Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry

Ric CH De Vos[1,2,7], Sofia Moco[1–3,7], Arjen Lommen[1,2,4,7], Joost JB Keurentjes[2,5,6], Raoul J Bino[1,2,5] & Robert D Hall[1,2]

[1]Plant Research International, Wageningen University and Research Centre (Wageningen-UR), PO Box 16, 6700 AA Wageningen, The Netherlands. [2]Centre for BioSystems Genomics, PO Box 98, Wageningen, The Netherlands. [3]Laboratory of Biochemistry, Wageningen-UR, The Netherlands. [4]RIKILT, Institute for Food Safety, Wageningen-UR, The Netherlands. [5]Laboratory of Plant Physiology, Wageningen-UR, The Netherlands. [6]Laboratory of Plant Genetics, Wageningen-UR, The Netherlands. [7]These authors contributed equally to this work. Correspondence should be addressed to R.C.H.d.V. (ric.devos@wur.nl).

Published online 5 April 2007; doi:10.1038/nprot.2007.95

Untargeted metabolomics aims to gather information on as many metabolites as possible in biological systems by taking into account all information present in the data sets. Here we describe a detailed protocol for large-scale untargeted metabolomics of plant tissues, based on reversed phase liquid chromatography coupled to high-resolution mass spectrometry (LC-QTOF MS) of aqueous methanol extracts. Dedicated software, MetAlign, is used for automated baseline correction and alignment of all extracted mass peaks across all samples, producing detailed information on the relative abundance of thousands of mass signals representing hundreds of metabolites. Subsequent statistics and bioinformatics tools can be used to provide a detailed view on the differences and similarities between (groups of) samples or to link metabolomics data to other systems biology information, genetic markers and/or specific quality parameters. The complete procedure from metabolite extraction to assembly of a data matrix with aligned mass signal intensities takes about 6 days for 50 samples.

INTRODUCTION

Metabolomics has emerged as a valuable technology for the comprehensive profiling and comparison of metabolites in biological systems, and a multitude of applications for human, microbial and plant systems have already been reported or predicted[1–9]. Plants are especially rich in chemically diverse metabolites, which are usually present in a large range of concentrations, and no single analytical method is currently capable of extracting and detecting all metabolites. Over the past decade, several methods suitable for large-scale analysis and comparison of metabolites in plant extracts have been established[2,5], including gas chromatography coupled to mass spectrometry (GC-MS)[10–16], direct flow injection-mass spectrometry (DFI-MS)[17–20], liquid chromatography-mass spectrometry (LC-MS)[21–26], capillary electrophoresis-mass spectrometry (CE-MS)[27] and NMR technologies[28,29]. LC-MS-based approaches are expected to be of particular importance in plants, owing to the highly rich biochemistry of plants, which covers many semi-polar compounds, including key secondary metabolite groups, which can best be separated and detected by LC-MS approaches[2,5,22–24,30–32]. Of the many semi-polar compounds not involved in primary metabolism, several have already been shown to have phenotypic/physiological importance. It is also mainly secondary metabolites that are attracting much attention from health, food and nutrition groups[5,26,33,34] owing to, for example, their resistance effects, antioxidant properties, and color and flavor characteristics. These and other so-called quality aspects of plant materials are generally not centered on individual metabolites but rather are related to a particular (balanced?) mixture of compounds from diverse, biochemically related and unrelated groups. As such, a metabolomics approach to help better understand the complexity of these mixtures, the components of which play the most important role, and how their biosynthesis is controlled, is likely to be of great future value and importance.

Commonly used plant metabolomics approaches and their advantages and limitations

Although NMR is in principle the most uniform detection technique and is essential for the unequivocal identification of unknown compounds, NMR-based metabolomics approaches still suffer from a relatively low sensitivity compared with MS. As yet, MS-based platforms are most widely used in plant metabolomics[2]. GC coupled to electron impact time-of-flight (TOF) MS was the first approach used in large-scale plant metabolomics[16], and a detailed protocol for sample extraction, derivatization and subsequent data analyses has recently been described[12]. This approach covers a large variety of nonvolatile metabolites, mainly those involved in primary metabolism, including organic and amino acids, sugars, sugar alcohols, phosphorylated intermediates (in the polar fraction of extracts), as well as lipophilic compounds such as fatty acids and sterols (in the apolar fraction). GC-(TOF)MS produces highly reproducible separation and fragmentation patterns of metabolites, which enables the development of common GC-TOF MS-based metabolite libraries[15,35]. Although CE-MS also enables good separation and detection of many polar primary metabolites[27], it is seldom used compared with GC-TOF MS. As most primary metabolites have commercially available standard compounds, both GC-TOF MS and CE-MS can produce quantitative data for hundreds of compounds involved in central metabolism.

The preferred method for analyzing semi-polar metabolites is LC-MS with a soft ionization technique, such as electrospray ionization (ESI) or atmospheric pressure chemical ionization (APCI), resulting in protonated (in positive mode) or deprotonated (in negative mode) molecular masses. Compounds detectable by LC-MS include the large and often economically important group of plant secondary metabolites such as alkaloids, saponins, phenolic acids, phenylpropanoids, flavonoids, glucosinolates, polyamines and derivatives thereof[22,23,26,30]. These compounds can be

effectively extracted with aqueous alcohol solutions and directly analyzed without derivatization. Depending on the type of column used, various primary metabolites including several polar organic acids and amino acids can be reliably analyzed using LC-MS[36]. Based on the high mass resolution of TOF-MS and Fourier transform-ion cyclotron resonance-MS (FTMS) instruments, enabling calculations of elemental formulae of detected ions, rapid DFI-MS approaches without any prior compound separation have been developed to compare metabolite fingerprints of crude plant extracts[17–20]. However, such direct injection approaches, irrespective of the resolution and accuracy of the mass spectrometer, may suffer from significant adduct formation and ion suppression phenomena upon ionization of complete crude extracts. Moreover by definition, direct injection methods cannot discriminate between the many molecular isomers. Therefore, most MS-based platforms in plant metabolomics perform at least some separation. LC preceding MS not only results in the detection of isomeric compounds, which are often abundantly present in plants, but also enables valuable structural information to be collected online, for example, MS/MS fragmentation patterns and UV–Vis absorbance spectra using photodiode array (PDA) detection[22–24,26,30,32,36]. It has been estimated that extensive LC in combination with high-resolution MS (e.g., TOF-MS) enables the detection of several hundreds of compounds in a single crude plant extract[22,24,25]. With continually improving tools for data acquisition, processing and mining, LC-MS will certainly grow in value for biochemical profiling and metabolite identification. Combining LC with ultra-high-resolution mass spectrometry such as FTMS[31,37] and other identification tools like LC-NMR-MS[38–40], as well as making use of improved separation technologies such as ultra-performance LC (UPLC) coupled to MS[41,42], will further improve our potential to identify metabolites and to provide an even more detailed metabolite profile of plant extracts.

Untargeted LC-MS for plant metabolomics

Compared with primary metabolites, the number of commercially available standards for secondary metabolites per plant species or tissue is still very limited. Consequently, metabolomics approaches based on analyses of compounds for which standards are available, which is common practice in GC-(TOF)MS-based metabolomics studying primary metabolism, would very much limit the great potential of LC-MS in plant research. Recent developments in processing software for unbiased mass peak extraction and alignment of LC-MS data, such as MetAlign[22,25,43,44], XCMS[41,45], MZmine[46] and Markerlynx[47], now offer possibilities for more holistic untargeted metabolomics approaches, which aim to gather information on as many metabolites as possible in each extract analyzed. In such untargeted approaches, mass peak identification using standards is not the primary step in data processing. In contrast, all analytical information present in the profiles is first transformed into coordinates on the basis of mass, retention time and signal amplitude. These coordinates are then aligned across all samples. By applying appropriate statistical and multivariate analysis tools, differential mass peaks or mass peaks correlating with a specific trait can be filtered out and identified to some degree by using accurate mass, MS/MS fragmentation and then confirmed with standards when available. Examples of such untargeted approaches in plant research are the comparison of secondary metabolites in roots and leaves of wild-type and mutant *Arabidopsis* (*Arabidopsis thaliana*) plants[24], studying metabolic alterations in fruits of a light-hypersensitive mutant of tomato (*Solanum lycopersicum*)[44], comparing tubers of potato (*Solanum tuberosum*) of different genetic origin and developmental stages[25], determining tissue specificity of metabolic pathways in tomato fruit[22], establishing gene-to-metabolite networks in *Catharanthus roseus*[26] and identifying quantitative trait loci (QTLs) controlling metabolite composition in *Arabidopsis*[43,48].

For our metabolomics approaches, we prefer to use the freeware MetAlign (http://www.metalign.nl and http://www.rikilt.wur.nl/UK/services/MetAlign+download) to process large LC-MS[22,25,43,44] as well as GC-MS[49] data sets, based on a number of features:

- compatibility with most mass spectrometry software such as Masslynx, Xcalibur, Chemstation, Agilent, Bruker and ANDI/netCDF formats and output in any of these formats as well as in Excel;
- compatibility with both LC and GC, and independent of mass spectrometer type (e.g., quadrupole-MS, TOF-MS, FTMS) or instrument maker;
- an easy interface for user-defined parameter settings;
- automated local noise calculation and mass-specific baseline corrections;
- capability to align up to hundreds of data sets.

Examples of using MetAlign for the comparison of ten to hundreds of LC-MS data files are available[22,25,43,44]. Although MetAlign converts accurate mass data into nominal masses, mainly for reasons of faster data processing, the masses of aligned signals can automatically be recovered using a script called MetAccure[22,25].

Considerations for tissue sampling and handling

Although no limitations regarding sample type are foreseen, except from a technical point of view, care must be taken in acquiring reproducible data. Sources of variation contributing to the total "noise" in subsequent statistical analyses are biological variation (e.g., variation in plant growth conditions, development), perturbations during and after tissue collection, and variation in tissue sampling for metabolite extraction including weighing errors. Metabolic conversions in tissues can be abolished by flash-freezing samples in liquid nitrogen immediately after harvest. Frozen samples should be fully homogenized into a fine powder in order to facilitate and standardize metabolite extraction. Nevertheless, each analysis provides only a single snapshot of the metabolic state of that sample without further information on biological variation or measurement errors. To estimate these variations, sufficient biological replicates and sufficient technical replicates from the same batch of tissue powder, respectively, need to be prepared and analyzed.

Considerations for metabolite extraction and LC-PDA-MS analyses

The extraction procedure is crucial for the detection of metabolites naturally occurring in the extracted tissues. Therefore, the extraction protocol should be reproducible and with high recovery and stability of most compounds, at least those of prime interest. We have tested a number of different solvents, such as methanol, ethanol and acetone, at different ratios of water versus organic solvent, for extraction efficiency, chromatographic behavior and extract stability. Acidified aqueous methanol at a final concentration of 75% methanol (v/v) and 0.1% formic acid (v/v) was the most suitable solvent for efficient extraction of a wide range of

compounds of our prime interest, mostly secondary metabolites, from different plant species and tissues[22,25,43,44]. Enzymes present in the sample should be inactivated by directly adding the solvent to frozen plant powder and mixing immediately. Extraction efficiency was tested using several (poly)phenolic compounds added to the frozen powder before extraction. At a solvent/sample ratio of 3 and a sonication time of 15 min, the recovery of all standards tested was higher than 90%. Sonication for up to 2 h did not significantly change the metabolite profile as compared with 15-min sonication. However, it is advised to check the extraction efficiency upon analysis of a completely different plant matrix or in case of main interest in specific key compounds.

The chromatographic conditions applied are always a compromise between metabolite resolution, retention time stability and sample throughput. In the standard protocol, we use a C_{18}-reversed phase microbore column with a relatively small particle size. This column was selected after testing different types of columns for their ability to retain and separate semi-polar compounds of our prime interest, including flavonoids and phenolic acids[22,43,44], alkaloids[22,25,44] and glucosinolates[43]. A gentle and continuous acetonitrile gradient of 45 min, followed by 15 min column washing and stabilization, resulted in adequate separation of many semi-polar compounds including isomeric forms (**Fig. 1**). We tend to use the same chromatographic conditions in our untargeted metabolomics work, in order to compare mass signals from different samples and to enable compound identification using LC-MS databases[22]. In most of our experiments, the LC-MS run itself is not the limiting factor in sample throughput. Instead, sample harvest, grinding, weighing and extraction, and finally data analyses usually take much more time. For large series of samples, for example, more than 300 extracts, steeper gradients with shorter run times may be useful in order to decrease total run time and therefore the chance of possible perturbations upon increasing analysis times. This might occur owing to (pre)-column deterioration or disturbances in the MS electronics or LC pump, thus introducing extra variation in the final data set. Thus, during analyses of an *Arabidopsis* recombinant inbred line (RIL) population consisting of 409 extracts including controls, we doubled the sample throughput by using a total run time of 30 min per extract[43]. However, speeding up the LC run time, with the same type of column, unavoidably results in an increased amount of co-eluting compounds and thus may lead to a loss of resolution of isomers and an increased ion suppression and adduct formation at the ionization source. We advise to start with the standard 60 min protocol as outlined below and, if needed, to modify the chromatographic conditions (gradient, column type) in such a way that at least the compounds of key interest are adequately separated and detected.

Upon starting up a new series of analyses, the chromatography is relatively unstable owing to (pre)column conditioning by the crude extracts themselves. To avoid suboptimal alignment resulting from this early-stage system instability, several "dummy" runs of extracts should be performed using identical conditions, before collecting the actual data. We routinely program the LC-MS software to inject and analyze repeatedly the first sample extract at least four times. Standard solutions should not be injected between crude extracts, as during analysis of these relatively clean samples, the column can partly be re-conditioned resulting in small retention shifts. To ensure constant and reproducible ionization, regularly check the actual pressure and supply of the nitrogen and argon gasses. In our system, we can check this pressure by comparing the intensity of the reference mass (lock mass; see below) over the samples. If the intensity of this mass signal is markedly changed in one or more samples, these samples should be reanalyzed within the same series.

Analyze extracts in a randomized order to avoid possible variation from time-dependent changes, for example, owing to slow deterioration of (pre-)column or ionization source. Owing to the high variability of metabolites present in crude extracts with respect to their chemical characteristics and intrinsic behaviors upon sample preparation, the use of a single internal standard to correct for variation in extraction and detection of all mass signals over the samples is of dubious value. Adding a series of internal standards, for example, each representing a different class of plant metabolites, may be a better option but may introduce ion suppression effects in the case of co-eluting compounds. Consequently, we recommend preparing a statistically relevant number of replicates from a homogenous (pooled) batch of material and analyzing these throughout the entire sample series, in order to estimate technical reproducibility and, if needed, to correct for this type of variation.

With our LC-QTOF MS system, we normally acquire data in centroid mode. In contrast to the continuum mode, in which the mass signal is represented by a Gaussian curve, the centroid mode projects each mass signal as accurate *m/z* value by on-the-fly mathematical transformation. Although relevant information on mass peak shape and purity may be lost upon centroiding, the raw data files are markedly reduced from about 500 Mb to a more useful size of about 10 Mb per sample (at a run time of 1 h and sampling rate of 1 scan per second). Especially upon analyzing and processing large series of extracts, and for storing and databasing thousands of raw data files gathered over years of analyses, acquiring data in centroid mode is the most practical option. In addition, by using a separate lock mass spray as reference and by continuously switching between sample and reference, the Masslynx software can automatically correct the centroid mass values in the sample for small deviations from the exact mass measurement[50], resulting in a mass accuracy of better than 5 p.p.m. generally.

This paper describes a detailed protocol for untargeted LC-MS-based metabolomics of large numbers of extracts. The standard procedure is schematized in **Figure 2** and consists of tissue sampling and extract preparation, LC-QTOF MS analysis using an ESI source, MetAlign-assisted mass peak extraction and alignment across samples, and identification of mass peaks selected by means of appropriate statistical filtering. In principal, the methodology described below is applicable to a wide range of plant species, tissues or products derived thereof.

MATERIALS

REAGENTS

- Acetonitrile, HPLC supra-gradient grade (Biosolve, cat. no. 01203502, CAS (75-05-8)) ! **CAUTION** Acetonitrile is harmful and highly flammable and should be handled in a fume hood
- Methanol absolute, HPLC supra-gradient grade (Biosolve, cat. no. 13683502, CAS (67-56-1)) ! **CAUTION** Methanol is toxic and highly flammable and should be handled in a fume hood

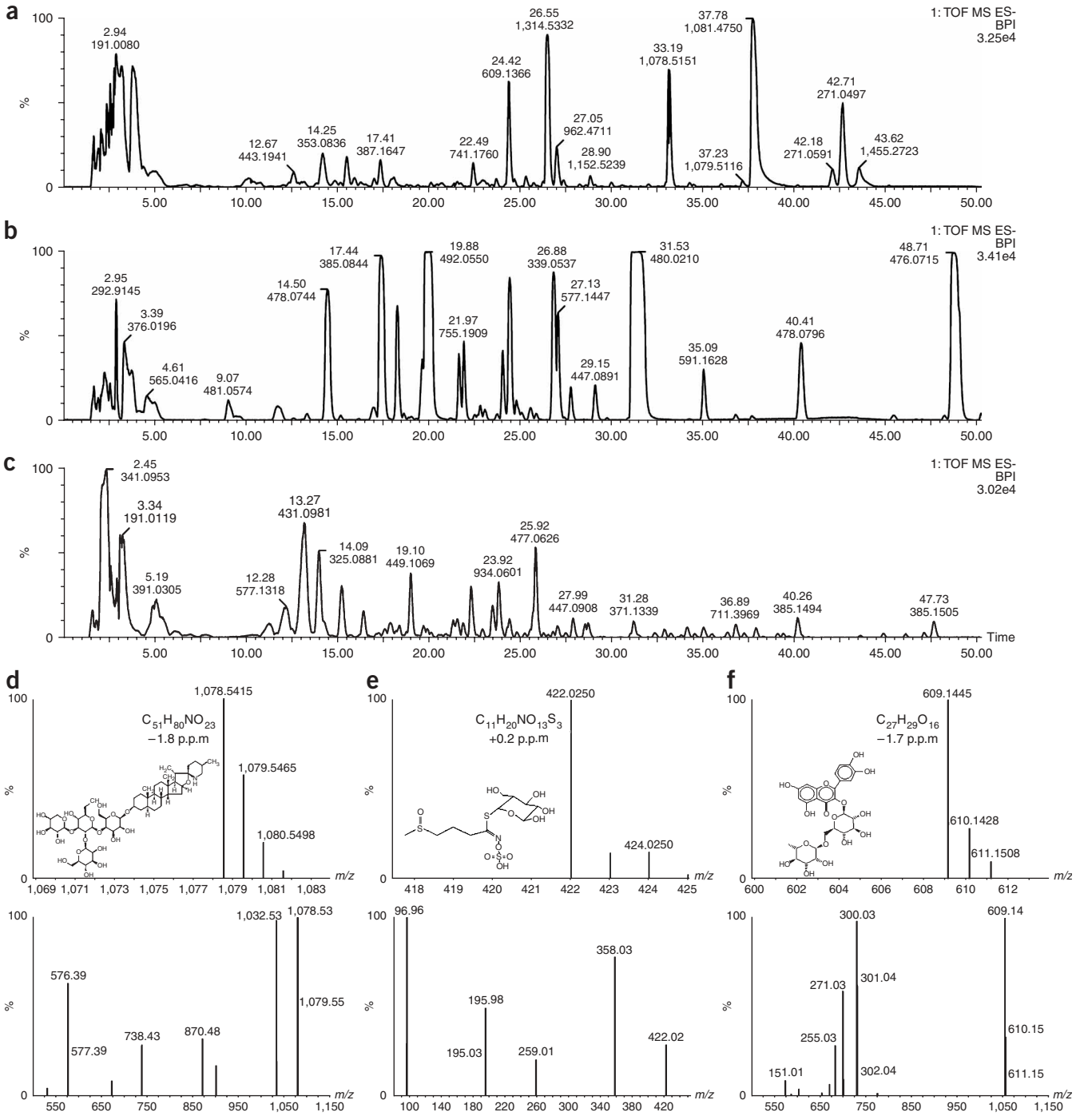

Figure 1 | LC-QTOF MS profiling of crude extracts from three different plant species. The upper panel shows typical ion chromatograms, obtained in ESI negative mode, of (**a**) tomato fruit, (**b**) *Arabidopsis* leaf and (**c**) strawberry fruit. Lower panels show detected accurate masses of $[M-H]^-$ ions and LC-MS/MS spectra of three compounds from different classes of secondary metabolites: (**d**) α-tomatine, an alkaloid, detected as formic acid adduct; (**e**) glucoiberin, a glucosinolate; and (**f**) rutin, a flavonoid.

- Formic acid (FA) for analysis, 98–100% (Merck-KGaA, cat. no. 1.00264.1000, CAS (64-18-6))
  ! CAUTION Formic acid is corrosive and volatile, and should be handled in a fume hood
- Leucine enkaphaline, ≥95% pure, isolated by HPLC (Sigma, cat. no. L9133, CAS (81678-16-2))
- Phosphoric acid p.a. 85% in water solution (w/v) (Acros, cat. no. 20114-0010, CAS (7664-38-2)) **! CAUTION** Phosphoric acid is corrosive and should be handled in a fume hood
- Ultrapure water (Elga Maxima, Bucks)
- Liquid nitrogen for freezing samples **! CAUTION** Liquid nitrogen is a low-temperature refrigerant and should be handled with protective glasses and protective gloves
- Liquid nitrogen for applying gas to mass spectrometer ionization source
- Argon 5.0, at least 99.999% pure, for applying gas to mass spectrometer collision cell
- Sample extraction solution (see REAGENT SETUP)
- HPLC mobile phase (see REAGENT SETUP)

- MS calibration solution (see REAGENT SETUP)
- Lock mass solution (see REAGENT SETUP)

EQUIPMENT

- Storage tubes or plastic bags resistant to liquid nitrogen, for example, polypropylene 50-ml tubes with screw cap (Greiner, cat. no. 210261) and Eppendorf microtest tubes, 12 ml glass tubes with screw caps (Omnilabo)
- IKA A11 basic grinder
- Pipettes and tips suitable for handling organic solvents (Microman, Gilson)
- Ultrasonic bath (Branson 3510)
- Single-use sterile and non-pyrogenic latex-free syringes, 0.01–1 ml Tuberkulin Omnifix-F (B.Braun Melsungen AG, cat. no. 9161406V)
- Single-use syringe filters free of polymers, such as Anotop 10 (diameter 10 mm, pore size 0.2 µm; Whatman, cat. no. 6809-1022) or Minisart RC4 (diameter 4 mm, pore size 0.2 µm; Sartorius, cat. no. 17821) ▲ CRITICAL Filters for MS analyses should be resistant to extraction solution (75% methanol + 0.1% FA) and free of polyethylene glycol or any other soluble polymer
- Crimp cap autosampler vials of 1–2 ml with aluminum crimp caps containing natural rubber/polytetrafluoroethylene septum
- Tecan Genesis Workstation with TeVacs vacuum filtration unit
- Protein filtration plates in 96 wells format (Captiva 0.45 µm; Ansys Technologies)
- Ninety-six-well plates with 700 µl glass inserts (Waters) and 96-square-well polytetrafluoroethylene-coated seal (Waters)
- Analytical column Luna C18(2), 2.0 mm diameter, 150 mm length, 100 Å pore size and spherical particles of 3 µm (Phenomenex)
- Pre-columns Luna C18(2), 2.0 mm diameter, 4 mm length (Security Guard, Phenomenex)
- PEEK in-line filter holder with PEEK frit 0.5 µm pore size (UpChurch Scientific)
- Alliance 2795 HT liquid chromatography system equipped with an internal degasser, sample cooler and column heater (Waters)
- Photodiode array detector 2996 (Waters)
- Quadrupole-time-of-flight Ultima V4.00.00 mass spectrometer equipped with an ESI source (Waters) and separate lock mass spray inlet
- Separate HPLC pump (e.g., Bromma 2150; LKB) for continuously pumping the lock mass solution at 10 µl min^{-1}
- PEEK tubings (Upchurch Scientific) for connecting the LC-PDA (125 µm inner diameter) and the lock mass pump (250 µm inner diameter) to the mass spectrometer
- PHD 4400 syringe pump (Harvard)
- Gastight glass syringe 0.1–1.0 ml (Hamilton-Bonaduz Schweiz, cat. no. 1001)
- Software: Masslynx data management software 4.0 (Waters), MetAlign (http://www.metalign.nl or http://www.rikilt.wur.nl/UK/services/MetAlign+download) and Microsoft Office Excel 2003. Optional: multivariate analyses software such as GeneMaths 2.01 (Applied Maths)

REAGENTS SETUP

Plant growth and sampling conditions Samples to be prepared for metabolomics studies should be as representative as possible for the genotype or tissues to be analyzed. For small plants like *Arabidopsis* seedlings, a combinatorial approach of controlled plant growth, pooling and replicate analyses can be used to minimize biological and experimental variation. For instance, in the large-scale metabolomics study in *Arabidopsis* RILs[43], seeds were sown on 10 ml 1/2 MS agar (2%) in 6 cm ∅ Petri dishes with a density of a few hundred seeds per dish. Dishes were placed in a cold room at 4 °C for 7 days in the dark to promote uniform germination and were then randomly placed in five blocks in a climate chamber where each block contained one replicate dish of each line. Growth conditions were 16 h light (30 W m^{-2}) at 20 °C and 8 h dark at 15 °C, at 75% relative humidity. After 6 days the lids of the Petri dishes were removed to ensure that seedlings were free of condensed water on the day of harvest. On day 7, at 7 h into light period, all seedlings were harvested within 2 h by submerging the complete Petri dish briefly in liquid nitrogen and scraping off the aerial parts with a razor blade. Finally, per line, material from two dishes was pooled to make one of the replicate samples and from the other three dishes to make the second. To obtain representative material from large plant tissues, such as fruits of tomato or apple, or tubers of potatoes, a representative "pie" segment was taken from at least five fruits or tubers per plant using a sharp knife. Segments were snap-frozen in liquid nitrogen and pooled per plant. Once harvested, plant material can be stored at −80 °C until further processing.

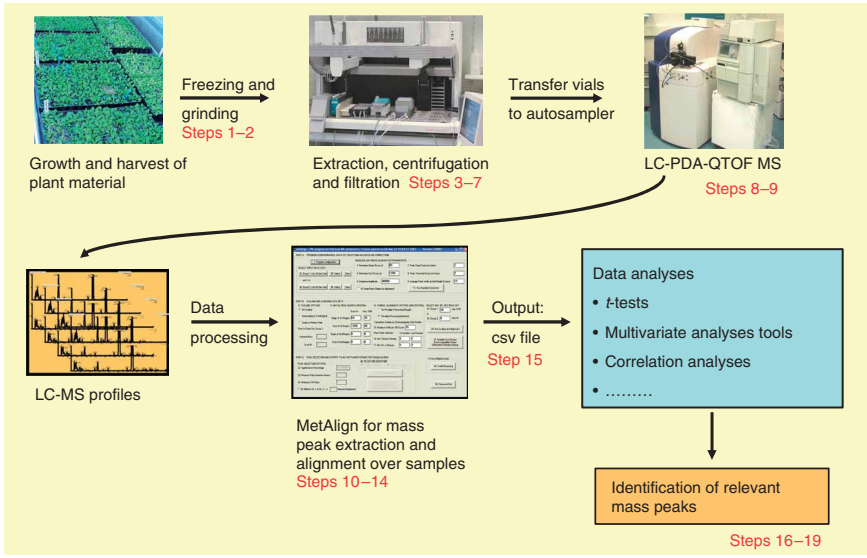

Figure 2 | Schematic overview of experimental setup and data flow for untargeted LC-QTOF MS-based metabolomics of plant materials. A detailed description of each step is given in PROCEDURE.

Sample extraction solution Prepare 99.875% methanol solution acidified with 0.125% (v/v) FA. ! CAUTION Methanol is toxic and highly flammable, whereas formic acid is corrosive. Both solvents should be handled in a fume hood.

HPLC mobile phase Two eluents are used as mobile phase; eluent A is 0.1% FA (v/v) in ultrapure water and eluent B is 0.1% FA (v/v) in acetonitrile. ! CAUTION Both methanol and acetonitrile are toxic and highly flammable, whereas FA is corrosive; all solutions should be handled in a fume hood. ▲ CRITICAL As the retention of some metabolites, especially alkaloids, is very sensitive to slight variations in the acidity of the mobile phase, always precisely add 0.1% (v/v) FA to both eluents and prepare sufficient eluents to analyze the entire sample series.

MS calibration solution To calibrate the mass spectrometer, freshly prepare about a 1 ml solution of phosphoric acid at a concentration of 0.05% (v/v) in 50% acetonitrile/ultrapure water and load into the gastight glass syringe. ! CAUTION Handle solvents in fume hood.

Lock mass solution Prepare a solution of leucine enkaphaline in 50% (v/v) acetonitrile/ultrapure water to obtain a final concentration of 0.1 µg ml^{-1}. Prepare sufficient solution for analysis of the complete series of samples. ! CAUTION Handle solvent in fume hood.

EQUIPMENT SETUP

LC-PDA-QTOF MS setup See **Boxes 1** and **2**. ▲ CRITICAL The LC-PDA system needs to be conditioned for a minimum of 1 h before use; the QTOF MS should be conditioned for a minimum of 2 h.

Data pre-processing and alignment We routinely program the MetAlign software to extract and align all mass signals having a signal-to-noise ratio of at least 3 (normally used as a threshold in analytical chemistry). The software performs the following processing steps: (i) mass data smoothing using a digital filter related to average peak width; (ii) local noise calculation as a function of retention time and ion trace; (iii) baseline correction of all ion traces and introduction of a threshold to obtain noise reduction; (iv) scaling and calculation and storage of peak maximum amplitudes; (v) between-chromatogram alignment using high signal-to-noise peaks common to all chromatograms; (vi) iterative fine alignment by including an increasing number of low-signal peaks; (vii) output of aligned data into a csv-file compatible with Microsoft Excel and most multivariate programs; and, finally and optional, (viii) significant difference filtering at user-defined thresholds and output of selected data back to the MS software platforms for visualization of differential chromatographic mass peaks. A picture of the MetAlign interface is given in

BOX 1 | LC-PDA-QTOF MS SETUP; CONDITIONING THE HPLC-PDA SYSTEM

1. Prepare mobile phase solvents, prime HPLC pump and tubing, and degas both solvents for at least 10 min using the in-line degasser of the Alliance 2795 HT
2. Install one PEEK in-line solvent filter between injection system and pre-column cartridge. Place two pre-columns in tandem in the cartridge, fix in front of the analytical column and place both columns in the column oven conditioned at 40 °C
3. Precondition column system by increasing the percentage of eluent A stepwise (starting at 100% eluent B) until the initial gradient conditions are reached
4. Program the inlet file according to the gradient settings given below. In the standard setup, we use relatively long chromatographic runs of 1 h, including column washing and re-conditioning, with a mobile phase flow of 0.19 ml min^{-1} into the analytical column (diameter of 2.0 mm). This flow rate corresponds to 1 ml min^{-1} on a 4.6-mm column, which is standard in most HPLC-UV/Vis applications. In the case of a large sample series, for example, more than 300 extracts, we consider the use of a 30-min run at a slightly higher flow rate, to lower the chance of possible perturbations

60 min run Flow rate 0.19 ml min^{-1}

Time (min)	%A	%B
0	95	5
45	65	35
47	25	75
52	25	75
54	95	5
60	95	5

30 min run Flow rate 0.20 ml min^{-1}

Time (min)	%A	%B
0	95	5
20	25	75
25	25	75
26	95	5
30	95	5

5. The PDA detector is placed between analytical column and the QTOF MS. Connect column outlet to flow cell of the PDA detector and switch on the detector. Program PDA to acquire data every second from 210 to 600 nm with a resolution of 4.8 nm. Wavelength range, scan rate and resolution can be adjusted according to LC runs times and research aims.
▲ CRITICAL Check HPLC pump for air bubbles and connections for leakage by verifying pressure stability
▲ CRITICAL Precondition PDA lamp, column oven temperature and analytical column for at least 1 h before starting sample analyses. Meanwhile, the mass spectrometer can be calibrated and checked for performance as described in **Box 2**.
6. Place the aqueous methanol extracts in trays inside the autosampler (20 °C) during the analysis series. Program the injection system to operate in sequential mode and to load the syringe with 5 µl of sample with 5 µl of air both before and after the sample. The injection needle is washed with 50% (v/v) methanol/water between injections

Figure 3. The parameters used for processing the 30-min LC-MS runs are shown in the figure itself; for the 60-min runs, the differing parameters are given in the legend. The software, examples and manual can be downloaded free of cost from http://www.metalign.nl or http://www.rikilt.wur.nl/UK/services/MetAlign+download/. It is recommended to carefully read the manual to become acquainted with the effect of the different parameters and how to optimize the settings. **Box 3** gives a summarized account of this information. Default parameters for some other MS systems can be found in the MetAlign manual.
LC-PDA-MS/MS setup If needed, mass signals can be further identified using LC-MS/MS. For this purpose, masses of interest are incorporated into a mass inclusion list (data-directed MS/MS). We perform LC-MS/MS on the QTOF Ultima with a scan time of 0.4 s and an interscan delay of 0.1 s. The collision energy profile is programmed to increase sequentially from 5, 10, 20 to 30 eV (ESI positive mode) or 10, 15, 30 to 50 eV (ESI negative mode). If these settings are insufficient to obtain MS/MS information for the masses of interest, the collision energy profile can be adjusted. ▲ CRITICAL In the case of random LC-MS/MS experiments, in which up to the eight highest intensity ions per survey scan can be automatically selected for MS/MS, use a mass exclusion list containing abundant eluent mass signals in order to prevent switching to MS/MS mode for these impurities.

PROCEDURE

Tissue sampling and extraction

1| Harvest a reproducible amount of tissue (leaf, roots, fruit, etc.) by rapid freezing in liquid nitrogen. Large plant parts such as tomato fruits or potato tubers should first be cut rapidly into representative smaller parts with a sharp knife before freezing. In the case of seeds or small seedlings (e.g., *Arabidopsis*), use 1.5- or 2.2-ml Eppendorf tubes; in the case of larger tissues, use 50-ml Greiner tubes or plastic bags that are resistant to liquid nitrogen.
! CAUTION To prevent storage tubes or bags from exploding, remove all liquid nitrogen by gently pouring off before closing and do not screw tube lids firmly!
■ PAUSE POINT frozen tissue can be stored at −80 °C for at least 1 year.

BOX 2 | LC-PDA-QTOF MS SETUP; CONDITIONING THE MS SYSTEM

Before each series of sample analyses, the mass spectrometer should be conditioned and calibrated to obtain good performance in terms of mass accuracy and resolution. In contrast to electron impact ionization, as used in most GC-(TOF)MS applications, detection sensitivity and mass spectra obtained by soft ionization LC-MS are completely dependent on the type of mass spectrometer, ionization source and chromatographic system used. The procedure and settings described here are for a QTOF Ultima with ESI source and the TOF tube in V-mode, in combination with the HPLC conditions described above

1. Connect the outlet of the PDA, with eluent flow of 0.19 ml min^{-1}, to the inlet of the mass spectrometer and set the capillary voltage at 2.75 kV, cone voltage at 35 V, source temperature at 120 °C and desolvation temperature at 250 °C. Use a cone gas flow of 50 liter h^{-1} and desolvation gas flow of 600 liter h^{-1}.

▲ **CRITICAL** Precondition MS for at least 2 h at these standard settings

2. Disconnect the eluent tubing from the MS and use the syringe pump to inject the phosphoric acid calibration solution directly into the ESI source, at an initial flow of 5 µl min^{-1}

3. Acquire data from *m/z* 80–1,500 at a scan rate of 0.9 s and an interscan delay of 0.1 s. A series of phosphoric acid cluster peaks should appear throughout the entire range of the mass spectrum.

▲ **CRITICAL** To obtain proper calibration and accurate mass calculations, none of the mass calibration peaks should exceed an intensity of 250 counts s^{-1} (in continuum mode) and the intensity of the clusters over the mass range should be as uniform as possible. Adjust pump flow, capillary voltage, cone voltage, desolvation gas flow and/or collision energy until criteria are fulfilled

4. Combine spectra of about 50 scans during acquisition at optimal settings in continuum mode, center the mass signals and check mass resolution of the machine for *m/z* 488.8772 (negative ionization mode) or 490.8918 (positive ionization mode). Mass resolution is calculated by dividing the *m/z* value of the centered mass signal by the mass difference at half height of the Gaussian-shaped mass peak in continuum mode, and should be better than 8,500 (with QTOF Ultima in V-mode); otherwise, re-tune instrument and repeat the procedure

5. Use the centered mass data for calibration of the instrument using a polynomial-5 fit.

▲ **CRITICAL** Mean residual mass deviation should be less than 1.5 p.p.m.; otherwise, adjust calibration settings.

6. Check calibration using leucine enkephalin as a standard. Inject the leucine enkephalin solution through the separate lock mass inlet into the ESI source and acquire data under MS conditions as used during sample analyses, but in continuum mode. Adjust flow to obtain a specific mass intensity of 250 counts s^{-1}. Collect and combine about 50 spectra and center the mass peak.

▲ **CRITICAL** The observed mass should be within 20 p.p.m. deviation of *m/z* 556.2767 in positive mode and 554.2619 in negative mode; otherwise, recalibrate instrument.

7. Reconnect the outlet of the PDA to the inlet of the mass spectrometer. Check the effluent from the LC system, including mobile phase, tubings, columns and PDA flow cell, by acquiring centroid data from *m/z* 80–1,500 under the exact conditions of sample analysis. Individual mass signals at initial gradient conditions should preferably be less than 200 counts per scan in negative mode or less than 500 counts per scan in positive mode, to prevent excessive ion suppression of sample compounds

8. Prepare MS method file to acquire mass data from *m/z* 80–1,500, at a scan rate of 0.9 s and an interscan delay of 0.1 s and in centroid mode.

▲ **CRITICAL** The range of masses to be detected in sample extracts should fall within the range of calibration masses. During sample analyses, the standard setting of collision energy is 10 eV in negative ion mode and 5 eV in positive ion mode. If needed for optimal ionization of key compounds, the collision energy may be adapted. The MS is programmed to switch from sample to lock spray every 10 s and to average two scans for lock mass correction (*m/z* 556.2767 in positive mode and 554.2619 in negative mode). The lock mass solution is used for online calibration of the mass accuracy during sample analysis[22,50].

▲ **CRITICAL** Adjust flow rate or concentration of the lock mass solution to obtain an intensity of about 500 counts per scan (in centroid mode) during LC-MS runs, to enable accurate mass calculation of as many compounds in the extracts as possible.

2| Homogenize the frozen tissue in liquid nitrogen into a fine powder using a pestle and mortar, but preferably use a ball mill (Retsch Mixer Mill MM 301 for *Arabidopsis*) or analytical mill (IKA A11 for larger tissues) that has been thoroughly pre-cooled with liquid nitrogen. Transfer homogenized powder into pre-cooled storage containers resistant to liquid nitrogen.

▲ **CRITICAL STEP** Take care that tissues stay well frozen during homogenization; discard any samples that start to thaw. If needed, carefully pour a small volume of liquid nitrogen onto the sample, let the nitrogen evaporate and continue homogenization.

■ **PAUSE POINT** Frozen powder can be stored at −80 °C for at least 1 year.

3| Weigh 100 mg of frozen *Arabidopsis* powder with an accuracy of more than 5% in a pre-cooled Eppendorf tube, or 500 mg in the case of larger amounts of tissue (e.g., tomato fruit or potato tuber) in a 10-ml glass tube with screw cap. Lower amounts can be used as well, but this is not advisable in view of the inherent relative higher weighing error using frozen material.

▲ **CRITICAL STEP** Take care that tissues stay fully frozen; discard any samples that start to thaw. Lyophilization of tissue is not recommended, unless for specific practical reasons, if the effects on the metabolite profile is unknown.

■ **PAUSE POINT** Frozen powder can be stored in tubes at −80 °C for at least 1 month.

4| Prepare extracts freshly at the beginning of a series of analyses. Add ice-cold sample extraction solution (99.875% methanol acidified with 0.125% FA) in a volume/fresh weight ratio of 3 to the tube containing the weighed frozen powder, close lid and immediately vortex for 10 s. Assuming a tissue water content of about 95%, this will result in a final concentration of 75% methanol and 0.1% FA. In the case of samples with highly variable water contents or lyophilized material, pure water

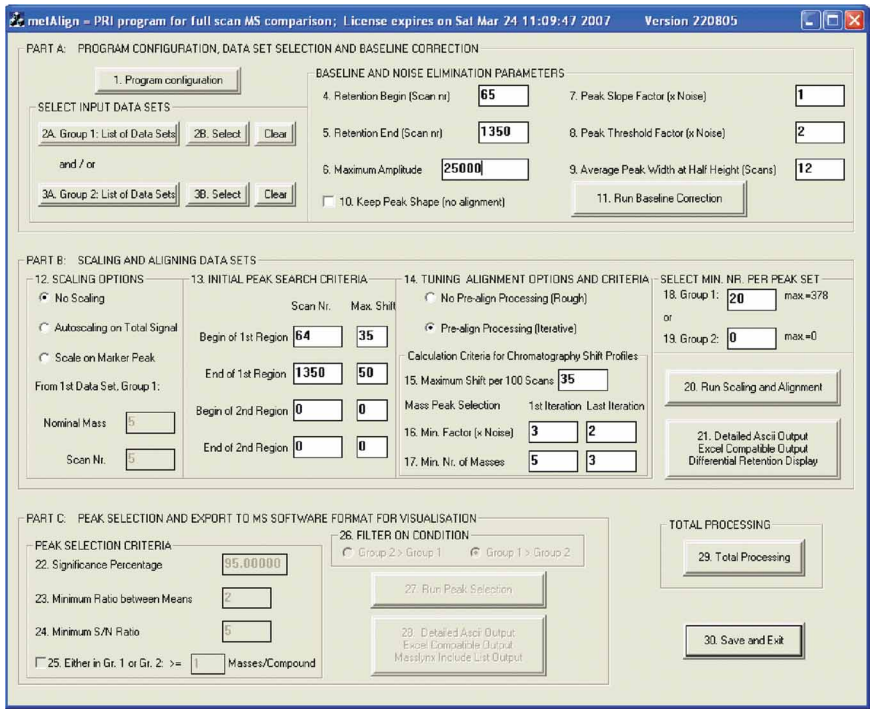

Figure 3 | Interface of MetAlign software used for untargeted processing of LC-QTOF MS data files. The program is divided into three parts: part A deals with program configuration, data selection, peak extraction and baseline correction; part B covers the actual alignment of extracted mass peaks and output of (mass peak intensity × samples)—data matrix; part C is used to identify and visualize chromatographic peaks that are statistically different between two groups of samples (optional). Parameter settings given in this figure correspond to the default values for processing of 30-min LC-MS runs. For 60-min LC-MS runs, the following default parameter settings are recommended: 4=70; 5=2,450; 8=3; 9=25; 13=69, 35 and 2,450, 35; and 16=10, 5. A short description of buttons and parameters is given in **Box 3**.

can be added to adjust each sample to a final solvent concentration of 75% methanol and 0.1% FA. Store extracts on ice until all samples are ready.

5| Sonicate each sample for 15 min at maximum frequency (40 kHz) continuously, in a water bath at room temperature (20 °C).

6| Centrifuge for 10 min at maximum speed (20,000*g* for Eppendorf tubes; 3,000*g* for glass tubes) at room temperature.

7| Filter the supernatant through a 0.2-µm PTFE filter using a disposable syringe into a 1.8-ml glass vial and close the vial with cap. In the case of large amounts of samples, use suitable 96-well filtration plates and a vacuum filtration unit. We use a TECAN Genesis Workstation 150 equipped with a four-channel pipetting robot and a TeVacS 96-well filtration unit. Pre-wash filtration plates (Captiva 0.45 µm, Ansys Technologies) at least three times with 700 µl of 75% methanol containing 0.1% FA. Dry bottom tips of the filters by blotting on filter paper. Place a 96-well plate with 700 µl glass inserts in the filtration unit under the pre-washed filtration plate. Load each well with 700 µl of extract and vacuum-filter two times for 20 s until dry. Carefully remove air bubbles trapped at the bottom of the inserts. Cover the plate with a 96-square-well PTFE-coated seal.
▲ **CRITICAL STEP** All filters used should be free of aqueous methanol-soluble polymers, such as polyethylene glycol.

LC-PDA-QTOF MS analysis

8| Place vials or 96-well plates in the autosampler conditioned at 20 °C.

9| Check for the presence of sufficient eluents, lock mass solution and nitrogen gas, and start sample series using the setup detailed in **Boxes 1** and **2**. Begin with at least four dummy injections to stabilize the LC-PDA-MS system. Check system performance and mass accuracy during these first runs. Deviations of observed known parent masses from their calculated masses should be less than 5 p.p.m. (at signal intensities similar to that of the local lock mass), otherwise recalibrate system.
? TROUBLESHOOTING
■ **PAUSE POINT** Raw data can be stored on hard disks, tapes, DVDs or other digital storage devices until further processing.

Pre-processing and alignment of LC-MS data

10| Configure MetAlign (see EQUIPMENT SETUP) and select the data to be processed (buttons 1–3, see **Box 3** for more details). The first sample selected with button 2B is used as the reference file in the actual alignment (part B, see **Fig. 3**). We recommend selecting the sample that has been analyzed in the middle of the entire LC-MS series as this reference file, to minimize the extent of retention profile correction between the first and last samples analyzed.

11| Perform a test baseline correction (part A, see **Fig. 3**) and alignment (part B, see **Fig. 3**) on only a few variable samples to check whether the default settings are suitable to extract and align mass peaks that are of specific interest (if any). Define parameters for peak extraction and noise (buttons 4–9, see **Box 3** for more details) and run baseline correction (button 11, see **Box 3** for more details). Manually inspect corresponding mass peaks in the beginning, middle and at the end of the baseline-corrected chromatograms and compare with the original raw data. If it is obvious that some mass signals from relatively broad chromatographic peaks are missing in the baseline corrected data, set parameter 9 (see **Box 3** for more details) at a slightly higher value and re-run baseline correction. On the other hand, if closely eluting peaks of compounds with similar (nominal) mass have been extracted as single peaks, lower the value at button 9.

BOX 3 | DESCRIPTION OF METALIGN BUTTONS AND PARAMETERS

A more detailed description can be found in the manual, which can be downloaded from http://www.metalign.nl or http://www.rikilt.wur.nl/UK/services/MetAlign+download/

Part A: Program configuration, data set selection and baseline correction

- Buttons 1–3 are used to define the data sets and to define folders and formats for input and output
- Parameters 4 and 5 (value in scans) refer to the region in the chromatogram, which should be processed. In particular, parameter 5 should be taken in an empty region of the chromatogram at the highest concentration of organic modifier in the gradient or at an earlier time point. This enables MetAlign to calculate a matrix of noise versus retention time versus mass. This noise matrix together with parameters 7 and 8 is then used as a basis to find real mass peaks
- Parameter 6 (value in ion counts of a single mass) is machine dependent and should be set at about 70% of the maximum value a detector can record, to be able to deal with artifacts owing to detector saturation. MetAlign creates artificial maxima at this value for all peaks above this value
- Parameters 7 and 8 (factor times local noise) are peak slope and threshold factors used to filter out peaks from noise
- Parameter 9 (value in scans) should be the average mass peak width at half height of non-saturated compounds. This parameter is used in determining the data smoothing (digital filter) as well as for a window in the alignment (see "14. Tuning Alignment Options and Criteria")
- Parameter 10 is "de-clicked" to indicate that the peak shapes should not be saved, which only in this mode is compatible with alignment; "clicked" keeps peak shapes and renders the output incompatible with alignment, but on the other hand is compatible with deconvolution algorithms from third party software
- Button 11 consecutively processes all data sets defined by buttons 1–3. It starts the noise estimation as a function of time and mass, the smoothing, maximum amplitude correction (if needed), baseline correction, noise elimination, peak picking and exporting of baseline-corrected peaks

Part B: Scaling and aligning data sets

- Button 12 provides different modes of scaling data sets. Options are (a) no scaling, (b) scaling on the basis of sum of all the amplitudes of the peaks picked and (c) scaling using a specific mass
- The parameters in "13. Initial Peak Search Criteria" provide the window (in +- the indicated scans) at a position (in scans) in the chromatogram in which a search for identical masses is carried out over all chromatograms. This window may vary with retention time; the parameters in 13 provide coordinates used for linear interpolation of the window size for the whole chromatogram
- The options in "14. Tuning Alignment Options and Criteria" determine if the rough or iterative alignment should be performed. In brief, the alignment is described as follows: in both modes of alignment, the window determined by "13. Initial Peak Search Criteria" is used to restrict searches for identical masses in different data sets. For the rough mode, the alignment finishes here. For the iterative alignment, this is the starting point for the first estimation of a retention shift profile for all data sets with regard to the first data set. For each time point in a retention shift profile, criteria (parameters 16 and 17) to calculate differences in retention times between files are on the basis of a minimum number of aligned masses present in all data sets, which are above a minimum amplitude (factor times noise) and occur in a chromatogram sub-window (of two times parameter 9). The next iteration will start from here. Using this first retention shift profile, the alignment is refined by doing bookkeeping on the differences in retention and automatically decreasing the parameters in "13. Initial Peak Search Criteria" to obtain a smaller search window throughout the chromatogram. The second alignment is then performed as described for the smaller retention corrected search window (13). Parameters 16 (number of masses) and 17 (factor times noise) are also automatically reduced and a new and better retention shift profile is calculated analogous to the first iteration. Iterations continue until the final values in parameters 16 and 17 are reached and the search window is two times the value of parameter 9 (average peak width). After finalizing the last iteration, incomplete mass peak sets spread over neighboring scans are combined in a fine-alignment process
- Parameter 15 restricts changes in retention time shifts between calculated points in a retention shift profile to a maximum value (in scans per 100 scans). This restriction is used after calculation of a retention shift profile and serves to filter out possible anomalies
- Parameters 18 and 19 are filters for aligned mass peaks, which indicate minimum completeness of aligned mass peak sets
- Button 20 starts the scaling and alignment of data obtained in part A
- Button 21 is used to obtain information on the alignment of masses. There are three options: (i) a normal ascii output, (ii) an excel-compatible CSV-file output, and (iii) a graphical display of the retention shift profiles of individual data sets with regard to the first reference file
- Button 29 executes the calculations under buttons 11, 20 and 28
- Button 30 exits the program saving the parameters set

Part C: Peak selection and export to MS software format for visualization (only applicable when comparing two groups of data)

- Parameter 22 is the significance percentage restriction when selecting differences between data in group 1 versus 2
- Parameter 23 restricts selection of differences between groups on the basis of the ratio in the means of individual aligned masses
- Parameter 24 restricts selection of differences between groups on the basis of the minimum amplitudes defined as a factor times noise, that is, it determines what is defined as present
- Parameter 25 is used to filter out peaks that are present in only one group. The extra edit box is a filter for this option. It determines the minimum number of masses that should be present for a "compound" that is present only in one group
- Parameter 26 is a condition. With this condition you conclude if peaks present in group 2 are larger than in group 1 or vice versa
- Button 27 executes part C and creates a selection of peaks on the basis of the parameters set (22–26)
- Button 28 gives similar output as described at button 21

12| Once peak extraction and baseline correction settings are satisfactory, run baseline correction for all samples. Note that baseline correction is the most time-consuming part of MetAlign and can take a few hours for 100 samples (depending on the configuration of the computer).

13| After baseline correction of the entire series, inspect retention shifts in the baseline-corrected data files of the reference sample and of the first and last samples of the entire data set. Set maximum shift at initial peak searching criteria (parameter 13, see **Box 3** for more details) according to default settings, or to a value at least a factor of 2 higher than visually observed retention shifts and higher than that set in parameter 9. In most experiments on related samples, we use the iterative alignment with parameters indicated in **Figure 3** and its legend (see also examples in the MetAlign manual).
? TROUBLESHOOTING

14| To prevent MetAlign outputting mass peaks that are detected in only one or a few samples, for example, owing to impurities present in one extract, it is recommended to increase parameter 18 (see **Box 3** for more details) to a value corresponding to the number of replicates or to relevant statistical units.

15| After running the alignment (button 20), create the data output file (button 21, see **Box 3** for more details).

Identification of relevant metabolites

16| Retrieve accurate masses of filtered mass peaks in the raw data file manually. Inspect absorbance spectra, recorded by the PDA detector, of compounds of interest.
? TROUBLESHOOTING

17| Perform additional LC-QTOF MS/MS fragmentation experiments for further identification. Enter selected masses into a mass inclusion list to ensure isolation in the quadrupole (data-directed MS/MS).

18| Predict the elemental composition of the mass peaks of interest from the accurate mass calculation, together with MS/MS fragmentation, isotopic patterns and, if possible, specific absorbance spectra.

19| Use the elemental formulae obtained to search the internet or commercially available compound databases (e.g., Database of Natural Products on CD-ROM) for possible candidates. As a first step to facilitate the query of LC-MS based plant metabolomics data, an open access database for identified semi-polar metabolites, currently mainly (poly)phenolic compounds, detected in tomato fruit has recently been developed[22] and can be searched at http://appliedbioinformatics.wur.nl/moto/. This database is derived using the protocol described here. However, in untargeted LC-MS, most of the elemental compositions detected in plant extracts are still unknown or reference compounds are not commercially available[22,24,25]. Therefore, many of the putatively annotated structures cannot yet be unambiguously identified without using NMR or other tools.

● **TIMING**

The timing of the procedure from tissue handling up to the final output for subsequent statistical analyses (matrix of intensity of aligned mass peaks versus samples) is schematized in **Figure 4**. For about 50 *Arabidopsis* samples, the sampling step, which includes grinding in liquid nitrogen using a ball mill and weighing of frozen tissues, can be completed in 2 days. However, for the same amount of samples from larger plant tissues such as tomato fruit and potato tubers, these activities usually take more time: about 4 days. Subsequent sample extraction, conditioning the LC-MS, extract analysis and mass peak alignment by MetAlign will take about 4 days for 50 samples, irrespective of the type and origin of tissue. Depending on the research question, much more time may be needed for further interpretation of the comprehensive metabolomics data set including statistical filtering and identification of relevant mass peaks.

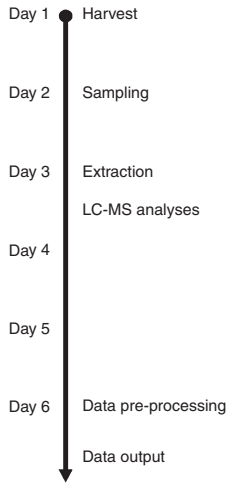

Figure 4 | Timing of standard procedure of untargeted LC-MS analyses, based on 50 *Arabidopsis* seedling samples and LC-MS analysis time of 1 h. For large plant tissues such as tomato fruits, the sampling step (including grinding and weighing) can take 4 days, resulting in a total time of 8 days for 50 samples.

? TROUBLESHOOTING

Major problems are not expected when applying this protocol if the advice given in the critical steps is adhered to. If during sample preparation the material is thawed, the material should be discarded. If the LC flow is stopped for any reason or the MS runs out of nitrogen gas, analyze at least four samples as dummies to re-stabilize the system. Upon malfunction of the

MS system, for example, sudden decrease in detector sensitivity, reset the instrument and test sensitivity and mass accuracy (re-calibrate if required). Meanwhile, the extracts can be stored at 4-10 °C for at least 1 week. Before re-running all samples, always sonicate vials or inserts to re-dissolve possible precipitates, and filter.

If, upon MetAlign processing, there seem to be insufficient land-mark peaks (i.e., mass signals common in all samples) for proper iterative alignment, a message will automatically be displayed. This can be the case if comparing highly unrelated samples ("apple and pears"). If such comparison is still essential for the research question, we recommend to lower parameters 16 and/or 17 or, alternatively, use the rough alignment tool at button 14 (see also **Box 3** and MetAlign manual).

With regard to accurate mass calculation, the mass accuracy of an ion detected by the QTOF-Ultima MS is in principle highest at signal intensities that are comparable to that of the local lock mass[22]. Thus, if in all samples the mass signal of interest is lower than about half the intensity of the lock mass, it is impossible to calculate its exact mass using this type of mass spectrometer. Lowering the lock mass intensity during analysis is not recommended, as this will prevent an accurate estimation of the lock mass itself. At low mass signals, it is difficult to obtain informative MS/MS fragmentation as well. Strategies to increase the mass signal, such as injecting higher sample volumes, analyzing in the opposite ionization mode, using a different ionization source (e.g., APCI) or post-column addition of ionization promoters (e.g., ammonium acetate), may be tested. Alternatively, the compound of interest can be concentrated or the sample can be re-analyzed by other instruments with higher mass accuracy and/or MS/MS capabilities at a low mass intensity range.

ANTICIPATED RESULTS

As this untargeted metabolomics protocol makes use of crude 75% aqueous methanol extracts of plants coupled to C_{18}-reversed phase LC and ESI-MS, the technique described is slightly biased toward semi-polar secondary metabolites. Nevertheless, within the same extracts, a number of primary metabolites, for example, several organic acids, nucleotides, amino acids, sugars and their phosphorylated forms, can be detected by this technique as well. However, as most of these primary metabolites are highly polar and usually co-elute with other compounds in the injection peak when using this type of columns, one should be aware that differences detected in the intensity of polar mass signals may result from differential degrees of ion suppression. Results on polar compounds obtained with this protocol should be checked with alternative LC systems[21,36] or other metabolomics techniques (e.g., GC-TOF MS, CE-MS).

As shown in **Figure 5**, the protocol described here enables highly stable chromatography and mass signal detection throughout analysis of large sample series. As the quality of MetAlign-assisted data alignment and untargeted sample comparison is higher with increasing reproducibility of chromatography, the maximum drift in retention time of (known) compounds over the sample series analyzed should be as small as possible and preferably less than 10 s (**Fig. 5a**). Larger retention shifts usually indicate column deterioration, trapped air bubbles or changes in eluent pH. Technical variation in relative quantification of mass signals between samples, which can be introduced at each step from 1 to 16 of PROCEDURE, can be calculated from the intensities of (known) mass peaks (**Fig. 5b**). The coefficient of variation in intensities between replicate samples should be less than 25% overall, and is usually less than 10% for the higher abundant signals[22]. In addition, technical reproducibility can be estimated by creating scatter plots of all mass peaks from replicate samples[25]. Upon adequate mass calibration and by using lock mass correction on-line, the accurate masses of ions detected are usually stable throughout large sample series (**Fig. 5c**). With the TOF resolution used and at a signal intensity that is comparable to that of the lock mass, the observed accurate mass of a compound of interest should be within 5 p.p.m. deviation from the calculated mass. In our laboratory, we use a script called

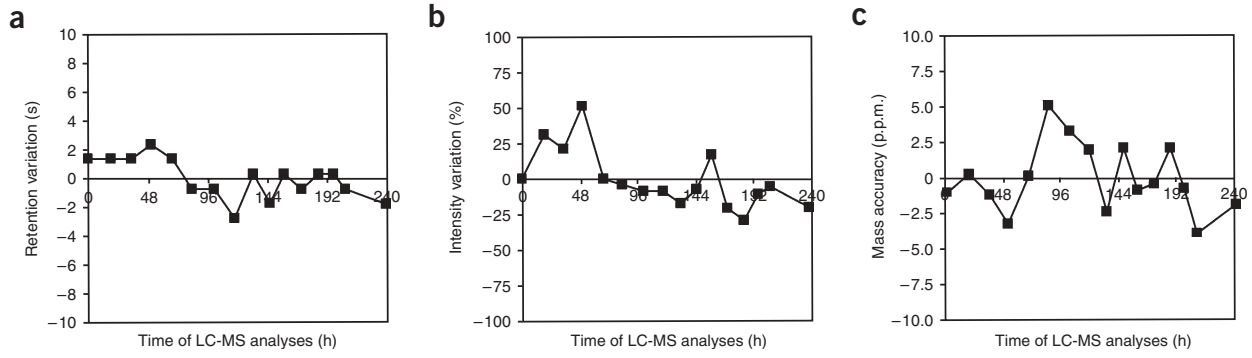

Figure 5 | Stability of the LC-QTOF MS system during 240 h continuous analyses of crude plant extracts (ESI negative mode). From a homogenous batch of *Brassica nigra* leaf tissue, 16 replicate extracts were prepared and analyzed throughout a series of 240 samples, using a run time of 1 h per sample. Variation between replicates in the detection of rutin (for identification, see **Fig. 1f**) is indicated. (**a**) Retention drift during analyses, expressed in seconds deviation from the mean retention time (23.195 min ± 1.3 s; n=16). (**b**) Variation in mass signal intensity (peak height calculated by MetAlign), expressed as percentage deviation from the mean intensity (1,721 ± 355 counts per scan, coefficient of variation=21%; n=16), versus time of analysis. Variation is the sum of all technical variation including weighing, extraction, LC-MS analysis and data processing. (**c**) Variation in accurate mass measurement, in p.p.m. deviation from the mean of accurate masses calculated on the top of chromatographic peaks. Scale of y axis: −10.0 to +10.0 p.p.m.

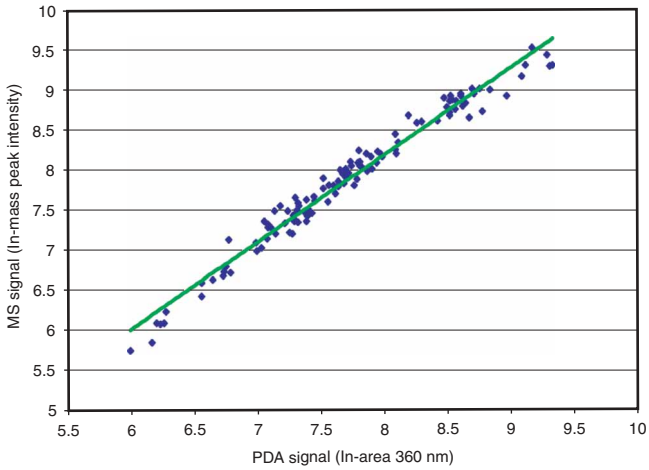

Figure 6 | Correlation between conventional LC-PDA analysis and untargeted LC-MS-based metabolomics with regard to detection of the flavonoid rutin (for identification, see **Fig. 1f**). Ripe fruits of 114 different tomato cultivars were analyzed by LC-PDA-QTOF MS in ESI negative mode, as described in this protocol. LC-PDA signals (peak areas at 360 nm) were subsequently extracted in a targeted manner using the QuanLynx tool of Masslynx, whereas LC-MS parent ion signals were retrieved in an untargeted manner using MetAlign. Ln-transformed data show high linear correlation (y=1.0937x with $r^2 = 0.972$; $P < 2.5 \times 10^{-7}$), indicating that the untargeted approach is equivalent to the targeted (conventional) LC-PDA approach.

MetAccure[22,25] to select scans within a user-defined intensity ratio of sample versus lock mass, to enable automated and correct accurate mass calculations. By calculating the mean values of observed accurate masses of compounds across all samples analyzed, mass accuracies of 2 p.p.m. or better can be obtained[22].

Reversed phase LC with PDA detection has been used for decades for quantitative analysis of many secondary metabolites in plants. As the analytical system described in this protocol consists of reversed phase LC coupled to both PDA and MS, the quality of the untargeted LC-MS data can be checked by comparing with LC-PDA data of the same samples (**Fig. 6**). After log transformation of both data, a significant and linear correlation should be achieved between a mass peak signal obtained by untargeted metabolomics and peak area obtained by conventional LC-PDA analysis. A low correlation may indicate significant ion

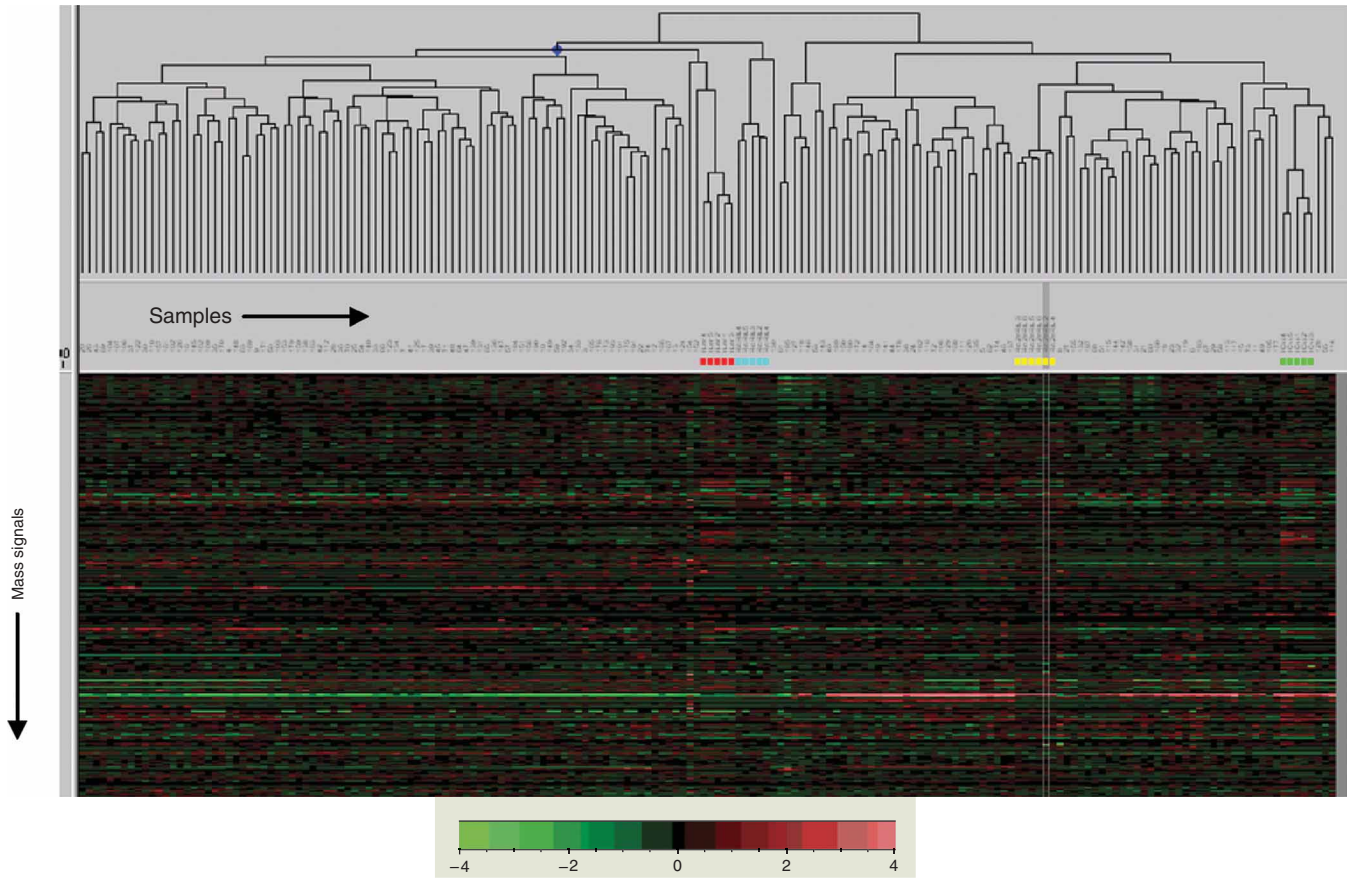

Figure 7 | Hierarchical clustering (Pearson correlation) of 180 *A. thaliana* genotypes consisting of a recombinant inbred line (RIL) population and their parents, based on untargeted metabolomics data. Samples were analyzed by LC-QTOF MS (30-min run), and 5,783 mass peaks, extracted and aligned by MetAlign, were loaded into GeneMaths software for multivariate analyses. Mass signal intensities (y axis) were nlog-transformed and standardized per raw average (each raw representing single mass peak), with color scale given in the lower panel (green indicates relatively low and red indicates relatively high intensity). Replicate samples are indicated with the same color on the sample key (x axis): yellow- and blue-colored samples are replicate analyses of two different samples each composed of a mixture of RILs, to check for LCMS reproducibility and alignment; green- and red-colored samples represent five biological replicates of the Ler and Cvi parents, respectively.

suppression, MS detector saturation or marked misalignments. However, correlations can only be established for compounds that show clearly separated PDA peaks in the chromatograms.

The aligned data sets can also be imported into software packages for large-scale multivariate or statistical analyses, such as GeneMaths[25] and MetaNetwork[48]. We recommend loading mass peak data as nlog-transformed values. We routinely use GeneMaths software to check the quality of the mass signal output from large-scale experiments, by applying principle component analysis and hierarchical clustering. In these multivariate approaches, replicate samples should cluster relatively closely, as compared to for example, different genotypes (**Fig. 7**), plant treatments or tissues, and the segregation of the scores should be according to the expected data structure[25] (if applicable).

ACKNOWLEDGMENTS The preparation of this paper and the work described herein was made possible through funding from the Centre for BioSystems Genomics (which is part of the Netherlands Genomics Initiative and The Netherlands Organisation for Scientific Research), Plant Research International (PRI) and the EU project META-PHOR (Food-CT-2006-03622). We thank Harry Jonker and Bert Schipper (PRI) and Jeroen Jansen (NIOO, Heteren, The Netherlands) for their excellent help in sample preparation and LC-PDA-QTOF MS analyses.

COMPETING INTERESTS STATEMENT The authors declare that they have no competing financial interests.

Published online at http://www.natureprotocols.com

1. Bino, R.J. *et al.* Potential of metabolomics as a functional genomics tool. *Trends Plant. Sci.* **9**, 418–425 (2004).
2. Hall, R.D. Plant metabolomics: from holistic hope, to hype, to hot topic. *New Phytol.* **169**, 453–468 (2006).
3. Jenkins, H. *et al.* A proposed framework for the description of plant metabolomics experiments and their results. *Nat. Biotechnol.* **22**, 1601–1606 (2004).
4. Sumner, L.W., Mendes, P. & Dixon, R.A. Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry* **62**, 817–836 (2003).
5. Dixon, R.A. *et al.* Applications of metabolomics in agriculture. *J. Agric. Food Chem.* **54**, 8984–8994 (2006).
6. Trethewey, R.N. Metabolite profiling as an aid to metabolic engineering in plants. *Curr. Opin. Plant Biol.* **7**, 196–201 (2004).
7. Saito, K., Dixon, R. & Willmitzer, L. *Plant Metabolomics* (Springer Verlag, Heidelberg, Germany, 2006).
8. Vaidyanathan, S., Harrigan, G.G., Goodacre, R. (eds.) *Metabolome Analyses: Strategies for Systems Biology* (Springer, New York, 2005).
9. Van der Greef, J., Stroobant, P. & Van der Heijden, R. The role of analytical sciences in medical systems biology. *Curr. Opin. Chem. Biol.* **8**, 559–565 (2004).
10. Fernie, A.R. Metabolome characterization in plant system analysis. *Funct. Plant Biol.* **30**, 111–120 (2003).
11. Fiehn, O. *et al.* Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* **18**, 1157–1161 (2000).
12. Lisec, J., Schauer, N., Kopka, J., Willmitzer, L. & Fernie, A.R. Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nat. Protoc.* **1**, 1–10 (2006).
13. Roessner, U. *et al.* Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* **13**, 11–29 (2001).
14. Roessner, U., Willmitzer, L. & Fernie, A.R. Metabolic profiling and biochemical phenotyping of plant systems. *Plant Cell Rep.* **21**, 189–196 (2002).
15. Schauer, N. *et al.* GC-MS libraries for the rapid identification of metabolites in complex biological samples. *FEBS Lett.* **579**, 1332–1337 (2005).
16. Fiehn, O. *et al.* Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* **18**, 1157–1161 (2000).
17. Aharoni, A. *et al.* Nontargeted metabolome analysis by use of Fourier transform ion cyclotron mass spectrometry. *Omics* **6**, 217–234 (2002).
18. Hirai, M.Y. *et al.* Elucidation of gene-to-gene and metabolite-to-gene networks in *Arabidopsis* by integration of metabolomics and transcriptomics. *J. Biol. Chem.* **280**, 25590–25595 (2005).
19. Overy, S.A. *et al.* Application of metabolite profiling to the identification of traits in a population of tomato introgression lines. *J. Exp. Bot.* **56**, 287–296 (2005).
20. Goodacre, R., York, E.V., Heald, J.K. & Scott, J.M. Chemometric discrimination of unfractionated plant extracts analyzed by electrospray mass spectrometry. *Phytochemistry* **62**, 859–863 (2003).
21. Jander, G. *et al.* Application of a high-throughput HPLC-MS/MS assay to *Arabidopsis* mutant screening; evidence that threonine aldolase plays a role in seed nutritional quality. *Plant J.* **39**, 465–475 (2004).
22. Moco, S. *et al.* A liquid chromatography-mass spectrometry-based metabolome database for tomato. *Plant Physiol.* **141**, 1205–1218 (2006).
23. Tolstikov, V.V., Lommen, A., Nakanishi, K., Tanaka, N. & Fiehn, O. Monolithic silica-based capillary reversed-phase liquid chromatography/electrospray mass spectrometry for plant metabolomics. *Anal. Chem.* **75**, 6737–6740 (2003).
24. von Roepenack-Lahaye, E. *et al.* Profiling of *Arabidopsis* secondary metabolites by capillary liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry. *Plant Physiol.* **134**, 548–559 (2004).
25. Vorst, O. *et al.* A non-directed approach to the differential analysis of multiple LC-MS-derived metabolic profiles. *Metabolomics* **1**, 169–180 (2005).
26. Rischer, H. *et al.* Gene-to-metabolite networks for terpenoid indole alkaloid biosynthesis in *Catharanthus roseus* cells. *Proc. Natl. Acad. Sci. USA* **103**, 5614–5619 (2006).
27. Sato, S., Soga, T., Nishioka, T. & Tomita, M. Simultaneous determination of the main metabolites in rice leaves using capillary electrophoresis mass spectrometry and capillary electrophoresis diode array detection. *Plant J.* **40**, 151–163 (2004).
28. Le Gall, G., Colquhoun, I.J., Davis, A.L., Collins, G.J. & Verhoeyen, M.E. Metabolite profiling of tomato (*Lycopersicon esculentum*) using ^{1}H NMR spectroscopy as a tool to detect potential unintended effects following a genetic modification. *J. Agric. Food Chem.* **51**, 2447–2456 (2003).
29. Ward, J.L., Harris, C., Lewis, J. & Beale, M.H. Assessment of H-1 NMR spectroscopy and multivariate analysis as a technique for metabolite fingerprinting of *Arabidopsis thaliana*. *Phytochemistry* **62**, 949–957 (2003).
30. Huhman, D.V. & Sumner, L.W. Metabolic profiling of saponins in *Medicago sativa* and *Medicago truncatula* using HPLC coupled to an electrospray ion-trap mass spectrometer. *Phytochemistry* **59**, 347–360 (2002).
31. Breitling, R., Pitt, A.R. & Barrett, M.P. Precision mapping of the metabolome. *Trends Biotechnol.* **24**, 543–548 (2006).
32. Verhoeven, H.A., de Vos, C.H., Bino, R.J. & Hall, R.D. Plant metabolomics strategies based upon quadrupole time of flight mass spectrometry (QTOF-MS). in *Plant Metabolomics—Biotechnology and Forestry* Vol. 57, pp. 33–48 (eds. Saito, K., Dixon, R.A. & Willmitzer, L.) (Springer-Verlag, Berlin, Heidelberg, 2006).
33. Beekwilder, J., Jonker, H., Meesters, P., Hall, R.F., van der Meer, I.M. & de Vos, C.H.R. Antioxidants in raspberry: on-line analysis links antioxidant activity to a diversity of individual metabolites. *J. Agric. Food Chem.* **53**, 3313–3320 (2005).
34. Hall, R.D., de Vos, C.H.R., Verhoeven, H.A. & Bino, R.J. Metabolomics for the assessment of functional diversity and quality traits in plants. in *Metabolome Analyses-Strategies for Systems Biology* (eds. Vaidyanathan, S., Harrigan, G.G. & Goodacre, R.) (Springer, New York, 2005).
35. Kopka, J. *et al.* GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics* **21**, 1635–1638 (2005).
36. Tolstikov, V.V. & Fiehn, O. Analysis of highly polar compounds of plant origin: combination of hydrophilic interaction chromatography and electrospray ion mass trap spectrometry. *Anal. Biochem.* **301**, 298–307 (2002).
37. Peterman, S.M., Duczak, N., Kalgutkar, A.S., Lame, M.E. & Soglia, J.R. Application of a linear ion trap/orbitrap mass spectrometer in metabolite characterization studies: examination of the human liver microsomal metabolism of the non-tricyclic anti-depressant nefazodone using data-dependent accurate mass measurements. *J. Am. Soc. Mass Spectrom.* **17**, 363–375 (2006).
38. Exarchou, V., Godejohann, M., van Beek, T.A., Gerothanassis, I.P. & Vervoort, J. LC-UV-solid-phase extraction-NMR-MS combined with a cryogenic flow probe and its application to the identification of compounds present in Greek oregano. *Anal. Chem.* **75**, 6288–6294 (2003).
39. Wilson, I.D. & Brinkman, U.A.T. Hyphenation and hypernation––the practice and prospects of multiple hyphenation. *J. Chromatogr. A* **1000**, 325–356 (2003).

40. Wolfender, J.L., Ndjoko, K. & Hostettmann, K. Liquid chromatography with ultraviolet absorbance-mass spectrometric detection and with nuclear magnetic resonance spectroscopy: a powerful combination for the on-line structural investigation of plant metabolites. *J. Chromatogr. A* **1000**, 437–455 (2003).
41. Nordström, A., O'Maille, G., Qin, C. & Siuzdak, G. Nonlinear data alignment for UPLC-MS and HPLC-MS based metabolomics: quantitative analysis of endogenous and exogenous metabolites in human serum. *Anal. Chem.* **78**, 3289–3295 (2006).
42. Laaksonen, R. *et al.* A systems biology strategy reveals biological pathways and plasma biomarker candidates for potentially toxic statin-induced changes in muscle. *PLoS ONE* **e97** (2006).
43. Keurentjes, J.J.B. *et al.* The genetics of plant metabolism. *Nat. Genet.* **38**, 842–849 (2006).
44. Bino, R.J. *et al.* The light-hyperresponsive *high pigment-2*^{dg} mutation of tomato: alterations in the fruit metabolome. *New Phytol.* **166**, 427–438 (2005).
45. Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* **78**, 779–787 (2006).
46. Katajamaa, M. & Oresic, M. Processing software for differential analysis of LC/MS profile data. *BMC Bioinformatics* **6**, 179.1–179.12 (2005).
47. Idborg, H., Zamani, L., Edlund, P., Schuppe-Koistinen, I. & Jacobsson, S.P. Metabolic fingerprinting of rat urine by LC/MS. Part 2. Data pretreatment methods for handling of complex data. *J. Chromatogr. B* **828**, 14–20 (2005).
48. Fu, J., Swertz, M.A., Keurentjes, J.J.B. & Jansen, R.C. MetaNetwork: a computational protocol for the genetic study of metabolic networks. *Nat. Protoc.* (in the press) DOI: 10.1038/nprot.2007.96 (2007).
49. Tikunov, Y. *et al.* A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. *Plant Physiol.* **139**, 1125–1137 (2005).
50. Wolff, J.C., Eckers, C., Sage, A.B., Giles, K. & Bateman, R. Accurate mass liquid chromatography/mass spectrometry on quadrupole orthogonal acceleration time-of-flight mass analyzers using switching between separate sample and reference sprays. 2. Applications using the dual-electrospray ion source. *Anal. Chem.* **73**, 2605–2612 (2001).