERVISED GROUPING APPROACH**
A method that does not require training with known data sets and that generates groups on the basis of the data structure in the experimental data.

---

In such studies, the actions of a small theoretical metabolic network can be studied on the basis of the strength of correlations between the metabolites that constitute these networks. This approach could therefore be used as an extension to the co-response method mentioned above. However, caution should be exercised here, as this field is still in its infancy and a lack of understanding of protein and transcript networks could lead to premature conclusions. A more comprehensive approach is to measure metabolites, proteins and/or mRNA from the same sample and to assess connectivity across different molecular entities. Preliminary studies have shown that discrimination analysis of *A. thaliana* ecotypes is possible using a combination of protein and metabolite levels[44]. Also, metabolite–transcript correlations from large data sets collected throughout development in wild-type and transgenic tubers engineered to have enhanced sucrose metabolism allows the identification of candidate genes for metabolic engineering[45]. In the latter case, the transcript levels of approximately 280 transcripts that showed reproducible changes with respect to control samples were systematically plotted against changes in metabolite levels of paired samples. A total of 517 out of the 26,616 possible pairs showed significant correlation (at the $P < 0.01$ level). Although some of these correlations were already known, most of the significant correlations were new and included the identification of several strong correlations between genes and nutritionally important metabolites. In the medical field, metabolite data are also beginning to be used within the broader context of systems biology (for example, see REFS 11,43). In addition, transcript profiling has been used in parallel with the metabolite profiling of a limited number of metabolites to aid the metabolic engineering of the production of medicinally important polyketides in *Aspergillus terreus*[46]. Further work is required before mechanistic links can be identified on the basis of such data sets. This would then facilitate the use of profiling data sets in more sophisticated network analyses.

### Concluding remarks

With respect to the technology of metabolite profiling, the number of metabolites that can be detected and quantified with current technologies represents the main limitation. In analogy to the isotope-coded affinity tag (ICAT) methodology from proteomics[47], the availability of a full complement of isotopically labelled standard substances could greatly aid metabolite quantification. In addition, further progress is required to determine the chemical identity of peaks that can be determined with metabolite-profiling methods.

The use of metabolic profiling as a diagnostic tool is, to a large extent, independent of the aforementioned limitations and, as a result, the field has developed more rapidly. By contrast, the application of metabolite profiling to gene-function analysis and the development of systems biology depends, to a large extent, on technological improvements. Given the fact that the phenotype of any biological system is largely determined by its metabolite composition, the future development of metabolite-profiling technologies is of crucial importance to biomedical research.

*Alisdair R. Fernie and Lothar Willmitzer are at the Department of Molecular Physiology, Max-Planck-Institute for Molecular Plant Physiology, Am Mühlenberg 1, 14476 Golm, Germany.*

*Richard N. Trethewey and Arno J. Krotzky are at metanomics GmbH and Co KGaA, and metanomics Health GmbH, Tegeler Weg 33, 10589 Berlin, Germany.*

*Correspondence to A.R.F. e-mail: fernie@mpimp-golm.mpg.de*
doi:10.1038/nrm1451

1. Harrigan, G. G. & Goodacre, R. (eds). *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Functional Analysis* (Kluwer Academic, Boston, 2003).
2. Fiehn, O. Metabolomics. The link between genotype and phenotype. *Plant Mol. Biol.* **48**, 155–171 (2002).
3. Kell, D. B. Metabolomics and systems biology: making sense of the soup. *Curr. Opin. Microbiol.* **7**, 296–307 (2004).
4. Roessner, U. *et al.* Metabolite profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* **13**, 11–29 (2001).
5. Fiehn, O. *et al.* Metabolite profiling for plant functional genomics. *Nature Biotech.* **18**, 1157–1161 (2000).
6. Halket, J. M. *et al.* Deconvolution gas chromatography mass spectrometry of urinary organic acids. Potential for pattern recognition and automated identification of metabolic disorders. *Rapid Commun. Mass Spectrom.* **13**, 279–284 (2003).
7. Aharoni, A. *et al.* Terpenoid metabolism in wild type and transgenic *Arabidopsis* plants. *Plant Cell* **15**, 2866–2884 (2003).
8. Swart, P. J. *et al.* HPLC-UV atmospheric-pressure ionisation mass-spectrometric determination of the dopamine-D2 agonist N-0923 and its major metabolites after oxidative metabolism by rat liver, monkey liver and human liver microsomes. *Toxicology Methods* **3**, 279–290 (1993).
9. Matuszewski, B. K., Constanzer, M. L. & Chavez-Eng, C. M. Strategies for the assessment of matrix effect in quantitative bioanalytical methods based on HPLC-MS/MS. *Anal. Chem.* **75**, 3019–3030 (2003).
10. Plumb, R. S. *et al.* Use of liquid chromatography/time-of-flight mass spectrometry and multivariate statistical analysis shows promise for the detection of drug metabolites in biological fluids. *Rapid Commun. Mass Spectrom.* **17**, 2632–2638 (2003).
11. Watkins, S. M. & German, J. B. Metabolomics and biochemical profiling in drug discovery and development. *Curr. Opin. Mol. Ther.* **4**, 224–228 (2002).
12. Aharoni, A. *et al.* Nontargeted metabolome analysis by use of Fourier transform ion cyclotron mass spectrometry. *OMICS* **6**, 217–234 (2002).

13. Soga, T. *et al.* Quantitative metabolome analysis using capillary electrophoresis mass spectrometry. *J. Proteome Res.* **2**, 488–494 (2003).
14. Nobeli, I., Krissinel, E. B. & Thornton, J. M. B. A structure-based anatomy of the *E. coli* metabolome. *J. Mol. Biol.* **334**, 697–719 (2003).
15. Hall, R. *et al.* Plant metabolomics: the missing link in functional genomics strategies. *Plant Cell* **14**, 1437–1440 (2002).
16. Roessner-Tunali, U. *et al.* Metabolic profiling of transgenic tomato plants overexpressing hexokinase reveals that the influence of hexose phosphorylation diminishes during fruit development. *Plant Physiol.* **133**, 84–99 (2003).
17. Walles, M. *et al.* Verapamil drug metabolism studies by automated in-tube solid phase microextraction. *J. Pharma. Biomed. Anal.* **30**, 307–319 (2002).
18. Kok, E. J. & Kuiper, H. A. Comparative safety assessment for biotech crops. *Trends Biotech.* **21**, 438–444 (2003).
19. Sauter, H., Lauer, M. & Fritsch, H. Metabolite profiling of plants — a new diagnostic technique. *Abstr. Pap. Am. Chem. Soc.* **195**, 129 (1988).
20. Allen, J. *et al.* High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nature Biotech.* **21**, 692–696 (2003).
21. Brindle, J. T. *et al.* Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1H-NMR-based metabonomics. *Nature Med.* **8**, 1439–1444 (2002).
22. Huhman, D. V. & Sumner, L. W. Metabolic profiling of saponins in *Medicago sativa* and *Medicago trunculata using* HPLC coupled to an electrospray ion-trap mass spectrometer. *Phytochemistry* **59**, 347–360 (2002).
23. Kose, F., Weckwerth, W., Linke, T. & Fiehn, O. Visualizing plant metabolomic correlation networks using clique-metabolite matrices. *Bioinformatics* **17**, 1198–1208 (2001).
24. 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).
25. Quackenbush, J. Computational analysis of microarray data. *Nature Rev. Genet.* **2**, 418–427 (2001).
26. Griffin, J. L. *et al.* NMR spectroscopy based metabonomic studies on the comparative biochemistry of the kidney and urine of the bank vole (*Clethrionomys glareolus*), wood mouse (*Apodemus sylvaticus*), white toothed shrew (*Crocidura suaveolens*) and the laboratory rat. *Comp. Biochem. Physiol. B* **127**, 357–367 (2000).
27. Kaderbhai, N. N., Broadhurst, D. I., Ellis, D. I., Goodacre, R. & Kell, D. B. Functional genomics via metabolic footprinting: monitoring metabolite secretion by *Escherichia coli* tryptophan metabolism mutants using FT-IR and direct injection electrospray mass spectrometry. *Comp. Funct. Genomics* **4**, 376–391 (2003).

28. Rashed, M. S. *et al.* Screening blood spots for inborn errors of metabolism by electrospray tandem mass spectrometry with a microplate batch process and a computer algorithm from automated flagging of abnormal profiles. *Clin. Chem.* **43**, 1129–1141 (1997).
29. Martzen, M. R. *et al.* A biochemical genomics approach for identifying genes by the activity of their products. *Science* **286**, 1153–1155 (1999).
30. Trethewey, R. N., Krotzky, A. J. & Willmitzer, L. Metabolic profiling: a Rosetta stone for genomics? *Curr. Opin. Plant Biol.* **2**, 83–85 (1999).
31. Raamsdonk, L. M. *et al.* A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature Biotech.* **19**, 45–50 (2001).
32. Fehr, M., Lalonde, S., Lager, I., Wolff, M. W. & Frommer, W. B. *In vivo* imaging of the dynamics of glucose uptake in the cytosol of COS-7 cells by fluorescent nanosensors. *J. Biol. Chem.* **278**, 19127–19133 (2003).
33. Barabasi, A. L. & Oltvai, Z. N. Network biology: understanding the cell's functional organisation. *Nature Rev. Genet.* **5**, 101–113 (2004).
34. Wagner, A. & Fell, D. A. The small world inside large metabolic networks. *Proc. R. Soc. Lond. B* **268**, 1803–1810 (2001).
35. Kamath, R. S. *et al.* Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421**, 231–237 (2003).
36. Kacser, H. & Burns, J. A. The control of flux. *Symposia Soc. Exp. Biol.* **28**, 65–104 (1974).
37. Fernie, A. R. *et al.* Metabolic profiling at the genome level. *Plant Animal Genome Abstr.* **XI**, W307 (2003).
38. Kitano, H. Perspectives on systems biology. *New Generation Comput.* **18**, 199–216 (2000).
39. Ideker, T., Galitski, T. & Hood, L. A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.* **2**, 343–372 (2001).
40. Edwards, J. S., Ibarra, R. U. & Palsson, B. O. *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nature Biotech.* **19**, 125–130 (1999).
41. Baliga, N. S. *et al.* Coordinate regulation of energy transduction modules in *Halobacterium* sp. analyzed by a global systems approach. *Proc. Natl Acad. Sci. USA* **99**, 14913–14918 (2002).
42. Davidson E. H. *et al.* A genomic regulatory network for development. *Science* **295**, 1669–1678 (2002).
43. Nicholson, J. K. & Wilson I. D. Understanding 'global' systems biology: metabonomics and the continuum of metabolism. *Nature Rev. Drug Discov.* **2**, 668–676 (2003).
44. Weckwerth, W. Metabolomics in systems biology. *Annu. Rev. Plant Biol.* **54**, 669–689 (2003).
45. Urbanczyk-Wochniak, E. *et al.* Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Reports* **4**, 989–993 (2003).
46. Askenazi, M. *et al.* Integrating transcriptional and metabolite profiles to direct the engineering of lovastatin-producing fungal strains. *Nature Biotech* **21**, 150–156 (2003).

47. Gygi, S. P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotech.* **17**, 994–999 (1999).
48. Stein, S. E. An integrated method for spectrum extraction and compound identification from GC/MS data. *J. Am. Soc. Mass Spectrom.* **10**, 770–781 (1999).
49. Wagner, C., Sefkow, M. & Kopka, J. Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. *Phytochemistry* **62**, 887–900 (2003).
50. Frenzel, T., Miller, A. & Engel, K. H. A methodology for automated comparative analysis of metabolite profiling data. *Eur. Food Res. Technol.* **216**, 335–342 (2003).
51. Duran, A. L., Yang, J., Wang, L. & Sumner, L. W. Metabolomics spectral formatting, alignment and conversion tools (MSFACTs). *Bioinformatics* **19**, 2283–2293 (2003).
52. Waisim, M., Hassan, M. S. & Brereton, R. G. Evaluation of chemometric methods for determining the number and position of components in high-performance liquid chromatography detected by diode array detector by diode array detector and on-flow 1H nuclear magnetic resonance spectroscopy. *Analyst* **128**, 1082–1090 (2003).
53. Lindon, J. C. HPLC-NMR-MS: past, present and future. *Drug Discov. Today* **8**, 1021–1022 (2003).
54. Meiler, J. & Will, M. Genius: a genetic algorithm for automated structure elucidation from 13C NMR spectra. *J. Am. Chem. Soc.* **124**, 1868–1870 (2002).

## Online links

**FURTHER INFORMATION**
AMDIS:
http://chemdata.nist.gov/mass-spc/amdis/overview.html
BRENDA enzyme database: www.brenda.uni-koeln.de
GenMAPP: www.genmapp.org
Kyoto Encyclopedia of Genes and Genomes:
http://www.genome.ad.jp/kegg/
MapMan: http://gabi.rzpd.de/projects/MapMan
Max-Planck-Institut für Molekulare Pflanzenphysiologie:
www.mpimp-golm.mpg.de
Metabolon: http://www.metabolon.com
metanomics: http://www.metanomics.de
Paradigm Genetics: http://www.paradigmgenetics.com
**Access to this interactive links box is free online.**