

TECHNIQUES FOR MOLECULAR ANALYSIS

Quality control for plant metabolomics: reporting MSI-compliant studies

Oliver Fiehn^{1,*}, Gert Wohlgemuth¹, Martin Scholz¹, Tobias Kind¹, Do Yup Lee¹, Yun Lu¹, Stephanie Moon² and Basil Nikolau²

¹Davis Genome Center, University of California, Davis, CA 95616, USA, and

²Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, IA 50011, USA

Received 22 October 2007; revised 26 November 2007; accepted 27 November 2007.

*For correspondence (fax +1 530 754 9658; e-mail ofiehn@ucdavis.edu).

Summary

The Metabolomics Standards Initiative (MSI) has recently released documents describing minimum parameters for reporting metabolomics experiments, in order to validate metabolomic studies and to facilitate data exchange. The reporting parameters encompassed by MSI include the biological study design, sample preparation, data acquisition, data processing, data analysis and interpretation relative to the biological hypotheses being evaluated. Herein we exemplify how such metadata can be reported by using a small case study – the metabolite profiling by GC-TOF mass spectrometry of *Arabidopsis thaliana* leaves from a knockout allele of the gene At1g08510 in the Wassilewskija ecotype. Pitfalls in quality control are highlighted that can invalidate results even if MSI reporting standards are fulfilled, including reliable compound identification and integration of unknown metabolites. Standardized data processing methods are proposed for consistent data storage and dissemination via databases.

Keywords: GC/MS, standard operating procedure, abiotic stress, wounding, BinBase, SetupX.

Introduction

Plant metabolism is known to be highly flexible in terms of both metabolite abundances and identities. By differentially quantifying metabolite data among biological samples that are differentiated by genetic, environmental, spatial or developmental parameters, it is envisioned that metabolomics data will yield information concerning the genetic, environmental, spatial or developmental control of metabolism (Fiehn, 2002). However, there are a number of technical and infrastructure issues that currently limit this vision.

There are many ways to conduct plant biochemical or physiological studies, and there are also a variety of valid methods and techniques to study metabolic alterations on a comprehensive scale. No single analytical technology or protocol facilitates the vision of comprehensive metabolite analysis (van der Greef *et al.*, 2004). Inherent in this technological barrier is the fact that metabolite concentrations span up to seven orders of magnitude (e.g. comparing abundance of plant hormones to transport compounds such as sucrose). Moreover, while core biochemical pathways are

certainly conserved across species, this is not true for the tremendous variety of secondary metabolites in plants, their pathway structures, regulatory mechanisms and spatial specializations. Due to this metabolic flexibility, metabolomes cannot easily be computed from genomes, and therefore the basic question of the extent of an organism's metabolome cannot be addressed as such. Consequently, the 'plant metabolome' is a poorly defined, variable entity that must be determined from empirical data. Repositories need to be developed that host the complement of plant metabolite data integrated with accurately curated metabolic information (Rhee *et al.*, 2003; Zhang *et al.*, 2005). Initial databases have been released (<http://lab.bcb.iastate.edu/projects/plantmetabolomics/>, <http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html>, http://fiehnlab.ucdavis.edu:8080/m1/main_public.jsp) to accomplish this goal.

While there is no single best way to conduct metabolomic studies, there are a number of pitfalls and known problems

that need to be carefully avoided, and detailed guidelines and practice protocols have been published previously (Fiehn, 2006; Lisec *et al.*, 2006). We present here some additional pitfalls and solutions for a number of problems encountered when using the select technique of GC/MS-based metabolite profiling, for example sample carry-over effects or matrix problems. Similar quality issues are relevant for other analytical methods such as LC/MS, but to maintain clarity, we focus here on the comparatively mature method of GC/MS. Other problems are independent of the analytical chemistry method, for example comprehensive and reproducible extraction of metabolites despite diverse structure and abundances, or difficulties in correctly annotating analytical signals as genuine plant metabolites and in storing and quantifying peaks that lack structural elucidation. Ultimately, a combination of metabolite-profiling techniques will need to be combined to yield a detailed overview of plant metabolism.

In order to enable researchers to re-use metabolomic data, repeat experiments or evaluate the validity of claims, it is mandatory that any metabolomic study provides clear descriptions of the biological design of the experiment and that all technical parameters are presented in great detail. Information about data origins is also called 'metadata'. Such metadata are usually reported in method sections, but may also be found as experimental design details in the body of plant research papers. It is often difficult to extract all the necessary information from a journal-published report regarding how a plant metabolomic study was actually performed. For instance, details on plant growth conditions that some authors may find unimportant may have been left out, such as details on the watering regimes or pathogen protection used in a greenhouse. Metabolism readily responds to environmental conditions. In order to enable comparisons across studies, sufficient details on environmental parameters need to be reported even for studies focusing on metabolic effects of genotype variations. Similarly, authors often only refer to technical methods by literature citations. This may be a reasonable option once procedures are standardized and validated to a great extent. However, currently metabolomics studies often involve a variety of methods, and therefore readers have to accumulate technical details from several sources of information. In addition, the actual use of technical methods sometimes deviates from previously published method descriptions, which may not be clear in the final reports. In summary, therefore, a lack of metadata on plant biology as well as missing information on technical methods make the final results and conclusions less useful and harder to reproduce.

The ever growing number of technical methods that are employed in a single plant metabolomic study requires that reporting structures be formalized. After a series of initial drafts, a consultation period and conference workshops (e.g. the 4th International Conference on Plant Metabolomics,

2006), the Metabolomics Society released a number of documents on proposed minimum standards for use in metabolomic reports (Fiehn *et al.*, 2007a), among them standards on plant biology context (Fiehn *et al.*, 2007b) and on reporting of methodological details of chemical analysis (Sumner *et al.*, 2007). These documents are open for criticism and continuing improvement using the open-access MSI website (<http://msi-workgroups.sourceforge.net/>). The number of parameters and the level of detail required to be given may change over time, but the principal idea is that all metadata are compiled and reported in a single document that would also typically be required by leading plant science journals, called 'minimal reporting standards'. These standards are called minimal because, for certain biological studies, it is important to gather further information, for example on the origin and storage history of plant seeds or on air circulation specifications in climate chambers that ultimately might have a confounding influence on metabolite levels. Even such compilations of 'minimal' metadata usually extend far beyond typical method sections (also called experimental procedures).

As yet, these 'minimal standards' lack defined and published instances of actual studies. We here outline how metadata for plant metabolomics should be reported using a small case study profiling a well-characterized Arabidopsis genotype. In addition, we present quality controls and methods on plant metabolite profiling by GC/MS that have been developed over the past 7 years, highlighting practical problems that need to be addressed when applying established methods to plant tissues for which these methods were not developed and validated. In addition, we exemplify how novel injection techniques can improve standard GC/MS-based metabolite profiling. We further outline how automatic data processing and the use of standardized methods support a GC/MS-based plant metabolomic database. While this database can so far only handle a specific GC/MS instrument, the algorithms and concepts for the data processing can and should be applied more widely to facilitate data exchange and comparison of data across studies.

Reporting experimental methods

The Metabolomics Standards Initiative has proposed a variety of parameters along the general metabolomic workflow. It is a common misconception that metabolomics approaches focus on a specific machine or utilize a specific best type of extraction procedure. This is generally not the case, as indicated by the vast set of methods and combination of these methods that are published in plant metabolomics. Instead, metabolomics workflows need to be described in detail, covering the plant biology context, chemical analysis methods, data processing methods, and the statistics and data interpretation methods.

An example of how such extended experimental procedures can be reported is given below, following point by point the recommendations outlined by the Metabolomics Standards Initiative. It is important to note that, by following these minimal reporting requirements, studies become 'MSI-compliant', but such compliance does not imply that the studies are validated studies or potentially more valuable with respect to scientific findings. MSI compliance only means that a sufficient amount of metadata is given to enable readers or users of the data to reproduce how the study was performed. However, the full complement of method details extends the space requirements allowed by many peer-reviewed journals. It is generally in accordance with MSI compliance to submit metadata as supplementary information in file formats, and to give only a brief overview over the most important parameters in the printed text.

To illustrate this reporting structure and provide an example of a good practice document, we present herein a small metabolomics study from an NSF2010 project that is exploring the use of metabolomics data to decipher functions of multiple genes whose functions are currently unknown. This study was designed as a means of validating the experimental approach, and hence used a T-DNA mutant stock that has undergone extensive characterization in previous studies (Bonaventure *et al.*, 2003). This *Arabidopsis* mutant stock is in the Wassilewskija (Ws) ecotype background. It carries a T-DNA disruption in the gene *At1g08510*, which codes for a fatty acyl-ACP thioesterase (FATB). The metabolomics experiment sought to determine the effect of the mutation on the metabolome in conjunction with an environmentally induced perturbation (plant wounding). All data generated by the NSF2010 consortium are publicly available and can be downloaded from a dedicated server (<http://www.plantmetabolomics.org>). GC-TOF data can also be downloaded directly from the participating laboratory (http://fiehnlab.ucdavis.edu:8080/m1/main_public.jsp). A number of technical details involving the data acquisition instrument and the database annotations are cutting-edge technology, and therefore may require additional investment. In such cases, we have added comments on the basic reasons why we chose these technologies and what substitutes could be used.

Plant context metadata

Plant biology parameters are separated into the physical object of the study (BioSource) and general growth conditions, followed by treatment and harvest parameters (Fiehn *et al.*, 2007b). Species names follow the National Center for Biotechnology Information (NCBI) taxonomy, and nomenclature for plant features must conform to the classifications and relationships given by the Plantontology consortium (<http://www.plantontology.org/>). Units must be specified but

are not yet standardized. For example, the amount of plant material used for the study is often given in terms of milligrams of fresh weight, but dry weight information might be preferable. Both alternatives are MSI-compliant, but 'fresh weight' comparisons would probably not be reasonable for drought-stress experiments. While in principle, MSI-compliant data can be submitted as flow text, it is more advantageous to present data in categories or tables as given below to facilitate comparisons between studies and to provide data inputs and outputs that are more easily machine-readable.

BioSource Species: Arabidopsis thaliana.

Genotype: Wassilewskija (Ws) and Wassilewskija (Ws) fatB knockout (At1g08510).

Organ: Leaf.

Organ specification: Rosette leaves, aerial portion.

Amount: 50 mg fresh weight per sample, of which 20 mg were used for extraction.

Growth Support: Fourteen to sixteen seeds were sown on 20–25 ml of sterile Murashige and Skoog basal salt mixture (MS medium) containing 0.1% w/v sucrose and 1× liquid vitamin solution (Sigma, <http://www.sigmaaldrich.com/>) containing 15 g l⁻¹ bacto agar (BD) in 100 × 100 × 15 mm square Falcon Petri dishes (Thermo Fisher Scientific; <http://www.thermofisher.com>). Seeds were arranged on the plates in a single horizontal line 1 cm from the top of the plate. Prior to sowing, seeds were sterilized by treating for 1 min at room temperature with 300 µl of 50% v/v ethanol; this solution was then removed and replaced by 300 µl of a solution consisting of 1% v/v Tween-20 (Thermo Fisher Scientific) and 50% v/v bleach (Clorox; <http://www.clorox.com>), and incubated at room temperature for 10 min. The seeds were then washed with three changes of 0.3 ml of sterile water. After sowing the seeds, the plates were wrapped using micropore tape (3 M Health Care; <http://www.3m.com>), and then stored horizontally for 4 days at 4°C in the dark. On the 5th day, plates were moved to the growth room, and held in a vertical position in Plexiglass holders for 12 days.

Location: Controlled-environment facility at Iowa State University, Nikolau laboratory.

Plot design: Each genotype and replicate were grown on individual plates and placed randomly in the Plexiglass holders.

Light period: 24 h day at 82 µmol m⁻² s⁻¹ (light source Sylvania; <http://www.sylvania.com>), F34CW/SS/ECO/RP).

Humidity: Day 100%, night 100%.

Temperature: Day 24°C, night 24°C.

Watering regime: No further watering, plates remained closed.

Nutritional regime: MS medium without further fertilizers.

Date of plant establishment: 25 September 2006.

Treatment Abiotic treatment: On 11 October 2006 at 11:30 a.m. (13th day after plant establishment), plates were uncovered in the growth room and leaves were wounded by piercing using an 18-gauge sterile needle. Plates were then re-covered with their lids and wrapped with micro-pore tape (3M Health Care). All plants, including the non-wounded plants, were exposed to the air during the wounding period. Plates were then placed vertically in Plexiglass holders under the conditions described above.

Dose: Ten punches.

Duration: 3 min wounding period; 2 h response period before harvest.

Harvest Date: 11 October 2006.

Time: 1:30 p.m.

Growth stage: Boyes 1.1–1.4.

Metabolism quenching method: Immersion in liquid nitrogen within 1 min after harvest.

Harvest method: Petri plates were opened and the aerial portions of the plants were cut.

Storage: -70°C for 1 day, then shipping on dry ice and storage at -80°C for 2 weeks.

Chemical analysis metadata

Following the recommendations of the Metabolomics Standards Initiative (Sumner *et al.*, 2007), chemical analysis metadata are separated into information on 'sample processing and extraction', 'chromatography', 'mass spectrometry', 'metabolite identification' and 'quality control'. Details given here refer to the *fatB* wounding study by GC/MS.

Sample processing and extraction Tissue processing: Frozen tissues were kept in 2 ml round-bottomed Eppendorf tubes equipped with one 3 mm diameter steel ball, and homogenized using a Retsch (<http://www.retech-us.com>) ball mill for 30 sec at 25 sec^{-1} .

Extraction: Ground tissue powder was kept in liquid nitrogen between homogenization and extraction. The extraction solvent was prepared by mixing isopropanol/ acetonitrile/water at the volume ratio 3:3:2 and degassing this mixture by directing a gentle stream of nitrogen through the solvent for 5 min. The solvent was cooled to -20°C prior to extraction. Randomly processing all samples of the study, 1 ml of cold solvent per 20 mg of ground tissue was added, vortexed for 10 sec, and shaken at 4°C for 5 min to extract metabolites and simultaneously precipitate proteins. After centrifugation at 12 800 *g* for 2 min, 90% of the supernatant was removed, taking care not to remove any residues from the pellet.

Extract concentration: The supernatant was separated into two equal aliquots and concentrated to dryness in a

Centrivap cold trap vacuum concentrator (<http://www.labconco.com>) at room temperature for 4 h.

Extract clean-up: In order to fractionate complex lipids and waxes, the residue was re-suspended in 500 μl 50% aqueous acetonitrile and centrifuged at 12 800 *g* for 2 min. The supernatant was transferred to a 1.5 ml Eppendorf tube and concentrated to dryness in a vacuum concentrator.

Extract storage: Dried extracts can be kept under nitrogen at -80°C for up to 4 weeks. In the study presented here, extracts were immediately derivatized for GC-TOF mass spectrometry.

Chromatography Sample preparation: A mixture of internal retention index (RI) markers was prepared using fatty acid methyl esters of C8, C9, C10, C12, C14, C16, C18, C20, C22, C24, C26, C28 and C30 linear chain length, dissolved in chloroform at concentrations of 0.8 mg ml^{-1} (C8–C16) or 0.4 mg ml^{-1} (C18–C30). Aliquots (2 μl) of this RI mixture were added to the dried extracts, then 10 μl of a solution of 20 mg ml^{-1} of 98% pure methoxyamine hydrochloride (CAS number 593–56-6, Sigma) in pyridine (silylation grade; Pierce; <http://www.pierce.net.com>) was added to protect aldehyde and ketone groups, and the mixture was shaken at 30°C for 90 min. Then 90 μl of N-methyl-N-trimethylsilyl-trifluoroacetamide with 1% trimethylchlorosilane (MSTFA/1% TMCS) (1 ml bottles; Pierce) was added for trimethylsilylation of acidic protons and shaken at 37°C for 30 min. The reaction mixture was transferred to a 2 ml clear glass auto-sampler vial with micro-insert (Agilent; <http://www.agilent.com>) and closed using a 11 mm T/S/T crimp cap (MicroLiter; <http://www.microliter.com>).

Auto-injector: A Gerstel automatic liner exchange system with a MPS2 dual rail multi-purpose sampler and two derivatization stations was used in conjunction with a Gerstel CIS cold injection system (Gerstel; <http://www.gerstelus.com>). For every 10 samples, a fresh multi-baffled liner was inserted (Gerstel catalog number 011711-010-00) using the Gerstel Maestro 1 software version 1.1.4.18. Before and after each injection, the 10 μl injection syringe was washed three times with 10 μl ethyl acetate. Each 1 μl sample was filled using 39 mm vial penetration at $1\text{ }\mu\text{l sec}^{-1}$ filling speed, injecting 0.5 μl at a $10\text{ }\mu\text{l sec}^{-1}$ injection speed at an initial temperature of 50°C which was ramped by $12^{\circ}\text{C sec}^{-1}$ to a final temperature of 250°C and held for 3 min. The injector was operated in split-less mode, opening the split vent after 25 sec. Samples were injected between 2–24 h after derivatization using randomized sequences controlled by the laboratory information management and database system, SetupX (Scholz and Fiehn, 2007).

Chromatography instrument: An Agilent 6890 gas chromatograph controlled using Leco ChromaTOF software version 2.32; <http://www.leco.com>.

Separation column: A 30 m long, 0.25 mm internal diameter Rtx-5Sil MS column with 0.25 μm 95% dimethyl/5% diphenyl polysiloxane film and an additional 10 m integrated guard column was used (Restek; <http://www.restek.com>).

Separation parameters: 99.9999% pure helium with built-in purifier (Airgas; <http://www.airgas.com>) was used at a constant flow of 1 ml min^{-1} . The oven temperature was held constant at 50°C for 1 min, and then ramped at 20°C min^{-1} to 330°C, and held constant for 5 min.

Mass spectrometry Instrument: A Leco Pegasus IV time-of-flight mass spectrometer controlled using Leco ChromaTOF software version 2.32, and operated by Yun Lu on 5 April 2007.

Sample introduction: The transfer line temperature between gas chromatograph and mass spectrometer was set to 280°C.

Ionization: Electron impact ionization at 70 V was employed, with an ion source temperature of 250°C.

Data acquisition: After 290 sec solvent delay, filament 1 was turned on and mass spectra were acquired at mass resolving power $R = 600$ from m/z 85-500 at 20 spectra per second and 1550 V detector voltage without turning on the mass defect option. Recording ended after 1200 sec. The instrument performed auto-tuning for mass calibration using FC43 (perfluorotributylamine) before starting analysis sequences.

Metabolite identifications First parameter: Retention index window ± 2000 U (around ± 2 sec retention time deviation).

Second parameter: Mass spectral similarity plus additional confidence criteria as detailed below (Data processing).

Metabolite library: Importantly, GC/MS peaks can only be annotated as identified according to MSI standards if at least two independent parameters are recorded and matched: mass spectra and retention index, in this instance. Hence, spectra that are just matched to library entries without retention time (or retention index) information cannot be considered identified. All signals that are exported by the BinBase database (Fiehn *et al.*, 2005) are reported by the quantification ion, a unique database identifier, retention index and the complete mass spectrum encoded as string. Database entries are named using the Fiehn library, which currently includes 713 unique metabolites and 1197 unique spectra, for which names, structure graphs and codes, retention indices and database references are available at <http://fiehnlab.ucdavis.edu/Metabolite-Library-2007/>. Mass spectra for the library itself are commercially available from the instrument manufacturers. Alternative libraries of GC/MS metabolite spectra are freely available at <http://csbdb.mpimp-golm.mpg.de/csbdb/>

gmd/msri/gmd_msri.html, and apply a different set of GC/MS conditions.

Quality controls: Daily quality controls were used. These comprised two method blanks (involving all the reagents and equipment used to control for laboratory contamination) and four calibration curve samples spanning one order of dynamic range and consisting of 31 pure reference compounds. These quality control calibration samples were injected at 0.5 μl injection volumes and a split ratio of 1/5 because potential problems with the stability of trimethylsilylated amino acids are first apparent in split mode. Intervention limits were established to ensure basic validation of the instrument for metabolite profiling.

Data processing

According to the recommendations of the Metabolomics Standards Initiative (Sumner *et al.*, 2007), metadata are required to detail how resulting GC-TOF mass spectrometry profiles were investigated. The details of this procedure are critical for achieving highly reliable and consistent results. Unbiased analysis seeks to (i) find all signals that can be distinguished from background noise or signals with very similar retention times, and then (ii) report signal intensities for all detected signals in all chromatograms. In subsequent steps, signals can be identified using mass spectral libraries and then investigated for differential regulation using statistics software. Each of these steps requires great diligence if data are to be stored and disseminated permanently. Details presented here involve elaborate algorithms in order to reduce the number of false-positive and false-negative peak detections and to avoid misidentification of metabolites. Freely available software programs such as AMDIS (Stein, 1999), mzmime (Katajamaa *et al.*, 2006) or SpectConnect (Styczynski *et al.*, 2007) accomplish parts of these processes but are not tailored towards a database approach.

File formats: Files were pre-processed directly after data acquisition and stored as ChromaTOF-specific *.peg files, as generic *.txt result files, and additionally as generic ANDI MS *.cdf files. It is recommended that such generic *.cdf files are made available for published data sets to facilitate data sharing.

Pre-processing details: ChromaTOF version 2.32 was used for data pre-processing without smoothing, with 3 sec peak width, baseline subtraction just above the noise level, and automatic mass spectral deconvolution and peak detection at signal/noise levels of 10:1 throughout the chromatogram. Apex masses were used for quantification. Resulting *.txt files were exported to a data server with absolute spectra intensities, and further processed using the BinBase algorithm (Fiehn *et al.*, 2005; see Figure S1). This algorithm used the settings: validity of

chromatogram (<10 peaks with intensity >10⁷ counts sec⁻¹), unbiased retention index marker detection (MS similarity >800 and exceeding thresholds for ion ratio abundances for high *m/z* marker ions), retention index calculation by 5th order polynomial regression. Spectra were cut to 5% base peak abundance, and matched to database entries from most- to least-abundant spectra using the following matching filters: retention index window ± 2000 U (equivalent to about ± 2 sec retention time), validation of unique ions and apex masses (unique ion must be included in apex masses and present at >3% of base peak abundance), mass spectrum similarity that must fit criteria dependent on peak purity and signal/noise ratios (Table S1), optional ion ratio settings to distinguish peaks with high similarity, and a final isomer filter (annotating the isomer spectrum with the closest RI fit). In any metabolomic report, it is important to state the confidence or quality thresholds that were used to identify signals as genuine (known) metabolites. Failed spectra were automatically entered as new database entries if signal-to-noise ratio >25, purity <1.0 and presence in the biological study design class was >80%. This filter ensured that (i) signals were reported that had never been detected previously in any other sample, but (ii) only signals are reported that can be assumed to be biologically relevant using relatively abundant and pure signals and ensuring that these are positively detected in most of the biological replicates.

All thresholds reflect settings for ChromaTOF version 2.32. Signal intensities were reported as peak heights using the unique ion as default, unless an alternative quantification ion was manually set in the BinBase administration software Bellerophon. A quantification report table was produced for all database entries that were positively detected in more than 80% of the samples of a study design class (as defined in the SetupX database). This procedure results in 10–30% missing values, which could be caused by true negatives (compounds that were below the detection limit in a specific sample) or false negatives (compounds that were present in a specific sample but that did not match quality criteria in the BinBase algorithm). A subsequent post-processing module was employed to automatically replace missing values from the *.cdf files using the open-access mzmime software (Katajamaa *et al.*, 2006) with the following parameters. For each positively detected metabolite, the average retention time was calculated. For each chromatogram and each missing value, the intensity of the quantification ion at this retention time was extracted by seeking its maximum value in a retention time region of ± 1 sec and subtracting the minimum (local background) intensity in a retention time region of ± 5 sec around the peak maximum. The resulting report table therefore did not have any missing values. Replaced values were labeled as 'low confidence' by color coding.

Statistics

The Metabolomics Standards Initiative has released minimum requirements for reporting data transformations and statistics (Goodacre *et al.*, 2007). Here we describe common data evaluation methods by separation into data transformation and statistics.

Data transformation: Result files were transformed by calculating the sum intensities of all structurally identified compounds for each sample (i.e. those signals that had been positively identified in the data pre-processing schema outlined above), and subsequently dividing all data associated with a sample by the corresponding metabolite sum. The resulting data were multiplied by a constant factor in order to obtain values without decimal places. Intensities of identified metabolites with more than one peak (e.g. for the syn- and anti-forms of methoxymated reducing sugars) were summed to only one value in the transformed data set. The original non-transformed data set was retained. The general concept of this data transformation is to normalize data to the 'total metabolite content', but disregarding unknowns that might potentially comprise artifact peaks or chemical contaminants. Caution must be exercised when comparing classes of samples that might have biologically very different metabolite concentrations (e.g. cold-acclimatized plants versus plants grown at room temperature). Alternative means of normalization may utilize the amount of material used for the analysis (e.g. the mass of the extracted tissue) and additional internal standards that would account for losses during sample preparation and injections.

Statistics: Statistical analyses were performed on all continuous variables using Statistica software version 7.1 (StatSoft; <http://www.statsoft.com>). Univariate statistical analysis for multiple study design classes was performed by breakdown and one-way ANOVA. *F* statistics and *P*-values were generated for all metabolites. Data distributions were displayed by box-whisker plots, giving the arithmetic mean value for each category and the standard error as box and whiskers for 1.96 times the category standard deviation to indicate the 95% confidence intervals, assuming normal distributions. Multivariate statistical analysis was performed by unsupervised principal component analysis (PCA) to obtain a general overview of the variance of metabolic phenotypes in the study, by entering metabolite values without study class assignments. In addition, supervised partial least-square (PLS) statistical analysis was performed, which requires information about the assigned study classes. Three plots were obtained for each PCA and PLS model. The first is a scree plot for the Eigen values of the correlation or covariance matrix. This is considered as a simple quality check and should have a steep descent with an increasing number of Eigen values. Second, 2D score

scatter plots were generated for at least the first three dimensionless principal components, and 3D plots are generated to better distinguish metabolic phenotypes. Third, loading plots were generated for each vector in PCA or PLS, showing the impact of variables on the formation of vectors. Metabolites near the coordinate center had no separation power; conversely, variables far away from the coordinate center were important for building PCA and PLS models. Variables that are located close to each other are strongly correlated.

Quality control measures in GC/MS-based metabolite profiling

Extraction parameters have a direct impact on metabolic data

The extraction conditions proposed above involved inactivation of enzymes by a -20°C cold protein precipitation step, concomitant with dissolving small molecule metabolites into an excess of a ternary solvent mixture of lipophilic and hydrophilic organic solvents with water. It is important to note that other extraction conditions can be applied, for example using stronger lipophilic solvents such as chloroform (Gullberg *et al.*, 2004; Weckwerth *et al.*, 2004) instead of isopropanol. Older protocols involved sequential use of hydrophilic and lipophilic solvents, e.g., hot aqueous methanol followed by a second chloroform extraction and a solvent-based fractionation schema (Fiehn *et al.*, 2000a; Roessner *et al.*, 2001). Quantitative results from metabolite profiles will differ according to the extraction protocol used. For example, if fractionation schemes are employed, semi-polar metabolites may be preferentially present in one or the other fraction, depending on the composition of the total matrix. However, metabolites may be found in different fractions depending on the plant species: *Arabidopsis* leaf extracts yield chlorophyll in the lipophilic phase, whereas a considerable fraction of chlorophyll is found in the polar phase in leaf extracts of *Cucurbita maxima* when performing fractionated extractions using methanol/water and methanol/chloroform protocols (Fiehn *et al.*, 2000a; Roessner *et al.*, 2001). It is a general problem for comparisons between extraction methods that no accepted benchmark profiles are available. While metabolomics means making compromises, each protocol has to be validated by the extent that known metabolites can be reproducibly and quantitatively extracted from the plant matrix. While some metabolites (such as glucose-6-phosphate) must be present in *Arabidopsis* extracts to validate a certain protocol, the presence of signals for unknown metabolites may always be explainable by presumed artifact formation or degradation. One way to overcome this problem is to assess recovery rates obtained by comparing external calibration curves to spiked standard additions of a select number of pure refer-

ence compounds. However, such recovery rates cannot be used as true surrogates for the accuracy of assessing *in vivo* concentrations, because external and internal calibration curves may vary due to extraction and matrix effects. Use of internal isotope-labeled compounds or plant material grown under isotope-labeled nutrient conditions may partly remedy this situation. A simpler way is to compare frequency distributions of the technical reproducibility of metabolite profiles as a discriminator between alternative extraction protocols. A non-normal distribution of precision values may be expected, with a high number of metabolites giving low error rates for repetitive analysis of technical replicates (<15% relative standard deviation) and a low number of compounds with high or very high errors (>40% technical errors). In general, the median technical reproducibility determined over all detectable metabolites should be smaller than 20% relative standard deviation. Despite similar values for overall precision, the frequency distribution of quantitative reproducibility may differ for different solvent mixtures (Figure 1). For technical replicate extractions of young soybean seeds, we found that a ternary extraction mixture of water–chloroform–methanol (Weckwerth *et al.*, 2004) yielded good quantitative precision for most analytes, whereas a water–isopropanol–acetonitrile mixture resulted in a bimodal distribution for soybean, with a range of compounds around 40–55% relative standard deviation. Such a bimodal frequency distribution was not found for extraction of *Arabidopsis* rosette leaves. Therefore, method validation studies must be carried out for every new plant organ that is subjected to metabolite-profiling studies. As a rule of thumb for such method development and validation studies, a list of expected metabolites should be generated for the target plant organ, including preliminary assessments of abundance and potential detectability by a specific metabolite-profiling technique. For example, ADP/ATP cannot be

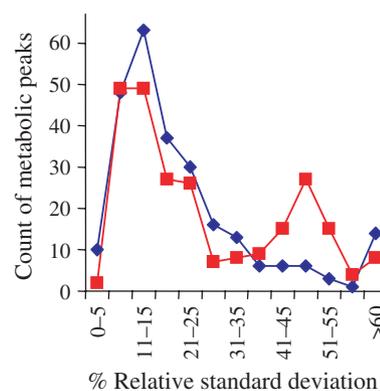


Figure 1. Precision of quantitative analysis for GC-TOF peaks reported by BinBase for young soybean seeds. Frequency distributions are given in increments of 5% relative standard deviations, comparing cold methanol/chloroform/water extractions (blue line) or the cold acetonitrile/isopropanol/water protocol (red line), using nine technical replicates each.

analyzed by GC–TOF mass spectrometry, whereas AMP levels are routinely determined. Special attention should be given to compounds that are prone to be degraded by residual enzymatic activity, heat or oxidation, for example sugar phosphates, cysteine and ascorbate.

Additionally, the choice of extraction solvent mixtures may be made in a rational design by calculating relative solubility. Here we employed predictions using *COSMOfrag* and the *COSMO-RS* software (Klamt *et al.*, 2002). Around 1000 compounds from various chemical classes were selected from the metabolic pathways database KEGG (<http://www.genome.jp/ligand>) database, and relative solubilities were calculated for 18 single solvents and ternary mixtures of methanol, chloroform and water. The worst solvent was dimethyl sulfoxide, which was incapable of dissolving around 80% of the selected substances. Calculations for pyridine showed excellent solubility for a number of carbohydrates, as also shown experimentally (Modi *et al.*, 2000), justifying its frequent use as a solvent for derivatizations in GC–TOF metabolite profiling. As expected, it was found that solvents such as diethyl ether, benzene or dichloroethane are strong but very selective solvents, and that solvent mixtures such as methanol:chloroform:water (3:1:1) can dissolve a very broad range of biologically active compounds and are superior to any solvent by itself.

The sample preparation method involves all steps including harvest, quenching of metabolism, storing tissues, homogenizing representative samples and extraction conditions. Each step should be carefully considered and not all potential pitfalls can be described here. Once the sample preparation method has been validated, details of the procedures must be documented in written form, resulting in a 'standard operating procedure'. However, an important part of the quality control process is to ensure that use of the protocol yields reproducible data sets, independently of the individual who carried out the sample preparations. Training of new laboratory members (or efforts involving partnering laboratories) must be documented by technical replicate analyses to be confident that the standard operating procedure is sufficiently clear and concise.

Quality of GC/MS-based metabolite profiles depends on rigorous control of injector conditions

Metabolomics poses a challenge to quality control: in GC/MS or LC/MS-based metabolomics, the instrument is physically exposed to the sample, unlike, for example, in NMR-based studies. Hence, sample constituents may remain inside the chromatograph or mass spectrometer, leading to cross-contamination or sensitivity losses, or otherwise adversely affecting data quality (Hajslova *et al.*, 1998). Using replicate washing steps eliminates contamination by the auto-sampler itself, but other parts of the instrument may be heavily exposed to the sample. For

example, it is common to use reverse-phase columns in liquid chromatography (HPLC): plant waxes will show severe and sometimes irreversible adsorption under standard HPLC conditions, such that obtaining high-quality quantitative metabolite profiles is very difficult. Similarly, waxes or membrane lipids are not volatile enough for GC/MS analysis, and would therefore be retained within the injector, either in the liner, the injector body or even within the first centimeters of the column. Over time, these contaminations result in severe problems that have an impact on data quality, from formation of catalytically active sites to absorption effects by pyrolytic particles in the liner.

One way to remove such involatile compounds is by lipid/polar fractionation schemes as described previously (Weckwerth *et al.*, 2004). However, apart from the problem of splitting semi-polar compounds into two fractions, sterols, plant hormones and free fatty acids are all found in the lipophilic fractions, which (due to time constraints) are often not analyzed in separate analytical runs. In addition, merging results from two separate data files is a challenge in itself. These problems can be avoided if liners are continuously exchanged using additional instrument hardware. Two automatic liner-exchange devices are commercially available, a direct thermo-desorption device (DTD) with crimp-capped liners that contain micro-inserts (ATAS; <http://www.atasgl.com>) (Denkert *et al.*, 2006; Meyer *et al.*, 2007; Sanz *et al.*, 2004), or a liner exchange/cold-injection system (ALEX-CIS, Gerstel) (David *et al.*, 2006). Both systems use cold injection followed by heat ramping of the injector to ensure that only volatile components reach the injector body and column. Cold-injection techniques prevent the violent explosions of injection drops found for classic hot injectors (Grob and Biedermann, 2000), and thus help to improve data quality by reduced degradation of thermolabile compounds. Involatile matrix components remain inside the micro-inserts or exchangeable liners, which are disposed of after use.

Automatic liner-exchange injectors facilitate the injection of plant extracts, including lipids such as sterols and free fatty acids. Without liner exchange, full metabolome injections in a hot split/split-less injector result in increasing signal intensities for low-abundance unsaturated free fatty acids, as observed in an analysis of blank controls (Figure 2) run subsequent to GC/MS analysis of derivatized plant extracts. Despite liner exchange, however, every injection into a GC/MS inlet system inevitably introduces dirt, which is deposited in the liner, septum, injector body, injector bottom plate and injection syringe, and ultimately the first few centimeters of the guard column. Even without automatic liner exchange, it has been recommended that liners are manually exchanged every 20 samples (Koek *et al.*, 2006). Matrix deposits in the liner and the injector system have detrimental consequences for the accuracy of quantitative analysis of amino acids. Using the standard protocol, acidic protons of all compounds are trimethylsilylated (TMS) to render the otherwise involatile

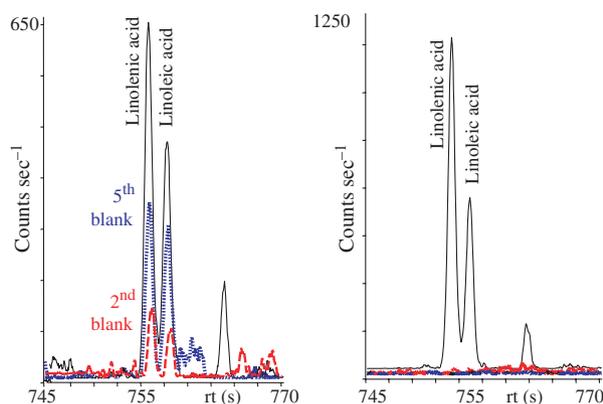


Figure 2. Rice leaf extracts including lipids injected into an Agilent hot split/splitless injector (left panel) and a Gerstel automatic liner exchange/cold injection system (right panel). Ion traces m/z 335 + 337 are given. The accumulation of matrix deposits without liner exchange causes increasing levels of background signals in intermittently injected blank controls. This cross-contamination reaches half the intensity of endogenous low-abundance free fatty acid levels after as few as five samples.

tile metabolites amenable to gas chromatography (Halket and Zaikin, 2003). It has been argued that quantitative amino acid analysis is less accurate by TMS derivatization and GC/MS analysis than by classic amino acid analysis, e.g. by phthalaldehyde derivatization and HPLC fluorescence detection (Noctor *et al.*, 2007). In other reports, formation of N-TMS groups was said to be dependent on the reaction time and temperature used during the derivatization (Gullberg *et al.*, 2004; Kanani and Klapa, 2007). While we cannot confirm the latter observation, all these reports clearly suggest that the signal intensities of the various TMS derivatives are not robust for several amino acids.

In our experience, this lack of robustness and the change in N-TMS derivatization status are directly related to the status of the injector and total matrix deposition. If the injector body, column, injector plate, liner and syringe are kept scrupulously clean, the intensity ratio of the various trimethylsilylated forms of amino acids can be kept constant. Only if dirt accumulates or if the sample itself contains too much matrix, will the relative peak ratios and overall abundance of TMS-derivatized amino acids be negatively affected. Figure 3 shows that this effect is not observed for compounds that only bear hydroxyl- or carboxyl moieties, such as sterols or sugars, and that some amino compounds are more affected than others. When comparing clean injector conditions to injector conditions for which dirt problems were recognized (Figure 3), amino acids such as asparagine, glutamate and serine are severely decreased, especially at lower concentrations, whereas putrescine or glycine are not affected. Other amino acids, such as alanine, urea, valine and isoleucine, are also negatively affected to a lesser extent. It is important to note that these effects are most prominent under split-injection conditions (here at

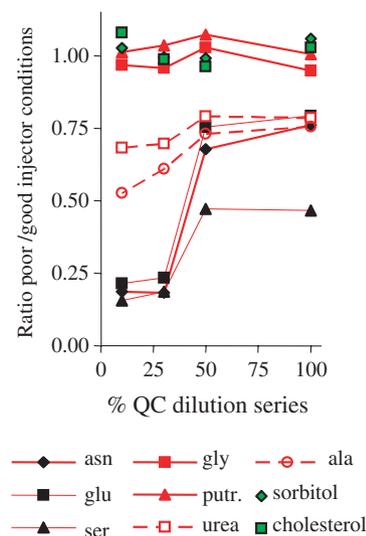


Figure 3. Normalized peak intensities for nine of the 31 quality control standards that are injected daily at four concentrations. The ratio of the results obtained under poor and good injector conditions is given.

split ratio 1/5), but are much less severe under split-less conditions. While the differences in adverse responses between structurally related amino acids cannot be readily explained, all observations suggest that there is a delicate balance between the formation and degradation of N-TMS bonds during the injection, in which the N-TMS formation is favored by clean and inert surfaces and by longer periods in the reactor (i.e. split-less conditions). In order to limit matrix effects caused by the sample matrix itself (e.g. complex lipids), clean-up before derivatization is recommended. Soluble metabolites should be re-dissolved in 50% aqueous acetonitrile to remove them from insoluble complex lipids and plant waxes that otherwise may hamper the N-TMS stability for some amino compounds. We have found that large amounts of sterols and free fatty acids could be completely recovered using this clean-up step, and that N-silylation was greatly improved for samples with high amounts of complex lipids.

Alternatively, amino acids could be separately analyzed using the far more stable tertiary butyldimethylsilyl derivatization (Fiehn *et al.*, 2000b; Glassop *et al.*, 2007). However, if adequate quality control measures are taken as discussed above, trimethylsilylation can still be used as a universal, mild and quick process for metabolite profiling, avoiding the need to double or triple the number of analytical runs per plant sample.

Data processing requires filtering of noisy and inconsistent signals

The main task of quality control in unbiased (non-target) metabolomics is to detect all metabolic signals and subse-

quently remove or combine redundant signals, filter out noisy peaks, select consistent peaks and annotate these in an unambiguous way as known and unknown metabolites as recommended by the MSI document on chemical analysis (Sumner *et al.*, 2007). A variety of commercial and free software tools are available that perform one or several of these tasks, but none of the tools are sufficient to annotate known metabolites in large and diverse datasets. Mostly, software tools assist in analyzing three-dimensional data files such as 'retention time * m/z value * intensity' by aligning chromatograms using similarity recognition (Smith *et al.*, 2006). However, while such tools are helpful in detecting 'biomarkers' of metabolic differences among related experimental classes in small-scale studies, alignments fail when comparisons between very different samples are attempted (such as different plant organs), or when studies with hundreds of samples are carried out that may comprise larger shifts in retention times. The single best way to robustly refer to locked retention times in GC/MS is by using a grid of marker compounds, the classic retention index. Alternatively, some GC/MS instrument software offers instrument control of carrier gas flow to adjust absolute retention times to a single marker compound. Hence, while peak picking and alignment software tools can result in MSI-compliant reports, comparisons of data sets across studies and data exchange between laboratories is much enhanced if internal retention markers are used. We suggest using C8–C30 fatty acid methyl esters because these compounds are absent in plants and generate abundant ions at high m/z values under electron impact mass spectrometry, which facilitates automatic recognition. Resulting retention indices can be converted to the community standard of Kovacs indices if required.

Quality control in data processing means that the incoming data are deconvoluted and filtered according to consistency criteria. Depending on the complexity of the samples, as well the peak detection and mass spectral deconvolution software that is used, between 300 and 800 spectra are reported per chromatogram. How many of these spectra represent authentic metabolic signals? While important advances have been made in the quality of this mass spectral deconvolution (Kind *et al.*, 2007; Stein, 1999), there will inevitably be false-positive and false-negative spectral reports: very-low-abundance compounds adjacent to very abundant peaks may be compromised in spectral quality, or the presence of two very closely co-eluting compounds may not be recognized, leading to export of only one spectrum that comprises ions of both peaks. It is not straightforward to filter these spectra into a unified and consistent table of quantified metabolites. No public repositories exist to date that would allow researchers to carry out the task of unambiguous peak detection and verification. A combination of mass spectral deconvolution and alignment-type data filtering meets some of the

necessary conditions to obtain unified and high-quality data-processing results from instrument-independent input files (Styczynski *et al.*, 2007). For *.peg file formats (as produced by Leco GC–TOF mass spectrometers), an algorithm has been published that removes noisy and inconsistent spectra using a multiple filter system and that annotates correct peaks even if mass spectral similarity is compromised (Fiehn *et al.*, 2005). In addition to the MSI criteria of retention index and mass spectral similarity matches, further peak metadata are exploited, such as peak abundance, mass spectral purity, unique ions and apex masses (Figure S1). Quality thresholds for peak annotations are given in Table S1.

Documentation and designs of the resulting database BinBase are available for free download at <http://fiehnlab.ucdavis.edu/staff/wohlgemuth/binbase/>. Open-source code will be available later this year. Importantly, the algorithm also recognizes novel signals that had never been detected in other samples, and enters these spectra into the database if they are detected in at least six samples and at least 80% of the samples of a class of the biological study design.

Naming signals as authentic metabolites must not be done on the basis of mass spectral matches alone, as required by the MSI document on chemical analysis. Instead, reference compounds or reference libraries must confirm matching for at least two independent parameters; for GC/MS, these are retention time and mass spectrum. Known artifact spectra such as column bleed, phthalates or polysiloxanes should be recognized and discarded from use in statistics. Compounds are identified using the Fiehn library of 713 unique metabolites as outlined above in order to annotate spectra with correct metabolite names and associated identifiers such as PubChem and KEGG numbers. In addition to mass spectra as encoded in the result files, the complete library is available as quadrupole and time-of-flight mass spectra from the corresponding instrument vendors. Result data sets are available from the SetupX and NSF2010 metabolomics databases (<http://www.plantmetabolomics.org> and http://fiehnlab.ucdavis.edu:8080/m1/main_public.jsp), including replacement of missing values using unprocessed *.cdf (ANDI MS) files as detailed above. For any reporting of metabolomic data, it is important to state the way in which missing values have been replaced so that users may be aware how confident the eventual statistical outputs are. While guess estimates for missing values can be computed, the approach proposed here is more favorable as it investigates the authentic data acquisition files instead of assigning putative values. It is important to note that the number of missing values not only depends on the samples and peak abundances, but first and foremost on the stringency of quality criteria: if lax criteria for retention index windows and mass spectral matching are used, the number of false-

positive matches will increase and falsely mask actual misannotations. Conversely, if the criteria are too strict, too many false-negative peak annotations will result. We suggest resolving this dilemma by adjusting peak criteria based on peak metadata (e.g. peak purity and peak abundance), by strictly annotating only one result entry per detected peak (avoiding double annotations), and by replacing the resulting missing data using the raw data instead of assigning values. Lower-confidence results are then labeled.

Presenting results of metabolite profiling by statistical comparisons

The metabolomic workflow is completed by statistical analysis of the quantitative data and presentation of the results in form of graphs, which support the subsequent biochemical or physiological interpretations. Pre-existing knowledge should be integrated as initial working hypotheses to facilitate the interpretation of differential regulation of metabolite levels. Discovery of unexpected metabolic changes may then advance our knowledge of regulation of metabolic pathways. We exemplify such hypothesis building by discussing background information on the current knowledge about the gene *At1g08510*, which encodes FATB, an acyl-ACP (acyl carrier protein) thioesterase. We wish to emphasize that the data concerning the *fatB* knockout allele presented herein are not sufficient to prove the lipid-signaling hypothesis presented below; this would require additional supporting evidence. Instead, this case is given here as an example indicating that the physiological relevance of gene disruptions may often only be revealed under testing under various environmental stress situations, such as physical wounding.

Acyl-ACP thioesterases catalyze hydrolysis of the thioester bond that links the acyl chain to the ACP carrier used in *de novo* fatty acid biosynthesis. Prior characterizations of the Arabidopsis *fatB* knockout mutant indicated that a change in growth rate was associated with a 40–50% decrease in the total amount of saturated fatty acids in various tissues compared to wild-type, specifically affecting the fatty acid composition of extra-plastidial phospholipids, sphingolipids, waxes in leaves and stems, and triacylglycerols in seeds (Bonaventure *et al.*, 2003). In cuticular waxes, the *fatB* mutation leads to an 80% reduction in hexadecane derivatives, concomitant with a reciprocal increase in unsaturated C18 fatty acids (Bonaventure *et al.*, 2004b). In further investigations, it was found that total fatty acid turnover (both synthesis and breakdown of lipids) was increased in order to maintain a steady export of palmitate from the plastids (Bonaventure *et al.*, 2004a). It was hypothesized that cellular sensors might exist that regulate lipid synthesis rates and plastidial palmitate export. It should be noted that, in these earlier studies, lipid analyses were performed after

transmethylations, hence obstructing information about the contents of free fatty acids. Moreover, these earlier studies targeted analysis specifically towards lipid constituents but disregarded potential impact on other metabolic pathways. For the *fatB* knockout mutant case study, a working hypothesis can thus be formulated that known changes in metabolite levels, such as the sphingolipid pools, might affect signaling pathways, specifically under stress situations. This impact on lipid-signaling pathways might subsequently influence unrelated metabolic pathways and lead to metabolic phenotypes that are more distinct under stress than under normal environmental conditions. Consequently, the study design included a simple physical stress imposed on both the *fatB* knockout mutant and its *Ws* ecotype parental genotype, using unstressed plants as negative controls for this difference in wounding response.

Because different statistical tests use specific data treatments and have a set of unique theoretical assumptions, we recommend that a variety of statistical tests be performed in order to verify metabolic differences. The methods and minimal requirements of statistical tests have been compared previously (Goodacre *et al.*, 2007). It is important to note that there is no single best statistical method or combination of methods, but the methods used should reflect the hypotheses and class design of the study. Nevertheless, both univariate and multivariate tests, as well as supervised and unsupervised tools, should be used. Supervised methods are multivariate tests that develop classification models based on *a priori* knowledge of the study design, and hence tend to emphasize differences between classes. Supervised methods comprise decision trees, linear discriminant analysis or regression methods such as partial least-square analysis (PLS). Unsupervised tools (such as the classic principal components analysis, PCA) are mainly used to compress data by linear combinations to derive principal vectors based on the variance inherent in the data set. Therefore, PCA graphs will resemble PLS plots if the biological variance in a data set is far greater than unrelated variance (e.g. noise), whereas PLS will focus on biologically significant variance that is useful for building model vectors that discriminate between classes. Supervised model building is at risk of data over-fitting, especially when a low number of samples and a high number of variables (here metabolites) are used. Univariate data analysis may be presented after analysis of variance (ANOVA) by significance thresholds (e.g. $P < 0.05$); however, data are often further validated by reducing the number of potential false positives using Bonferroni adjustments (i.e. by dividing the P value by the overall number of tested variables to obtain stricter P values).

The finding of general differences in metabolic responses should be exemplified by multivariate and univariate data analysis (Figure 4). The overall extent of differences in metabolic phenotypes can be assessed by multivariate

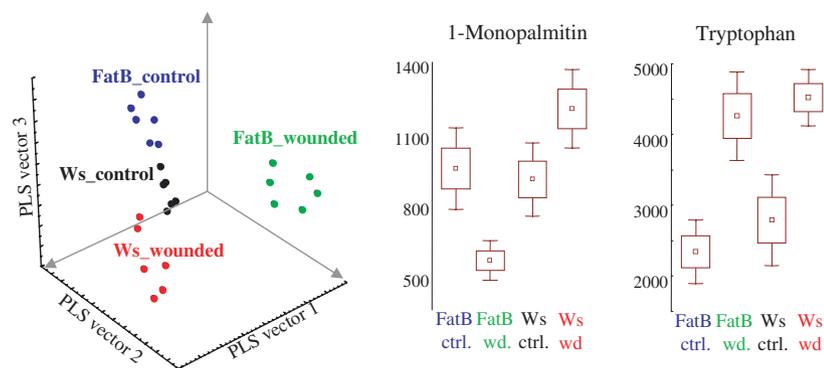


Figure 4. Example of statistical representation of genotype and treatment effects on Arabidopsis metabolism. Left panel: partial least-square multivariate statistics on overall metabolic phenotypes of *fatB* wounded plants and Wassilewskija controls. Right panels: examples of metabolites responding to wounding in a generic and a *FATB*-specific way, using univariate analysis of variance.

statistics such as PLS. The PLS vector with the highest score distinguished the *fatB* mutant under wound response from its wild-type counterpart (Figure 4, left panel), indicating that the physiological roles of the *fatB* gene involve control of wound responses. Vector 2 emphasized generic differences between control and wound responses for both genotypes, whereas vector 3 best separated the *fatB* knockout mutant from the wild-type phenotypes under both treatments. Therefore, each parameter of the study design could be dissected and led to specific metabolic differences that discriminated the metabolic phenotypes.

Univariate statistical analysis is used to investigate changes for individual compounds. A total of 369 metabolic signals were reported by BinBase after final data processing, of which 64% were found to be significantly altered in the *fatB* knockout allele compared to the wild-type after wounding ($P < 0.05$, one-way ANOVA). In comparison, only 11% of the metabolites were found to be significantly altered under control conditions. In accordance with the unaltered export rates of palmitate in *fatB* mutants (Bonaventure *et al.*, 2004a), we found that concentrations of endogenous free fatty acids (both unsaturated and saturated) remained unchanged under control conditions, including those of the potential precursors for membrane lipids such as monopalmitin and monostearin. However, upon wounding stress, the metabolic response in the *fatB* mutant greatly differed from that of the wild-type control (Figure 4, right panel). Free palmitate levels were only slightly decreased upon wounding in the *fatB* knockout mutant, while short-chain free fatty acids such as lauric and capric acids, as well as monoacylglycerols, were significantly reduced by 30–66% 2 h after wounding in the mutant, but not in the Ws control genotype. Further metabolic responses in the *fatB* knockout mutant included up-regulation of the hydroxylated short-chain fatty acid 2-hydroxyvaleric acid. Low-abundance odd-chain free fatty acids such as pelargonate, but also free stearate and the C16/C18 ratio, remained unchanged under all conditions. In addition to these specific differences, a number of wound responses were found to be similar between both genotypes, for example up-regulation of amino acids and increased levels of ascorbate. Therefore,

the *fatB* gene may be assumed not to be the only regulatory gene in wound responses.

It should be noted that use of so-called box-whisker plots (Figure 4, right panel) gives more information than classic bar diagrams because mean values, error bars, quartiles and outliers are compared. Therefore, the use of box-whisker plots rather than bar diagrams is encouraged. Due to the number of detected peaks, not all metabolites can be displayed on graphs. Some may be mentioned in the text while other metabolic changes might be left unspecified. The graph presentations, tables or text discussions of metabolic differences are usually targeted to support a certain interpretation that authors wish to convey. However, it is important that data are not presented in a way that distorts the wealth of detected differences. We therefore present here a plot of tryptophan levels (Figure 4, right panel) in addition to the lipid focus of the study to support the notion that some wounding responses were generic and unaltered in the *fatB* knockout genotype compared to its Ws genetic background, and we mention a variety of other metabolite classes for which significant differences were observed. Additionally, all significant differences should be given at least in the form of processed data, for example in supplementary tables. For our test case, we go one step further by allowing scientists to download all processed and original data from <http://www.plantmetabolomics.org> or http://fiehnlab.ucdavis.edu:8080/m1/main_public.jsp to verify statistical differences that are claimed here or to perform further tests.

As stated above, this example data set can be interpreted as indicating that metabolic remodeling in the wound response is severely affected in the *fatB* mutant, and this largely outweighs the metabolic differences between the genotypes under control conditions. This observation supports the hypothesis that control of physiological stress conditions (such as the wounding response) may be a functional role for this gene; however, the lack of supplementary lines of evidence (e.g. time-course studies, gene expression or studying putative signal receptors and signal molecules) does not allow us to draw further conclusions on the validity of the lipid-signaling hypothesis (Bonaventure *et al.*, 2004a,b).

Conclusions

The value of metabolomic data is directly proportional to the quality of the associated metadata. Thus, metadata that define parameters associated with the source of biological materials, the protocols for metabolite extraction, identification and quantification, and the procedures for data analysis and interpretation, are essential for integrating metabolomics datasets. Uniformity in such metadata will provide a means of querying complex datasets using a specific hypothesis that can be assessed from the collected datasets (e.g. stress associated or organ-specific metabolites). Queries could then compare studies across species to distinguish species-specific from generic responses. In essence, the usefulness of metabolomics depends on the quality of the acquired data as well as the level of detail used in describing the studies.

The examples given here should show that quality control extends far beyond automatic routines, such as the internal calibration of a mass spectrometer. Quality control is instead an attitude of diligence that seeks to optimize procedures. Here we have focused on several aspects of the metabolomic workflow, specifically on GC-TOF metabolite profiling. We have outlined a good-practice document, including consistent data-processing algorithms and database queries. However, we have not covered all pitfalls in a comprehensive manner, and more observations and improvements will alter our standard operating procedures and algorithms. The usefulness of plant metabolite profiling will benefit from establishing open access MSI-compliant repositories.

In a similar manner as presented here for GC-TOF metabolite profiling, quality controls and procedures should be developed for LC/MS-based approaches, direct-infusion MS lipid fingerprints and NMR. The Metabolomics Society has established an initiative to foster work in this direction, and MSI documents are publicly available at the initiative's public web page (<http://msi-workgroups.sourceforge.net/>). Comments on any of the suggested minimal reporting standards are welcome, and may be sent to msi-workgroups-feedback@lists.sourceforge.net.

Acknowledgments

This work presented here has been mainly funded by research grant 5R01ES13932 from the US National Institute of Environmental Health Sciences and grant MCB-0520140 from the US National Science Foundation. Additional contributions and projects by corporate sponsors Agilent (Santa Clara, CA), Leco (St Joseph, MI), Monsanto (St Louis, MO) and DuPont-Pioneer (Wilmington, DE) helped in improving the protocols, technologies and databases. We are grateful for the solubility calculations performed using COSMOfrag by Andreas Klamt (COSMOlogic GmbH and Co. KG, Leverkusen, Germany). We further appreciate the diligence of all the contributors with regard to the minimum reporting standards within the Metabolomics Standards Initiative. We acknowledge the contribution by Katayoon Dehesh (University of California at Davis),

who provided seeds of the *fatB* knockout mutant first described by Bonaventure *et al.* (2003).

Supplementary Material

The following supplementary material is available for this article online:

Figure S1. The BinBase algorithm for annotation of GC-TOF mass spectra.

Table S1. Spectrum similarity criteria in BinBase.

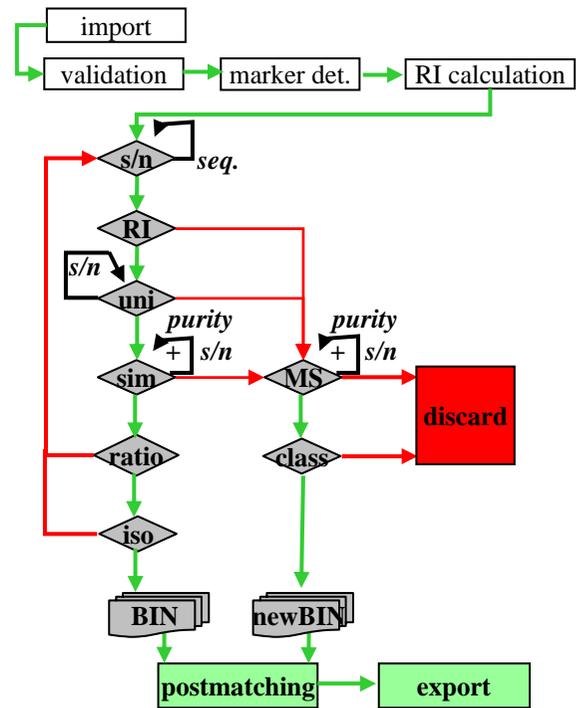
This material is available as part of the online article from <http://www.blackwell-synergy.com>.

Please note: Blackwell publishing are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

References

- Bonaventure, G., Salas, J.J., Pollard, M.R. and Ohlrogge, J.B. (2003) Disruption of the FATB gene in Arabidopsis demonstrates an essential role of saturated fatty acids in plant growth. *Plant Cell*, **15**, 1020–1033.
- Bonaventure, G., Ba, X.M., Ohlrogge, J. and Pollard, M. (2004a) Metabolic responses to the reduction in palmitate caused by disruption of the FATB gene in Arabidopsis. *Plant Physiol.* **135**, 1269–1279.
- Bonaventure, G., Beisson, F., Ohlrogge, J. and Pollard, M. (2004b) Analysis of the aliphatic monomer composition of polyesters associated with Arabidopsis epidermis: occurrence of octadecacis-6,cis-9-diene-1,18-dioate as the major component. *Plant J.* **40**, 920–930.
- David, F., Devos, C., Joulain, D., Chaintreau, A. and Sandra, P. (2006) Determination of suspected allergens in non-volatile matrices using PTV injection with automated liner exchange and GC-MS. *J. Sep. Sci.* **29**, 1587–1594.
- Denkert, C., Budczies, J., Kind, T., Weichert, W., Tablack, P., Sehoul, J., Niesporek, S., Konsgen, D., Dietel, M. and Fiehn, O. (2006) Mass spectrometry-based metabolic profiling reveals different metabolite patterns in invasive ovarian carcinomas and ovarian borderline tumors. *Cancer Res.* **66**, 10795–10804.
- Fiehn, O. (2002) Metabolomics – the link between genotypes and phenotypes. *Plant Mol. Biol.* **48**, 155–171.
- Fiehn, O. (2006) Metabolite profiling in Arabidopsis. In *Methods in Molecular Biology* (Salinas, J. and Sanchez-Serrano, J.J., eds). Totowa, NJ: Humana Press, pp. 439–447.
- Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, R.N. and Willmitzer, L. (2000a) Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* **18**, 1157–1161.
- Fiehn, O., Kopka, J., Trethewey, R.N. and Willmitzer, L. (2000b) Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Anal. Chem.* **72**, 3573–3580.
- Fiehn, O., Wohlgenuth, G. and Scholz, M. (2005) Setup and annotation of metabolomic experiments by integrating biological and mass spectrometric metadata. *LNCS*, **3615**, 224–239.
- Fiehn, O., Robertson, D., Griffin, J. *et al.* (2007a) The Metabolomics Standards Initiative (MSI). *Metabolomics*, **3**, 175–178.
- Fiehn, O., Sumner, L.W., Rhee, S.Y., Ward, J., Dickerson, J., Lange, B.M., Lane, G., Roessner, U., Last, R. and Nikolau, B. (2007b) Minimum reporting standards for plant biology context information in metabolomic studies. *Metabolomics*, **3**, 195–201.

- Glassop, D., Roessner, U., Bacic, A. and Bonnett, G.D. (2007) Changes in the sugarcane metabolome with stem development. Are they related to sucrose accumulation? *Plant Cell Physiol.* **48**, 573–584.
- Goodacre, R., Broadhurst, D., Smilde, A.K. et al. (2007) Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics*, **3**, 231–241.
- van der Greef, J., Stroobant, P. and van der Heijden, R. (2004) The role of analytical sciences in medical systems biology. *Curr. Opin. Chem. Biol.* **8**, 559–565.
- Grob, K. and Biedermann, M. (2000) Video-taped sample evaporation in hot chambers simulating gas chromatography split and splitless injectors – II. Injection with band formation. *J. Chromatogr. A*, **897**, 247–258.
- Gullberg, J., Jonsson, P., Nordstrom, A., Sjostrom, M. and Moritz, T. (2004) Design of experiments: an efficient strategy to identify factors influencing extraction and derivatization of *Arabidopsis thaliana* samples in metabolomic studies with gas chromatography/mass spectrometry. *Anal. Biochem.*, **331**, 283–295.
- Hajslova, J., Holadova, K., Kocourek, V., Poustka, J., Godula, M., Cuhra, P. and Kempny, M. (1998) Matrix-induced effects: a critical point in the gas chromatographic analysis of pesticide residues. *J. Chromatogr. A*, **800**, 283–295.
- Halket, J.M. and Zaikin, V.G. (2003) Derivatization in mass spectrometry – 1. Silylation. *Eur. J. Mass Spectrom.*, **9**, 1–21.
- Kanani, H.H. and Klapa, M.I. (2007) Data correction strategy for metabolomics analysis using gas chromatography–mass spectrometry. *Metab. Eng.* **9**, 39–51.
- Katajamaa, M., Miettinen, J. and Oresic, M. (2006) MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics*, **22**, 634–636.
- Kind, T., Tolstikov, V., Fiehn, O. and Weiss, R.H. (2007) A comprehensive urinary metabolomic approach for identifying kidney cancer. *Anal. Biochem.* **363**, 185–195.
- Klamt, A., Eckert, F., Hornig, M., Beck, M.E. and Burger, T. (2002) Prediction of aqueous solubility of drugs and pesticides with COSMO-RS. *J. Comput. Chem.* **23**, 275–281.
- Koek, M.M., Muilwijk, B., van der Werf, M.J. and Hankemeier, T. (2006) Microbial metabolomics with gas chromatography/mass spectrometry. *Anal. Chem.* **78**, 1272–1281.
- Lisec, J., Schauer, N., Kopka, J., Willmitzer, L. and Fernie, A.R. (2006) Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nat. Protocols*, **1**, 387–396.
- Meyer, R.C., Steinfath, M., Lisec, J. et al. (2007) The metabolic signature related to high plant growth rate in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA*, **104**, 4759–4764.
- Modi, A.T., Streeter, J.G. and McDonald, M.B. (2000) Relative efficiency of ethanol and pyridine as extractants of low molecular weight carbohydrates from soybean seed axes. *Seed Sci. Technol.* **28**, 193–200.
- Noctor, G., Bergot, G.L., Mauve, C., Thominet, D., Lelarge-Trouverie, C. and Prioul, J.L. (2007) A comparative study of amino acid measurement in leaf extracts by gas chromatography–time of flight–mass spectrometry and high performance liquid chromatography with fluorescence detection. *Metabolomics*, **3**, 161–174.
- Rhee, S.Y., Beavis, W., Berardini, T.Z. et al. (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res.* **31**, 224–228.
- Roessner, U., Luedemann, A., Brust, D., Fiehn, O., Linke, T., Willmitzer, L. and Fernie, A.R. (2001) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell*, **13**, 11–29.
- Sanz, J., Soria, A.C. and Garcia-Vallejo, M.C. (2004) Analysis of volatile components of *Lavandula luisieri* L. by direct thermal desorption–gas chromatography–mass spectrometry. *J. Chromatogr. A*, **1024**, 139–146.
- Scholz, M. and Fiehn, O. (2007) SetupX – a public study design database for metabolomic projects. *Pacific Symposium on Bio-computing*, **12**, 169–180.
- Smith, C.A., Want, E.J., O’Maille, G., Abagyan, R. and Siuzdak, G. (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* **78**, 779–787.
- 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.
- Styczynski, M.P., Moxley, J.F., Tong, L.V., Walther, J.L., Jensen, K.L. and Stephanopoulos, G.N. (2007) Systematic identification of conserved metabolites in GC/MS data for metabolomics and biomarker discovery. *Anal. Chem.* **79**, 966–973.
- Sumner, L.W., Amberg, A., Barrett, D. et al. (2007) Proposed minimum reporting standards for chemical analysis. *Metabolomics*, **3**, 211–221.
- 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.
- Zhang, P., Foerster, H., Tissier, C.P., Mueller, L., Paley, S., Karp, P.D. and Rhee, S.Y. (2005) MetaCyc and AraCyc. Metabolic pathway databases for plant research. *Plant Physiol.* **138**, 27–37.



Supplemental table S2

	s/n > 250	s/n > 25	s/n < 25
purity < 0.1	800	700	600
purity < 1.5	700	600	500
purity > 1.5	600	500	500