Metabolite Profiling in Arabidopsis

Oliver Fiehn
Max-Planck-Institute of Molecular Plant Physiology
Postfach, 14476 Potsdam/Golm
Email: fiehn@mpimp-golm.mpg.de , phone: +49-331-5678-216, fax -250

Key words: mass spectrometry, GC-MS, metabolomics, data mining

Abstract

Metabolite profiling is the multiparallel relative quantification of a mixture of compounds or compound classes using chromatography and universal detection technologies (GC-MS, LC-MS). In this respect it is an extension of classical single target methods from which it can be distinguished by its broader view on profiling major biochemical events. This broader scope of analysis outweighs the disadvantages by taking compromises in method development and the reduced accuracy for specific metabolites. This chapter exemplifies the strategies in metabolite profiling of polar compounds by GC-MS. It gives experimental details on the basic steps: harvest, homogenization, extraction, fractionation, concentration, derivatization, data acquisition, raw data processing and result data tranformation.

1. Introduction

Metabolite profiling is an analytical method for relative quantitation of a number of metabolites from biological samples ***(1)***. Commonly, these samples have been garnered from a specific tissue or a part of a tissue of interest, but, depending on the biological question, also either from a larger mixture of different organs (such as whole-shoots) or conversely on a micro scale from single cells or purified organelles. Metabolite profiling is distinguished from other analytical procedures by its scope:

(a) *Target analysis* is constrained to one or a very few target compounds (such as phytohormones). Such targets are usually quantified in an absolute manner using calibration curves and/or stable isotope labeled internal standards.

(b) *Metabolite profiling* restricts itself to a certain range of compounds or even to screening a pre-defined number of members of a compound class. Within these constraints, a single analytical platform may be sufficient. Examples might be the analysis of xantophyll cycle intermediates by liquid chromatography/diode array UV detection (HPLC-UV), or sugars, hydroxy – and amino acids by fractionation and gas chromatography/mass spectrometry (GC-MS), or membrane lipid profiling by HPLC-MS/MS. Quantification in metabolite profiling is usually carried out relative to comparator samples, such as positive and negative controls.

(c) *Metabolomics* seeks for a truly unbiased quantitative and qualitative analysis of all biochemical intermediates in a sample. It must not be restricted by any physiochemical property of the metabolites, such as molecular weight, polarity, volatility, electrical charge, chemical structure and others. Since there is currently no single technology available that would allow such comprehensive analysis, metabolomics is characterized by the use of multiple techniques and unbiased software. Metabolomics also uses relative quantification. In addition, it must include a strong focus on *de novo* identification of unknown metabolites whose presence is demonstrated.

(d) *Metabolite fingerprinting* is different from the other three approaches in that it does not aim to physically separate individual metabolites. Instead, spectra from full sample extracts are acquired by a single instrument (such as ^{1}H-NMR). Spectra are then compared by multivariate statistics in order to find spectral regions that discriminate samples by their biological origin. In some instances, these regions may again point

towards specific metabolites; in general, however, one dimensional methods are restricted in resolving complex mixtures.

Metabolite profiling therefore stands between classical *target analysis* and cutting-edge *metabolomics*. Since it often aims at chemically very different compounds, there are various methods published for acquiring metabolite profiles from a certain tissue. Each procedure will be a compromise between several parameters such as compound stability, solubility, influence of the cellular matrix, time needed to carry out the protocol, available devices for tissue harvesting, homogenization, extraction, fractionation, submission to analytical instruments, raw data analysis and statistics. For example, a protocol found to be well suited for the analysis of lipophilic leaf membrane lipids and cuticle waxes will be very much different from one that aims at hydrophilic sugars and amino acids. Any protocol on metabolite profiling will therefore have validation criteria that are different from target analysis: (a) the reproducibility of the method is far more important than the absolute recovery. (b) the robustness and achievability of a method is more important than its absolute precision. (c) the comprehensiveness of a method is more important than the inclusion of a certain metabolite that is missed. (d) the overall dynamic range for the majority of compounds is more important than the detection limit for a specific substance. (e) the ability to include important known key metabolites is more important than the detection of unidentified peaks that might be biochemical side-products of enzymes with low substrate specificity.

In this respect, the most important validation criterion for a metabolite profiling protocol is the exact definition of its scope ***(2)***, and here, the identity and analytical reproducibility for each selected compound in a given biological matrix. In this chapter, the acquisition of metabolite profiles for hydrophilic and semi-polar compounds is presented. It is based on the use of inexpensive and robust technologies (such as GC/quadrupole MS) that is found in many biological laboratories, and not on more sophisticated, pilot-type instrumentation. The basic steps in the process **(Fig. 1)** can be summarized as:

1. Design randomized plots for plant growth, with an experimental design according to the biological question.
2. Harvest and weigh the plant tissue quickly and immediately freeze it in liquid nitrogen.
3. Homogenize the plant tissue in the frozen state.

4. Extract the tissue in a comprehensive and mild way concomitant with enzyme inactivation and add internal standards.
5. Fractionate the extract into a polar and a lipophilic fraction.
6. Dry down the polar fraction.
7. Derivatize the polar fraction by first adding methoxyamine.pyridine, and then a trimethylsilylating agent.
8. Analyze the derivatized sample by GC-MS.
9. Process the raw GC-MS data.
10. Normalize and transform the result data, and perform statistical evaluations.

The basic theoretical considerations behind this process are quite simple: the measured metabolite levels should reflect the *in vivo* state. Therefore, any artifactual formation by chemical processes, or any post-harvest biochemical activity, must be prevented. Biochemical inactivation can be ensured by coagulation of enzymes, either using heat-shock or cold-shock methods, with the help of organic solvents such as chloroform or acetonitrile that force protein precipitation. Conversely, chemical artifact formation depends on the stability of each specific compound and is therefore hard to predict. Generally, any harsh treatment of the metabolome mixture should be avoided. Instead, conditions for extraction, storage, chemical derivatization and analysis should be as mild as possible (and, on the other hand, as comprehensive and universal as possible.)

2. Materials

2.1 Instruments and preparations for harvesting plant organs or tissues

A digital camera is helpful for recording the state of plant tissues or organs prior to harvesting. A balance with a precision of ± 0.1 mg is needed at the harvest site in order to allow an accurate fresh weight determination of plant tissues. For separating the tissue from the plant, scissors or other appropriate devices are needed, and tweezers for picking up the collected tissue. Round-bottomed, uncolored and pre-labeled 2 mL micro centrifuge tubes, equipped with 5 mm i.d. metal balls are set aside to collect each individual sample. Dewars with liquid nitrogen will ensure the immediate arrest of biological activity after weighing of the samples. A notebook in electronic or paper format is needed to keep track on sample identity numbers, developmental stage, weights, day, time, duration of harvesting, and physiological parameters like humidity, light fluence rate, light period (day length) and

temperature at harvest.

2.2 Homogenisation and extraction

Homogenisation is carried out by a ball mill (available e.g. from Retsch, Germany). For preparing the extraction mixture, degassing devices (such as vacuum/ultrasonic bath, or pure argon or nitrogen gas bombs) and a liquid cooling system must be available. A freshly prepared, chilled (-15°C) and degassed mixture of chloroform, methanol and water is prepared at a ratio of 1 : 2.5 : 1 (v/v/v). For each solvent, the highest quality (e.g. >99% ultra-pure HPLC-MS gradient grade purity) is used and stored at room temperature in the dark. Volumes are measured using calibrated pipettes whose accuracies are subjected to quality control routines at least once every six months. An ice bath and liquid nitrogen dewars are used for temporarily storing samples during the process. A centrifuge (rcf 20.800) for micro centrifuge tubes is used for fractionation. Extraction is performed in a micro centrifuge tube shaker.

2.3 Derivatisation

A speed vacuum concentrator (e.g. Heraeus, Germany) is used for drying extracts to complete dryness. A mixture of 20 mg/mL of methoxyamine.HCl in pyridine (p.a. quality) is freshly prepared. N-methyl-N-trimethylsilyltrifluoroacetamide (MSTFA) is used from freshly opened 1-mL bottles. Derivatisations are carried out in micro centrifuge tube thermoshakers the temperatures of which can be set to 28°C and 37°C.

2.4 Mass spectrometric analysis

GC-MS analysis is carried out on a quadrupole mass spectrometer equipped with autosampler and electron impact ionization. Samples must be injected in randomized order or appropriate block designs. For each injection sequence, the analysis of quality control samples is a prerequisite (e.g. reagent blanks, method blanks, reference compound mixture, reference sample). Standard glass liners with glass wool are used in splitless or split mode. Low bleeding injector septa are preferable. Standard 10 µL injection needles (e.g. Hamilton) are mounted into the autosampler. Chromatography is carried out on a 30 m long, 0.25 mm I.D. and 0.25 µm (5%-Phenyl)-methylpolysiloxane column with an additional 10 m integrated guard column. The GC oven must be temperature programmable up to 350 °C. The quadrupole mass spectrometer must having minimum scanning duty cylces of at least 2 s^{-1} and a mass range of at least 83-500 Da. Raw GC-MS data files are transferred to servers.

Long-term data safety is ensured by back up routines on DVD's or by mirrored server space. Data analysis is carried out on office personal computers using a GC-MS software that is able to carry out multi-target analysis, including compound identity checks based on mass spectral and retention index matching. The software must be capable of quantitation by area and height on user defined ion traces. Examples for such software are Agilent's ChemStation or ThermoFinnigan's old MassLab software.

3. Methods

3.1 Harvesting (see Note 4.1)

1. Take digital pictures of the plants of interest, indicating the plants that are going to be harvested.
2. Cut the organs of interested with scissors or a cork borer.
3. Collect the plant organ with tweezers and weigh it. The range, which is validated with respect to the amount of 1 mL extraction solvent, is 10-50 mg FW.
4. Store the organ in a labeled 2 mL micro centrifuge tube that contains a metal ball.
5. Close the tube and place it into liquid nitrogen.

3.2 Homogenization (see Note 4.2)

1. Pre-chill two ball mill tube holders in liquid nitrogen
2. Quickly take out six samples from liquid nitrogen and place them into tube holders (three samples per holder)
3. Fix tuber holders in ball mill and mill at 25 s^{-1} for 30 s
4. Quickly take out the six samples and return them to liquid nitrogen
5. Repeat steps 1-4 until all samples are homogenized

3.3 Extraction and fractionation (see Note 4.3)

1. Take out homogenized samples one by one and add 1 mL of cold extraction solvent (-15°C, degassed) to each ***(3)***.
2. Add internal standards, e.g. U-^{13}C-Sorbitol (200 ng per vial) for normalization.
3. Shake the samples in batches of 10 for 5 min in a 4°C room. When taking out the samples, place them in an ice bath.
4. Centrifuge samples at 20.800 rcf for 2 min.

5. Collect the liquid supernatant of each sample and store in a clean micro centrifuge tubes. The metal balls can be re-used after cleaning. The cell debris pellet can be discarded.
6. Repeat steps 1-5 until all samples are extracted.
7. For storage, samples must be deoxygenated with a gentle stream of nitrogen or argon gas for 1 min prior to tube closure. Tubes can then be stored in the dark at -80°C for about four weeks.
8. Add 400 µL of pure water to each sample, and vortex for 10 s.
9. Centrifuge samples at 20.800 rcf for 2 min.
10. Collect the upper phase of each sample (a mixture of water and methanol, the 'polar phase') and store it in a new micro centrifuge tube. The lower phase (a mixture of chloroform and methanol, 'lipophilic phase') can be used for lipid metabolite profiling or discarded.
11. Close the polar phase sample tubes with spare tube caps that have been punched with 50 µm O.D. needles.
12. Dry the polar phase samples in a speed vacuum concentrator to complete dryness and remove the punched tube caps.
13. For storage, deoxygenate samples with a gentle stream of nitrogen or argon gas for 1 min before closing the tubes. Tubes can then be stored in the dark at -80°C for at least four weeks.

3.4 Derivatization (see Note 4.4)

1. Take out dried samples from store and allow them to warm up to room temperature for at least 15 min before opening.
2. Add 20 µL of methoxyamine solution (20 mg/mL in pyridine) to each sample, and immediately close each sample.
3. Shake samples for 90 min at 28°C.
4. Centrifuge samples at 14,000 min^{-1} for 30 s.
5. Add 180 µL silylating agent MSTFA to each sample, and immediately close each sample after methoxyamine addition.
6. Shake samples for 30 min at 37°C.
7. Transfer sample reaction solutions to glass vials suitable for the GC-MS autosampler. Immediately close each sample with crimps that contain a teflon rubber seal. Wait two hours before injecting the first sample into the GC-MS.

3.5 Data acquisition by GC-MS (see Note 4.5)

1. The mass spectrometer must be tuned according to the manufacturer's manuals for optimal parameters for ion lenses, detector voltage and other settings. Usually, this can be performed in autotune operation.
2. Change or clean the liner every 50 samples.
3. Check that manufacturere's recommended maintenance routines have all been carried out.
4. Inject 1 µL of each sample in splitless or split mode, depending on the metabolite concentrations and eventual signal-to-noise ratios in the GC-MS profiles. Injection temperature is set to 230°C. Injection programs have to include syringe washing steps before and after the injection, a sample pumping step for removal of small air bubbles, and an air buffer for complete sample removal during injection.
5. Separate metabolites using a GC temperature ramping program. Reasonable values are: GC start conditions at 80°C, 2 min isothermal, ramp with 5°C/min up to 330°C, 5 min isothermal, cool down to initial conditions. The ion source should be turned off during the solvent delay.
6. Detect metabolites by setting the ion source filament energy to 70 eV. Scan a mass range of at least 83-500 Da, or 40-500 Da, if low mass-to-charge (m/z) fragment ions are to be recorded. At least two scans per second should be recorded in full scan mode.
7. Transfer raw GC-MS profile chromatograms to a server station.

3.6 Data analysis (see Note 4.6)

1. For raw data processing, use appropriate software. First choice is the GC-MS manufacturer's software. For data deconvolution, the freely available software AMDIS is recommended (*http://chemdata.nist.gov/mass-spc/amdis/*)***(4)***.
2. Define target peaks that are to be included in the metabolite profiles.
3. Define optimal peak finding thresholds and quantification ion traces for each target compound. Peak identifications have to be carried out by matching retention indices and mass spectral similarity against a user-defined metabolite library. **(Fig. 2)**
4. Quantify metabolite peaks by area of target ion traces. Export result peak tables of all chromatograms to a data base or a PC office table calculation software (e.g. MS Excel 6.0).
5. Organize peak area results in a matrix of metabolites *vs.* chromatograms.

6. Count the number of detected metabolites per chromatogram. In case one or a few chromatograms show an unexplainable large deviation in the number of detected peaks, check the chromatograms visually and delete them from the result table.
7. For each target metabolite, count the number of chromatograms in which the metabolite could be positively identified. In case one or a few metabolites show an unexplainable large deviation in the number of positive peak findings, check the chromatograms visually, especially for the thresholds that were used for peak finding. Delete the metabolites from the result table that have lots of negative peak findings (missing values).
8. For each chromatogram, divide all peak areas by the area of the internal standard (e.g. U-^{13}C-sorbitol) and the sample weight. Log10 transform all data to down weight outliers and ensure a more Gaussian-type frequency distribution.
9. Calculate univariate statistics (e.g. t-test in Excel, ANOVA in MatLab).
10. Calculate multivariate statistics. Often, such calculations do not accommodate missing values for metabolites so suitable strategies must be employed for dealing with such occurrences. The results of multivariate statistics from two strategies have to be compared: (a) calculations that were carried out on all metabolites that had no missing values. (b) calculations that were carried on all metabolites after replacement of missing values.

4. Notes

4.1 Harvesting

Metabolite profiling starts with the experimental design of plant growth and harvest. It is a rather inexpensive technique, compared to proteomics or transcript expression analysis. Therefore, larger numbers of individual analysis can be carried out allowing rigid statistical assessments of the quantitative results. This allows the adequate addressing of the issue of natural biological variability, which usually contributes more to the standard deviation of metabolite mean levels than technical errors. In this respect, the first issue to consider is the randomization of plant growth and the accuracy of controlling and monitoring physiological conditions. Plants should be grown in complete randomization or in appropriate block designs, e.g. latin square. Although biological variation has been known for a great length of time, it is often underestimated due to the tradition in other branches of biology aiming at average values and qualitative

yes/no results ***(5)***. For example, metabolite levels vary dramatically over different zones of a leaf, and also between young, growing leaves and mature, fully expanded leaves. Depending on the biological question underlying a study, pooling strategies have to be designed to counteract this variability. For example, all rosette leaves of a single plant might be pooled and compared to other plants, or, alternatively, small disks of many individual plants could be collated and analyzed in batches. Similarly, plants grown in climate chambers or greenhouses should be grown in plots that are adopted from agronomical field trials. The reasoning is that microclimate conditions in climate chambers are not uniformly distributed, especially with respect to air circulations that lead to location effects by differences in individual plant transpiration rates.

4.2 Homogenization

For comprehensive extraction, a complete breakdown of plant cell walls is needed to ensure that the extraction solvent reaches all metabolites, independent of subcellular organization. Plant leaves include some 30 different cell types, all of which may have different internal metabolite levels. Homogenization therefore enables randomization over organs, which is a prerequisite for repeatability of analysis. For the scheme presented here, milling of frozen Arabidopsis leaves with a ball mill is proposed. However, for other plants or other organs, ball milling may not be sufficiently disruptive. Then, other devices may be used, e.g. an Ultra-Turrax that disrupt tissues with small rotating razor blades. Generally, also a simple pestle and mortar may be used (under liquid nitrogen). However, the throughput of this procedure is believed to be too low for preparing enough samples for accurate statistics. Still, if a ball mill or an Ultra-Turrax is not available, pestle and mortar are definitely regarded as valid tool for homogenization.

If you are using a ball mill, certain pitfalls should be avoided: firstly, it is wise to have more than one dewar available. Use one for chilling the tube holders, one for the samples that have not been homogenized, and one for samples that are already done. If you don’t want to lose your order of samples, you may also use paper boxes with separating inserts. The boxes would then be placed on dry ice and be filled with liquid nitrogen. The most important thing during homogenization is that samples must not thaw. For this reason, each ball mill tube holder is not filled with five but only with two or three samples: it is simply quicker and more convenient. If tubes are milled at a too

higher frequency (>25 s^{-1}) the balls might disrupt the micro centrifuge tube caps resulting in lost samples!

4.3 Extraction

The extraction is the most critical step in metabolite profiling. As stated above, it compromises between the demands for best possible metabolite recoveries, total metabolic comprehensiveness and time needed to perform the extraction. The protocol suggested here puts more emphasis on comprehensiveness and throughput, and less on recovery. For example, each sample might be extracted two or three times instead of a single extraction step. As given by Nernst's law, a duplicate or triplicate extraction would give a better recovery and accuracy. However, more time would be needed per sample. Using the proposed single extraction method, repeatability was found to be around 10% CV for most compounds. This is ensured by a well defined ratio of solvent volume to fresh weight of tissue (50 : 1). Comprehensiveness of extraction is given by the simultaneous use of highly polar solvents (water) and highly hydrophilic solvents (chloroform), with methanol being the mediator to avoid phase separation. Note, that chloroform will diffuse through micro centrifuge tube plastic faster than methanol or water will, if samples are stored for prolonged periods at –80°C. Chloroform/methanol at low temperatures will also help precipitating proteins, thereby ensuring the integrity of the metabolic composition. As important as stopping any enzymatic activity is avoiding oxidation. Solvents will contain huge amounts of oxygen if they are not degassed by vacuum/ultrasonicator or by bubbling inert gases through it (argon or nitrogen are most convenient). If deoxygenation is performed by gas exchange, great care must be taken to use ultrapure gases and clean bubble tips (e.g. rinsed Pasteur pipettes). Somewhat less important is the avoidance of light: some metabolites, such as catecholamines, will decompose if exposed to light for too long. For *Arabidopsi*s leaves, polar metabolite profiles will tell you whether you have carried out the protocol in a safe way, avoiding biochemical or physical alterations of metabolite compositions: (a) Low abundant cysteine must be present. It will disappear if there is oxygen left in the solvents. The accuracy of redox state preservation may also be checked by comparing the ratio of ascorbate/dehydroascorbate using this protocol and target assays for these compounds. (b) Low abundance glucose-6-phosphate and fructose-6-phosphate must be present. These compounds will disappear if enzymatic activity is not immediately stopped, and also through prolonged high temperature exposure (if heat shock enzyme

inactivation is used). (c) Classical compounds such as fumarate, malate, citrate, serine, threonine, aspartate, glutamate, glucose and sucrose tend to represent the most abundant peaks in Arabidopsis profiles. If any of these are missing (or in low abundance in comparison to others), the analysis has gone badly wrong.

When drying samples in a speed vacuum concentrator, caution should be taken to avoid sample losses, spilling or cross-contamination due to boiling retardation. This is the reason why extracts are dried with puntured plastic tube caps.

4.4 Derivatisation

The most critical point is to avoid any water or moisture during derivatization. Especially the silylating step is highly vulnerable. Problems can be detected through occurrence and abundance of polysiloxanes in GC-MS chromatograms. Such degradation (hydrolysation) products are recognized by their typical spectra with abundant ions m/z 221 and m/z 281. Generally, it is not needed to perform the derivatisation in completely dry athmosphere. However, water condensation due to early opening of sample tubes after cold storage must be avoided, as well as storage of derivatized samples in refrigerators or freezers. If, through bad luck, samples cannot be injected after derivatization, they should be stored in the dark at room temperature. After injections, sample vials will have received an amount of water, and seals may have been compromised. Re-analysis of samples that have already been used is not recommended, therefore. Temperatures and times of derivatization steps can be kept flexible, because they again present a compromise between completeness of reaction, time and efforts needed to perform the reactions, and breakdown of certain compounds (e.g. chemical conversion of glutamine to oxoproline). Pyridine serves as catalyst in the methoximation procedure which protects carbonyl moieties. It does not seem to be replaceable by other aprotic polar solvents. The volume ratios of pyridine.methoxyamine to MSTFA are again flexible: we here propose a ratio of 1:9, but other ratios like 1:2 or 1:1 have also been reported in the literature. Generally, lower amounts of pyridine will give better peak shapes for early eluting metabolites if splitless injections are carried out.

4.5 GC-MS

Take care to randomize your injection sequence: you must not inject the samples in the order of your underlying biological question, because there might always be subtle machine drifts that would obscure statistical analysis. Problems that are seemingly

attributed to the machine are usually neither the gas chromatograph nor the mass spectrometer ***(6)***: in >80% of cases the injection is to blame. Here, problems may occur due to dirt ('matrix') injected into the liner, the injector body and the first centimeters of the column. As for the column, problems may be recognized by decreasing intensity of trisaccharides. Shortening the column by 10 cm will help, but take care to readjust the total length of the column in the GC-MS front end, since this value will be taken to adjust the gas flow. With respect to the liners, glass wool will prevent the majority of non-volatile matrix constituents reaching the injector body or the column. However, ultimately also the injector body itself will become contaminated which will pyrolyse and build up spots with catalytic or adsorption properties, disabling high quality GC-MS runs. The level of contamination in the injector body is related to the type of liner used, and also to the presence of so-called cold spots in the specific injector type. A prolonged heating at 330°C for 6 hours (without column!) will cure this problem in most instances. If you plan long sample sequences which may even involve column changes, you need to refer to retention indices instead of retention times. Retention indices are calculated from retention times of internal marker peaks: usually alkanes are added to the samples to serve as retention anchor points. In the protocol presented above, an unusual scan range of 85-500 Da is proposed for mass spectrometric detection. Reasons for this choice are found in the properties of silylated compounds, which often have characteristic ions between 100-370 Da. Additionally, at m/z 79, bleeding of pyridine may infer mass spectra of low boiling compounds. For almost all peaks, m/z 73 is the most abundant ion – although this is helpful for lower limits of detection for pure compounds, m/z 73 does not have any selectivity power in metabolic profiles.

4.6 Data analysis

Metabolite profiles normally result in complex chromatograms that contain numerous overlapping peaks. For ensuring routine peak identification in high throughput operations (i.e. without manual inspection of chromatograms), mass spectral deconvolution is mandatory, especially for low abundant metabolites that might co-elute with abundant major peaks ***(Fig. 2)***. Deconvolution software will also suggest model ions that best discriminate a peak from its co-eluting neighbor compounds: hence, such model ions are already a good choice for calculating peak areas. Defining thresholds for peak finding is a difficult task. If the thresholds are set too high, a lot of peaks will not be taken as metabolite targets although these are actually present in the chromatograms:

such instances are called false negatives and would result in missing values in the resulting experiment data matrix. *Vice versa*, if the thresholds for mass spectrum matching, retention index windows, abundance and peak widths are set too low, peaks might be falsely taken as true target metabolites although these are actually be absent from the chromatograms (false positives). In any way, as much meta-information about a peak should be acquired as possible to ensure correctness of peak annotations in metabolite profiling data-bases. Arguably, the resulting data matrix will still contain missing values. Certain statistical tools such as principal component analysis will require complete data matrices without missing values, therefore these must be filled. Whatever is filled into these missing value cells in an automatic mode will carry a larger error than truly detected targets that passed all peak finding thresholds. Therefore, multivariate statistical analysis should be done twice: first with all metabolites that do not contain a single missing value, and secondly also including those metabolites for which empty cells have been replaced. In literature, there are many ways proposed of how to best perform such replacements. For less sophisticated approaches, a simple rationale might be to replace missing values by the arithmetic mean of each line (i.e. wild type control line, mutant 1 etc.). For cases in which a target metabolite is positively detected in less than 20% of the cases or in no chromatogram of the corresponding line at all, it may be suspected that it is indeed not present in this line and the actual peak findings are false positives. In this case, entering half of the detection limit might be a sensible way. In any case, the better way to replace missing values is to investigate the chromatograms one by one and replace the empty cells in the data matrix manually.

Last, the protocol proposed here suggests using the log10 transformation for down weighting outliers and transforming the data matrix to a more normal frequency distribution. In transcript microarray experiments, often the natural logarithm is taken, but there are no theoretical considerations that clearly vote for one or the other alternative. However, in any case it must not be forgotten that data need to be re-transformed when *x*-fold average values are to be computed, e.g. in mutant/wild type line comparisons.

Acknowledgments

Leaf extracts from an Arabidopsis cold stress experiment were kindly provided by Dana Wiese and Dr. Dirk Hincha, Max-Planck Institute of Molecular Plant Physiology, Potsdam,

Germany. The author thanks Dr. Gareth Catchpole for proofreading and manuscript corrections.

Fig.1. Scheme of the process of metabolite profiling. (1) Statistical design of plant growth, (2) harvest into liquid nitrogen, (3) homogenization, (4) extraction, (5) fractionation, (6) concentration to dryness, (7) derivatization, (8a) transfer to GC-MS vial, (8b) GC-MS data acquisition, (9) raw data processing, (10) data matrix transformation and statistics.

Fig.2. Example of differential analysis of two Arabidopsis thaliana accessions under cold stress situation by GC-MS metabolite profiling for polar leaf extracts.
Total ion chromatogram for overview purposes **(A):** Solid line: *Cvi*, dotted line: C24. Peak #1= internal standard, #2 fumarate, #3 serine, #4 threonate lactone, #5 threonine. Ion trace chromatograms **(B):** Quantification of peak #4 at m/z 247 (solid line) is achievable despite co-elution of abundant peak #5, which can be quantified at m/z 320 (dotted line). Mass spectrum at retention time 504.5 s **(C):** Metabolite identification before (upper panel) and after (lower panel) peak deconvolution. Automated peak finding and identification of low intensity peak #4 as threonate lactone (lower panel) is impossible without mass spectral deconvolution, due to co-elution of the highly abundant peak #5 (threonine) at this retention time (upper panel).

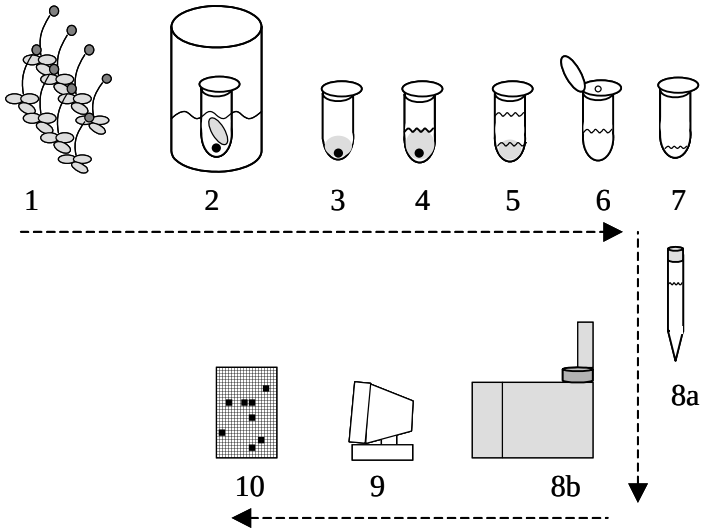

Fig.1. FIEHN

480

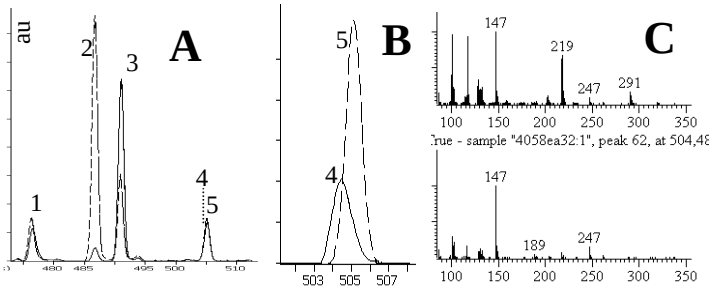

Fig.2. FIEHN

485 **References**

1. Fiehn, O. (2002) Metabolomics – the link between genotypes and phenotypes. *Plant Mol. Biol.* **48**, 155-171.
2. Krull, I.S. and Swartz, M. (1999) Analytical method development and validation for the academic researcher. *Anal. Lett*. 32,1067-1080.
3 . Weckwerth, W., Wenzel, K. and Fiehn, O. (2004) Process for the integrated extraction, identification and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks. *Proteomics* **4**, 78-83.
4 Stein, S.E. (1999) An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *J. Am. Soc. Mass. Spectrom.* **10**, 770-781
5 Molloy, M.P., Brzezinski, E.E., Hang, J., McDowell, M.T. and van Bogelen, R.A. (2003) Overcoming technical variation in quantitative proteomics. *Proteomics* **3**, 1912-1919
6. Oehme, M. (1998) Practical Introduction to GC-MS Analysis with Quadrupoles, 1st ed. Hüthig, Heidelberg, Germany.