FSA G03012

Technical Report

“Development of unified data models and data pre-processing strategies and the generation of meaningful, standardised statistical analyses of metabolome variability in crop plants”

PI - Prof. John Draper, IBERS, Aberystwyth University, July 2008

Table of Contents

G03012: Non-Technical Executive Summary

PI Prof. John Draper, Aberystwyth University

Background

In a previous co-ordinated research programme (G02) the Foods Standards Agency commissioned 12 collaborating research teams in the UK, Germany and France to develop technology that would allow the substantial equivalence of genetically manipulated crop versus traditionally bred crops to be determined. A major conclusion emerging from the G02 programme was that an assessment of overall chemical composition of food raw materials provided the most comprehensive, cost effective and reproducible method to assess substantial equivalence. Overall chemical composition was evaluated using techniques which assay the total metabolite content (i.e. all small, non-polymer molecules such as sugars, amino acids, vitamins, fatty acids) in crude plant tissue extracts. This combination of analytical methods is referred to collectively as **metabolomics** technology. The metabolomics technology platform depends on a combination of analytical chemistry techniques to produce chemical 'fingerprints/profiles' of food raw material samples which are then investigated using powerful data modelling techniques. The G02 programme produced data for potato, wheat, barley, Arabidopsis and tomato at a time when no widely applied standards for experimental procedure had been developed in the emerging research field of metabolomics.

Rationale

The large amount of data collected in the G02 programme was not necessarily comparable as no prior arrangements were defined to ensure or enhance comparability of data from different projects. An initial survey of the G02 data sets revealed that potato had provided a common crop plant for metabolomics technology development in several research centres (Aberystwyth University, Scottish Crops Research Institute, Institute of Food Research) and thus it was agreed from the outset that the G03012 project would be centred on potato tubers. The overall aim of the G03012 project was to first evaluate the structure of these disparate G02 data, developed on different instruments using a range of procedures, and then to determine whether unified data models and standardised data analysis procedures could be debeloped in order that metabolomics technology could be used to generate meaningful and durable descriptions of food raw material composition for any future safety or quality assessments.

One major output would be a validated data generation and pre-processing strategy for production of a standardised check list of ‘expected’ metabolites bespoke for any crop (in our case potato tubers). As part of this process a survey of external literature and database information describing the chemical composition of potato tubers was undertaken as a comparator. A pilot database was to be developed (ARMeC: Aberystwyth Repository of Metabolite Characteristics) which could hold such information in a unified and searchable format. A robust statistical methodology to generate meaningful metrics for describing metabolite baselines in food raw materials was also to be validated within the G03012 project.

Approach and Objectives

The G03012 project was derived from the merger of two contracts initially put out for tender by the FSA. As the G02 data structure was evaluated in the initial phases of the G03012 project the research objectives were consolidated into eight major interacting themes with the research lead in different subtasks split between Aberystwyth_Biology, Aberystwyth_Computer Science and SCRI:

1) Collation of metabolite profiling/fingerprinting data from selected FSA G02 projects based on potato.

2) Analysis and summary of external literature and database information describing the chemical composition of potato tubers.

3) Development of a pilot database (ARMeC) to accommodate information describing the individual chemical components of a food raw material.

4) Development of unified data models for G02 metabolomics data.

5) Development of standardised procedures for data generation and pre-processing of data tables to allow comparison and validation of metabolite profiles and fingerprints generated in different laboratories.

6) Development of standardised statistical and data modelling software in a single package.

7) Validation of metrics derived from statistical tests that can be used to compare metabolite composition in food raw materials

8) Generation of a baseline description of metabolite composition in potato tubers.

Outcomes and Key Results

A major output was the development of standardised methods for chemical fingerprinting based on Mass Spectrometry (MS) which was published as part of a series of invited articles in the Nature Protocols Journal. Additionally, contributions from G03012 staff to international collaborations helped to deliver standardised guidelines for generation and processing of Nuclear Magnetic Resonance (NMR) fingerprinting data. Data tables representing metabolite profiles of potato tubers developed in 3 independent laboratories were successfully aligned and standardised procedures for metabolite peak identification and quantification were refined. A major output of G03012 was a ‘checklist’ of metabolite peaks that could be detected in potato tubers using mass spectrometric profiling together with the development of the ARMeC database to hold information on each chemical.

A substantial assessment of external literature and food-related databases revealed that information on the chemical compositional of potato tubers was extremely disparate; importantly biological materials and analytical techniques were often not described adequately, the list of metabolites measured differed dramatically and a bewildering range of concentration metrics were used. We concluded that data unification was essentially impossible and under such circumstances there was little value in implementing a new database to accommodate such data.

The data modelling routines used effectively in the G02006 project had been written in several software languages and could only be operated by experienced data miners. A major output of the G03012 project was the conversion of all data analysis software into a common package and the development of tutorials and example workflows that allowed biologists, with little training in statistical routines, to become proficient in data mining. All software and tutorials are now available on the web and their use described in a commissioned Nature Protocols article. In a demonstration project utilising both metabolome fingerprint and metabolite profiling data it was demonstrated that the standardised data mining procedures could be used to compare the total chemical composition of potato genotypes. The ultimate output of the G03012 project was therefore validated procedures for standardised generation and interrogation of metabolomics data which can be used to provide an overview of chemical composition of any food raw material.   This technology is expected to have great utility for the development of food composition databases in future studies investigating safety, quality and provenance aspects of food raw materials.

G03012: Technical Summary of Achievements, Outputs, Main Conclusions & Recommendations

PI Prof. John Draper, Aberystwyth University

"Development of unified data models and data pre-processing strategies and the generation of meaningfull, standardised statistical analyses of metabolome variability in crop plants"

Project Background and Technical Report Overview

Background

In a previous co-ordinated research programme (G02) the Foods Standards Agency commissioned 12 collaborating research teams in the UK, Germany and the Netherlands to develop technology that would allow the substantial equivalence of genetically manipulated crops versus traditionally bred crops to be determined. A major conclusion emerging from the G02 programme was that metabolomics data provided a comprehensive assessment of the overall chemical composition of food raw materials. The metabolomics technology platform depends on a combination of analytical chemistry techniques to produce chemical 'fingerprints/profiles' of food raw material samples which are then investigated using powerful data modelling techniques. The G02 programme produced metabolomics data for potato, wheat, barley, Arabidopsis and tomato at a time when no widely applied standards for experimental procedure had been developed in the emerging research field of metabolomics. These data were not necessarily comparable as no prior arrangements were defined to ensure or enhance comparability of data from different projects.

Genesis of G03012 project

In an initial call for proposals the FSA advertised for tenders for two research requirements, G03R0003 and G03R0002. In principle G03R0003 was to collate metabolomics data from the G02 projects and attempt to unify data structures in a suitable database, whilst G03R0002 was to analyse the data and develop standardised procedures for data modelling to allow compositional comparisons to be made between food raw materials. After receiving all tenders the FSA invited a team based in Aberystwyth (headed by Prof. John Draper) with collaborators in SCRI to develop a merged proposal covering both project objectives, with less emphasis on database development and subsequent omission of CSL as a collaborating partner. With specific guidance from the FSA at a meeting in Birmingham on 9th November 2005 the G03012 project was developed that focused on potato tubers. The merged proposal was accepted by the FSA in the last week of November 2005 with a start date of 1st December 2005. This exceedingly tight timeframe for project initiation was only possible because several key researchers from the G02 projects where available immediately in Aberystwyth and SCRI.

These new organisational arrangements requested by the FSA had two knock on effects acknowledged from the outset by all parties. Firstly, it had to be accepted that certain objectives would be delayed and re-scheduled as there was a need to hire further specialised staff at both sites. Secondly, as the data collation and data analysis elements of the merged project were essentially to run in parallel the project milestones relating to data set integration and analysis could only be planned effectively when the G02 data sets had been obtained and their overall structure adequately assessed. The amended objectives and milestones/deliverables were finalised at a FSA review meeting on 28th June 2007 and for the purpose of future reporting the G03012 project was re-organised into 9 major elements (as indicated below) that are used also in the full Technical Report.

G03012 Revised Objectives

1. Compiling and assessment of data from G02 programme

2. Compiling and assessment of external information on composition of potato tubers

3. Standardise procedures for manual labeling of metabolite peaks in GC-MS profiles

4. Assessment of software for automated alignment and annotation of GC-MS and LC-MS data

5. Development of methodology for standardized generation and pre-processing of ESI-MS metabolite fingerprint data

6. Develop a pilot database for annotation of nominal mass ESI-MS data

7. Dealing with missing values in metabolomics data and analysis of ranges of metabolite values and development of statistical models of metabolite distributions in conventional crops

8. Refinement of standardised multivariate statistical methods for both describing and comparing metabolite composition of food raw materials

9. Understanding the output of standardised multivariate methods to analyse metabolomics data from the G02 programme.

G03012 Rationale

An initial survey of the G02 data sets revealed that potato had provided a common crop plant for metabolomics technology development in several research centres (Aberystwyth University, Scottish Crops Research Institute, Institute of Food Research) and thus it was agreed from the outset that the G03012 projects would be centred on potato tubers. The overall aim of the G03012 project was to first evaluate the structure of these disparate G02 data sets, developed on different instruments using a range of procedures, and then to determine whether unified data models and standardised data analysis procedures could be developed in order that metabolomics technology could be used to generate meaningful and durable descriptions of food raw material composition for any future safety or quality assessments.

One major output would be a validated data generation and pre-processing strategy for production of a standardised check list of 'expected' metabolites bespoke for any crop (in our case potato tubers). As part of this process a survey of external literature and database information describing the chemical composition of potato tubers was undertaken as a comparator. A pilot database was to be developed (ARMeC: Aberystwyth Repository of Metabolite Characteristics) which could hold such information in a unified and searchable format. A robust statistical methodology to generate meaningful metrics for describing metabolite baselines in food raw materials was also to be validated within the G03012 project.

Outcomes and Key Results

A major output was the development of standardised methods for chemical fingerprinting based on Mass Spectrometry (MS) which was published as part of a series of invited articles in the *Nature Protocols* Journal. Additionally, contributions from G03012 staff to international collaborations helped to deliver standardised guidelines for generation and processing of Nuclear Magnetic Resonance (NMR) fingerprinting data. Data tables representing metabolite profiles of potato tubers developed in 3 independent laboratories were successfully aligned and standardised procedures for metabolite peak identification and quantification were refined. A major output of G03012 was a 'checklist' of metabolite peaks that could be detected in potato tubers using mass spectrometric profiling together with the development of the ARMeC database to hold information on each chemical.

A substantial assessment of external literature and food-related databases revealed that information on the chemical composition of potato tubers was extremely disparate; importantly biological materials and analytical techniques often were not described adequately, the list of metabolites measured differed dramatically and a bewildering range of concentration metrics were used. We concluded that data unification was essentially impossible and under such circumstances there was little value in implementing a new database to accommodate such data.

The data modelling routines used effectively in the G02006 project had been written in several software languages and could only be operated by experienced data miners. A major output of G03012 was the conversion of all data analysis software into a common package and the development of tutorials and example workflows that allowed biologists, with little training in statistical routines, to become proficient in data mining. These software and tutorials now are available on the web and their use described in a commissioned *Nature Protocol* article. In a demonstration project utilising using both metabolome fingerprint and metabolite profiling data it was demonstrated that the standardised data mining procedures could be used to compare the total chemical composition of potato genotypes.

Risk assessment in relation to novel foods, including new genetically manipulated crop varieties, demands a strategy which at a minimum is:

- Comprehensive in its analytical 'reach' in order to survey as much chemistry as possible
- Clear about the normal range and expected variability of chemical composition in current crop varieties
- Explicit about how compositional data from novel food raw materials are compared to existing foods
- Reproducible in other laboratories

The ultimate output of the G03012 project was effectively validated procedures for standardised generation and interrogation of metabolomics data which can be used to provide an overview of chemical composition of any food raw material. ***The analytical SOPs and software resources generated by the G03012 are freely available as web resources and supported by comprehensive tutorials and example analyses which can be used 'off the shelf' to support the safety assessment of novel foods in the future.***
A major output of the G03012 project was the combining of validated analytical chemistry techniques with appropriate validated data analysis techniques specifically to address the issue of comparing either novel GM foods to their progenitors, or, to make a comprehensive compositional analysis *de novo* of a food crop species new to the UK. ***Thus the GO3012 project demonstrated that 'first pass' screening by metabolite fingerprinting (ESI-MS or NMR) is a cost-effective and reproducible method to rapidly compare novel foods to their progenitors, as long as signal intensity/resolution thresholds are similar in any instrument used to make measurements***. A particular benefit of fingerprinting methods is that they are generally high throughput and, importantly, data pre-processing and alignment are trivial and can be automated. If any unexpected differences are found then these can be followed up by more targeted analysis using metabolite profiling methods such as GC-MS/LC-MS with the caveat that data alignment procedures need to be agreed beforehand when different instruments (e.g. in different laboratories) are used. ***The G03012 project demonstrated that, even after exhaustive analysis, annotation in metabolite GC-MS profiles was only successful for around 40% of peaks (i.e. approximately 60% remained unknowns), however a sub-set of expected peaks could be identified using any instruments which should be used as the basis of future rational comparisons.***
A further major conclusion from the GO3012 project is that food raw materials can indeed be compared in a standardised way (in terms of both analytical chemistry technology and statistical analysis) as long as there has been a previous comprehensive analysis of the food crop species in question. ***For example a major finding from the analysis of potato, Arabidopsis and grass varieties/ecotypes was the surprisingly large extent of compositional diversity within the germplasm representative of a particular plant species.*** This was likely only to increase when the same genotypes were grown in different geographical regions. A key message here was that a metabolite 'range' analysis is certainly a good starting point for any comparison and so genotype diversity in composition needs to be assessed beforehand. ***The G03012 projected offered SOPs***

for making this assessment based on metabolite fingerprinting and discrimination/clustering analyses. Following on from this a second major conclusion concerned answering the question "what should any novel food be compared to?" ***We came to the conclusion that the best universal comparator would essentially be a 'mastermix' population of extracts representative of crop species overall chemical diversity.*** **It is expected this approach can be scaled up in the future to develop a resource for querying the extent of any compositional difference between any novel food crop genotypes and existing foods currently encountered by consumers in the UK.**
A final key contribution of the G03012 project to food safety assessment procedures was the validation of a strategy to determine specific thresholds of 'similarity' between samples, below which the food raw material should be considered effectively identical when examined by the specific analytical technique in use. ***The G03012 project demonstrated that only by promoting the use of specific validated quantitative measures of similarity (such as model margins, AUC or Eigen values) can the relative scale of any compositional differences between two types of food raw material can be assessed rationally.*** In addition to food safety this technology validated in the G03012 project is also expected to have great utility for the development of food composition databases in future studies investigating, quality and provenance aspects of food raw materials.

When the G03012 project was nearing completion the PI was informed by the FSA that we had been nominated for a formal research quality assessment by UKAS. This review (carried out on 12th February 2008) concluded that the project had been carried out to a high standard and the full report is presented as an Appendix to the final report.

Report Organisation

The full technical report is comprised of independent contributions from 5 different research teams, each of which combined efforts to achieve the project's overall objectives. The individual report sections are different in style and content depending on whether they needed to provide extensive literature overviews, or SOPs of experimental procedures or whether they were required to display database designs or data analysis results.

The full reports and Appendices are provided as a web archive. The present technical report summary re-iterates the main ***Requirements*** of the FSA G03012 project and then summarises the main **Achievements**, providing links to relevant sections in the full technical report. All **Outputs** for each Requirement are indicated and links to PDFs of publication or URLs for web sites are provided.

As most of the sections are interactive with others it was deemed appropriate to present the main ***Conclusions*** and ***Recommendations*** of the G03012 project in this Technical Summary.

Requirement (1) “*Compile metabolomics data relating to potato tubers from G02 programme participating laboratories*”

Achievements (1)

- G02 data archive developed and both analytical chemistry and meta-data field information clarified in conjunction with collaborating laboratories. (DATA FROM THE G02 PROGRAMME)
- Specific problems particularly with meta-data ontology were identified.
- All the identified issues contributed to international efforts aiming to standardize ontology in metabolomic database strategies which resulted in an invited contribution to the journal *Metabolomics*. (Assunta Sansone *et al.* (2007))

Conclusions/recommendations (1)

- In the future it would be beneficial to give laboratories generating ‘omics’ data some idea of the minimal information required to allow others to understand the experiment and data.
- This need not be too prescriptive and it is suggested that the FSA encourage researchers it funds to be aware of (and perhaps contribute to) international standardization efforts for metabolomics for at least the next 3-5 years, until minimal reporting requirements are consolidated. (see http://msi-workgroups.sourceforge.net/ for information on the Metabolomics Standards Initiative).

Outputs (1)

- Assunta Sansone *et al.* (2007).

Requirement (2) “*Assess structure of G02 data and identify data sets suitable to study for experiments aiming to develop unified data models for metabolomics*”

Achievements (2)

- Three large GC-MS datasets and two large ESI-MS datasets describing potato tuber metabolite content were identified for deeper comparison.
- In contrast, only a single NMR data set and single small LC-MS data set were identified, providing no scope for assessment of data integration.
- G03012 staff developed interactions with international efforts concerning NMR data modelling and spearheaded development of international standards for NMR data.
- NMR data standardization strategy research resulted in an invited contribution to the journal *Metabolomics*. (Rubtsov *et al*. (2007))

Conclusions/recommendations (2)

- The Aberystwyth conclusions relating to standardisation of GC-MS and NMR experimental and meta-data data was broadly in line with developing international recommendations.

- Further work will be required to validate approaches for LC-MS data as there were not sufficient data sets derived from G02 projects. With the advent of UHPLC and data alignment software LC-MS profiling experiments are becoming more and more common and it is very important that realistic standards are developed in the future.

Outputs (2)

- Rubtsov *et al*. (2007).

Requirement (3) ***"Review the structure of external literature and database information describing metabolite content of potato tubers and assess scope for future data integration"***

Achievements (3)

- An extensive literature and external database review was conducted and major problems in data heterogeneity, data sparcity and lack of experimental detail regarding data generation were identified. (Section 2.3.5)
- One major output was a list of metabolites that had been measured (together with reviewed literature) in potato tubers. (potato tuber metabolites)
- Further detailed analysis of data in literature allowed the development of a semi-quantitative scale to describe metabolite content in potato tubers.

Conclusions/recommendations (3)

- It was concluded in conjunction with ArMet database designers (see *Requirement* 4) that, although there was scope for data integration between the metabolomics analytical & experiment meta-data and external literature/database information relating to food composition, the effort involved in the current limited project would not be warranted in terms of the quality of information generated.
- In future it would be worthwhile to review the type and quality of compositional information for a much wider range of food raw materials and develop further the concept of a unified approach to represent this disparate information.
- It is recommended that issues of compositional differences between different cultivars and environmental effects (e.g. geographical source of crop, post harvest storage, age of crop) are recognized more fully and more information provided on the sources of food raw materials and extraction and measurement techniques used to describe a typical food class.
- There would need to be at least a national or EU consensus from institutions planning to use food composition information on the structure and user requirements of any future archive.

Outputs (3)

- Metabolite 'check list', associated literature and semi-quantitative metabolite concentration data contributed to new implementation of the Aberystwyth Repository of

Metabolite Characteristics (ARMeC) database which is publicly available and focuses specifically on potato tubers. (Section 2.5)

- Scope for future publication identified by SCRI for *Trends in Plant Science* highlighting the problems of representation of raw food material composition in the literature.
- Overy *et al*., 2008

Requirement (4) “***Assess the scope for the further modification of ArMet database to hold data relating to both metabolomics fingerprinting/profiling experiments from G02 programme and information from external literature/databases”***

Achievements (4)

- Architecture for a new implementation of ArMet was designed to accommodate new data models and modified experimental meta-data fields (Section 2.4)
- Decision made with agreement of the FSA not to attempt to populate with information from G02 projects or external literature on metabolite content (see *Requirement* 3).

Conclusions/recommendations (4)

- See Conclusions: see Requirement (3).

Outputs (4)

- Concepts provided background for Hardy and Jenkins, (2007) and Hardy and Taylor (2007).

Requirement (5) “***Standardise procedures for manual annotation of metabolite peaks in GC-MS profiles”***

Achievements (5)

- A detailed analysis of the structure of the GC-MS data produced on three different instruments (each in a different institution) demonstrated that there was clear scope for the development of unified GC-MS data models. (Appendix 3A)
- Importantly, many structurally uncharacterized compounds could be aligned using a data model based on retention time window and characteristic masses, rather than fully deconvolved spectra. (Section 3.5; Appendix 3B)

Conclusions/recommendations (5)

- The number of measured peaks in a peak table from any GC-MS instrument will always depend on its sensitivity and resolution but in general terms the use of common types of columns, ionisation systems and protocols provides an excellent possibility that, with care, data can be reproduced between laboratories.

- There is certainly a core set of ‘expected’ peaks (both structurally characterized and unknown) than can be recorded and aligned in any GC-MS profile data which will provide scope for future data integration.
- It is recommended that this set of core peaks could be extracted from data generated in any competent laboratory and provide a common feature set for any future comparison of food raw material composition.
- This list of peaks generally covers (and in most cases is greater in number than) the list of **targeted** metabolites recorded in the better quality external databases representing food composition. Any key nutritionally-relevant metabolites not easily monitored by a standardized GC-MS profile could be measured by an appropriate targeted analysis method.

Outputs (5)

- No specific outputs solely from G03 project but concepts helped provided background for Jenkins *et al*., (2007).
- A check list of measured metabolites in potato tubers was developed and the information used to create a table in the ARMeC database (see Requirement (3)).

Requirement (6) *“Assessment of software for automated alignment of GC-MS and LC-MS data”*

Achievements (6)

- A range of software was accessed and a research assistant was trained in their operation.
- The characteristics of more than 20 software packages were examined in detail (Section 4.2)
- Two ‘open platform’ packages (Metalign and XC-MS) were eventually chosen to assess functionality in detail using small LC-MS and GC-MS data sets.
- The output of the automated analysis was compared to the bespoke instrument vendor software for peak detection and spectral deconvolution.
- The analysis revealed that both packages required substantial amounts of computing time to detect many of the strong well-resolved peaks found using instrument software and additionally highlighted many hundreds of mass signals with characteristics of metabolite peaks which would normally be considered machine ’noise’. (Section 4.3)

Conclusions/recommendations (6)

- It was concluded that manual intervention by expert users was still required to sensibly parameterize all software for adequate peak finding and data alignment in both LC-MS and GC-MS raw data.
- Total automation was not possible, especially for larger data sets (> 30 chromatograms) in which peak alignment becomes a problem.

- Out of all the packages the XC-MS software performed the most robustly but it had less ‘push button’ operations and did require researchers to learn some basic R (language for computational statistics) command line programming.
- As XC-MS works directly with data in the R environment. Ultimately this will lead more likely into fully automated methods for data processing and subsequent modeling. Such solutions should be encouraged.
- Note that recent external publications have seen increasing use of XC-MS as a preferred solution.

Outputs (6)

- The result of data processing assessments was presented by Prof. Draper as part of a workshop on data mining at the NUGO Metabolomics Workshop in December 2007.

Requirement (7) ***“Develop standardized procedures for the generation of metabolite fingerprints using LC-MS”***

Achievements (7)

- An extensive literature review was undertaken to determine the range of methodologies used to generate LC-MS global fingerprints (Section 5.3).
- Standardized ESI-MS protocols were developed which can be used on both ion trap and ToF instruments (LTQ, LCT). (Section 5.4)
- A detailed SOP was developed and is accompanied by a comprehensive illustration of anticipated results (Section 5.5)
- QA procedures were developed based on monitoring ‘mastermix‘ samples interspersed with experimental samples. (Section 5.3.5)
- Standardized approaches for data pre-processing (e.g. base line corrections, outlier detection and normalization) were developed. (Section 5.5.2)
- The ESI-MS fingerprinting strategy formed the basis of an invited contribution to *Nature Protocols* (Beckmann *et al*., 2008)

Conclusions/recommendations (7)

- It was concluded that by far the majority of LC-MS fingerprinting experiments used electrospray ionization (ESI) in both ionisation modes.
- Sample introduction by direct injection produced more variability than either flow infusion or use of a nano-spray device (e.g. Advion Nanomate). It is recommended that the controlled, stable infusion with a Nanomate-type spray device be the method of choice wherever possible.

- Instruments (both quads and ion traps) used ranged from those with simple quadrupoles (Q) detectors, through to instruments with more accurate Time of Flight (TOF) dectors, hybrid detectors (Q-ToF) or ultra high accuracy FT-ICR-MS detectors. As predicted we found there was always a tradeoff between sensitivity and mass resolution and thus in most cases an optimal actual mass accuracy for fingerprinting was well below the maximum theoretically possible for every instrument.
- It is recommended that nominal mass fingerprinting on the very stable linear ion traps is used for first pass, high throughput studies to identify mass 'bins' containing important discriminatory signals. Unlike instruments with high mass accuracy there is no requirement to align signals and so there is great scope for data integration in the future.
- Instruments with ToF or FT-ICR-MS detectors can subsequently be used to analyze pooled samples representing the biological classes under study to identify and annotate explanatory signals if deeper analysis is required, for example to define the chemical differences.
- Flow infusion ESI-MS methodology can be used to quickly determine whether substantial compositional differences exist between different batches of food raw materials. It is noted that no methodology is fully comprehensive in terms of metabolite coverage and so it is recommended that further non-targeted profiling techniques (such as GC-MS) and targeted analyses for any **expected** nutritionally relevant metabolites should be carried out on any new raw food material matrix.

Outputs (7)

- Beckmann *et al*., 2008

Requirement (8) ***"Develop a pilot database for annotation of nominal mass ESI-MS data"***

Achievements (8)

- A comprehensive review was made of the problems associated with annotation of ESI-MS data. (Section 6.3)
- Collaboration between researchers at all 3 sites allowed the design of a database housing information on potato tuber metabolite characteristics: Aberystwyth Repository of Metabolite Characteristics (ARMeC).
- In addition to standard physical information derived from other comprehensive databases such as KEGG, the pilot database had links to associated literature generated in *Requirement* 3 and used actual GC-MS metabolomics data from *Requirement* 5.
- Actual measured ESI-MS spectra were developed for a range of metabolites when pure standard chemicals were available
- Query and output tools were developed to interpret ESI-MS fingerprinting data.
- It was demonstrated that with suitable sample replication (minimum of 12-16 samples) a correlation analysis of 'explanatory' nominal mass *m/z* could link signal clusters to individual metabolites and metabolic pathways already known to be present in potato.

Further GC-MS analysis showed that many of these annotations were in fact correct (Section 6.5).

- Lower intensity 'orphan' signals often dominated in many fingerprints and generally showed poor correlation which meant it was difficult to narrow down annotation possibilities.
- ARMeC is the first database of its kind designed:
    - to store mass spectrometry spectra of chemical standards ionized in a matrix containing salt levels similar to that of the sample to be extracted
    - to provide MS/MS fragmentation data for future identification confidence
    - to take into account the fact that many ESI-MS ionization products are salt adducts or neural loss fragments and allow direct searching for both species
    - to use an 'intelligent spreadsheet' to link together potentially related ionization products on the basis of atomic mass
    - to allow automated *m/z* signal annotation in the R-computing environment, providing scope for future automation of data mining
    - A detailed SOP explaining how to use the database was developed which is accompanied by a comprehensive illustration of anticipated results (Section 6.5)
    - The ARMeC signal annotation strategy formed the basis of an invited contribution to *Nature Protocols* (Overy *et al*., 2008).

Conclusions/recommendations (8)

- ARMeC was constructed manually by individually accessing data fields from other databases such as KEGG for physical information and re-drawing structure in a common format. Users reported a small but significant frequency of human-derived errors (particularly in molecular formulae and occasionally structure). It is recommended that in the future alternative automated methods be used to generate information from Mol. Files archived in the large databases.
- Experience gained from both construction and testing of the ARMeC database provided a clear steer that *m/z* signal annotation databases in the future **must** take into account the diversity of ionization products generated in LC-MS experiments.
- A significant number of ESI-MS signals in certain mass bins remained unequivocally non-annotated either because many metabolites share the same nominal mass, or because there remain a large number of metabolites still to be identified in potato tubers that do not have entries in ARMeC.
- There is an urgent need to make species-specific databases more comprehensive.
- Annotation optimisation will be possible on a much wider range of *m/z* with alternative analytical approaches using accurate mass information (e.g. FT-ICR-MS or more recently Orbitrap mass analysers).

Outputs (8)

- Overy *et al*., 2008

Requirement (9) ***"Develop standardized univariate and non-supervised multivariate methods for describing & comparing metabolite composition in G02 GC-MS data"***

Achievements (9)

- A review was made of literature discussing the problems of describing the overall metabolite content of raw food materials based on metabolite data (e.g. units used, missing values, replication). (Section 7.3)
- An assessment was made of freely available software packages that could be used for univariate analysis of the GC-MS profile data derived from the G02 programme. All worked with few problems.
- Missing data distributions were evaluated and the effect of various data infilling strategies on peak intensity distributions was examined. It was shown that simple imputation methods provided easy methods to replace zeros values and the improvement in modelling characteristics was generally significant (Section 7.5)
- A range of approaches were tested to describe signal behaviour in GC-MS data and to determine the best-fit distribution pattern. This analysis demonstrated clearly that most metabolites had a near-normal distribution (Section 7.4.4)
- An analysis was made of the effects of data transformation on univariate behaviour and log standardisation in most cases improved the normality of signal distributions. (Section 7.4.1)
- Peak table feature content, intensity distributions and effects of experimental factors on non-supervised multivariate behaviour were examined. (Section 7.4.3)

Conclusions/recommendations (9)

- There are many excellent standard packages for univariate analysis of metabolite data and there is no requirement to develop further software.
- Many of the univariate approaches to describe metabolite composition of potatoes described previously in the report from the G02006 project are still valid, including ANOVA and various types of metabolite concentration range analyses.
- The intrinsic dependencies in all three data sets hinder their deciphering and conceal real sources of variation. In fact since the G02 programme reported the standard of data produced either at SCRI, Golm or Aberystwyth University has been incrementally improved by the production of standard operating procedures, such as those presented in the series of *Nature Protocol* articles produced during the G03012 project. Thus, experimental factors should no longer be major sources of unwanted variation.
- The finding of a statistical distribution that fits individual metabolite distribution in all three data sets is a useful tool since it can detect whether a new sample is within the expected range in a new sample or not. This was the suggestion reached previously in the G02006 project.

- PCA (PCO) plots are a useful tool to analyse all compounds simultaneously and visualise overall sample differences/similarities. New individual samples can be added to the plot to see where, in relation to overall variability, they will be placed.
- PCA (PCO) plots can also be used to understand partitioning of overall variation by colouring sample points according to levels of each known experimental factor (meta-data). These conclusions confirmed observations reported previously in the G02006 project.

Outputs (9)

- Almost all of the observations made here using univariate and non-supervised multivariate data analysis methods validated the approaches previously suggested to compare the composition of food raw material in the output from the G02006 project and the G02 programme. Thus there was no further opportunity to publish further papers.
- With the validation of these methods using several data sets from different laboratories several of the routines (log transformation, missing value imputation, ANOVA, and PCA) have been incorporated into the standardized data pre-processing and analysis software package (FIEMSpro) developed in the G03012 project (see Requirement (10))

Requirement (10) ***"Develop standardized multivariate methods for comparing metabolite composition of food raw materials"***

Achievements (10)

- The literature on multivariate data analysis and particularly the tools that have been used since the G02006 project reported were reviewed. Special emphasis was placed on seeking recent literature concerned with comparing the overall metabolite content of raw food materials based on metabolomics fingerprint and profile data. (Section 8.3)
- The literature review compared various visions of a data processing and analysis 'pipeline' and a series of common steps were identified which are used to treat the majority of metabolomics data sets. These routines encompassed both tools used for raw data pre-processing (e.g. data integrity detection [batch problems, false positives/binning issues, base line correction], data normalisation, outlier detection) and data mining tools for sample classification and feature selection.
- The software routines reported on previously in the G02006 project were supported by a combination of commercial statistical packages, web-accessible freeware and in-house algorithms written in several different computer languages (e.g. C++, Fortran, 'R'). A key achievement of the G03012 project has been to re-write all the software code for a single platform ('R') which is designed specifically to support statistical tools. (Section 8.4)
- The data mining tools were implemented sucessfully as a single web-accessible workflow (FIEMSpro) which, after parameterisation to deal with the structure of the data table, was able to process and analyse **automatically** the G02 GC-MS and ESI-MS potato tuber

data. FIEMSpro is the first free validated software package in the R environment dedicated for automated analysis of metabolomics data mining.
- After its first implementation FIEMSpro was validated by biologists and analytical chemists who had little experience in data mining. Tutorials were written to describe how to use the software which are accessible at the following URL (http://users.aber.ac.uk/jhd).

Conclusions/recommendations (10)

- All of the selected data mining tools approved for analysing potato tuber data at the end of the G02006 project proved to be robust as part of automated workflows when tested by a range of operators with different data sets.
- A **free** automated workflow for data analysis in a single software package provides a major advance in the standardisation of data treatment and data analysis in experiments using metabolomics approaches.
- The R environment is increasingly the platform of choice for statistical computation. Although the use of FIEMSpro does demand that the operator lean how to use command line programming, the code that needs to be modified is fairly simple. The experience with novice data miners (i.e. biologists and analytical chemists) was very positive and most individuals were competent within a couple of days.
- It was concluded that data sets with any unusual confounding experimental factors are rapidly detected using the standard automated FIEMSpro workflow. If unusual outlier distrubutions or unexpected partitions in the data are visualised at the stage of PCA then the investigator can seek advice from a professional statistician or machine learning expert.

Outputs (10)

- The strategy for automated metabolomics data analysis (using ESI-MS data as an example) was published as an invited *Nature Protocol* (Enot et al., 2008)
- The FIEMSpro software and tutorials are now freely-available on an Aberystwyth University web site (http://users.aber.ac.uk/jhd) and are updated regularly as any reported bugs are resolved or new routines developed.
- Enot et al., 2008

Requirement (11) “*Validate statistical tests and supervised data modelling output metrics used to describe differences in chemical composition of biological materials*”

Achievements (11)

- All of the data sets selected for analysis from the G02 programme compared extracts of stored potato tubers which represent a physiologically ‘quiet’ sample type with a simple organ structure and little effect from environmental variability. Thus all data pre-processing and data analysis routines were also tested for robustness using a range of other data sets including *Arabidopsis* leaf material which display photoperiodicity and shading effects and pathogen (rice blast) challenged *Brachypodium distachyon* (a model grass species) which displays dynamic temporal changes in metabolome as disease progresses. These data, particularly the latter, proved much more challenging and so material from a rice blast infected *B. distachyon* leaf tissue experiment was chosen as an example data set to test automated data mining software.

- As one desired outcome was to pilot methods for comparison of large numbers of similar genotypes, experiments were also carried out using the existing flow infusion electrospray (FIE-MS) fingerprint data representing the chemical composition of 5 non-transgenic potato varieties from the G02 programme and new GC-MS profile data representing the same cultivars generated during G03012. These potato data provided the opportunity to validate the conclusions of the FIE-MS fingerprint modelling by examining the behaviour of validated metabolite peaks in the GC-MS data. Further 'in house' data generated on the FSA G02 project describing the chemical composition of a range of *Arabidopsis* genotypes provided a larger genotype collection with which to evaluate the reproducibility of the methodology.
- Multivariate approaches to compare metabolite composition of potatoes is described in the previous G02006 report and a preferred selection of data analysis methods and proposed modeling output 'similarity' measures had already been made. In the present G03012 project these output measures were validated using a much wider range of data representing typical experiments (described above).
- For sample classification/discrimination it was shown that several supervised multivariate methods for data analysis, including Linear Discriminant Analysis (LDA), Random Forest Decision Trees (RF), and Support Vector Machine (SVM), performed efficiently with a wide range of data sets. (Enot and Draper 2007)
- A baseline for a 'significant difference' was established by establishing the maximum average values when a least two near isogenic cultivars/varieties grown under similar environmental conditions are compared (Section 9.3.2)
- Model validation by boostrapping, feature class label permutation or cross-validation revealed that classification accuracy alone was generally not a sufficient measure of food raw material compositional similarity, unless a very high number of sample replicates (e.g. > 36) were available. (Section 9.4)
- Several classification model 'similarity' measures (*e.g.* 'margins') and 'distance' measures (*e.g.* 'Eigen values') proved robust in a range of circumstances. (Section 9.2)
- Specifically these similarity measures could be used to calibrate the minimum number of sample replicates required to develop an optimal model (Section 9.4)
- Considerable emphasis also was placed on evaluating supervised data mining procdures for their ability to select data features that are explanatory of differences between food raw materials classes. We specifially chose to analyse *Arabidopsis* mutants in which the biochemical differences between mutants and wild type lines was well established. Data features were ranked for their significance in the discrimination of mutant lines from wild type material.The ranked lists were then examined to determine when metabolite signals unrelated to the biochemical lesion began to populate the lists, thus esablishing a potential threshold for significance (Enot *et al.*, 2006).
- The assessment of AUC values from LDA, RF and SVM tests were all useful and a score of > 0.9 usually was a reasonable level for significance cut off. However, it was demonstrated that Random Forest ranking by feature Importance Score was by far the

most stringent test for significance and a threshold score of 0.03 was suitable in most analyses (Section 9.5).

- The evaluation of model metrics for comparison of the composition of food raw materials were the subject of 2 invited papers for a special edition of the journal *Metabolomics*. (Enot *et al.*, 2007)(Enot and Draper 2007)
- With regards to LC-MS fingerpinting data (see Requirement (8)) we were able to develop methods to interpret the ranked list of *m/z* signals to provide a rapid assessment of which types of metbolite species were potentially responsible for the compositional differences between food raw materials. These potential annotations were confirmed by targeted GC-MS analysis of the same sample extracts (Overy *et al.*, 2008)
- Methods were piloted that in the future might allow the large scale comparison of the chemical composition of potato cultivars based on ranking of explanatory ESI-MS features derived from multivariate classification. These methods were based on the assumption that a 'mastermix' population of randomly combined extracts representative of the species genetic diversity could provide a common comparator that was equally similar to all varieties (Beckmann *et al.*, 2007) (Section 9.5).
- Examples of the use of the chosen data mining tools in different steps in the evaluation of compositional similarity between plant materials have been generated and illustrated with example results (see Outputs (11)).

Conclusions/recommendations (11)

- A range of data analysis methods have been used in the literature to assess the compositional similarity of food raw materials and the majority do have value if used correctly for the intended purpose. It is easy to be influenced by the output of discrimination tests when examining visual outputs such as grouping in scores plots from LDA or PLS-DA where classes can appear well separated even if differences between them are effectively non-existent. Such visualizations need to be accompanied by robust, cross-validated mathematical measures of similarity.
- Compositional difference measures based on simple classification accuracies should be treated with caution unless models are thoroughly validated using previously unseen data. It was concluded that food raw materials should only be considered compositionally similar following metabolome analysis if similarity measures fell below specific thresholds of significance in at least 3 statistical tests.
- Defining how similar food raw materials are relative to each other is dependent on how much natural variability exists for any particular crop. In many tests margin values < 0.2; Eigen values in first DF <2 and AUC values < 0.8 generally represent an adequate threshold for significant differences.
- Feature selection is a major problem and it is recommended that data mining methods that retain all variables in final models and which transparently rank each variable individually are chosen. Random Forest modelling was demonstrated to be the most robust in nearly all experiments and would be a method of first choice.
- Feature selection using data mining methods which incorporate significant dimensionality reduction routines, such as PCA or PLS, can present problems if variable intensities do not show a normal distribution in the full data set (e.g. it is not uncommon that some classes contain large numbers of missing values for certain less intense peaks/*m/z* signals). Thus some features can be selected by an apparent high rank in PC1 loading plots or PLS-

DA contribution scores which are clearly not significant to the biological problem but which have skewed intensity distributions.

- A general rule is that features should only be designated as significant if they are highlighted by at least 3 independent algorithms.
- It is recommended that the output of supervised data mining methods such as AUC values or RF Importance Scores should also be confirmed using significance measures from univariate tests such as ANOVA F-values and Welch T-test *p*-values.
- *P*-values should always be corrected for false discovery rate when multiple biological classes are compared in larger experiments.

Outputs (11)

- These experiments validated the FIEMS-pro software and provide examples of anticipated results:
    - Enot *et al.*, 2006
    - Enot *et al.*, 2007
    - Beckmann *et al.*, 2007
    - Enot and Draper 2007

Publications Generated in Association with the G03012 Project (G03012 PI, Co-PIs and research staff underlined)

Enot, D.P., Beckmann, M., Overy, D. & Draper, J. (2006) Predicting interpretability of metabolome models based on behaviour, putative identity, and biological relevance of explanatory signals. *Proc Natl Acad Sci USA* **103**, 14865-14870.[1, 2]

Beckmann, M., Parker, D., Enot, D.P., Duval, E. & Draper, J. (2008) High throughput, non-targeted metabolite fingerprinting using nominal mass Flow Injection Electrospray Mass Spectrometry**.** *Nature Protocols,* **3,** 486-504.

Enot, D.P. & Draper, J. (2007) Statistical measures for testing substantial equivalence of GM plant genotypes in a multivariate context. *Metabolomics* **3***,* 349-355. [1, 2]

Beckmann, M., Enot, D.P., Overy, D.P. & Draper, J. (2007) Representation, comparison and interpretation of metabolome fingerprint data for total composition analysis and quality trait investigation in potato cultivars. *Journal of Agricultural and Food Chemistry* **55**, 3444-3451.[2]

Enot, D.P., Beckmann, M. & Draper, J. (2007) Detecting a difference - assessing generalisability when modelling metabolome fingerprint data in longer term studies of genetically modified plants. *Metabolomics* **3**, 335-347.[1,2]

Enot, D., Lin, W., Beckmann, M., Parker, D. Overy, D., & Draper, J. (2008) Pre-processing, classification modelling and feature selection using Flow Injection Electrospray Mass Spectrometry (FIE-MS) metabolite fingerprint data. *Nature Protocols,* **3,** 446-470.

Overy, D.P., Enot, D.P., Tailliart, K., Jenkins, H., Parker, D., Beckmann, M. & Draper, J. (2008) Explanatory signal interpretation and metabolite identification strategies for nominal mass FIE-MS metabolite fingerprints. *Nature Protocols,* **3**, 471-485.

Helen Jenkins, Manfred Beckmann, John Draper and Nigel Hardy (2007) GC-MS Peak labelling under ArMet. In: *Concepts in Plant Metabolomics,* Nikolau, Basil J.; Wurtele, Eve Syrkin (Eds.), Springer, ISBN: 978-1-4020-5607-9.[2]

Rubtsov, D.V., Jenkins, H., Ludwig, C., Easton, J., Viant, M.R., Günther, U., Griffin, J.L. & Hardy, N. (2007) Proposed reporting requirements for the description of NMR-based metabolomics experiments. *Metabolomics*, **3**, 223-229.

Susanna Assunta Sansone, Daniel Schober, Helen J Atherton, Oliver Fiehn, Helen Jenkins, Philippe Rocca-Serra, Denis V Rubtsov, Irena Spasic, Larisa Soldatova, Chris Taylor, Andy Tseng and Mark R Viant. (2007) Metabolomics Standards Initiative – Ontology Working Group – Work in Progress. *Metabolomics*, **3**, 249-256.

Nigel W Hardy and Chris F Taylor (2007[1]) A roadmap for the establishment of standard data exchange structures for metabolomics. *Metabolomics*, **3**, 243-248. [3]

Nigel Hardy and Helen Jenkins (2007) Reporting Standards in Metabolomics, Jens Nielsen and Michael C Jewett (Eds), In: Topics in Current Genetics (Series: Volume 18), Springer, pp 53-73. [3]

David Parker, Manfred Beckmann, David P Enot, David P Overy, Zaira Caracuel Rios, Martin Gilbert, Nicholas Talbot and John Draper (2008) Rice blast infection of *Brachypodium distachyon* as a model system to study dynamic host/pathogen interactions. *Nature Protocols* **3**, 435-445[4]

[1] Joint output with Met-RO programme

[2] Generated under G02006 and published since

[3] Joint output with EU METAPHOR programme

[4] Joint output with BBSRC project with G03012 staff providing data analysis

G03012 Final Report - Section 1

Helen Jenkins/Nigel Hardy

Compilation and assessment of metabolomic data (Objective 01)

1.1 BACKGROUND TO EXPERIMENTAL SECTION

An important initial activity within the project was to collate and evaluate metabolomic analytical data from selected G02 programme contractors. During discussions with the FSA to merge two earlier proposals drafted in response to the original tenders it was agreed that the data sets were to concern only potato (and tomato for comparison where needed) and were to provide support for the wider aims of the new project in developing enhanced data models and standardized data processing.

1.2 OVERALL OBJECTIVES

The main objective was to identify appropriate data sets which was helped by provision of final reports for all of the FSA G02 contractors. The objectives for this section of the project can be summarized as follows:

- Read G02 reports and identify projects with metabolomics data relating to potato (and tomato)
- Contact contractors and visit to discuss and record data structure.
- Collate appropriate data in a simple web archive for analysis in the remainder of the project

1.3 DATA FROM THE G02 PROGRAMME (Task 01.01)

Following receipt of the final reports the projects G02001 (potato), G02005 (tomato) and G02006 (potato) were identified as the possible sources. Further investigation and visits to SCRI, RHUL and IFR identified the following potential data sets.

G02001	GC-MS (Aqueous extraction and derivitisation)
	GC-MS (Non-polar extraction and derivitisation)
	LC-MS (SCRI set 1)
	LC-MS (SCRI set2)
	LC-MS (IFR)
	NMR
G02005	GC/MS
G02006	GC/TOF-MS
	GC/Qudrupole-MS
	FIE-MS

This final list was shorter than anticipated when the project was proposed but to be adequate to support the aims of the project.

To establish in more detail the structure of these data sets a questionnaire was prepared. Cooperating sites were offered the opportunity to fill in the form or to deliver equivalent information in another format. The questionnaire is provided in Appendix 1.1. The important aspects to be captured were broadly two. Firstly the extent of procedural meta-data available is important for comparability of data. The enquiry was based around the developing ARMeC structure (see Section 6 of the G03012 report) and emerging metabolomics standards. Questions were expected to elicit general free text responses. Secondly the detailed computer format was sought so that potential conversion problems could be established. The developing consolidated responses are shown in Appendix 1.2. These show that in general there would be no major technical difficulties in transforming data (though it would be a manual task requiring a good understanding of the technologies) and that the metadata could be made to conform well to the emerging standards. This is discussed further in Section 2.4.

Phase 2 of compilation was characterised in the original proposal as “data collection”. An simple and adequate infostructure was developed, as planned, using the Forrest web publishing framework (http://forrest.apache.org/) as part of the project web site developed under project Objective 5. This proved adequate and appropriate for the data collected. It is on-line at http://g03r0002.dcs.aber.ac.uk/ with suitable restricted access.

As the project developed, the wider collection of data was not seen as contributing to the main objectives. In particular, the NMR and LC-MS data sets from IFR were dropped because there were no sets with which to compare them (18 month report to FSA). Data integration (as identified in the project proposal) was also abandoned at the 18 month review since the development of a new ArMeT implementation for a compositional database was seen as unproductive (Section 2.4 below) and new ArMeT implementation to hold only a rather small amount of G02 data would not be useful.

G03012 Final Report - Section 2

Matthieu Vignes/Nigel Hardy/David Overy

Report on a search of published literature and accessible databases describing compositional analysis of crop plants (Task 01.02)

> "Science and engineering, building, agriculture and even cookery depend on accurate weighing and measuring…" - from a sign at Albert and Victoria Museum in London. This sentence contains a first justification for our wish to have an accurate view of the content of crops for further research whatever the final goals are.

2.1 BACKGROUND TO EXPERIMENTAL SECTION

Compositional analysis of food is of outstanding importance and there is huge interest in many quarters regarding the quality of food eaten daily. Consumers want to be given information on the composition of the products they are buying. Information provided on packaging can be in various forms; it can be energy content, proteins/sugar/fat contents or even more detailed information on some vitamins or certain amino acids for example. Raw food material compositional analysis is also a crucial activity for plant breeders and nutritionists in order to give detailed qualitative and quantitative information on raw materials destined for human or animal nutrition. Our interest lies in such detailed description of the raw elements that are found in food materials.

The level of compositional information that we are interested in is the individual metabolite. The term *metabolites* usually depicts small molecules involved in the metabolism of living organism, either as an intermediate or a product of a reaction. In other terms they are the participants of the complete set of chemical reactions that occurs in living cells. They are involved in processes which form the basis of life, allowing cells to grow and reproduce, maintain their structures, and respond to their environment. Catabolic reactions yield energy, an example being the breakdown of food in cellular respiration. Anabolic reactions, on the other hand, use this energy to construct components of cells such as proteins and nucleic acids. The *metabolome* is thus the network of reactions involving metabolites. Products (outputs) from one enzymatic chemical reaction are often reactants (inputs) of another. The analysis of such activity of metabolites is called metabolomics. Namely, it is the study of metabolite profiles in a system (cell, tissue, or organism) under a given set of conditions.

We will distinguish *primary* and *secondary metabolism*. The former one encompasses all the chemical reactions and interactions of a biological system involving those compounds which are formed as a part of the normal anabolic and catabolic processes which result in assimilation, respiration, transport, and differentiation. These processes take place in most -if not all- cells of the organism. Common examples of primary compounds are sugars, amino acids, nucleotides *etc.* A compound is classified as a secondary metabolite if it does not seem to directly function in the processes of growth and development. Although secondary compounds are a normal –in the sense that they appear in the vast majority if not every living organisms- part of the metabolism of an organism, they are often produced in specialized cells, and tend to be more complex than primary compounds. Examples of secondary compounds include antibiotics or plant chemical defenses such as alkaloids and steroids; many such compounds provide plant-derived food products with specific aroma and taste characteristics.

In the G03 project we are mainly interested in the composition of the potato plant (*Solanum tuberosum*) and more particularly in tuber compositional analysis. Potato is the 4th crop worldwide behind wheat, rice and maize. It is used in food, feed and biotechnological applications. The potato plant is relatively easy to modify genetically and its large tuberous storage organs constitute an interesting host for recombinant protein production. We may also consider some information derived from other *Solanaceae* like tomato plants (*Solanum lycopersicum* which has a simpler genome).

Whilst writing the G03012 proposal a cursory search for literature and databases revealed the presence of a large body of information on the composition of raw food materials. Thus in the present section, in order to provide a comparator for metabolomics data we undertook a formal search for data on potato tuber chemical composition with the aim of examining data format and determining the possibility of data integration in the future. Our secondary focus was to report values found either in the literature or in devoted databases for metabolites in potatoes and while doing so to develop a list of metabolites that can be measured in potato. Information arising from other species will only be considered if they have some added value as far as potato is concerned. However, we need to keep in mind that just like for any plant (or more generally living organisms), the chemical composition of potatoes varies with variety, soil type, area where grown, cultural practices, maturity, method of harvesting, storage environment and other factors. Clearly it is impossible to make any meaningful comparisons on the composition of potato tubers unless most of the factors listed above are under control.

2. 2 OVERALL OBJECTIVES

The major aim of this section is to evaluate external data structure and assess the possibility of developing methods for converting this information into a common format. One important activity was to assess how much information actually relates to direct measurements of metabolites in the data available to us at present. A key question was to know how scientists who work within the context of food research describe the raw material in terms of its overall composition. When this is understood we can evaluate the scope for integration of compositional information and, if appropriate, develop a new implementation of the ArMeT database developed in the G02006 project. Ultimately, when this data has been retrieved and converted (if worthwhile) in a unified format, there may be opportunities to enhance/improve the quality of the data.

We will examine as many as possible of such external databases because we would like to retrieve a broad list of quantitative information on metabolites and because we would like this information to be accurate. This would kill two birds with one stone: in the first place, it will allow us to have an idea of the list of metabolites covered by existing databases and lastly we would like to be able to compare the relative concentration in common databases with GC- and LC-MS profiles the G03 project is analysing. Ultimately we intend to determine whether it is possible to propose a database structure to provide homogeneous access to data on potato composition by assessing the scope for unification of all external data on quantitative measurements on individual metabolites and integration with metabolomics data sets.

The objectives can be summarised as follows:

- Develop a summary of accessible current databases and literature with data relating to potato tuber composition.

- Assess the presentation and quality of the data on compositional analysis of raw potato tubers. *It is important to concentrate on raw material but if the information is only available on cooked or processed tubers this should commented upon.*

- Report if data is linked to the original experiments producing them and if there are publications related to them. *Develop a bibliography*

- Evaluate the kinds of standard methods used to identify metabolites as well as the format of given values. *Is there any indication of statistical accuracy given?*

- Gather common information in a unique repository containing representative qualitative information and quantitative aspects on metabolites in potato (and tomato when relevant for example). *The structure of the data included will depend on the framework initially proposed by retrieved sources of information.*

- Develop with analytical chemists a list of potato tuber metabolites that should be potentially detected by metabolomics technology.

2.3 INTRODUCTION

Some sources of information for compositional analysis of crops were previously (sometimes partially) known. The strategy to make the list of them complete and to explore this list to the greatest possible extent started with simple query inspecting available literature and first known databases. Queries were carried out either in general search engines or in more specialized databases mentioning keywords of interest (‘metabol*’, ‘potato’ -or ‘tomato’-, ‘compositional (analysis)’, ‘peak’, ‘measurement’. Names of metabolites of interest were added (e.g. ‘(beta/phenyl)alanine’, ‘fructose’, ‘valine’, ‘(iso)leucine’, ‘citrate’, ‘(oxo)proline’, phosphate) when there was a need to constrain retrieved results. Mining the list of retrieved sites gave various leads. We tried to extract useful information therein and we will comment upon this in the present report.

Generally speaking, it was found that that individual information sources on metabolites were much sparser and poorer than expected. However, the assembly of information provided by the numerous databases and documents found did provide a useful repository overall.

In the present section of the GO3012 project we will present the information we were able to extract and explain how this information will deciphered and assembled into the components of the targeted external compositional database (ARMeC; see Section 2.5 and Section 6 of the G03012 report).

2.3.1 Exploration of “Food Composition Databases”:

According to the website (http://www.nutrition.org.uk/) of the British Nutrition Foundation there are over 150 food composition tables and electronic databases already in existence worldwide. Since the same food might be listed in a great number of these databases, in some circumstances, different nutrient values will be given for the ‘same’ raw material. This can be for a variety of reasons: genuine differences between the nutrient composition of the foods selected for sampling (*e.g.* selenium content of cereal strongly depends upon selenium content of the soil), different sampling techniques might have been used or different technologies (recent vs. modern) or different methods may have been used to analyze the data. This has been a hindrance to unification of such data. However, we try here to report the consistency of the quality and quantity of the data in food composition tables for potato (and possibly tomato or other crops). One would then be able to have an accurate and reliable background for a genuine study on metabolite concentrations in potato tubers for example. It could be a good starting point for other projects as well.

The first site that proposes a quality database is the United States Department of Agriculture (under the Agricultural Research Service). The current actual release is the 19th and it seems to provide an annual update. The web address is http://www.ars.usda.gov/main/site_main.htm?modecode=12-35-45-00. A search on "raw potato" gives the choice between unspecified, red, white or russet potato "flesh and skin" (*Solanum tuberosum*) and with sweet potato either only leaves or unprepared (*Ipomoea batatas*). There is a choice of scale of analysis thereafter: 100g, cup, large, medium or small potato for example. We reported any number given here in percentages. We assume that these are mean values given since number of points (measures certainly) and standard error are reported. In this database major constituents with numerical summaries are given (water, ashes, sugar, energy,…) as well as less information (e.g. presence, absence) on minerals, vitamins, lipids, amino acids (18 of them) and 'others'. Values for the above cited different categories of potato can be quite different. As an example, the values for some amino acids between the basic potato and the red cultivar can double and the beta-carotene level is four times higher in red potato. Data are also associated with varying numbers of replicates (2 to 11). Sometimes the number of data points is 0. We assume that it corresponds to a value inferred from another source than the genuine one considered by the USDA. An example of retrievable information is given in Appendix 2.1. A simplified comparison between some databases is provided in Table 2.1 and as a comparison to the USDA web site an extract of the Finnish web site (FINELI**)** when searched for potato tuber composition is presented in Appendix 2.2.

We had in mind then to find similar database in different similar National Institutions. The French AFSSA (French Food Safety Agency, warning in French!) was quite disappointing with only some pages related to constituents of different food products (*cf.* Ciqual). But there was no real database. It was similar on the Ministère de l'Agriculture et de la Pêche (only in French as well). The only available information is advice on healthy diets!

Agriculture and Agri-Food (Canada) did not provide better results than the USDA database. A Canadian Nutrient File (http://www.hc-sc.gc.ca/fn-an/nutrition/fiche-nutri-data/index_e.html) on Health Canada exists and gives compositional information on food eaten in Canada. It just seems to be equally informative as the American database and measurements are often identical to those in the USDA database. It suggests a direct copy. This observation raises the more general issue that the sources of the data need to be made clearer if it is a possibility that some measures are inferred from others'.

New Zealand organisations (AgResearch, Crop & Food Research,…) are mainly contract based research and development companies. However some of them are state-owned so it was worth having a look. Crop & Food Research produce a New Zealand Food Composition Database (http://www.crop.cri.nz/home/products-services/nutrition/foodcompdata/index.jsp). It is less comprehensive than the USDA database as far as the sample we were able to see is concerned. AgResearch was only able to give a short presentation on their website. Anyway it doesn't seem to be relevant because the nearest theme we were able to determine was enhancement of meat food for livestock.

Food Standards Australia/New Zealand in under its Food Composition Programme (http://www.foodstandards.gov.au/monitoringandsurveillance/foodcompositionprogram/) provides information either on nutrients or foods. A version NUTTAB2006 (http://www.foodstandards.gov.au/monitoringandsurveillance/nuttab2006/) has been released recently and merges, replaces and update several older published items. The sight of it is very encouraging in our scope. They distinguish between cultivars (Coliban, Desiree, Sebago…) that might be common in Australia and/or New-Zealand. A choice of processed or indigenous food is proposed as well. References and experimental procedure are readily commented upon. The only drawback is that values do not seem to be always very accurate (*e.g.* organic acids for potato) and no measurement of error is provided.

The Danish Institute for Food and Veterinary Research offers a good Food Composition Database at http://www.foodcomp.dk/fcdb_default.asp. English is available. Their links section is quite comprehensive towards other National organisations. The database is complementary to the USDA database. A search on potato can return different compositional analysis for new, autumn and old potato. Means values, number (of experiments?) and sometimes a range for nutrients are given. Some other information are reported (ripe or not, imported or from Denmark). However, the format adopted for presentation makes it sometimes not very precise (or unclear to us at least).

In fact there are many such local database supported by national governments. But to agree with our scope, they are not very detailed. At best they give relative contents of a list of thirty to forty metabolites in different food products. Most of them are gathered under the aegis of the FAO (see http://faostat.fao.org/site/291/default.aspx for example). The International Network on Food Data Systems (http://www.fao.org/infoods/index_en.stm) can give a list of tables and databases developed regionally. Moreover it aims at centralizing as much as possible the activity upon food analysis on a worldwide scale. Historically, the first reported composition table is dated 1818. We have 2 main works on Food Composition Tables for international use (1st edition 1949, 2nd 1953) published by the FAO. But their focus is often too broad. It will be at best good background information for a list of common metabolites. It will require some treatment to organize these data to provide a good framework for reasonable ranges of compositional values for metabolites for potato. Many points were impossible to be clarified in term sof the the information presented: assessment of reliability of the databases, ways of identifying cultivars, list of metabolites to be considered for example.

Name	Internet address	Comments	Information given	Kind of numerical values
USDA's National Nutrient Database for Standard Reference	http://www.ars.usda.gov/Main/site_main.htm?modecode=12-35-45-00	Foundation of most food and nutrition databases in the US, used in food policy, research and nutrition monitoring, relatively comprehensive for a general food database with raw and prepared food.	Choice between cv. ('Russet') or type ('White'). Raw or processed food. Sources are given but not always obvious.	Choice for scale. Gives mean values, number of points and SE (when available).
Canadian nutrient file	http://www.hc-sc.gc.ca/fn-an/nutrition/fiche-nutri-data/index_e.html	Access to analysis of various food from raw to included in recipe. Seems to be a copy of the USDA database.	*Cf* USDA but sources are cited more explicitely	*Cf* USDA
FSANZ's NUTTAB2006 (Australia and New-Zealand)	http://www.foodstandards.gov.au/monitoringandsurveillance/nuttab2006/	Paper, online or electronic format proposed. It is an updated version of a database that reflects food available in Australia.	Sources are cited and experimental processes are explained. Choice over a few varieties of potatoes.	Only means (guess) and no sign of a number of independent experiments or SE values.
Danish Food Composition Databank	http://www.foodcomp.dk/fcdb_default.asp	Analysis of food coumpounds, relatively precise and accurate. Sources are quoted.	Choice between season of potato production for raw material. Sources are explicitely given with links.	Mainly mean values are given, neither number of replicates nor SE.
NCC Food and Nutrient Database	http://www.ncc.umn.edu/products/database.html	Database maintained by a Minesota University department, sample available upon request. Many interesting nutrients.	Only able to look at boiled potato with or without skin. The accuracy seems acceptable and the list of analyzed nutrients is given.	Compositional value for main metabolites (~40-50) and other food oriented values (energy…)
Fineli	http://www.fineli.fi/index.php?lang=en	Information about Finnish food composition maintained by the National Public Health Institute of Finland with average nutrient concentration	Processed and raw food (only 1 for potato). Sources could be given but none provided for 'potato unpeeled'.	Guess mean values are given but no further indication than 'content'.
Tabela de Brasileira de Composição de Alimentos	http://www.fcf.usp.br/tabela/	Well, very similar to other national databases. Not many nutrients. But sorry only in Portuguese…So I only looked at "batata" (potato).	Not many fields apart from basic information (energy, total of proteins, lipids, sugar…). A reference is given for each	Certainly mean values; that's all!

			nutrient.	
New Zealand Food Composition Database	http://www.crop.cri.nz/home/products-services/nutrition/foodcompdata/	A database devoted to food prepared and eaten in New Zealand. It is not free and really food-consumption oriented	We were only able to look at the sample page. It seems to be too general public oriented.	Certainly mean values (perhaps only one measure) for the proposed sample size.
NutriBase	http://nutribase.com/homepage.shtml	Commercial software for "nutrition and fitness" mainly proving USDA and Canadian databases	Not able to test it. It computes compositional information for entered recipes with regards USDA and Canadian food databases.	N.A.
Nutrigenie's Nutrition 2005	http://nutrigenie.biz/	Commercial software including analysis of food products.	Similar to the previous one, suited to analyse a diet rather than providing compositional analysis of raw food.	N.A.
INFOODS tables	http://www.fao.org/infoods/directory_en.stm	Report that gathers most of the above (and others) databases (national-regional, reference-user,…) providing details (format, gross content,…) for each.	Links towards many regional databases referenced by the UN's FAO under INFOODS.	Really depends on the selected databases. Can be very sparse (cf. Brazilian database above) or more complete.

Table 2.1: Some databases providing values for metabolite content for a great variety of food.

To sum up, major issues to overcome until now are: the accurate labelling of individual metabolites (see perhaps `http://msi-workgroups.sourceforge.net/` or Greenfield's *Food composition data: production, management and use*), the missing information on many metabolites under study and the lack of references on methods used to derive the data or even on the sample used. From the sources above, it was possible to find information on amino acids. However, the INFOODS proposed Standards do not include the full list of around 100 expected common metabolites decribed in potato tuber (see Section 3 of the G03012 report) given by our data (e.g. malate). Nevertheless, it is more complete than any results we were able to retrieve (e.g. it had values on inositol!)

There are several relevant journals which often contain articles on the subject of food composition. We can quote: *Nutrition Journal* (published by BioMed Central), *Journal of Food Composition and Analysis* (official publication of INFOODS) and *Trends in Food Science and Technology* (publication of the European Federation of Food Science and Technology –EFFoST- and the International Union of Food Science and Technology - IUFoST), *Plant Science* (edited by Elsevier), *Phytochemistry* (official journal of the Phytochemical Society of Europe and the Phytochemical Society of North America). From such journals we produced an extensive list of retrieved articles which are presented in the bibliographic file which is to be found after the list of general references for Sections 1 and 2.

In addition to journal articles there were conferences reports, often in book formats: US National Nutrient Databank Conference or International Food Data Conference for example. A useful book collecting many methods on proper food data handling is *Food composition data: production, management and use*, H. Greenfield and D.A.T. Southgate, 2nd edition, FAO Rome 2003. More generally, literature proposed on the INFOODS webpages seems relevant. But we will tackle this kind of information later in the report.

2.3.2 Databases devoted to Potato

Although it was a straightforward idea to look directly for databases giving metabolite contents of potato, it is only briefly presented here. The reason is quite simple: there aren't many of them and they do not provide what we expect in terms of quantitative data. They are summarized in the following table:

Name	Internet address	Comments
European Cultivated Potato Database	http://www.europotato.org/	Announced as a resource for potato breeders, scientists and farmers. It lists a list of cultivars and their physical features (aspect, size, environment grown…). The advanced search procedure allow to search for (rare) pieces of information on metabolites (*e.g.* tuber glycoalkaloids but available only for 2 varieties among ~3000?!). Note that many indicators are not numerical values.
British Potato Council	http://www.potato.org.uk/	The aim of this site is to promote the Potato industry in Britain. It provides information on weather condition, market prices, economical trends, diseases…
World Potato Congress	http://www.potatocongress.org/	Gathers diverse kinds of information for people interest in potato: events, forums, links…
Centro Internacional de la Papa	http://www.cipotato.org/	Scientific research and related activities on potato, sweet potato and other root and tuber crops aimed at reducing poverty and achieve food security in developing countries and to improve management of natural resources.
Ikisan portal, potato part	http://www.ikisan.com/cache/ap_Potato.shtml	Many different information on the history, the morphology or the farming aspects of growing potatoes.
GMO Safety / potato	http://www.gmo-safety.eu/en/potato/	Information on GM potatoes and the impact/safety of modifications on potato analysis made in Germany. It has 3 sections: new compounds, starch content and resistance to diseases.

Table 2.2: Some potato devoted websites

It is appropriate to mention here that some “old” books devoted to potato give a relatively accurate estimation of tuber composition. The advantage is obvious: experiments and material are quite well explained. The major difficulty is to know how to retrieve information from these old texts (no electronic form exists). There is a question to know if the reported experiments are precise enough according to our aims. We can quote: *The Potato Crop,* 2nd edition (1992) edited by Paul Harris, Grazyna Lisinska and Waclaw Leszczynski’s *Potato Science and Technology* (1989) and *Potatoes: Production, storing and Processing*, 2nd edition (1979) by Ora Smith as useful reference books for example. One exception is the Ikisan website that provides a good background on potato (and more generally on several crops or farming issues). Ikisan is an agricultural portal providing public information on the basis of exchanging knowledge. This data might be worth including in our external database for compositional data. Its drawback is it lacks scientific background with many values cited without references.

To finish the present document, we report below the last clues we have been following up before giving an overview of the work that has been done and achievable targets we could set ourselves according to the available external data on compositional analysis on potato.

2.3.3 Further sources of information relating to potato tuber composition

A simple concept is to look for information specific to a targeted metabolite. For example Ellin Doyle gives a report of free Asparagine level in foods (available online at http://www.wisc.edu/fri/briefs/asparagine1102.pdf). It has a full section on Potato among other foods, mainly devoted to raw material. It gives references to sources and some comments on the conditions that lead to numerical values. The content of such a report is very rich. The major drawback is the time it would consume to check that the original sources still exist, to retrieve them and to incorporate them into any integrated database.

Another direction would be to explore databases on so called metabolomics data. As an example, the Golm Metabolome Database gathers information on profiles or Mass Spectra libraries (*e.g.* Retention Index). It has a public access policy. Information therein may be very relevant. References to (metabolite profiling) experiments are made very clear. However, the relevance of such semi-quantitative data and its linkage and interpretation related to our study on quantitative measures on food chemical components are not obvious yet. Data are available on several species: *Arabidopsis thaliana, Arabidopsis halleri, Cucurbita maxima, Cucurbita pepo, Kigelia Africana*, *Lotus japonicus*, *Markhamia acuminata*, *Medicago truncatula, Newbouldia laevis, Nicotiana tabacum, Ovis aries*, *Spathodea campanulata,* and include potato (*Solanum tuberosum*). Sometimes, there is a difference made upon the sampled organ (leaf, petiole, root, seedling…). The problem here would be to relate measurements (e.g. relative ratios) to quantitative compositional values. In essence the main value of such web sites is the fact that they provide libraries of profiles and retention times specific to compounds or samples; this kind of ‘check list’ data is very similar to the GO2 data we will be analysing in the G03012 project. Other organizations provide different forms of data. Table 2.3 lists some examples of repository for metabolomics data:

Name	Internet address	Comments
Max Plank Institute's Metabolite Mass Spectra Library	http://www-en.mpimp-golm.mpg.de/02-instUeberInstitut/04-instRessources/webbasedRsrc/metaboliteMSL/index.html	Library of available metabolites for sample analysis.
The Golm Metabolome Database	http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html	Public access to custom mass spectra libraries and metabolite profiling experiments.
Spectral Database for Organic Compounds (SDBS)	http://www.aist.go.jp/RIODB/SDBS/cgi-bin/cre_index.cgi?lang=eng	Spectral database system for organic compounds, which includes EI-MS, FT-IR, ^{1}H NMR, ^{13}C NMR, a laser Raman, and ESR spectrum.
Fiehn lab metabolite listing	http://fiehnlab.ucdavis.edu/compounds	List of identified metabolites with chemical/physical background, synonyms list and links towards KEGG
KNApSAck	http://prime.psc.riken.jp/?action=metabolites_index	Tool for combined analysis of (natural) metabolites and organisms. The entry can also be made according to molecular weight or chemical formula. It focuses on plants. References are provided for the metabolite-organism association.
ToMet	http://tomet.bti.cornell.edu/	List of metabolites for tomato with links towards KEGG and references to experiments leading to their identification within the Tomato Nutrient Project.
Life Sciences Institute's Crop Composition Database	http://www.cropcomposition.org/	Compilation of crop analyses from a number of companies engaged in agricultural life sciences and standardized through ISLI. Only on corn, cotton and soybeans
Human Metabolome Database	http://www.hmdb.ca/	Freely available database containing information on metabolites found in the human body. At the moment, there are approximately 2500 metabolites.

Table 2.3: Short list of metabolite-oriented databases.

The compound database of the Fiehn laboratory gives a list of identified metabolites that can be used for further experimental identification. No compositional analysis is given, only links towards KEGG database and synonyms list in addition to physical/chemical properties. It is not really a database, rather a list. The laboratory also proposes a mass spectrometric database (BinBase) on its webpage.

KNApSAck gives lots of indications on either metabolites or specific organisms. It provides links between these 2 features as well. It doesn't give other numerical values on metabolites than molecular weight however. References are provided on the context of the metabolite.

The Human Metabolome Database provides much of what we are looking for as an external database but is only linked to human measurements. A description of each metabolite is provided with a list of known synonyms. Many links towards relevant databases (ByoCyc – dataset metaCyc-, GeneAtlas, KEGG, Pfam, PubChem, PubMed, SwissProt, UniProt…) are proposed to find information on corresponding spectra (experimental or predicted, ^{1}H, ^{13}C NMR) relevant sequences, structures (*cf.* proteins) and even corresponding enzymes. Normal ranges of values within several tissues are provided. Last but not last, it gives a list of some references relative to all the information for a metabolite.

In the same way, an interesting alternative database is the one developed at Cornell University as part of the Tomato Nutrient Project by Zhangjun Fei: ToMet. The goal is to correlate metabolite changes with expression data for a specific sort of experiments (backcross recombinant inbred lines). A list of metabolites is given (nearly 60) and stored in different categories. Obviously this database gives information for tomato. Few of the metabolites are shared by SCRI's initial data on potato. The reason is very simple: genuine and interesting tomato metabolites are rarely the same as potato ones. The 'Tools' section allows searching for values of a metabolite according to a paper (whose values are present in the database). An interesting distinction is that no value is given, the data is missing and if a "0" is reported, it means that the value is below the detection level. See Appendix 2.4 for an example. Other features are available such as multiple search of metabolites and correlated search with expression data values. The structure of the database is worth having a look at.

We may also quote the Life Sciences Institute's Crop Composition Database, but it contains only data on corn, cotton or soybeans.

Another possibility is to do some *literature mining* in the metabolomics research fields. Relevant articles were sought within Google and its Scholar and Books versions, Pubmed and ISI Web of Knowledge portal with links to the three tools used: Web of Science, BIOSIS and CAB Abstract (authentification is required for the these latter browsers). From these complementary search engines, we followed some interesting authors to check their lab web pages. The literature references are reported in the Bibliography at the end of Section 2, whilst Appendix 2.10 contains a spreadsheet of the retrieved data.

As an example of an article reporting the normal range of values for important nutritional (and anti-nutritional) factors in tubers we could suggest. "Compositional analysis of tubers from insect and virus resistant potato plants", Rogan et al. (*J. Agric. Food Chem.* 2000); this paper also provides further links to different values from other sources. It gives its own measured values as well. An example of such information is given at Appendix 2.3.

An idea of values to expect for important metabolites can be found in OECD's "Consensus document on compositional considerations for new varieties of Potatoes: key food and feed nutrients, anti-nutrients and antioxidants" (http://www.olis.oecd.org/olis/2002doc.nsf/LinkTo/env-jm-mono(2002)5).
Just remind here we quoted general books on potato that have compositional analysis in their pages at the end of the "devoted to potato" section of the present report. One of the first articles on the use of GC-MS for simultaneous metabolite analysis, Roessner et al.'s "Simultaneous analysis of metabolites in potato tuber by gas chromatography–mass spectrometry" (*The Plant Journal* 2000) derives quantitative determination of metabolite concentrations in developing potato tubers.

A standard example of papers reporting genetic engineering is Diretto et al.'s "Metabolic engineering of potato tuber carotenoids through tuber-specific silencing of lycopene epsilon cyclase" (*BMC Plant Biology* 2006). They are useful in that they contain normal range of wild type plants metabolites for which they want to "enhance" the content of the plant. Many comparable articles have been published recently. Such documents are highly valuable because they give the full protocol to achieve their measurements. Some online databases (many of them not quoted in the present report) don't…We also noticed that the unit to report compositional analysis (*e.g.* nmol/g or µmol/g) can vary. It might depend on the analytical method chosen.

We may also have in mind to look at metabolic networks available. We can quote the following examples:

Name	Internet address	Comments
KEGG	http://www.genome.jp/kegg/pathway.html	Pathway maps representing our knowledge on the molecular interaction and reaction networks for metabolism, genetic interaction, diseases…
MetaCyc	http://metacyc.org/	Database for experimentally elucidated pathways with links to genes, proteins and metabolites.
Drastic	http://www.drastic.org.uk/	A resource for the analysis of signal transduction in plants. The chart on metabolic pathways of the diseased potato is very informative.

Table 2.4: Online resources for information on metabolites in biological networks

Basically, these databases give the network of reactions involving metabolites as compounds (reagents or products). Sometimes the information is not explicitly made available for potato (KEGG, MetaCyc). They are useful because they prove that if a reaction has been observed in an organism, the involved compounds have been at last temporarily present in it. The main issues here are to know whether a compound has really been observed in the organism (in which organ then) and whether it is a stable compound (and not just an intermediary compounds in the network that is produced and then immediately consumed by another reaction).

2.3.4 Scope for developing a common format for potato content

We have now reviewed many different sources of data for external metabolite content of potato. An original goal of the G03012 project ideally was to produce common standard concentration ranges for metabolites in potato tubers derived from the external data; the question is how to make the best use of these different sources of information?   Given the structure of the data we propose two different measures for each metabolite. The first one will be an *incidence* measure and the second one will provide an *intensity* measurement for those metabolites that show a significant incidence. Let us clarify this piece of information. The incidence was defined as the frequency of related occurrence of the metabolite under study in our references pool as compared to the number of times it was searched for. So two values will be necessary to estimate this frequency: “B”, the number of references that admit they would like to detect the metabolite and “A”, the number of references reporting that they eventually detected it. The incidence will simply be defined as I=A/B. The higher the incidence, the most probable the metabolite is present within potato. However the correct estimation of “I” will face two major issues: the difficulty to accurately evaluate “A” and of course the honesty of the reference to admit it was unsuccessful at detecting the metabolite (especially if they think it should be present). The problem to correctly assess “A” and “B” will rely on the ability of the person to detect this information within a database documentation (when available), an article, a book…and to judge if a reference can be fairly evaluated as reliable. In general we expect these measures to be underestimated. In the case where the incidence test has proven the presence of the metabolite in potato, we would like a numerical estimation of it. This is the subject of the next paragraph.

Ideally we would like to retrieve a distribution of metabolite content within potato tubers. However it soon became apparent that this was simply not achievable, the first reason being that this distribution strongly depends on the cultivar. For example, the dry matter can vary from 13.1 to 36.8% (Lisinska 1989)! Even if this factor is set with certitude, com position measurements can depend on many other factors. As an example, the size of the tuber has a major impact on sugar (glucose, fructose and sucrose) content. Table 2.5 from Linsinska 1989 reports that means values for the three above-cited sugars are 0.164, 0.027 and 0.341% of Fresh Weight for a large tuber and 0.412, 0.129 and 0.171% of FW for small ones?! There is also the issue of homogenization of adopted unit. Some references give values on Dry Weight, whereas other do on Fresh Weight. Some report the metabolite abundance in mol/g while others chose to report it as mg/g…Hence prior careful data handling is necessary.

Given the complexity of the compositional information we decided to try initially to retrieve values for as many cultivars as possible and record factors that hinder data integration. From this information we attempted to derive also a simplified measure of numerical value and at the same time comment on these values where appropriate. Essentially, were possible we would have liked to generate metabolite concentration values that represented a quintile distribution (possibly just the median, minimum and maximum) over the biggest set of values that seem reasonably comparable. This is what we called an intensity measurement for a given metabolite in potato. A simpler solution would be to point at references that we identified as having valuable quantitative information for the queried metabolite(s) and to propose to a user a small well chosen set of representative values for potato tubers content as regards these metabolites.

2.3.5 Concluding remarks on external data

We have reported and briefly described here all the sources of information for compositional analysis of metabolites within potatoes that we were able to find and we think are relevant to the FSA G03012 project. Merging values retrieved from the literature or available online could give birth to a database comparable to the HMDB in its content and form but devoted to potato. However after a detailed examination of all the information on potato metabolites gleaned from external sources it became clear that developing a simple method for transforming concentration measures into a unified format was more or less impossible without consulting in detail the original papers and in many cases corresponding with original authors. Thus populating an extended implementation of ArMet with this data was not seen as a productive exercise (see section 2.4). As an alter native we established a simple scoring system to provide the best relative estimates of metabolite incidences and intensity measures as follows:

0 not present
1 not detectable
2 identified,(very) small amount
3 present in stressed conditions
4 present in GM plants
5 traces
6 variable amount (depending on storage, bio availability,...) but important component
7 important component

All the data collected from external sources is summarised the spreadsheets attached as Appendix 2.10. Instead of trying to get this information into ArMet our attempt to collate external information and relate also to metabolomics data culminated in the development of a pilot database to contain information on individual metabolites (ARMeC) for the G03012 project described in Section 6 of the present report. Specifically the software engineering for ARMeC based on the analyses of external compositional information from this section is presented in Section 2.5. Further considerations related to evaluating the scope for extension of the ArMet database to contain whole experiments derived from external literature describing potato tuber composition are presented in the next section.

2.4 DEVELOPMENT OF AN ArMet IMPLEMENTATION TO EXPERIMENTAL DATA ON POTATO TUBER COMPOSITION FROM EXTERNAL DATA

The ArMet framework for metabolomics data was developed under project G02006. It was the starting point for collection of G02 data from other projects and for the representation of external compositional data. Between the end of G02006 and the start of G03R012 and continuing under the latter project, the ArMet design was modified in the light of feedback from publication[1], standards initiatives [2,3], additional publications[4] and input from other fields[5]. Development also took place under the BBSRC MeT-RO project. Input from emerging G03 requirements was important here.

These developments resulted in a version 2 of ArMet which is described in detail in Appendix 2.6. The major conceptual design improvement is to ensure a cleaner division between materials and protocols applied to the material. Earlier designs had moved in this direction but a number of anomalies still existed. This change is broadly in line with the FuGE design[5] where, in that high level and "generic" materials and protocols to describe actions are the only concepts for the main backbone of the data structure. ArMet retains specializations for particular types of material and protocol to ensure completeness and coherence of the metadata for the metabolomics pipeline. A major example of this is chemical analysis where this can be modeled as a protocol applied to transfer biological samples into data. This significantly simplifies the data structure but clearly it is a special type of protocol (a Chemical Analysis protocol) which must be applied and that cannot be applied elsewhere.

Alongside the greater emphasis on protocols comes enhancement of protocol representation. In particular standard protocols and data on the application of a protocol (at a time and place by an experimentalist) must be distinguished and implemented. Protocol representation is more sophisticated in ArMet 2 (Appendix 2.7, Section 2.2 and diagram).

A major emerging requirement was to represent a 'set' of material. In early versions of ArMet and in particular in G02006, lack of such structures was noted. Ways of identifying all samples from a planting or all samples put through a machine were of great practical value. In ArMet 2 any type of material can be grouped into a set. These sets are called "studies". Care was taken to ensure that a mapping from ArMet 1 to ArMet 2 was possible (Appendix 2.6, Section 5). This was to ensure both compatibility of data sets and as a demonstration that all concepts representable in version 1 could be represented in version 2.

Appendix 2.7 provides a complete draft design for a G03 database based on ArMet version 2. This is informed by the data collection exercises (Sections 1 and 2.3). Full implementation and population of this design was not seen as profitable for the project (18 month report to FSA) and was abandoned.

Prospects for a G03 implementation from a technical perspective were good. The data harvested from external sources (Section 2.3) were structurally suitable for representation. The major problem was one of completeness. External sources would not always provide full "metabolomcs reporting" details. This would require two types of modification in an implementation. First, ArMet is prescriptive. Almost all "fields" are mandatory. This is an easy restriction to remove and a database with the potential to store complete metadata but the permission to accept incomplete sets would be possible. The second change would be to allow dummy entries in structurally necessary fields where data are absent. The metabolomics pipline is detailed in sample growth and preparation. Food reports often lack such detail and dummy placeholders would be necessary. These two changes would alter interpretation of queries from the database. Searching for value "x" does not return record with values other than "x" but in addition it does not return those which might have been "x". Missing values clearly require careful treatment in analyses.

Turning from data structures to data values two areas of difficulty should be highlighted. First, measurement of metabolite content in food is reported as absolute or relative, in a wide range of units and to varying accuracies. Harvesting from publications and databases for comparison would require that these be reduced to common units. That is not an entirely automatable process. ArMet permits declaration of units as well as values but, particularly for literature references clear understanding of what is being reported is required to ensure true comparability. This would require time consuming curation by suitable experts and the project did not have that resource or see significant benefit from the exercise.

The second issue in data values is chemical species description. Food composition tables often report at a rather abstract level ("sugars, total"; "starch, total" – see Section 2.3 above). They also often report using common or ambiguous names for chemicals. Both of these issues can be tackled using an ArMet type structure. Section 6 of the G03012 report discusses issues of chemical identity, synonyms and of chemical ontologies. Thus, as has been investigated within ARMeC, unambiguous identification can be achieved. The cost however would again be careful and expert curation. Chemical ontologies, for example, MESH in PubChem (http://pubchem.ncbi.nlm.nih.gov/), ChEBI (http://www.ebi.ac.uk/chebi/) or the MetaCyc (http://metacyc.org/) provide the ability to characterize chemicals at levels higher than distinct species. They provide hierarchies and in some cases multiple hierarchies to show groupings of similar structures. Thus, again with careful and expensive curation, reports could be "translated" to values where, for example, α-D-glucose could automatically be related to "sugars".

These possibilities are attractive from a bioinformatics perspective. From the perspective of G03R012 however the result would have been both expensive to curate (see above) and **extremely sparse.** Direct comparisons would be very few. Issues of chemical species identification, preparation and analysis methods and missing values would mean that very few close and instructive comparisons could be made. For this reason ArMet re-implementation and population was abandoned after discussion with the FSA.

2.5 DEVELOPMENT OF DATABASE ARCHITECTURE FOR ARMeC – ABERYSTWYTH RESPOSITORY FOR METABOLITE CHARACTERISTICS

With an examination of both G02 data (Section 1) and external information on metabolite content (Section 2.3) completed it was agreed with the FSA (18 month report) that there was little value in attempting to populate a new implementation of ArMet with these data as outlined in Section 2.4. However, during the course of these studies it became apparent that (with the partial exception of the Human Metabolome Database) that no databases had sufficient utility to integrate archive information from external literature AND also accommodate information from ongoing metabolomics studies that could help to both annotate metabolite signals (see Section 3 of the G03012 report) and biologically interpret data. In a collaboration between computer scientists, analytical chemists and biologists we therefore set out to develop a pilot database which was to contain data pertaining to the features of individual metabolites – the Aberystwyth Repository of Metabolite Characteristics (ARMeC) fields of information. The development of ARMeC from an analytical chemistry perspective is outlined in Section 6 of the G03012 report and was published as a Nature Protocol by Overy et al (2008) [6]. In the present section we outline the design development and software engineering which resulted in ARMeC.

Development proceeded in two stages. Version 1 was developed during the middle phase of the project, based on the initial spreadsheet based system, and provided support for analytical work in the later part of the project. At the end of the project, a Version 2 was developed which a) retains the functionality of Version 1; b) improves on the data design and web interface in the light of experience from Version 1 and c) adds structures for incorporating food composition data.

2.5.1 ARMeC Version 1

Systems analysis for this tool began with existing data and procedures which had been developed under a BBSRC ISIS project by Dr David Overy. It resulted in the data design provided in Appendix 2.8. The design centres around the list of metabolites with simple chemical data. A list of adducts are also represented. An adduct is associated with a metabolite in one or two ways: it may be a predicted adduct from the metabolite and/or a measured adduct from the metabolite. Either type of association may be on nominal or on accurate mass. Collection of these data is described under Section 6.4.3 and this relatively simple structure reflects the possible information.

When adducts are measured experimentally using MS^n, a fragment tree can be collected. This provides additional identification evidence and ARMeC provides for storing this as a recursive structure. The relational model, reflected in relational database systems is known to be poor at representing arbitrary tree-like data. ARMeC uses a standard design for accommodating trees which is contrived and inefficient but adequate for requirements here.

Metabolites are attested by external references. This is done in two ways: general and source specific. General references include literature reporting discovery of a metabolite and external databases such as KEGG and PubChem which record its existence and give access to additional information. Source-specific references record identification of the metabolite in a particular type of biological sample. This should specify taxonomy - species would be the typical classification. The identification might have been in samples of one or more organs or types of tissue. These aspects are represented as “BiologicalSource” in the design. It is the basis of allowing ARMeC to hold data on more than one species but to permit species-specific querying. The initial data set included associations to identify some metabolites as distinctive of major organism groups (plants, mammals, fungi, bacteria). This was catered for in preliminary plans but is not now seen as a useful concept.

External references of either type can be to literature, web resources or to a person (equivalent to a *pers. comm.).* Web references are typically to databases and *pers. comm.* is clearly of lesser value but necessary at least for interim data.

Synonyms are a pervasive problem in metabolomics data analysis. Common metabolites will typically have many synonyms and some names can refer to more than one chemical species. They have no value in identification. Their value is to humans in recognising a metabolite when results are produced and in searching for metabolites. Manual curation of a list of synonyms for each metabolite is supported.

In evaluating metabolomic data sets, recognising that candidate metabolites are, or are not, in the same metabolic pathway can be of value. Initial data used KEGG pathways and the model supports a list of pathways containing metabolites and permits metabolites to be in any number of pathways.

The logical design (Appendix 2.9) was implemented under the Oracle® 9*i* database management system (DBMS). The design permits implementation under any standards compliant DBMS - Oracle was chosen on grounds of availability and staff experience. Access to the database is via a Web interface. This was build using Java and J2EE facilities. Specifically servlets and Java Server Pages (JSP) were used. These are served using Tomcat through an Apache 2 server at http://www.armec.org.

For the implementation, a System Specification and Transaction analysis was undertaken[7]. This evaluated loads and search requirements to establish that that the platform was adequate. A number of optimisations were suggested. Transaction analysis suggested pages necessary in the Web interface and access rights were considered. All querying and browsing is made public.

A physical design was then produced (Appendix 2.8). This is a design targeted at a particular DBMS (Oracle) and taking into account optimisations. Figure 2.1 shows the table structure and constraints for the implementation. A number of points should be made. The authority subtypes (literature, online, contact) are implemented as a single table since this supports faster searching. Optimisations have been made and indexes set on a number of fields which the transaction analysis identified as common search keys. The most substantial design issue concerns the implementation of BiologicalSource. Each biological source must characterise a taxon, the anatomical description of the sample and (to support the higher level taxononomic groups) a so called class. These values must be controlled. They are controlled by externals terms. The taxon and class must be NCBI taxonomy identifiers. This permits flexibility, since any taxonomic level could be used. The values of class are limited to the 4 groups (plants, mammals, fungi, bacteria) by design. Taxon is not directly limited to species. In practice it is limited to two species: *Solanum tuberosum* and *Arabidopsis thaliana*. The sample is controlled by the Plant Ontology (http://www.plantontology.org/)[8]. This includes terms for organs and tissues and is currently the best vocabulary for describing material. The tables Taxon_T, Class_T and Sample_T are local caches of these external authorities to provide constraint and support for menu-driven choices in the Web interface.

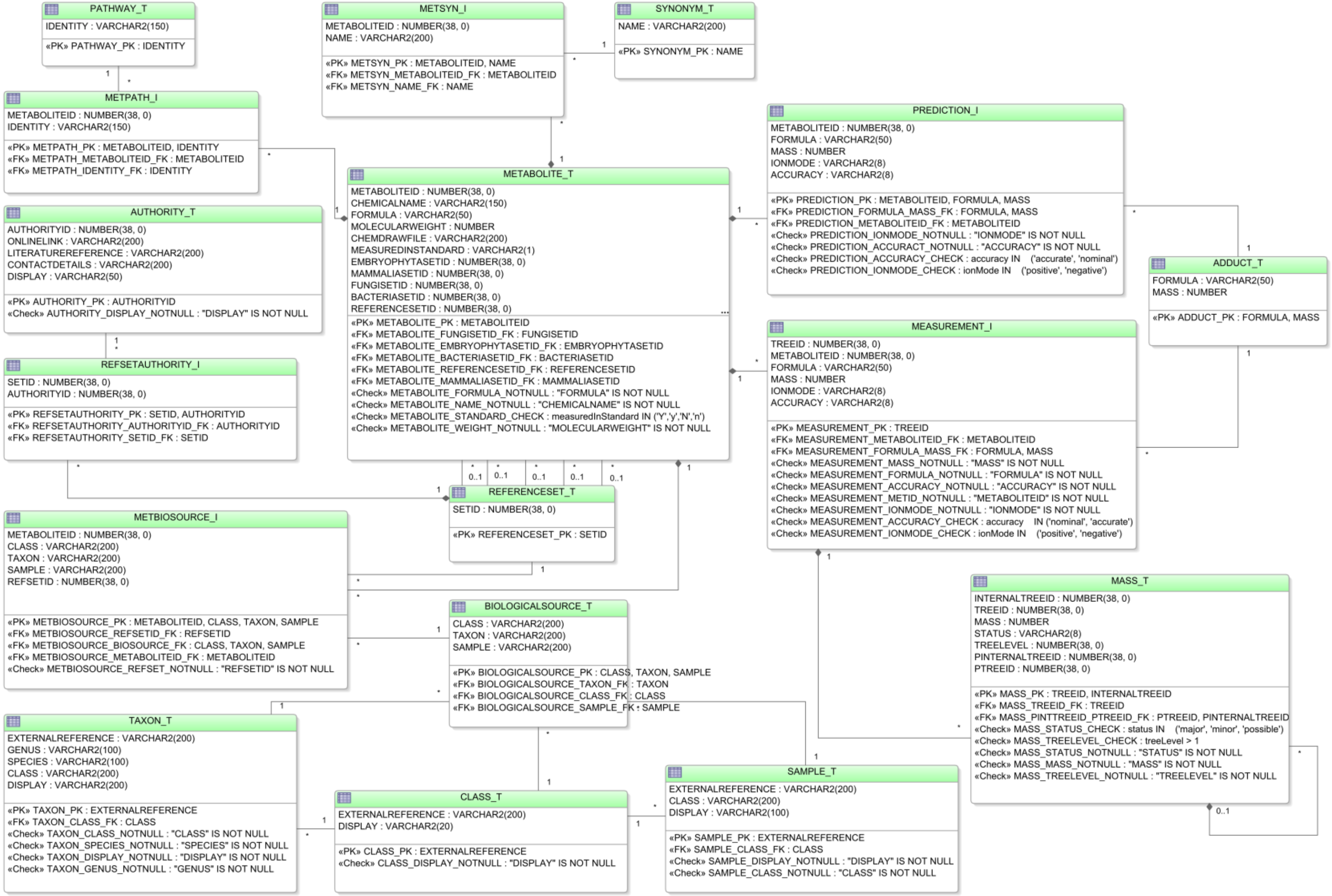

PATHWAY_T
IDENTITY : VARCHAR2(150)
«PK» PATHWAY_PK : IDENTITY
METPATH_I
METABOLITEID : NUMBER(38, 0)
IDENTITY : VARCHAR2(150)
«PK» METPATH_PK : METABOLITEID, IDENTITY
«FK» METPATH_METABOLITEID_FK : METABOLITEID
«FK» METPATH_IDENTITY_FK : IDENTITY
AUTHORITY_T
AUTHORITYID : NUMBER(38, 0)
ONLINELINK : VARCHAR2(200)
LITERATUREREFERENCE : VARCHAR2(200)
CONTACTDETAILS : VARCHAR2(200)
DISPLAY : VARCHAR2(50)
«PK» AUTHORITY_PK : AUTHORITYID
«Check» AUTHORITY_DISPLAY_NOTNULL : "DISPLAY" IS NOT NULL
REFSETAUTHORITY_I
SETID : NUMBER(38, 0)
AUTHORITYID : NUMBER(38, 0)
«PK» REFSETAUTHORITY_PK : SETID, AUTHORITYID
«FK» REFSETAUTHORITY_AUTHORITYID_FK : AUTHORITYID
«FK» REFSETAUTHORITY_SETID_FK : SETID
METBIOSOURCE_I
METABOLITEID : NUMBER(38, 0)
CLASS : VARCHAR2(200)
TAXON : VARCHAR2(200)
SAMPLE : VARCHAR2(200)
REFSETID : NUMBER(38, 0)
«PK» METBIOSOURCE_PK : METABOLITEID, CLASS, TAXON, SAMPLE
«FK» METBIOSOURCE_REFSETID_FK : REFSETID
«FK» METBIOSOURCE_BIOSOURCE_FK : CLASS, TAXON, SAMPLE
«FK» METBIOSOURCE_METABOLITEID_FK : METABOLITEID
«Check» METBIOSOURCE_REFSET_NOTNULL : "REFSETID" IS NOT NULL
TAXON_T
EXTERNALREFERENCE : VARCHAR2(200)
GENUS : VARCHAR2(100)
SPECIES : VARCHAR2(100)
CLASS : VARCHAR2(200)
DISPLAY : VARCHAR2(200)
«PK» TAXON_PK : EXTERNALREFERENCE
«FK» TAXON_CLASS_FK : CLASS
«Check» TAXON_CLASS_NOTNULL : "CLASS" IS NOT NULL
«Check» TAXON_SPECIES_NOTNULL : "SPECIES" IS NOT NULL
«Check» TAXON_DISPLAY_NOTNULL : "DISPLAY" IS NOT NULL
«Check» TAXON_GENUS_NOTNULL : "GENUS" IS NOT NULL
METSYN_I
METABOLITEID : NUMBER(38, 0)
NAME : VARCHAR2(200)
«PK» METSYN_PK : METABOLITEID, NAME
«FK» METSYN_METABOLITEID_FK : METABOLITEID
«FK» METSYN_NAME_FK : NAME
SYNONYM_T
NAME : VARCHAR2(200)
«PK» SYNONYM_PK : NAME
METABOLITE_T
METABOLITEID : NUMBER(38, 0)
CHEMICALNAME : VARCHAR2(150)
FORMULA : VARCHAR2(50)
MOLECULARWEIGHT : NUMBER
CHEMDRAWFILE : VARCHAR2(200)
MEASUREDINSTANDARD : VARCHAR2(1)
EMBRYOPHYTASETID : NUMBER(38, 0)
MAMMALIASETID : NUMBER(38, 0)
FUNGISETID : NUMBER(38, 0)
BACTERIASETID : NUMBER(38, 0)
REFERENCESETID : NUMBER(38, 0)
«PK» METABOLITE_PK : METABOLITEID
«FK» METABOLITE_FUNGISETID_FK : FUNGISETID
«FK» METABOLITE_EMBRYOPHYTASETID_FK : EMBRYOPHYTASETID
«FK» METABOLITE_BACTERIASETID_FK : BACTERIASETID
«FK» METABOLITE_REFERENCESETID_FK : REFERENCESETID
«FK» METABOLITE_MAMMALIASETID_FK : MAMMALIASETID
«Check» METABOLITE_FORMULA_NOTNULL : "FORMULA" IS NOT NULL
«Check» METABOLITE_NAME_NOTNULL : "CHEMICALNAME" IS NOT NULL
«Check» METABOLITE_STANDARD_CHECK : measuredInStandard IN ('Y','y','N','n')
«Check» METABOLITE_WEIGHT_NOTNULL : "MOLECULARWEIGHT" IS NOT NULL
REFERENCESET_T
SETID : NUMBER(38, 0)
«PK» REFERENCESET_PK : SETID
BIOLOGICALSOURCE_T
CLASS : VARCHAR2(200)
TAXON : VARCHAR2(200)
SAMPLE : VARCHAR2(200)
«PK» BIOLOGICALSOURCE_PK : CLASS, TAXON, SAMPLE
«FK» BIOLOGICALSOURCE_TAXON_FK : TAXON
«FK» BIOLOGICALSOURCE_CLASS_FK : CLASS
«FK» BIOLOGICALSOURCE_SAMPLE_FK : SAMPLE
CLASS_T
EXTERNALREFERENCE : VARCHAR2(200)
DISPLAY : VARCHAR2(20)
«PK» CLASS_PK : EXTERNALREFERENCE
«Check» CLASS_DISPLAY_NOTNULL : "DISPLAY" IS NOT NULL
SAMPLE_T
EXTERNALREFERENCE : VARCHAR2(200)
CLASS : VARCHAR2(200)
DISPLAY : VARCHAR2(100)
«PK» SAMPLE_PK : EXTERNALREFERENCE
«FK» SAMPLE_CLASS_FK : CLASS
«Check» SAMPLE_DISPLAY_NOTNULL : "DISPLAY" IS NOT NULL
«Check» SAMPLE_CLASS_NOTNULL : "CLASS" IS NOT NULL
PREDICTION_I
METABOLITEID : NUMBER(38, 0)
FORMULA : VARCHAR2(50)
MASS : NUMBER
IONMODE : VARCHAR2(8)
ACCURACY : VARCHAR2(8)
«PK» PREDICTION_PK : METABOLITEID, FORMULA, MASS
«FK» PREDICTION_FORMULA_MASS_FK : FORMULA, MASS
«FK» PREDICTION_METABOLITEID_FK : METABOLITEID
«Check» PREDICTION_IONMODE_NOTNULL : "IONMODE" IS NOT NULL
«Check» PREDICTION_ACCURACT_NOTNULL : "ACCURACY" IS NOT NULL
«Check» PREDICTION_ACCURACY_CHECK : accuracy IN ('accurate', 'nominal')
«Check» PREDICTION_IONMODE_CHECK : ionMode IN ('positive', 'negative')
ADDUCT_T
FORMULA : VARCHAR2(50)
MASS : NUMBER
«PK» ADDUCT_PK : FORMULA, MASS
MEASUREMENT_I
TREEID : NUMBER(38, 0)
METABOLITEID : NUMBER(38, 0)
FORMULA : VARCHAR2(50)
MASS : NUMBER
IONMODE : VARCHAR2(8)
ACCURACY : VARCHAR2(8)
«PK» MEASUREMENT_PK : TREEID
«FK» MEASUREMENT_METABOLITEID_FK : METABOLITEID
«FK» MEASUREMENT_FORMULA_MASS_FK : FORMULA, MASS
«Check» MEASUREMENT_MASS_NOTNULL : "MASS" IS NOT NULL
«Check» MEASUREMENT_FORMULA_NOTNULL : "FORMULA" IS NOT NULL
«Check» MEASUREMENT_ACCURACY_NOTNULL : "ACCURACY" IS NOT NULL
«Check» MEASUREMENT_METID_NOTNULL : "METABOLITEID" IS NOT NULL
«Check» MEASUREMENT_IONMODE_NOTNULL : "IONMODE" IS NOT NULL
«Check» MEASUREMENT_ACCURACY_CHECK : accuracy IN ('nominal', 'accurate')
«Check» MEASUREMENT_IONMODE_CHECK : ionMode IN ('positive', 'negative')
MASS_T
INTERNALTREEID : NUMBER(38, 0)
TREEID : NUMBER(38, 0)
MASS : NUMBER
STATUS : VARCHAR2(8)
TREELEVEL : NUMBER(38, 0)
PINTERNALTREEID : NUMBER(38, 0)
PTREEID : NUMBER(38, 0)
«PK» MASS_PK : TREEID, INTERNALTREEID
«FK» MASS_TREEID_FK : TREEID
«FK» MASS_PINTTREEID_PTREEID_FK : PTREEID, PINTERNALTREEID
«Check» MASS_STATUS_CHECK : status IN ('major', 'minor', 'possible')
«Check» MASS_TREELEVEL_CHECK : treeLevel > 1
«Check» MASS_STATUS_NOTNULL : "STATUS" IS NOT NULL
«Check» MASS_MASS_NOTNULL : "MASS" IS NOT NULL
«Check» MASS_TREELEVEL_NOTNULL : "TREELEVEL" IS NOT NULL
1
*
0..1

Figure 2.1: The implemented tables for ARMeC Version 1

2.5.2 ARMeC Version 2: Restructuring and Extension

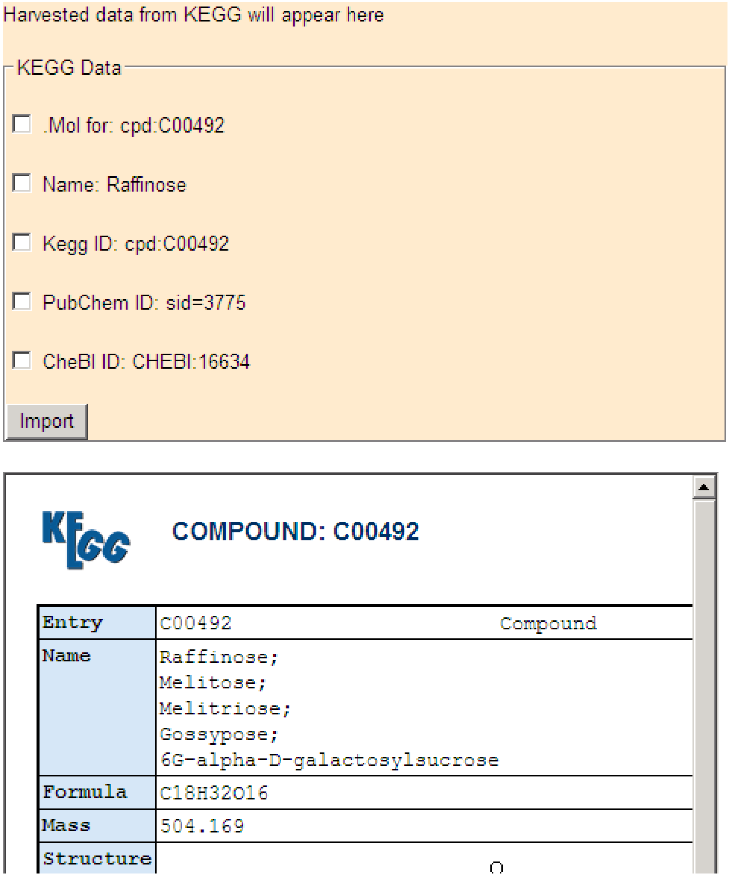
ARMeC Version 2 concentrates soley on potato tuber data and was designed specifically for the G03012 project.. Data curation efforts centre here and all new database design work has worked on the assumption of fixing the BiologicalSource values to *Solanum tubnerosum* and "tuber". This is most apparent in the user interface where those values are the only choices available. The underlying structure remains capable of supporting multiple species and the database could be extended in that way.

The fundamental change from Version 1 is to base the metabolite list on chemical structure. Databases based on names have no chemical meaning. They suffer from the problem of synonyms and they do not handle novel structures well. It may be argued that IUPAC names are both unique and can be generated for novel compounds. They are, however, not widely used and cannot be easily used to generate other data. A chemical structure is the fundamental concept of a metabolite. The new design therefore is a collection of structures. These are represented in the MOL format. Other formats could have been chosen, but MOL is an open format widely used and interpretable by a wide range of software. All chemical data about a metabolite in ARMeC is generated from MOL. In Version 1, curation involved entry of the canonical formula and molecular weight and of an image. This was time consuming but, more importantly, error prone. The IUPAC INChi code provides a unique and computer intelligible representation of structure. In ARMeC it is calculated (alone with SMILES and, potentially, other structural representations) and used for database purposes. Most notably it is used to prevent duplicate entries (i.e. it is a primary key). While much data is calculated from the MOL file, this is not done "on the fly" as the database is queried. On addition of a new metabolite, all derived values are calculated once. We believe that this approach to metabolite reporting is an important way forward for metabolomics and it will be taken forward elsewhere, particularly in the development of the ArMet system.

Addition of a new metabolite only therefore requires a MOL file (a name can also be provided but this is only for display). Novel metabolites can thus be entered. Any metabolite useful to ARMeC (for adduct searching) must have a known structure; many packages will generate MOL files and ARMeC provides a Web file upload facility. Most typically metabolites will be known in external databases which typically provide MOL files that can be harvested.

Currently, ARMeC permits automatic harvesting from KEGG. Other sources could be incorporated, but manual download can be used for them.

The harvesting interface is web-based. In a split screen, ARMeC is seen on the left while other web pages can be browsed on the right. A web page which can be harvested causes additional facilities to appear which mean that the curator can select items to be transferred to ARMeC. This includes MOL data. A clip from a screen dump is shown in Figure 2.2 to illustrate this.



Harvested data from KEGG will appear here

KEGG Data

□ .Mol for: cpd:C00492

□ Name: Raffinose

□ Kegg ID: cpd:C00492

□ PubChem ID: sid=3775

□ CheBI ID: CHEBI:16634

COMPOUND: C00492

Entry	C00492	Compound

Name	Raffinose;
	Melitose;
	Melitriose;
	Gossypose;
	6G-alpha-D-galactosylsucrose	

Formula	C18H32O16	
Mass	504.169	
Structure		

Figure 2.2 A partial screen dump to illustrate data harvesting.

The table structure and constraints for Version 2 are shown in Figure 2.3 (changed tables are distinguished by the colour of the header. Links to external references have been simplified and in particular one class of link has been removed. Links to external databases were via the referencing system. Now they are considered attributes of the metabolite. Fields for important sources (URLs) have been added to the metabolite table. These are not required data and they are under curator control. Curators can decide on these links perhaps even if the structure represented externally is not quite the same as the ARMeC one, but the association is believed to be correct. Links can be harvested. Figure 2.3 shows opportunity for 3 to be harvested from KEGG, including its own which would be sensible if the MOL is imported. They can also be manually entered. These can be used to provide outward browsing links from ARMeC. Harvesting extends to synonyms. The synonym structure has been simplified and this permits periodic automatic harvesting from all known external sites to make the ARMeC list a superset of them all. The links to external authorities has been simplified. This is not a very sophisticated part of the system and could be significantly improved to permit seamless access to on-line bibliographic information.

Pathway information in Version 1 was related to metabolites. It was manually curated, largely from KEGG. In Version 2 it is related to species and sample. This permits linkage to species specific databases. In particular, SolCyc (http:solcyc.sgn.cornell.edu) provides the PotatoCyc data set in the style of BioCyc (http://biocyc.org/). The SPPATHWAY_T table (Figure 2.3) holds data to identify and externally access pathways while SPMETPATH_I associates a metabolite in a particular type of sample to a pathway. Pathway data can be harvested from websites. A curator, perhaps when adding new metabolite, can decide which sites to attempt to harvest.

A major new aspect to Version 2 is the incorporation of compositional data from external sources (See Section 2.3). Simple, textual extracts from cited sources for the estimation of metabolite content in a species can be added. For a given species and sample type, when a reference is added to ARMeC, a small section text can be added which can then be displayed with the metabolite record. A measure of overall occurrence is reported in Section 2. This is associated not with a citation, but with the curator’s overall assessment of occurrence in the sample type. That value can therefore be stored in N_BIOSOURCE_T (Figure 2.3).

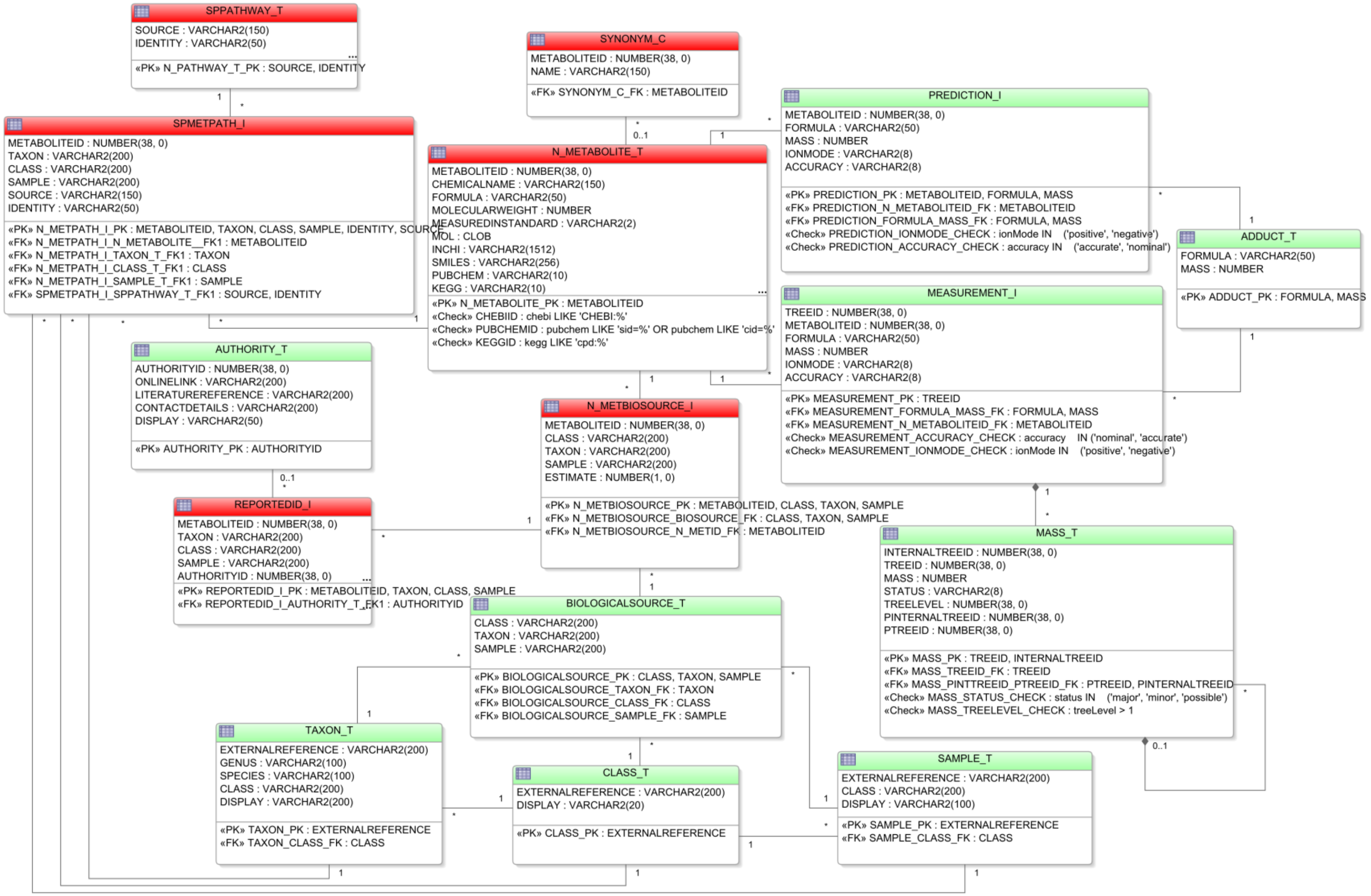

SPPATHWAY_T
SOURCE : VARCHAR2(150)
IDENTITY : VARCHAR2(50)
«PK» N_PATHWAY_T_PK : SOURCE, IDENTITY
SPMETPATH_I
METABOLITEID : NUMBER(38, 0)
TAXON : VARCHAR2(200)
CLASS : VARCHAR2(200)
SAMPLE : VARCHAR2(200)
SOURCE : VARCHAR2(150)
IDENTITY : VARCHAR2(50)
«PK» N_METPATH_I_PK : METABOLITEID, TAXON, CLASS, SAMPLE, IDENTITY, SOURCE
«FK» N_METPATH_I_N_METABOLITE__FK1 : METABOLITEID
«FK» N_METPATH_I_TAXON_T_FK1 : TAXON
«FK» N_METPATH_I_CLASS_T_FK1 : CLASS
«FK» N_METPATH_I_SAMPLE_T_FK1 : SAMPLE
«FK» SPMETPATH_I_SPPATHWAY_T_FK1 : SOURCE, IDENTITY
SYNONYM_C
METABOLITEID : NUMBER(38, 0)
NAME : VARCHAR2(150)
«FK» SYNONYM_C_FK : METABOLITEID
N_METABOLITE_T
METABOLITEID : NUMBER(38, 0)
CHEMICALNAME : VARCHAR2(150)
FORMULA : VARCHAR2(50)
MOLECULARWEIGHT : NUMBER
MEASUREDINSTANDARD : VARCHAR2(2)
MOL : CLOB
INCHI : VARCHAR2(1512)
SMILES : VARCHAR2(256)
PUBCHEM : VARCHAR2(10)
KEGG : VARCHAR2(10)
«PK» N_METABOLITE_PK : METABOLITEID
«Check» CHEBIID : chebi LIKE 'CHEBI:%'
«Check» PUBCHEMID : pubchem LIKE 'sid=%' OR pubchem LIKE 'cid=%'
«Check» KEGGID : kegg LIKE 'cpd:%'
PREDICTION_I
METABOLITEID : NUMBER(38, 0)
FORMULA : VARCHAR2(50)
MASS : NUMBER
IONMODE : VARCHAR2(8)
ACCURACY : VARCHAR2(8)
«PK» PREDICTION_PK : METABOLITEID, FORMULA, MASS
«FK» PREDICTION_N_METABOLITEID_FK : METABOLITEID
«FK» PREDICTION_FORMULA_MASS_FK : FORMULA, MASS
«Check» PREDICTION_IONMODE_CHECK : ionMode IN ('positive', 'negative')
«Check» PREDICTION_ACCURACY_CHECK : accuracy IN ('accurate', 'nominal')
ADDUCT_T
FORMULA : VARCHAR2(50)
MASS : NUMBER
«PK» ADDUCT_PK : FORMULA, MASS
AUTHORITY_T
AUTHORITYID : NUMBER(38, 0)
ONLINELINK : VARCHAR2(200)
LITERATUREREFERENCE : VARCHAR2(200)
CONTACTDETAILS : VARCHAR2(200)
DISPLAY : VARCHAR2(50)
«PK» AUTHORITY_PK : AUTHORITYID
MEASUREMENT_I
TREEID : NUMBER(38, 0)
METABOLITEID : NUMBER(38, 0)
FORMULA : VARCHAR2(50)
MASS : NUMBER
IONMODE : VARCHAR2(8)
ACCURACY : VARCHAR2(8)
«PK» MEASUREMENT_PK : TREEID
«FK» MEASUREMENT_FORMULA_MASS_FK : FORMULA, MASS
«FK» MEASUREMENT_N_METABOLITEID_FK : METABOLITEID
«Check» MEASUREMENT_ACCURACY_CHECK : accuracy IN ('nominal', 'accurate')
«Check» MEASUREMENT_IONMODE_CHECK : ionMode IN ('positive', 'negative')
N_METBIOSOURCE_I
METABOLITEID : NUMBER(38, 0)
CLASS : VARCHAR2(200)
TAXON : VARCHAR2(200)
SAMPLE : VARCHAR2(200)
ESTIMATE : NUMBER(1, 0)
«PK» N_METBIOSOURCE_PK : METABOLITEID, CLASS, TAXON, SAMPLE
«FK» N_METBIOSOURCE_BIOSOURCE_FK : CLASS, TAXON, SAMPLE
«FK» N_METBIOSOURCE_N_METID_FK : METABOLITEID
REPORTEDID_I
METABOLITEID : NUMBER(38, 0)
TAXON : VARCHAR2(200)
CLASS : VARCHAR2(200)
SAMPLE : VARCHAR2(200)
AUTHORITYID : NUMBER(38, 0)
«PK» REPORTEDID_I_PK : METABOLITEID, TAXON, CLASS, SAMPLE
«FK» REPORTEDID_I_AUTHORITY_T_FK1 : AUTHORITYID
MASS_T
INTERNALTREEID : NUMBER(38, 0)
TREEID : NUMBER(38, 0)
MASS : NUMBER
STATUS : VARCHAR2(8)
TREELEVEL : NUMBER(38, 0)
PINTERNALTREEID : NUMBER(38, 0)
PTREEID : NUMBER(38, 0)
«PK» MASS_PK : TREEID, INTERNALTREEID
«FK» MASS_TREEID_FK : TREEID
«FK» MASS_PINTTREEID_PTREEID_FK : PTREEID, PINTERNALTREEID
«Check» MASS_STATUS_CHECK : status IN ('major', 'minor', 'possible')
«Check» MASS_TREELEVEL_CHECK : treeLevel > 1
BIOLOGICALSOURCE_T
CLASS : VARCHAR2(200)
TAXON : VARCHAR2(200)
SAMPLE : VARCHAR2(200)
«PK» BIOLOGICALSOURCE_PK : CLASS, TAXON, SAMPLE
«FK» BIOLOGICALSOURCE_TAXON_FK : TAXON
«FK» BIOLOGICALSOURCE_CLASS_FK : CLASS
«FK» BIOLOGICALSOURCE_SAMPLE_FK : SAMPLE
TAXON_T
EXTERNALREFERENCE : VARCHAR2(200)
GENUS : VARCHAR2(100)
SPECIES : VARCHAR2(100)
CLASS : VARCHAR2(200)
DISPLAY : VARCHAR2(200)
«PK» TAXON_PK : EXTERNALREFERENCE
«FK» TAXON_CLASS_FK : CLASS
CLASS_T
EXTERNALREFERENCE : VARCHAR2(200)
DISPLAY : VARCHAR2(20)
«PK» CLASS_PK : EXTERNALREFERENCE
SAMPLE_T
EXTERNALREFERENCE : VARCHAR2(200)
CLASS : VARCHAR2(200)
DISPLAY : VARCHAR2(100)
«PK» SAMPLE_PK : EXTERNALREFERENCE
«FK» SAMPLE_CLASS_FK : CLASS

Figure 2.3: The implementation tables for ARMeC Version 2

REFERENCES - Sections 1 & 2

(N.B. an extensive Bibliography of external literature related to potato metabolite composition is presented at the end of this section.

[1] Helen Jenkins, Nigel Hardy, Manfred Beckmann, John Draper, Aileen R Smith, Janet Taylor, Oliver Fiehn, Royston Goodacre, Raoul J Bino, Robert Hall, Joachim Kopka, Geoffrey A Lane, B Markus Lange, Jang R Liu, Pedro Mendes, Basil J Nikolau, Stephen G Oliver, Norman W Paton, Sue Rhee, Ute Roessner-Tunali, Kazuki Saito, Jørn Smedsgaard, Lloyd W Sumner, Trevor Wang, Sean Walsh, Eve Syrkin Wurtele and Douglas B Kell. A proposed framework for the description of plant metabolomics experiments and their results. *Nature Biotechnology* **22(**12):1601-1606 (2004) Dec. doi:10.1038/nbt1041

[2] John C Lindon, Jeremy K Nicholson, Elaine Holmes, Hector C Keun, Andrew Craig, Jake T M Pearce, Stephen J Bruce, Nigel Hardy, Susanna-Assunta Sansone, Henrik Antti, Par Jonsson, Clare Daykin, Mahendra Navarange, Richard D Beger, Elwin R Verheij, Alexander Amberg, Dorrit Baunsgaard, Glenn H Cantor, Lois Lehman-McKeeman, Mark Earll, Svante Wold, Erik Johansson, John N Haselden, Kerstin Kramer, Craig Thomas, Johann Lindberg, Ina Schuppe-Koistinen, Ian D Wilson, Michael D Reily, Donald G Robertson, Hans Senn, Arno Krotzky, Sunil Kochhar, Jonathan Powell, Frans van der Ouderaa, Robert Plumb, Hartmut Schaefer, Manfred Spraul. Summary recommendations for standardization and reporting of metabolic analyses. *Nature Biotechnology* **23**:833-838 (2005) doi:10.1038/nbt0705-833

[3] Oliver Fiehn, Bruce Kristal, Ben Van Ommen, Lloyd W. Sumner, Susanna-Assunta Sansone, Chris Taylor, Nigel Hardy, Rima Kaddurah-Daouk Establishing Reporting Standards for Metabolomic and Metabonomic Studies: A Call for Participation OMICS: A Journal of Integrative Biology 10(2):158-163 (2006). doi:10.1089/omi.2006.10.158

[4] Irena Spasic, Warwick Dunn, Giles Velarde, Andy Tseng, Helen Jenkins, *Nigel W Hardy*, Stephen G Oliver and Douglas B Kell MeMo: a hybrid SQL/XML approach to metabolomic data management for functional genomics *BMC Bioinformatics* **7**:281 (2006) doi:10.1186/1471-2105-7-281

[5] Andrew R Jones, Michael Miller, Ruedi Aebersold, Rolf Apweiler, Catherine A Ball, Alvis Brazma, James DeGreef, ***Nigel Hardy***, Henning Hermjakob, Simon J Hubbard, Peter Hussey, Mark Igra, Helen Jenkins, Randall K Julian Jr, Kent Laursen, Stephen G Oliver, Norman W Paton, Susanna-Assunta Sansone, Ugis Sarkans, Christian J Stoeckert Jr, Chris F Taylor, Patricia L Whetzel, Joseph A White, Paul Spellman and Angel Pizarro The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics. *Nature Biotechnology* **25**(10):1127-1133 (doi:10.1038/nbt1347.

[6] Overy, D.P., Enot, D.P., Tailliart, K., Jenkins, H., Parker, D., Beckmann, M. & Draper, J. (2008) Explanatory signal interpretation and metabolite identification strategies for nominal mass FIE-MS metabolite fingerprints. *Nature Protocols*, **3**, 471-485.

[7] Jenkin, H. G03 Metabolite Library System Specification and Transactions Analysis. G03R012 Internal report. http://sirius.dcs.aber.ac.uk:9095/g03r0002/docs/design/MetaboliteLibrary/index.html

[8] Shulamit Avraham et al. The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. Nucleic Acids Research 2008 **36**:D449-D454; doi:10.1093/nar/gkm908

Appendix 2.1 USDA extract describing raw potato tuber (flesh and skin)

Refuse: 0%
Scientific Name: *Solanum tuberosum*
NDB No: 11352 (Nutrient values and weights are for edible portion)

Nutrient	**Units**	**Value per 100 grams**	**Number of Data Points**	**Std. Error**
Proximates				
Water	G	79.34	11	0.368
Energy	kcal	77	0	
Energy	kj	321	0	
Protein	g	2.02	11	0.034
Total lipid (fat)	g	0.09	11	0.009
Ash	g	1.08	11	0.097
Carbohydrate, by difference	g	17.47	0	
Fiber, total dietary	g	2.2	5	
Sugars, total	g	0.78	0	
Sucrose	g	0.17	11	0.021
Glucose (dextrose)	g	0.33	11	0.047
Fructose	g	0.27	11	0.029
Lactose	g	0.00	11	0
Maltose	g	0.00	11	0
Galactose	g	0.00	5	
Starch	g	15.44	11	1.448
Minerals				
Calcium, Ca	mg	12	11	1.368
Iron, Fe	mg	0.78	11	0.04
Magnesium, Mg	mg	23	11	0.452
Phosphorus, P	mg	57	11	0.88
Potassium, K	mg	421	11	11.942
Sodium, Na	mg	6	73	0.76
Zinc, Zn	mg	0.29	11	0.003
Copper, Cu	mg	0.108	11	0.016
Manganese, Mn	mg	0.153	11	0.026
Selenium, Se	mcg	0.3	299	0.031
Vitamins				
Vitamin C, total ascorbic acid	mg	19.7	141	0.769
Thiamin	mg	0.080	11	0.006

Riboflavin	mg	0.032	11	0.013
Niacin	mg	1.054	10	0.04
Pantothenic acid	mg	0.296	11	0.036
Vitamin B-6	mg	0.295	11	0.148
Folate, total	mcg	16	11	3.324
Folic acid	mcg	0	0	
Folate, food	mcg	16	11	3.324
Folate, DFE	mcg_DFE	16	0	
Vitamin B-12	mcg	0.00	0	
Vitamin B-12, added	mcg	0.00	0	
Vitamin A, IU	IU	2	0	
Vitamin A, RAE	mcg_RAE	0	0	
Retinol	mcg	0	0	
Vitamin E (alpha-tocopherol)	mg	0.01	2	
Vitamin E, added	mg	0.00	0	
Tocopherol, beta	mg	0.00	2	
Tocopherol, gamma	mg	0.00	2	
Tocopherol, delta	mg	0.00	2	
Vitamin K (phylloquinone)	mcg	1.9	11	0.66
Lipids				
Fatty acids, total saturated	g	0.026	0	
4:0	g	0.000	0	
6:0	g	0.000	0	
8:0	g	0.000	0	
10:0	g	0.001	6	
12:0	g	0.003	6	
14:0	g	0.001	18	
16:0	g	0.016	59	
18:0	g	0.004	59	
Fatty acids, total monounsaturated	g	0.002	0	
16:1 undifferentiated	g	0.001	37	
18:1 undifferentiated	g	0.001	59	
20:1	g	0.000	0	
22:1 undifferentiated	g	0.000	0	
Fatty acids, total polyunsaturated	g	0.043	0	
18:2 undifferentiated	g	0.032	59	
18:3 undifferentiated	g	0.010	59	

18:4	g	0.000	0	
20:4 undifferentiated	g	0.000	0	
20:5 n-3	g	0.000	0	
22:5 n-3	g	0.000	0	
22:6 n-3	g	0.000	0	
Cholesterol	mg	0	0	
Phytosterols	mg	5	0	
Amino acids				
Tryptophan	g	0.032	6	
Threonine	g	0.075	7	
Isoleucine	g	0.084	7	
Leucine	g	0.124	7	
Lysine	g	0.126	7	
Methionine	g	0.033	7	
Cystine	g	0.026	6	
Phenylalanine	g	0.092	7	
Tyrosine	g	0.077	6	
Valine	g	0.117	7	
Arginine	g	0.095	9	
Histidine	g	0.045	9	
Alanine	g	0.064	6	
Aspartic acid	g	0.506	6	
Glutamic acid	g	0.347	8	
Glycine	g	0.062	6	
Proline	g	0.074	8	
Serine	g	0.090	6	
Other				
Alcohol, ethyl	g	0.0	0	
Caffeine	mg	0	0	
Theobromine	mg	0	0	
Carotene, beta	mcg	1	11	0.57
Carotene, alpha	mcg	0	11	0
Cryptoxanthin, beta	mcg	0	11	0
Lycopene	mcg	0	11	0
Lutein + zeaxanthin	mcg	8	10	2.357

Appendix 2.2 Extract of FINELI Database

Potato, unpeeled, average, raw

- Food number: 28913
- Scientific name: *Solanum tuberosum*
- Food type: Food
- Processing method: No treatment
- Ingredient class: Potato
- Food use class: Cooked potatoes
- Edible proportion: 85%
- Special diets

Total energy content is divided into

fat, total	🟥	2%
protein, total	🟦	10%
carbohydrate, available	🟩	86%
alcohol		0%
organic acids, total	▮	2%
sugar alcohols		0%

Nutrient values / 100 g.

Nutrient factor	Content	Unit	Method	Acquisition	Reference
Macro-components					
energy, calculated	264 (63)	kJ (kcal)	calculated as recipe	value created within host-system	
carbohydrate, available	13.2	g	calculated as recipe	value created within host-system	
fat, total	0.2	g	calculated as recipe	value created within host-system	
protein, total	1.6	g	calculated as recipe	value created within host-system	
alcohol	0	g	calculated as recipe	value created within host-system	
Carbohydrate Components					
organic acids, total	0.5	g	calculated as recipe	value created	

				within host-system	
starch, total	12.7	g	calculated as recipe	value created within host-system	
sugars, total	0.5	g	calculated as recipe	value created within host-system	
sucrose	0	g	calculated as recipe	value created within host-system	
lactose	0	g	calculated as recipe	value created within host-system	
fructose	0.3	g	calculated as recipe	value created within host-system	
sugar alcohols	0	g	calculated as recipe	value created within host-system	
fibre, total	0.9	g	calculated as recipe	value created within host-system	
fibre, water-insoluble	0.9	g	calculated as recipe	value created within host-system	
polysaccharides, non-cellulosic, water-soluble	0.3	g	calculated as recipe	value created within host-system	
glucose	0.3	g	calculated as recipe	value created within host-system	
maltose	0	g	calculated as recipe	value created within host-system	

Fat					
fatty acids, total, calculated as TAG equivalents	< 0.1	g	calculated as recipe	value created within host-system	
fatty acids, total	< 0.1	g	calculated as recipe	value created within host-system	
fatty acids, total saturated	< 0.1	g	calculated as recipe	value created within host-system	
fatty acids, total monounsaturated cis	< 0.1	g	calculated as recipe	value created within host-system	
fatty acids, total polyunsaturated	< 0.1	g	calculated as recipe	value created within host-system	
fatty acids, total trans	0	g	calculated as recipe	value created within host-system	
fatty acid 18:2 cis,cis n-6 (linoleic acid)	29	mg	calculated as recipe	value created within host-system	
fatty acid 18:3 n-3 (alpha-linolenic acid)	25	mg	calculated as recipe	value created within host-system	
fatty acid 20:5 n-3 (EPA)	0	mg	calculated as recipe	value created within host-system	
fatty acid 22:6 n-3 (DHA)	0	mg	calculated as recipe	value created within host-system	
cholesterol (GC)	0.3	mg	calculated as recipe	value created within host-system	

sterols, total	4.4	mg	calculated as recipe	value created within host-system	
Minerals					
sodium	0.9	mg	calculated as recipe	value created within host-system	
salt	2.2	mg	calculated as recipe	value created within host-system	
potassium	425.0	mg	calculated as recipe	value created within host-system	
magnesium	20.4	mg	calculated as recipe	value created within host-system	
calcium	4.8	mg	calculated as recipe	value created within host-system	
phosphorus	38.3	mg	calculated as recipe	value created within host-system	
iron, total	0.6	mg	calculated as recipe	value created within host-system	
zinc	0.3	mg	calculated as recipe	value created within host-system	
iodide (iodine)	0.9	µg	calculated as recipe	value created within host-system	
selenium, total	0.5	µg	calculated as recipe	value created within host-system	

Vitamins					
vitamin A retinol activity equivalents	0.5	µg	calculated as recipe	value created within host-system	
vitamin D	0	µg	calculated as recipe	value created within host-system	
vitamin E alphatocopherol	<0.1	mg	calculated as recipe	value created within host-system	
vitamin K, total	0.88	µg	calculated as recipe	value created within host-system	
vitamin C (ascorbic acid)	8.5	mg	calculated as recipe	value created within host-system	
folate (HPLC)	19.7	µg	calculated as recipe	value created within host-system	
niacin equivalents, total	0.5	mg	calculated as recipe	value created within host-system	
riboflavine	0.03	mg	calculated as recipe	value created within host-system	
thiamin (vitamin B1)	0.18	mg	calculated as recipe	value created within host-system	
vitamin B-12 (cobalamin)	0	µg	calculated as recipe	value created within host-system	
vitamers pyridoxine (hydrochloride)	0.10	mg	calculated as recipe	value created within host-system	

carotenoids, total	57.7	µg	calculated as recipe	value created within host-system	

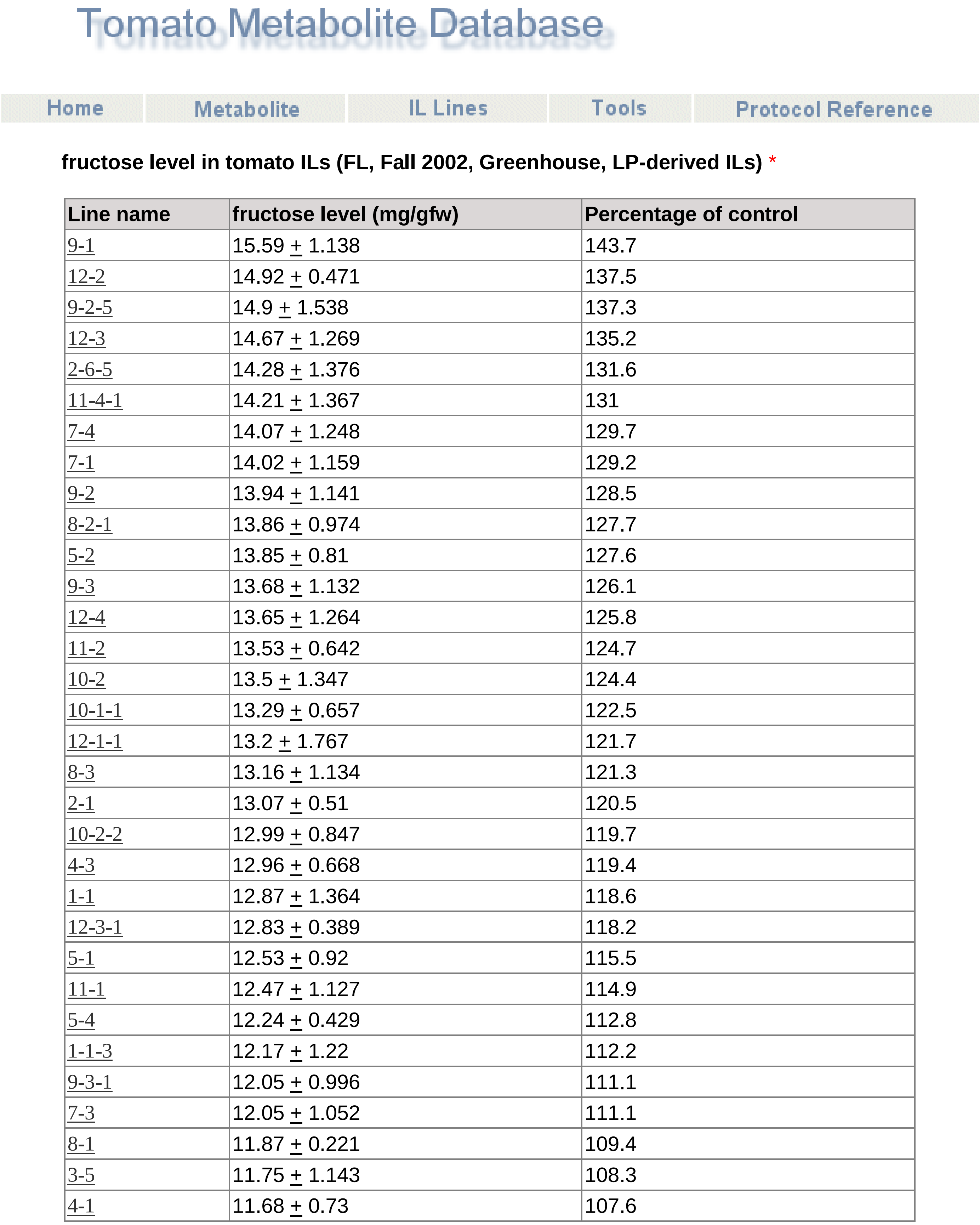
Appendix 2.3 Table from “Compositional analysis of tubers from insect and virus resistant potato plants”, Rogan et al. (*J. Agric. Food Chem.* 2000)

Table 7. Vitamin, Mineral, and Amino Acid Composition of NewLeaf Y Shepody and Shepody Potato Tubers[a]

	NewLeaf Y SE clone									
	SEMT15-02			SEMT15-15			Shepody control			
		range			range			range		
component (mg/200 g of FW)	mean	max	min	mean	max	min	mean	max	min	literature range[b]
vitamin B_6	0.56	0.62	0.49	0.50	0.72	0.32	0.52	0.62	0.40	0.26-0.82
niacin	4.55	5.05	4.14	4.78	5.86	3.98	4.43	5.15	3.73	0.18-6.2
copper	0.41	0.61	0.20	0.48	1.10	0.23	0.39	0.53	0.20	0.03-1.4
magnesium	53.13	67.16	48.17	56.95	90.00	47.85	54.22	65.54	48.95	22.5-110
potassium	1097.24	1326.78	996.82	1135.14	1634.40	971.46	1162.01	1259.30	1105.92	700-1250
aspartic acid	919.24	1151.50	614.88	994.27	1404.00	702.24	1001.97	1324.80	670.72	677-1476
threonine	185.54	220.50	142.13	202.22	278.64	157.47	183.00	225.77	138.75	102-214
serine	191.19	231.77	147.67	201.53	286.56	149.18	187.80	230.18	141.06	125-255
glutamic acid	865.48	1073.10	665.28	976.81	1173.60	856.52	966.06	1181.28	773.12	583-1207
proline	171.46	232.06	127.01	181.36	272.16	130.34	164.45	201.48	118.78	89-366
glycine	165.64	196.97	133.06	179.23	249.84	147.90	161.84	184.92	133.12	92-195
alanine	149.36	171.99	117.94	163.20	219.60	133.53	146.35	172.22	119.30	87-238
cystine	78.86	89.99	71.57	83.82	108.72	75.54	76.36	86.55	66.56	96-185
valine	219.00	271.68	186.48	249.09	346.32	223.44	225.81	247.84	200.70	196-363
methionine	75.21	85.28	63.00	83.49	105.84	72.35	72.13	83.90	55.30	57-100
isoleucine	159.52	207.72	129.53	183.88	259.20	159.60	164.08	187.15	137.22	119-238
leucine	303.91	363.37	227.30	331.62	460.80	252.17	291.80	359.35	213.50	171-346
tyrosine	147.49	170.93	128.02	170.91	228.24	151.62	150.98	160.63	137.22	114-236
phenylalanine	200.41	239.98	161.28	226.08	315.36	193.12	201.51	227.98	165.38	138-272
histidine	84.53	93.96	72.58	94.17	128.16	81.93	87.05	96.60	76.29	33-117
lysine	276.80	318.09	231.34	304.06	410.40	265.47	274.90	314.64	225.79	154-342
arginine	220.30	259.21	173.88	250.47	339.84	200.56	241.84	314.09	171.52	175-362
tryptophan	43.23	49.30	38.56	46.09	67.18	36.27	42.75	48.96	34.51	29-70

[a] Samples were collected from Island Falls, ME, and two sites in Canada (Hartland, NB; Summerside, PE). Plots were replicated four times at Hartland, NB. Plots were not replicated at the other two locations. Values presented represent the mean calculated from all six values.[b] For vitamins and minerals, reported by Storey and Davis (1978) and Lisinska and Leszczynski (1989); for amino acids, reported by Talley et al. (1984). Fresh weight concentration for literature range was determined by assuming that potatoes are composed of ~75% water. All values are reported as mg/200 g of FW.

Appendix 2.4 Tomato Metabolite database: information on fructose levels



fructose level in tomato ILs (FL, Fall 2002, Greenhouse, LP-derived ILs) *

Line name	fructose level (mg/gfw)	Percentage of control
9-1	15.59 ± 1.138	143.7
12-2	14.92 ± 0.471	137.5
9-2-5	14.9 ± 1.538	137.3
12-3	14.67 ± 1.269	135.2
2-6-5	14.28 ± 1.376	131.6
11-4-1	14.21 ± 1.367	131
7-4	14.07 ± 1.248	129.7
7-1	14.02 ± 1.159	129.2
9-2	13.94 ± 1.141	128.5
8-2-1	13.86 ± 0.974	127.7
5-2	13.85 ± 0.81	127.6
9-3	13.68 ± 1.132	126.1
12-4	13.65 ± 1.264	125.8
11-2	13.53 ± 0.642	124.7
10-2	13.5 ± 1.347	124.4
10-1-1	13.29 ± 0.657	122.5
12-1-1	13.2 ± 1.767	121.7
8-3	13.16 ± 1.134	121.3
2-1	13.07 ± 0.51	120.5
10-2-2	12.99 ± 0.847	119.7
4-3	12.96 ± 0.668	119.4
1-1	12.87 ± 1.364	118.6
12-3-1	12.83 ± 0.389	118.2
5-1	12.53 ± 0.92	115.5
11-1	12.47 ± 1.127	114.9
5-4	12.24 ± 0.429	112.8
1-1-3	12.17 ± 1.22	112.2
9-3-1	12.05 ± 0.996	111.1
7-3	12.05 ± 1.052	111.1
8-1	11.87 ± 0.221	109.4
3-5	11.75 ± 1.143	108.3
4-1	11.68 ± 0.73	107.6

7-4-1	11.67 ± 0.549	107.6
10-1	11.65 ± 0.773	107.4
8-3-1	11.63 ± 0.63	107.2
7-5-5	11.57 ± 1.282	106.6
4-2	11.31 ± 1.009	104.2
4-3-2	11.27 ± 0.592	103.9
10-3	11.27 ± 0.39	103.9
1-1-2	11.26 ± 0.735	103.8
6-4	11.26 ± 1.398	103.8
8-1-1	11.25 ± 0.507	103.7
1-4	11.01 ± 0.433	101.5
8-1-5	10.97 ± 0.592	101.1
2-2	10.94 ± 0.572	100.8
5-3	10.91 ± 1	100.6
m82	10.85 ± 0.659	100
1-4-18	10.85 ± 0.988	100
12-4-1	10.72 ± 0.82	98.8
12-1	10.7 ± 0.192	98.6
8-2	10.65 ± 0.791	98.2
11-3	10.55 ± 0.769	97.2
2-1-1	9.89 ± 1.186	91.2
9-3-2	9.83 ± 0.786	90.6
4-1-1	9.7 ± 0.709	89.4
1-3	9.59 ± 0.441	88.4
3-1	9.3 ± 1.34	85.7
6-2	9.2 ± 0.799	84.8
5-5	8.91 ± 0.333	82.1
9-2-6	8.43 ± 0.045	77.7
2-6	8.2 ± 1.056	75.6
7-5	8.15 ± 0.734	75.1
9-1-2	7.9 ± 0.832	72.8
3-4	7.87 ± 1.053	72.5
11-4	7.85 ± 0.662	72.4
6-1	7.66 ± 0.577	70.6
3-2	7.43 ± 0.421	68.5

*** Note**

- A value of zero means the fructose concentration was below the level of dection
- No measure of fructose was performed on lines not shown in the table

Appendix 2.5 Example of typical information retrievable from the literature

(a) as text:

Table 6 *Carotenoid content of* dxs-*transformed Desiree tubers compared with controls*

	Carotenoid levels (μg g^{-1} dry weight)							
Sample	Total	Neo[a]	Vio	Ant	Lut	Zea	Esters	Phyt
Wild-type DT	3.6±0.1	0.4±0.1	0.8±0.1	0.3±0.0	0.5±0.0	0.1±0.0	1.2±0.2	0.4±0.1
Empty vector DT	4.7±0.4	0.4±0.0	1.2±0.1	0.5±0.1	0.5±0.1	0.1±0.0	1.6±0.3	0.4±0.1
*dxs*1 DT	6.5±0.1	0.3±0.0	1.4±0.1	0.1±0.0	0.8±0.0	0	0.9±0.1	3.0±0.6*
*dxs*2 DT	7.0±0.9	0.2±0.0	1.3±0.3	0.1±0.0	0.9±0.2	0	1.2±0.4	3.2±0.4*
Wild-type MT	4.1±1.2	0.3±0.1	1.2±0.5	0.3±0.1	0.3±0.1	0	1.7±0.6	0.4±0.1
Empty vector MT	2.1±0.6	0.2±0.0	0.4±0.1	0.1±0.0	0.1±0.0	0	0.9±0.4	0.4±0.1
*dxs*1 MT	6.8±0.9	0.4±0.1	1.0±0.1	0.7±0.3	1.2±0.1	0.2±0.1	0.8±0.1	2.5±0.1*
*dxs*2 MT	3.5±0.5	0.2±0.0	0.8±0.3	0.5±0.2	0.5±0.2	0.1±0.0	0.5±0.1	0.9±0.2

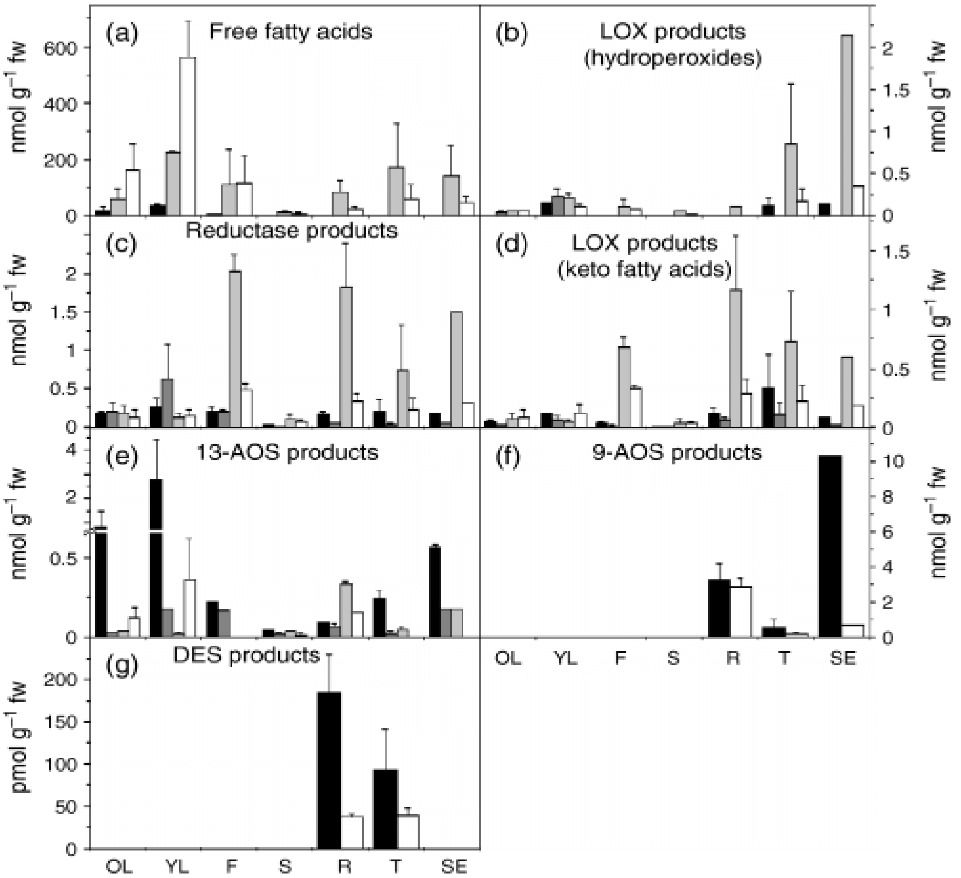
Values are means of the determinations from three tubers ±SEM. Values marked by an asterisk are significantly different at *P*=0.05.

[a] Neo, neoxanthin; Vio, violaxanthin; Ant, antheraxanthin; Lut, lutein; Zea, zeaxanthin; Phyt, phytoene; DT, developing tuber; MT, mature tuber.

(from: Wayne L. Morris, Laurence J. M. Ducreux, Peter Hedden, Steve Millam, and Mark A. Taylor, Overexpression of a bacterial 1-deoxy-D-xylulose 5-phosphate synthase gene in potato tubers perturbs the isoprenoid metabolic network: implications for the control of the tuber life cycle, J. Exp. Bot., 57(12): 3007-3018, 2006)

(b) as graphs:

Figure 6. Oxylipin profiles of various potato organs. Oxylipins were extracted from various potato organs and quantified as described in Experimental procedures. (a) Free fatty acids: black, roughanic acid (16:3); gray, linoleic acid (18:2); white, *α*-linolenic acid (18:3). (b) LOX pathway products (dioxygenase activity, fatty acid hydroperoxides): black, 13-HPODE; dark gray, 13-HPOTE; light gray, 9-HPODE; white, 9-HPOTE. (c) Reductase pathway products: black, 13-HODE; dark gray, 13-HOTE; light gray, 9-HODE; white, 9-HOTE. (d) LOX products (peroxidase activity, keto fatty acids): black, 13-KODE; dark gray, 13-KOTE; light gray, 9-KODE; white, 9-KOTE. (e) 13-AOS pathway products: black, 12-OPDA; dark gray, JA; light gray, *α*-ketol derived from 13-HPODE; white, *α*-ketol derived from 13-HPOTE. (f) 9-AOS pathway products: black, *α*-ketol derived from 9-HPODE; white, *α*-ketol derived from 9-HPOTE. (g) DES pathway products: black, colneleic acid; white, colnelenic acid. OL, old leaves; YL, young leaves; F, flowers; S, stems; R, roots; T, tubers; SE, sprouting eyes. The results are representative of two independent experiments.



(from: Michael Stumpe, Cornelia Göbel, Kirill Demchenko, Manuela Hoffmann, Ralf B. Klösgen, Katharina Pawlowski, Ivo Feussner, Identification of an allene oxide synthase (*CYP74C) that leads to formation of α*-ketols from 9-hydroperoxides of linoleic and linolenic acid in below-ground organs of potato The Plant Journal 47 (6):883–896, 2006) .

List of Appendices for G03012 Report Sections 1 & 2

available on accompanying CD

Section 1

Appendix 1.1: G02 Data Selection

Appendix 1.2: G02 data catalogue

Section 2

Appendix 2.6 ArMet Re-design

Appendix 2.7 G03 ArMet design details

Appendix 2.8 G03 Metabolite Library: ARMeC logical design

Appendix 2.9 G03 Metabolite Library: ARMeC physical design

Appendix 2.10 Spreadsheet summarising external data relating to metabolite content of potato tubers

G03012 Final Report - Section 3

Tom Shepherd/Manfred Beckmann/Derek Stewart

Development of standardized procedures for manual labeling of metabolite peaks in GC-MS profiles (Task 02.01)

3.1 BACKGROUND TO EXPERIMENTAL SECTION

During the G02 programme the Institutes comprising the different consortia developed a number of protocols for metabolite profiling of their respective target crop species using a number of complementary analytical techniques including gas chromatography-mass spectrometry (GC-MS). Each organization developed and applied their own protocols for all steps of the analytical process from sample preparation through to chromatographic separation of metabolites using different technology platforms and alignment, deconvolution and annotation of data using various peak annotation software (NIST, Wiley, Binbase). Each Institute generated its own lists of metabolites detected in their target crop species with various levels of annotation and degrees of identification from known metabolites to total unknowns. Several considerations arose from this, relating to how the different technology platforms and methods for peak picking and data extraction influence the final outcome. In order to allow for the comparison to and mining of data models between Institutes, the establishment of a standardized set of criteria to describe output generated from metabolomics experiments is essential, especially from experiments generated on different technologies.

3.2 OVERALL OBJECTIVES

Work carried out focused upon defining a set of criteria for the standardised description and annotation of metabolite peaks resolvable by GC-MS profiling in potato from G02 projects from three Institutes (SCRI, UWA, and Golm).

3.2.1 Summary of objectives

- Review on GC-MS metabolite profiling and reiterate main issues arising from G02 programme related to possible **integration** of data sets from different laboratories. Note that methods/SOPs for generation and pre-processing of GC-MS data are already available in G02 reports from SCRI ad Aber which can be referred to.

- Compare spreadsheets containing GC-MS data describing composition of potato tubers from SCRI, Golm and Aber.

- Develop a standardized method for the description of a GC-MS peak which can be generated from all three instruments (LECO, Agilent and Thermo kit).

- Compare suggested methodology retrospectively (not metadata) to recommendations in Metabolomics Special Issue on standard reporting of analytical chemistry.

- Align data from all three instruments and generate a list of metabolites that would be expected to be measurable in a GC-MS profile of potato. Put information into ARMeC – i.e. develop a field indicating that compounds are measurable by GC-tof –MS.

3.3 METABOLITE PROFILING USING GAS CHROMATOGRAPHY MASS SPECTROMETRY (GC-MS)

3.3.1 Introduction to GC-MS

Gas chromatography-mass spectrometry (GC-MS) is a technique in which chromatographic separation of complex mixture of analytes occur in the gas phase by passing them through a long narrow-bore tube coated with a thin chemically bonded layer of a reactive polymer (stationary phase), in a flow of inert carrier gas (typically Helium). The analytes interact differentially with the stationary phase via ionic or Van der Waals type processes, migrating through the column at different rates, becoming spatially separated (resolved) with time. The analytes pass through a heated connection from the GC to the mass spectrometer (MS). In the ion source of the MS the analyte molecules are bombarded with high energy electrons with an industry-standard acceleration voltage of 70 eV. The collision process abstracts an electron from the molecule resulting in the formation of a radical cation (molecular ion) along with energy transfer to the molecular ion. Subsequent fragmentation of the molecular ion results in formation a series of daughter ions and neutral fragments the composition of which is dependent solely on the type and arrangement of atoms and types of bonding present within the original molecule. Further fragmentation of daughter ions may occur generating a cloud of ions and fragments. Only ions carrying a positive charge are then transferred to the mass analyser for characterisation. Four technologies are available for measurement of the mass to charge ratio (m/z) of the ions, namely sector instruments, quadropoles, ion traps and time of flight (TOF) instruments. The great majority of GC-MS based metobolomics studies use either the quadrupole or TOF technology platforms and the alternatives will not be considered further here.

3.3.2 Use of GC-MS in metabolomics studies

The basic approach for use of GC-MS for simultaneous analysis of metabolites in metabolomics studies of plant materials, as exemplified by potato, was described by Roessner *et al*., in 2000[1] and was used to phenotype genetically and environmentally modified plants[2].

The introduction of GC-TOF-MS analysis for analysis of highly complex plant extracts was hailed as a significant technological advance since it offered unparalleled ability for quantification and detection of several hundred metabolites in a single sample[3,4,5]. Subsequently use of the more sensitive TOF instrumentation has become the technology of choice in many laboratories[6-13]. The use of GC-MS in metabolomics studies has been the subject of a number of recent reviews[14, 15, 16].

3.3.3 Sample extraction and isolation and chemical derivatisation

Once metabolites have been extracted they must be chemically transformed to increase their volatility and to reduce their polarity prior to analysis by GC-MS. Although numerous methods have been developed over the years for targeted analysis of specific classes of metabolite by GC-MS, specific methodologies have arisen for analysis of the wider spectrum of metabolites encountered during metabolomics based studies. Most current methods are variations on a protocol which involves conversion of reducing sugars and other metabolites with reactive carbonyl groups into their respective methyl oximes by reaction with methoxyamine hydrochloride, prior to trimethylsilylation using *N*-methyl-*N*-trimethylsilyltrifluoroacetamide (MSTFA), in which acidic protons are exchanged for trimethylsilyl (TMS) groups[1, 11, 16-18] The oximation step locks reducing sugars in their open chain forms and prevents decarboxylation of α-ketoacids, whereas silylation reduces the polarity and increases the volatility and thermal stability of the derivatives. Oximation reactions of this type are sensitive to the effects of reaction time and temperature, as reported for oximation reactions using hydroxylamine hydrochloride[19] and methoxyamine hydrochloride[11] Therefore, to ensure the reaction goes to completion, experimental conditions should be thoroughly optimised, as otherwise significant amounts of unoximated sugars will also be present. Although some users of the methodology report that conditions were optimised there are relatively few published descriptions of this[11]. A drawback of the use of TMS derivatives is their moisture sensitivity, therefore they must be formed and handled under anhydrous conditions. The use of *tert*-butyldimethylsilyl (TBDMS) dervitives as an alternative to TMS has some advantages, the derivatives being more hydrolytically stable. In addition the increase in mass along with a propensity to have abundant fragment ions corresponding to loss of the *tert*-butyl group (*m/z* M-57) can help with structural assignment of metabolites[20, 21]. However, the increase in molecular mass associated with TBDMS derivatives renders them unsuitable for analysis of metabolites with multiple derivatisation sites such as carbohydrates, since they may not pass through the chromatography column or may exceed the mass range of the MS (typically 800-1000). Replacement of methyloxime derivatisation with ethyloxime derivatisation gives a more modest increase in mass (14 amu) and has proven useful for identification of the uncommon plant disaccharides inulobiose and levanbiose in transgenic potato[6, 15].

A further factor which should be considered with regard to the derivatisation methods is the stability of individual metabolites and derivatives under the reaction conditions. Known examples of metabolite breakdown and transformation under derivatisation conditions are conversion of arginine to ornithine by reaction with the silylating reagents BSTFA[21a] and MSTFA[16]. and oxidative degradatation of ascorbate and dehydroascorbate to 2,3-diketogulonic acid[11].

3.3.4 Mass spectral characteristics

For a given analyte mass spectrometers operating under electron impact (EI) conditions with an industry-standard acceleration voltage of 70 eV produce mass spectra with generally similar ion ratios, irrespective of the specific instrument design or manufacturer. On this basis, each separate component within a complex mixture potentially gives a unique and characteristic mass spectrum dependent solely on the type and arrangement of atoms and types of bonding present within the molecule giving rise to a statistically averaged series of fragmention processes and products.

In practice, several different but structurally related analytes, for example the hexose sugars, may give rise to essentially identical mass spectra, and these cannot be resolved on the basis of mass spectral characteristics alone.

Ideally, positive identification of a metabolite relies on isolation of a pure mass spectrum of that component. However, it is usually necessary to deconvolute signals arising from the component from the general chemical background, for example GC column bleed, and more significantly, from other analytes from which they are not chromatigraphically resolved. Various approaches can be taken for signal deconvolution. Software programs such as the freely available AMDIS examine changes in the abundance of individual ions over the whole mass range (typically 30-1000) in the time domain, and aggregate ions showing the same profiles for changes in ion abundance with time. This in theory can eliminate noise and other common background signals and provide a list of ions which constitute a library-searchable reconstructed mass spectrum for individual analytes, and provide information regarding the position of coeluting peak apices. In practice non sample-related fluctuations in signal levels can interfere with measurement of concentration-dependent variations in analyte ion abundance, particularly at low analyte concentrations, giving rise to ion mismatching and generation of false peak indications. Consequently the results of signal deconvolution require further filtering and verification[15].

A pure mass spectrum is one in which the ratios of the ion abundances to one another remain consistent with changes in analyte concentration over a chromatographic peak. Any factor which skews the measured ion abundance reduces the purity of the peak, and can result in misidentification of the analyte when compared with reference samples. Quadrupole instruments take a finite period of time to scan through their full mass range in which time the analyte concentration may have changed. Faster data sampling rates reduce this effect; modern quadrupoles achieve scan rates in the order of 10 scans s^{-1}, whereas TOF instruments can attain displayed rates of up to 500 spectra s^{-1}. The scan process also filters out a large proportion of all ions generated during ionisation since only ions of a specific m/z value are detected at any point in time. In comparison, ion separation and detection in TOF instruments is almost continuous, therefore each data point acquisition potentially carries much more information content. Consequently TOF instrument not only acquire data much faster than scanning instruments, they are also inherently more sensitive.

3.3.5 Retention characteristics

Mass spectral data alone is insufficient for characterisation and identification of analytes in complex mixtures by GC-MS. In addition some form of GC retention data is required as a second point of identity, usually presented initially as a retention time (R_t), corresponding to the peak apex for the specific metabolite. In addition, a retention index (RI) can be calculated for each component, relative to retention index markers included with each analysis[10, 11]. The RI value latter is very useful for inter-comparability of data since it accommodates local variations in analytical conditions provided the same type of GC column stationary phase is used and the analysis is done under similar conditions.

3.3.6 Retention indices and retention standards

Values for RI are interpolated from the retention characteristics of suitable RI marker compounds. These may be specific metabolites present in relatively high abundance within the sample. A more rigorous approach first described by Kovats[23] is to use exogenous retention time (Rt) standards, commonly long-chain alky compounds such as alkanes or fatty acid methyl esters (FAMES)[24] . These are run with analytical samples, either by addition of the chosen compounds to each sample, or by running reference samples (e.g blanks) containing the Rt standards regularly within sample batches. Each Rt standard is assigned a RI corresponding to its carbon chain length x 100 (e.g. Hexadecane or hexadecanoic methyl ester is given the value 1600). A sufficient number of Rt standards, with differing chain lengths (typically separated by C_3 to C_6 spacing), are selected to cover the duration of the chromatographic separation. Values of RI for metabolites within samples are calculated by linear interpolation based on retention times between adjacent Rt standards. Usually 6 to 8 retention standards are sufficient to characterize the retention characteristics. Some software tools such as the publicly available AMDIS and some commercial products such as ChromaTOF^{TM} (LECO, St Joseph, MI, USA) automatically calculate RI values although the latter can only use the data file formats produced by the company's GC-TOF-MS instruments. AMDIS on the other hand handles a wide range of data file formats. Other MS data analysis software packages such as Xcalibur^{TM} (Thermo Fisher, Hemel Heampstead, UK) use specific metabolites as retention markers within their data processing methods to correct for retention time variations, although they do not calculate and report RI values[11]. A further advantage of the use of retention markers, particularly when included in analytical sample blanks, is that they can help identify the development of problems with the chromatographic separation. For example, minor degradation of chromatographic columns can lead to peak splitting within a discrete region of the chromatogram, which can be readily seen in the chromatography of one or more of the RI markers. Such splitting can occur in chromatographic regions where several metabolites with similar mass spectra elute, and consequently can lead to difficulties in correct peak identification with the possibility of misidentification of split peaks as different metabolites.

3.4 COMPARISON OF SPREADSHEETS CONTAINING GC-MS DATA DESCRIBING COMPOSITION OF POTATO TUBERS FROM SCRI GOLM AND ABERYSTWYTH

Compositional data for metabolites present in potato tubers of potato cultivar Desiree from three separate GCMS studies conducted under G02006 at the Scottish Crop Research Institute (SCRI; SCRI 2002), University of Wales, Aberystwyth (UWA; Aber 2002) and the Max Planck Institute, Golm (MPI; Golm 2003) were compared.

Although these data were chromatographed, aligned and deconvoluted using different technology platforms and annotated using various peak annotation software (NIST, Wiley, Binbase), a metabolite list of both conserved primary and secondary metabolites were correspondingly identified by all three institutes (Appendix 3A).

The datasets were representative of similar methodologies developed and optimized at each centre, though they were all based on the method described by Roessner *et al*.[1] Furthermore, the data was representative of the differing analytical capabilities of instrumental platforms based on use of time of flight (TOF) and quadrupole (Quad) mass analyser technolgies within the GC-MS instrumentation used at the three sites (Table 3.1). The initial intention was to compare output from one Quad (UWA) and two TOF (SCRI, Golm) instruments, all products of different manufacturers. However the acquisition by SCRI of a quadrupole MS coupled to the same type of GC used on their TOF instrument provided a further opportunity for a comparison between TOF and Quad technologies. Furthermore the SCRI instruments used a Programmable Temperature Vapourising Injector (PTV) for sample introduction on their Thermo GCs whereas UWA and Golm used a hot injection technique on the Agilent GCs coupled to their respective MS. The main difference between the two approaches is that PTV injections are carried out at lower temperatures (*ca* 130°C) than the more conventional hot injections (*250 to* 280°C). The perceived benefit of lower injection temperatures is that the analystes are subject to less thermal shock, and are consequently less likely to undergo thermal degradation in the injector.

Table 3.1. Details of GC-MS instrumentation used at SCRI, UWA and Golm

Site	**MS Technology**	**Gas Chromatograph Manufacturer**	**Mass Analyser Manufacturer**
SCRI	Time Of Flight Quadrupole	Thermo Fisher Thermo Fisher	Thermo Fisher Thermo Fisher
UWA	Quadrupole	Agilent Technologies	Agilent Technologies
GOLM	Time Of Flight	Agilent Technologies	LECO Instruments

In Appendix 3A, the metabolite lists generated for potato by SCRI, UWA and Golm are presented in order of elution. When the lists were compared it was possible to determine the numbers of detected metabolites (Table 3.2, including those with known identities, those with putative identities and those classified as unknowns which were common to two or more of the lists (Table 3.2) identified for each instrument (Appendix 3B). In general more metabolites were measured using TOF instruments than quadrupoles, which is consistent with the greater sensitivity and faster data acquisition capabilities of the former over the latter. However less than 75% of the metabolites listed for each TOF were common to both, and the commonality between Quads was less than 71%. When the Quads were compared with TOF instruments, the greatest overlap in metabolite lists occurred between instruments using the same GC (SCRI vs SCRI; UWA vs Golm). This suggests that factors other than just sensitivity and data acquisition rates influence the specificity of metabolite detection, but that other factors relating to the GC component of the systems have a significant impact. All three sites used GC columns with similar but non-identical non-polar stationary phases. Under similar sample loading conditions this should not give rise to changes in metabolite specificity, bur rather may cause minor differences in the order of metabolite elution, as is indeed the case. It is more likely that such discrimination is related to the difference in injection technique mentioned previously. The 'core' metabolites common to peak lists in Table 3.1 are listed in Appendix 3Band are typical of those previously published for potato tubers.[1, 2, 10-13, 22] Among the metabolites more readily detected using TOF instrumentsation are phosphorylated sugars, several sugar acids and alcohols, numerous unidentified polysaccharides, isomers of chlorogenic acid, pantothenic acid, and the purine degradation product allantoin.

Table 3.2 Comparison of numbers of metabolites common to metabolite lists generated using quadrupole instruments at SCRI and UWA and TOF at SCR and Golm

	Metabolites (total)	SCRI (Quad)	SCRI (TOF)	GOLM (TOF)	UWA (Quad)
Metabolites (total)		80	114	109	72
SCRI (Quad)	80	-	80	71	51
SCRI (TOF)	114	80	-	75	53
GOLM (TOF)	109	71	75	-	57
UWA (Quad)	72	51	53	57	-
Common to all	46				

3.5 STANDARDISED METHOD FOR DESCRIPTION OF A GC-MS PEAK GENERATED FROM THE DIFFERENT MANUFACTURERS GC-MS INSTRUMENTATION (LECO, AGILENT, THERMO)

Specific information can be defined which serves to identify and characterize a given metabolite (Figure 3.1). For each component, information falls into two categories, based on chromatographic retention characteristics and mass spectrometric (MS) characterisation. For each metabolite it is possible, in principle, to extract from the raw data a characteristic mass spectrum. Ions with a specific mass (m/z), or in some cases, several such ions, are used in conjunction with retention information for identification and quantification of the metabolite during initial data analysis using the appropriate data processing method. The primary and ancillary descriptors are shown diagrammatically in Figure 3.1. Additional descriptors may also be used, e.g., molecular formulae and molecular weight (M).

3.5.1 Method for describing GC-MS peaks

The following criteria were established as forming the basis for a standardised method for description of a GC-MS peak irrespective of technology platform used for data acquisition.

Having conducted any necessary peak picking and signal cleanup/deconvolution:

1. Identify the Rt of the peak apex (in seconds).

2. Calculate a retention index (RI) based either on addition of exogenous Rt standards (alkanes, FAMES) or using selected marker metabolites within the analytical sample. The RI compounds used should be specified.

3. Identify ions which can serve to characterise the component, and report their absolute and relative abundances. These should include up to 20 qualifier ions preferably to cover the observed mass range of the component but which should include the most abundant ions. It may be advantageous to include molecular ions and/or those corresponding to low mass losses (e.g. *m/z* M-15 etc) if they are considered to have diagnostic value. Identify ion(s) to be used for quantification purposes.

4. Characterise a level of identification.

- Confirmed identification: by comparison with the chromatographic and mass spectral properties of reference standards
- Putative identification: compounds for which no standards are available, but for which there is strong physicochemical evidence and/or match with entries mass spectral libraries and databases
- Unknown identification but which exhibit mass spectral properties suggesting that they belong to a specific class or group of metabolites.
- Unknowns identification.

Although it may not be possible to identify unknown metabolites it should be possible to characterise them on the basis of Rt/RI and mass spectra including the m/z and abundance of the 20 qualifier ions.

Appropriate nomenclature should be used ranging from common names, IUPAC names and laboratory specific identifier for unknowns. In the latter case this could perhaps be based on RRI values.

3.5.2 Comparison of criteria for peak description with the recommendations of the Chemical Analysis Working Group (CAWG) of the Metabolomics Standards Initiative

The Chemical Analysis Working Group (CAWG) of the Metabolomics Standards Initiative has made a number of recommendations regarding minimum reporting standards for chemical analysis[30]. Although there is still an on-going debate as to what constitutes valid metabolite identification, on the basis of current practice in the metabolomics literature four levels of identification were defined.

1. Identified compounds.
2. Compounds with putative annotation based on physicochemical properties and/or similarity of mass spectra with mass spectral libraries, but without matching data for chemical reference standards.
3. Putative characterisation as belonging to a specific class of compound on the basis of mass spectral and physicochemical properties.
4. Unidentified compounds which can be differentiated and quantified on the basis of their mass spectral data but which remain uncharacterised.

For positive identification of a compound the standard requirement of provision of a minimum of two independent and orthogonal points of identity as required for publication in the chemical literature has been adopted. In the context of GC-MS based studies these include Rt or RI values with a mass spectrum, and accurate mass in combination with tandem MS and/or isotopic composition. Such requirement suffice for non-novel compounds, however for novel metabolites a higher degree of rigour is required with at least an additional orthogonal point of identity. Ideally, the traditional approach of isolation, purification and physicochemical characterisation (elemental analysis, NMR absorption spectra etc.) should be followed.

The CAWG also made a number of recommendations regarding naming of metabolites, and for known metabolites their recommendation was to report one chemical name (common or IUPAC) along with one structural code for which the recommendation is to use the IUPAC International Chemical Identifer (InChI: http://inchi.info/). For first naming of novel metabolites the IUPAC nomenclature rules should be followed, preferably with issue of structural codes and submission of novel structures to the PubChem compound identifier (CIC; http://pubchem.ncbi.nlm.nih.gov/). For unknown compounds the recommendation is to report a combination of Rt, RI and/or prominent ions in the mass spectrum along with MS-MS where available. A similar approach which also includes associated metadata for the methodology use to generate the unknown has been proposed by Bino et al.[31]

These recommendations are very similar to the methodology for peak description we have proposed above.

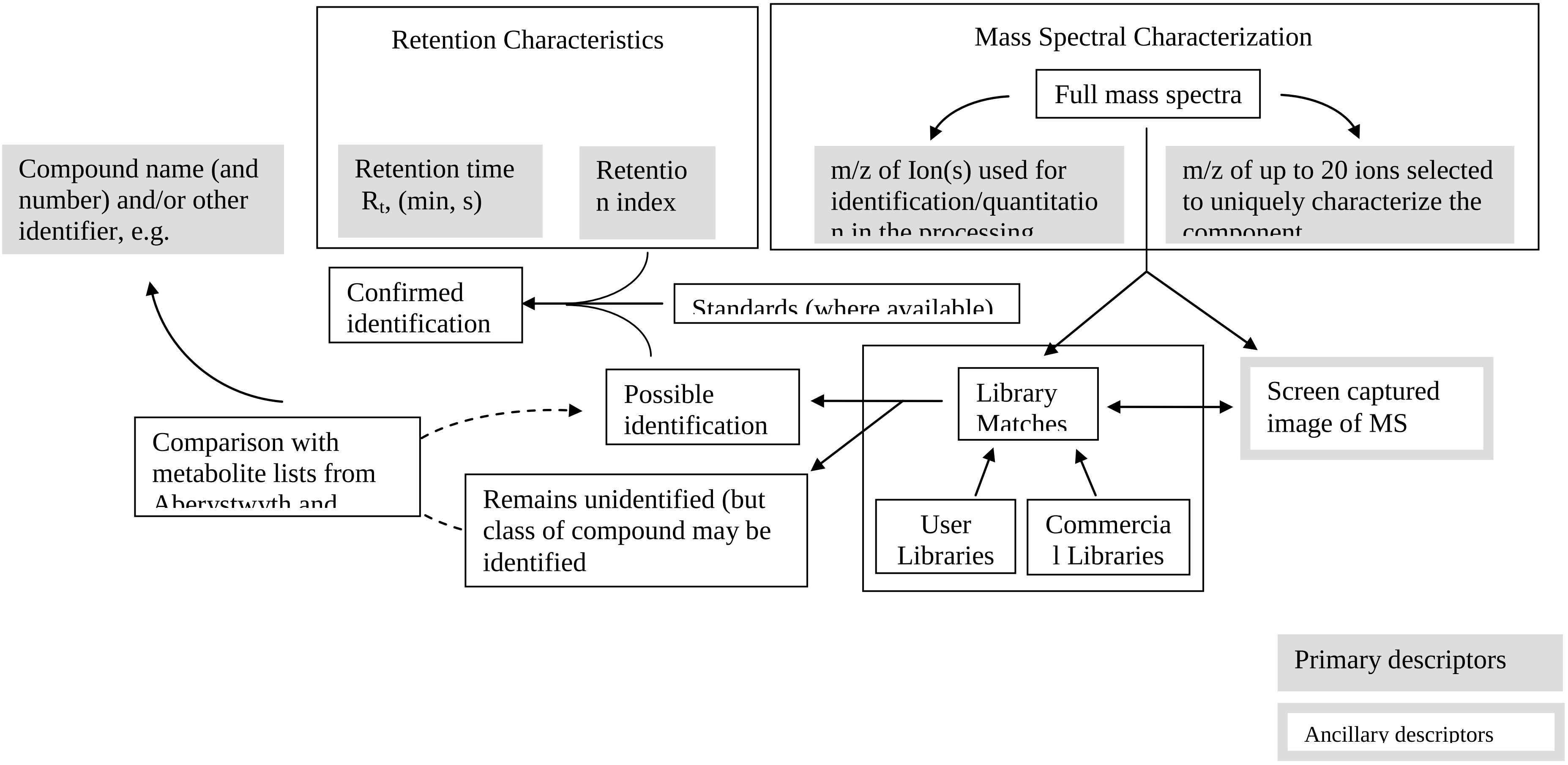

Retention Characteristics
Mass Spectral Characterization
Full mass spectra
Compound name (and number) and/or other identifier, e.g.
Retention time R_t, (min, s)
Retentio n index
m/z of Ion(s) used for identification/quantitatio n in the processing
m/z of up to 20 ions selected to uniquely characterize the component
Confirmed identification
Standards (where available)
Possible identification
Library Matches
Screen captured image of MS
Comparison with metabolite lists from Aberystwyth and
Remains unidentified (but class of compound may be identified
User Libraries
Commercia l Libraries
Primary descriptors
Ancillary descriptors

Figure 3.1 Descriptors used for characterisation of metabolites following GC-MS analysis

3.6 ALIGNMENT OF DATA FROM THE DIFFERENT INSTRUMENTS AND DEFINE A LIST OF METABOLITES THAT WOULD BE EXPECTED TO BE MEASURABLE IN THE GC-MS PROFILE OF POTATO

In Appendix 3A, the metabolite lists generated for potato by SCRI, UWA and Golm are presented in order of elution, and have been aligned as closely as possible. The approach taken to co-alignment was as follows.

1. Align major and minor metabolites of known or putative identity. This is straightforward.
2. Align major metabolites of unknown identity. This is also relatively easy; some unknowns in potato are present in relatively high abundance and are easy to distinguish.
3. Attempt to align minor metabolites of unknown identity. This is difficult but use of 20 ion lists to identify unique qualifier ions can help.

There is a high level of co-alignment for the majority of the metabolites with known identities. A number of metabolites show variation in elution order between the three lists, which is not unexpected and reflects differences between the non polar stationary phase chemistries of the GC columns used at each of the three sites. It is well known that subtle differences in the composition of essentially similar GC column stationary phases can lead to changes in the elution order of analytes.

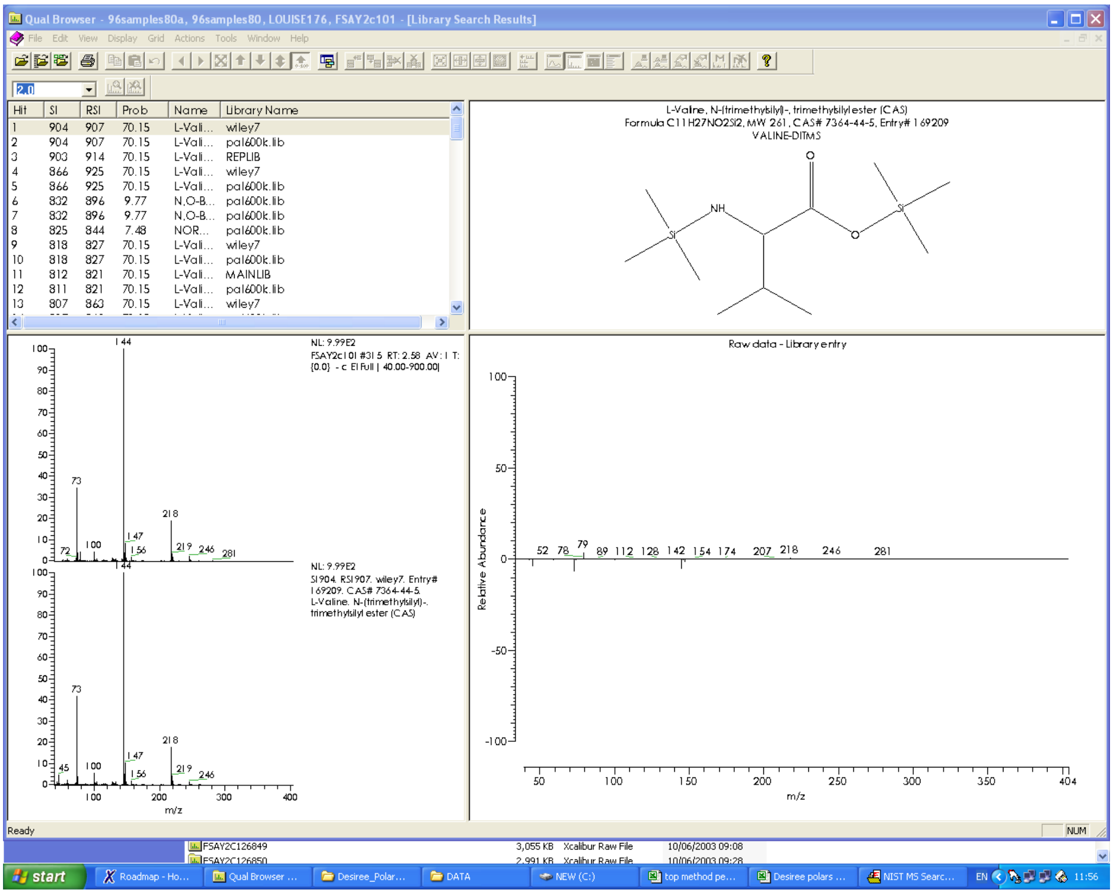
Two lists of core metabolites which would be expected to be measurable in a GC-MS profiling experiment using potato tubers are shown in Appendix 3B. The column on the left lists 65 metabolites which should be detected using TOF instrumentation, whereas that on the right lists 41 metabolites which should be detectable using a quadrupole instrument. Both lists include multiple derivatives of some metabolites either having different numbers of TMS substituents (asparagine, glutamine and allantoin) or different positional isomers (methyl oximes of some reducing sugars). With expected improvements in instrumental sensitivity as new products become available from the instrument manufacturers, and wider use is made of techniques such as GCxGC-MS, it is likely that the range of expected core metabolites will widen in both cases.

In addition to use of mass spectral libraries (user and commercial) for compound characterisation we found that generation of a hard copy database of screen captured images of mass spectra of analytes and reference standards was very useful for helping with metabolite recognition and alignment. Examples of some screen captures of extracted MS information are shown in Figures 3.2-3.4. Figure 3.2 shows a library match for a metabolite positively identified as the amino acid L-Valine, which has been confirmed by analysis of a standard. Figure 3.3 shows MS of two previously unidentified components, which have now been tentatively identified as derivatives of allantoin, a degradation product of purines, following screening with the Palisade 600K library. Figures 3.4 & 3.5 shows how the MS of a minor component forming part of a co-eluting mixture has been extracted following subtraction of the MS of a nearby compound which has a similar MS to the other component in the mixture.

Extracted MS from analytical samples and standards such as those illustrated above can also be incorporated into local user-generated MS libraries

3.7 CONCLUSIONS

At the end of the G02006 project GC-MS analysis had been performed on similar extracts representing a number of common potato cultivars. Although the analyses were performed on different instruments in 3 different laboratories we have shown that by and large a core set of metabolites can be aligned between these data sets with good reproducibility. This information will be useful in the future in order to develop a 'gold standard' list of expected metabolites to be measured in any new potato varieties in any screens for unusual concentration distributions.



Qual Browser - 96samples80a, 96samples80, LOUISE176, FSAY2c101 - [Library Search Results]

Hit	SI	RSI	Prob	Name	Library Name
1	904	907	70.15	L-Vali...	wiley7
2	904	907	70.15	L-Vali...	pal600k.lib
3	903	914	70.15	L-Vali...	REPLIB
4	866	925	70.15	L-Vali...	wiley7
5	866	925	70.15	L-Vali...	pal600k.lib
6	832	896	9.77	N,O-B...	pal600k.lib
7	832	896	9.77	N,O-B...	pal600k.lib
8	825	844	7.48	NOR...	pal600k.lib
9	818	827	70.15	L-Vali...	wiley7
10	818	827	70.15	L-Vali...	pal600k.lib
11	812	821	70.15	L-Vali...	MAINLIB
12	811	821	70.15	L-Vali...	pal600k.lib
13	807	863	70.15	L-Vali...	wiley7

L-Valine, N-(trimethylsilyl)-, trimethylsilylester (CAS)
Formula C11H27NO2Si2, MW 261, CAS# 7364-44-5, Entry# 169209
VALINE-DITMS

NL: 9.99E2
FSAY2c101 #315 RT: 2.58 AV: 1 T: {0.0} - c EI Full | 40.00-900.00]

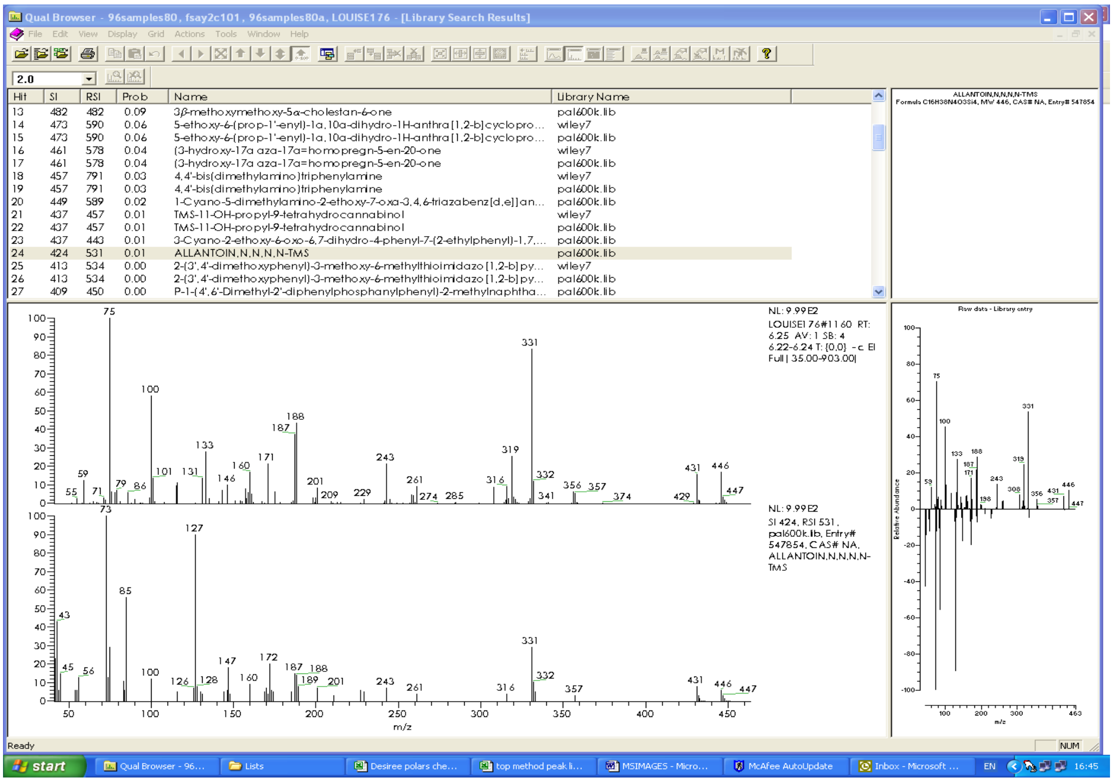
NL: 9.99E2
SI 904. RSI 907. wiley7. Entry# 169209. CAS# 7364-44-5. L-Valine. N-(trimethylsilyl)-. trimethylsilyl ester (CAS)

Raw data - Library entry

Ready

FSAY2C126849	3,055 KB	Xcalibur Raw File	10/06/2003 09:08
FSAY2C126850	2,991 KB	Xcalibur Raw File	10/06/2003 09:28

Figure 3.2 Screen capture of MS of L-Valine following library search against the Wiley 7 library



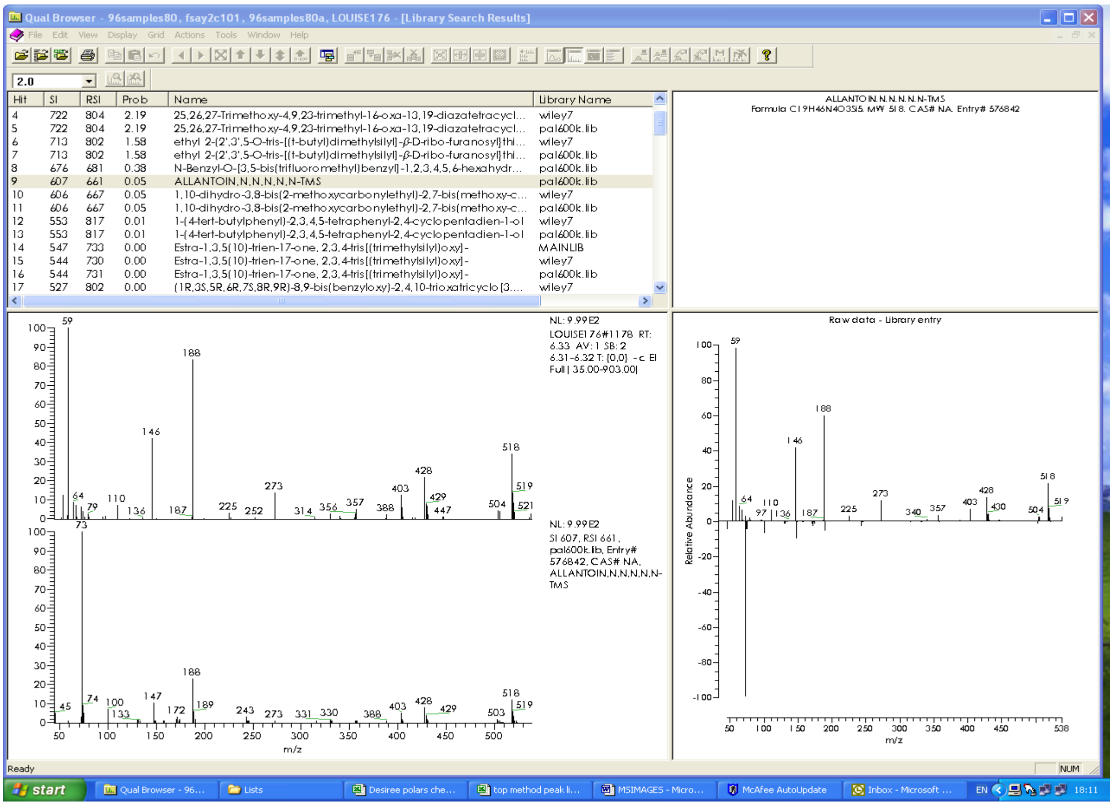
Qual Browser - 96samples80, fsay2c101, 96samples80a, LOUISE176 - [Library Search Results]

Hit	SI	RSI	Prob	Name	Library Name
13	482	482	0.09	3β-methoxymethoxy-5α-cholestan-6-one	pal600k.lib
14	473	590	0.06	5-ethoxy-6-(prop-1'-enyl)-1a,10a-dihydro-1H-anthra[1,2-b]cyclopro...	wiley7
15	473	590	0.06	5-ethoxy-6-(prop-1'-enyl)-1a,10a-dihydro-1H-anthra[1,2-b]cyclopro...	pal600k.lib
16	461	578	0.04	(3-hydroxy-17a aza-17a=homopregn-5-en-20-one	wiley7
17	461	578	0.04	(3-hydroxy-17a aza-17a=homopregn-5-en-20-one	pal600k.lib
18	457	791	0.03	4,4'-bis(dimethylamino)triphenylamine	wiley7
19	457	791	0.03	4,4'-bis(dimethylamino)triphenylamine	pal600k.lib
20	449	589	0.02	1-Cyano-5-dimethylamino-2-ethoxy-7-oxa-3,4,6-triazabenz[d,e]]an...	pal600k.lib
21	437	457	0.01	TMS-11-OH-propyl-9-tetrahydrocannabinol	wiley7
22	437	457	0.01	TMS-11-OH-propyl-9-tetrahydrocannabinol	pal600k.lib
23	437	443	0.01	3-Cyano-2-ethoxy-6-oxo-6,7-dihydro-4-phenyl-7-(2-ethylphenyl)-1,7,...	pal600k.lib
24	424	531	0.01	ALLANTOIN,N,N,N,N-TMS	pal600k.lib
25	413	534	0.00	2-(3',4'-dimethoxyphenyl)-3-methoxy-6-methylthioimidazo[1,2-b]py...	wiley7
26	413	534	0.00	2-(3',4'-dimethoxyphenyl)-3-methoxy-6-methylthioimidazo[1,2-b]py...	pal600k.lib
27	409	450	0.00	P-1-(4',6'-Dimethyl-2'-diphenylphosphanylphenyl)-2-methylnaphtha...	pal600k.lib

ALLANTOIN,N,N,N,N-TMS
Formula C16H38N4O3Si4, MW 446, CAS# NA, Entry# 547854

NL: 9.99E2
LOUISE176#1160 RT: 6.25 AV: 1 SB: 4 6.22-6.24 T: {0,0} - c EI Full | 35.00-903.00]

NL: 9.99E2
SI 424, RSI 531, pal600k.lib, Entry# 547854, CAS# NA, ALLANTOIN,N,N,N,N-TMS



Qual Browser - 96samples80, fsay2c101, 96samples80a, LOUISE176 - [Library Search Results]

Hit	SI	RSI	Prob	Name	Library Name
4	722	804	2.19	25,26,27-Trimethoxy-4,9,23-trimethyl-16-oxa-13,19-diazatetracycl...	wiley7
5	722	804	2.19	25,26,27-Trimethoxy-4,9,23-trimethyl-16-oxa-13,19-diazatetracycl...	pal600k.lib
6	713	802	1.58	ethyl 2-(2',3',5-O-tris-[(t-butyl)dimethylsilyl]-β-D-ribo-furanosyl}thi...	wiley7
7	713	802	1.58	ethyl 2-(2',3',5-O-tris-[(t-butyl)dimethylsilyl]-β-D-ribo-furanosyl}thi...	pal600k.lib
8	676	681	0.38	N-Benzyl-O-[3,5-bis(trifluoromethyl)benzyl]-1,2,3,4,5,6-hexahydr...	pal600k.lib
9	607	661	0.05	ALLANTOIN,N,N,N,N,N-TMS	pal600k.lib
10	606	667	0.05	1,10-dihydro-3,8-bis(2-methoxycarbonylethyl)-2,7-bis(methoxy-c...	wiley7
11	606	667	0.05	1,10-dihydro-3,8-bis(2-methoxycarbonylethyl)-2,7-bis(methoxy-c...	pal600k.lib
12	553	817	0.01	1-(4-tert-butylphenyl)-2,3,4,5-tetraphenyl-2,4-cyclopentadien-1-ol	wiley7
13	553	817	0.01	1-(4-tert-butylphenyl)-2,3,4,5-tetraphenyl-2,4-cyclopentadien-1-ol	pal600k.lib
14	547	733	0.00	Estra-1,3,5(10)-trien-17-one, 2,3,4-tris[(trimethylsilyl)oxy]-	MAINLIB
15	544	730	0.00	Estra-1,3,5(10)-trien-17-one, 2,3,4-tris[(trimethylsilyl)oxy]-	wiley7
16	544	731	0.00	Estra-1,3,5(10)-trien-17-one, 2,3,4-tris[(trimethylsilyl)oxy]-	pal600k.lib
17	527	802	0.00	(1R,3S,5R,6R,7S,8R,9R)-8,9-bis(benzyloxy)-2,4,10-trioxatricyclo[3....	wiley7

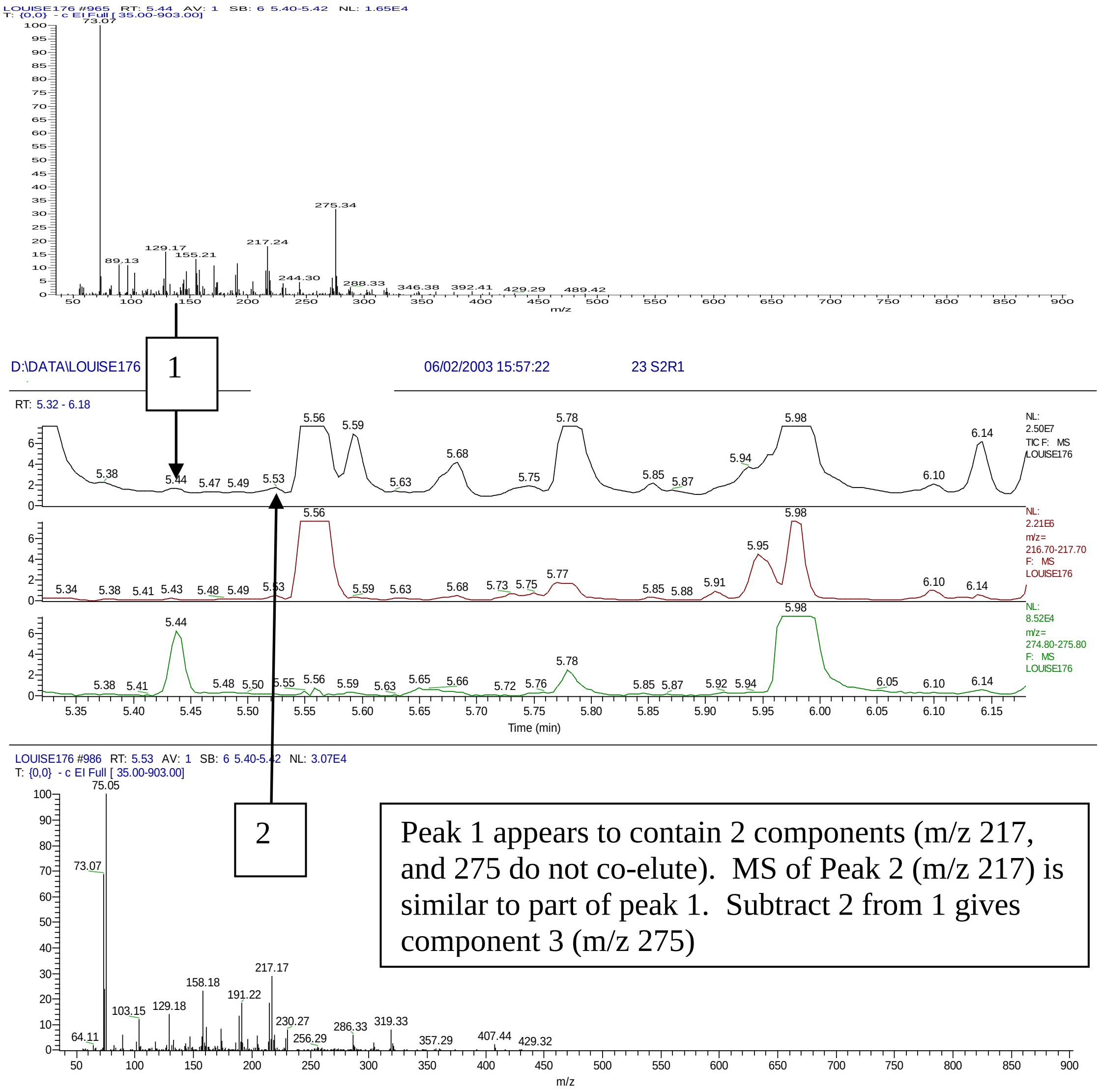
ALLANTOIN.N.N.N.N.N-TMS
Formula C19H46N4O3Si5. MW 518. CAS# NA. Entry# 576842

NL: 9.99E2
LOUISE176#1178 RT: 6.33 AV: 1 SB: 2 6.31-6.32 T: {0,0} - c EI Full | 35.00-903.00]

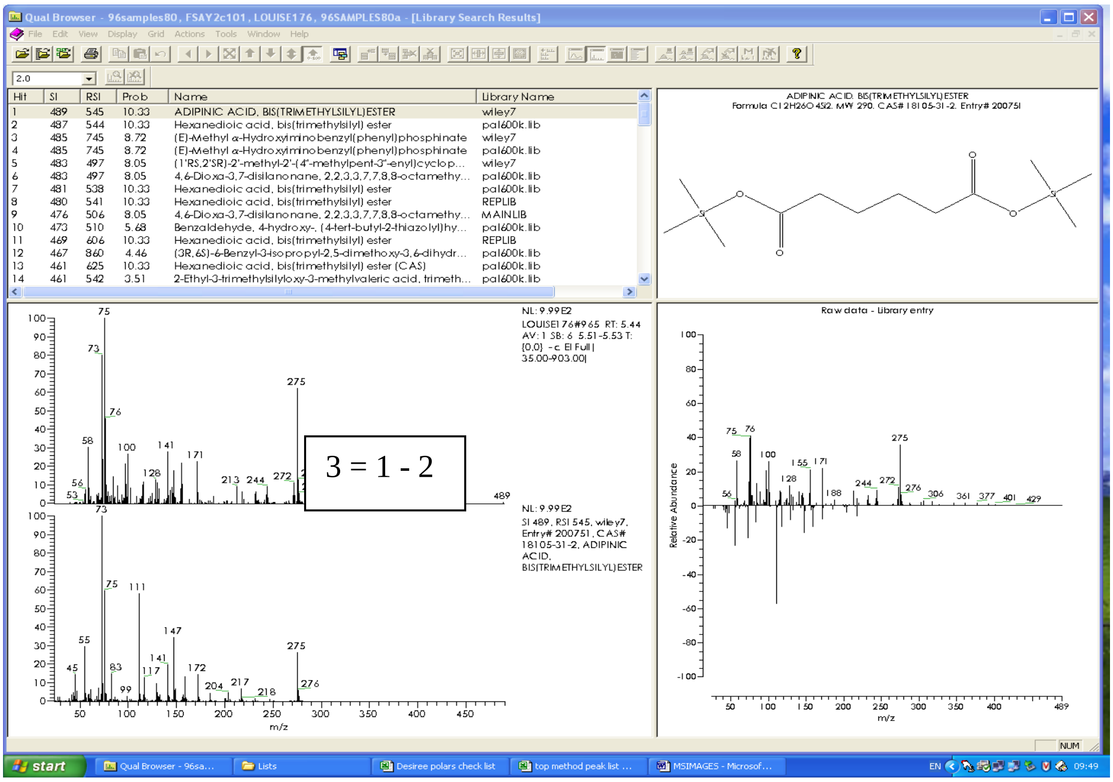
NL: 9.99E2
SI 607, RSI 661, pal600k.lib, Entry# 576842, CAS# NA, ALLANTOIN,N,N,N,N,N-TMS

Figure 3.3 Screen capture of MS of two unknowns following library search against the Palisade 600K library. Preliminary identification as tetra and penta TMS derivatives of Allantoin.

Figure 3.4



Peak 1 appears to contain 2 components (m/z 217, and 275 do not co-elute). MS of Peak 2 (m/z 217) is similar to part of peak 1. Subtract 2 from 1 gives component 3 (m/z 275)



Qual Browser - 96samples80, FSAY2c101, LOUISE176, 96SAMPLES80a - [Library Search Results]
File
Edit
View
Display
Grid
Actions
Tools
Window
Help
2.0

Hit	SI	RSI	Prob	Name	Library Name
1	489	545	10.33	ADIPINIC ACID, BIS(TRIMETHYLSILYL)ESTER	wiley7
2	487	544	10.33	Hexanedioic acid, bis(trimethylsilyl) ester	pal600k.lib
3	485	745	8.72	(E)-Methyl α-Hydroxyiminobenzyl(phenyl)phosphinate	wiley7
4	485	745	8.72	(E)-Methyl α-Hydroxyiminobenzyl(phenyl)phosphinate	pal600k.lib
5	483	497	8.05	(1'RS,2'SR)-2'-methyl-2'-(4"-methylpent-3"-enyl)cyclop...	wiley7
6	483	497	8.05	4,6-Dioxa-3,7-disilanonane, 2,2,3,3,7,7,8,8-octamethy...	pal600k.lib
7	481	538	10.33	Hexanedioic acid, bis(trimethylsilyl) ester	pal600k.lib
8	480	541	10.33	Hexanedioic acid, bis(trimethylsilyl) ester	REPLIB
9	476	506	8.05	4,6-Dioxa-3,7-disilanonane, 2,2,3,3,7,7,8,8-octamethy...	MAINLIB
10	473	510	5.68	Benzaldehyde, 4-hydroxy-, (4-tert-butyl-2-thiazolyl)hy...	pal600k.lib
11	469	606	10.33	Hexanedioic acid, bis(trimethylsilyl) ester	REPLIB
12	467	860	4.46	(3R,6S)-6-Benzyl-3-isopropyl-2,5-dimethoxy-3,6-dihydr...	pal600k.lib
13	461	625	10.33	Hexanedioic acid, bis(trimethylsilyl) ester (CAS)	pal600k.lib
14	461	542	3.51	2-Ethyl-3-trimethylsilyloxy-3-methylvaleric acid, trimeth...	pal600k.lib

ADIPINIC ACID, BIS(TRIMETHYLSILYL)ESTER
Formula C12H26O4Si2 MW 290 CAS# 18105-31-2 Entry# 200751

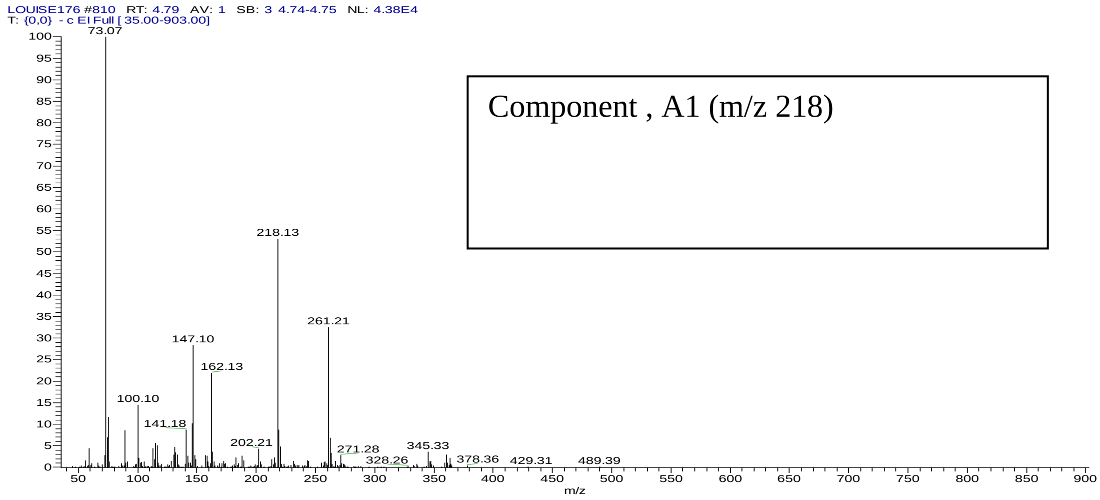
NL: 9.99E2
LOUISE176#965 RT: 5.44
AV: 1 SB: 6 5.51-5.53 T: {0,0} - c EI Full | 35.00-903.00|

NL: 9.99E2
SI 489, RSI 545, wiley7, Entry# 200751, CAS# 18105-31-2, ADIPINIC ACID, BIS(TRIMETHYLSILYL)ESTER

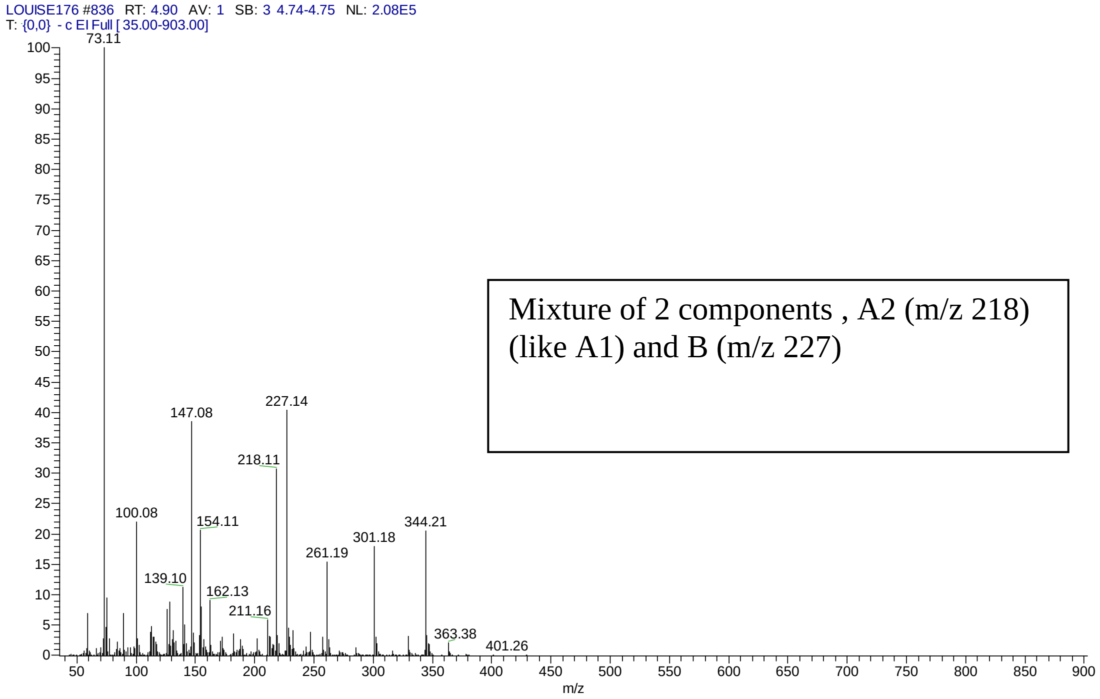
Raw data - Library entry

3 = 1 - 2

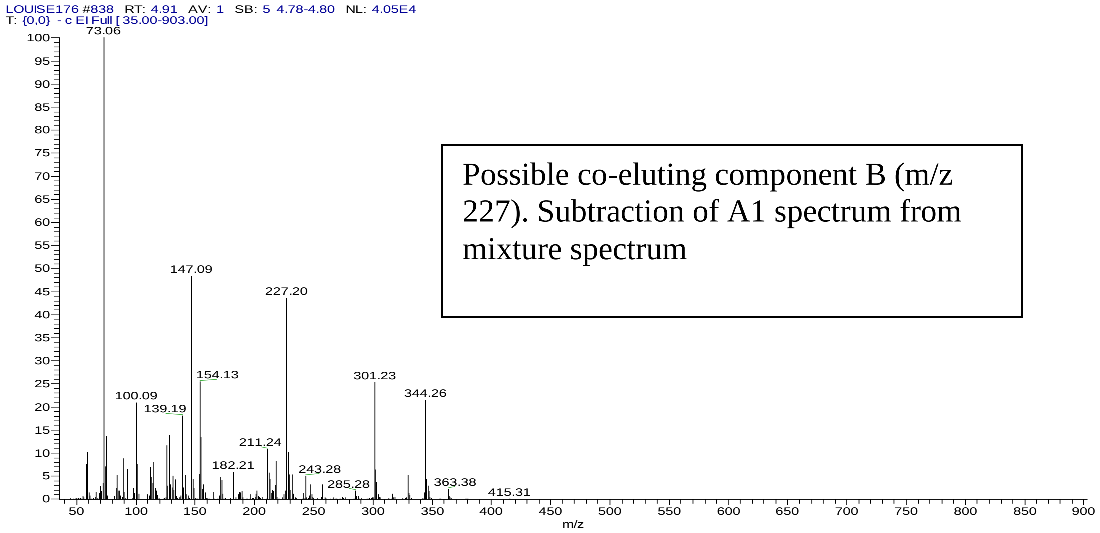
Figure 3.5



Component , A1 (m/z 218)



Mixture of 2 components , A2 (m/z 218) (like A1) and B (m/z 227)



Possible co-eluting component B (m/z 227). Subtraction of A1 spectrum from mixture spectrum

REFERENCES

1. Roessner, U., Wagner, C., Kopka, J., Trethewey, R.N. & Willmitzer, L. Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant Journal* **23**, 131-142 (2000).

2. Roessner, U., Luedemann, A., Brust, D., Fiehn, O., Linke, T., Willmitzer, L. and Fernie, A.R. Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* **13,** 11-29 (2001).

3. Weckworth, W., Tolstikov, V. & Fiehn, O. Metabolomic characterisation of transgenic potato plants using GC/TOF and LC-MS analysis reveals silent metabolic phenotypes. *Proceedings of the 49th ASMS Conference on Mass Spectrometry and Allied Topics*, American Society of Mass Spectrometry, Chicago, pp. 1-2 (2001).

4. Weckworth, W., Loureiro M.E., Wenzel, K. & Fiehn, O. Differential metabolic networks unravel the effects of silent plant phenotypes Proc. *Natl. Acad. Sci. U S A* **101**, 7809-7814 (2004).

5. Weckworth, W., Wenzel, K. & Fiehn, O. 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 (2004).

6. Catchpole, G.S. et al. Hierarchical metabolomics demonstrates substantial compositional similarity between genetically modified and conventional potato crops. *Proc. Natl. Acad. Sci. U S A* **102**, 14458-14462 (2005).

7. Fiehn, O. Metabolomics - the link between genotypes and phenotypes. *Plant Molecular Biology* **48**, 155-171 (2002).

8. Lisec, J., Schauer, N., Kopka, J., Willmitzer, L. & Fernie, A.R. Gas chromatography mass spectrometry–based metabolite profiling in plants. *Nature Protocols* **1**, 387-396 (2006).

9. Taylor, J., King, R.D., Altmann, T. & Fiehn, O. Application of metabolomics to plant genotype discrimination using statistics and machine learning. *Bioinformatics* **18**, S241-S248 (2002).

10. Wagner, C., Sefkow, M. & Kopka, J. Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. *Phytochemistry* **62**, 887-900 (2003).

11. Shepherd, T., Dobson, G., Verrall, S. R., Conner, S., Griffiths, D. W., McNicol, J. W., Davies, H. V. & Stewart D. Potato metabolomics by GC-MS: what are the limiting factors? *Metabolomics* **3**, 475-488 (2007).

12. Shepherd, T., Dobson, G., Marshall, R., Verrall, S. R., Connor, S., Griffiths, D. W., Stewart, D. & Davies, H. V. Profiling of Metabolites and volatile flavour compounds

from solanum species using gas chromatography-mass spectrometry. In Nikolau B. L & Wurtele, E.S. (eds) *Concepts in Plant Metabolomics*, Springer, Dordrecht, The Netherlands, 208-219 (2007). .

13. Dobson, G., Shepherd, T., Marshall, R., Verrall, S. R., Connor, S., Griffiths, D. W., McNicol, J. W., Stewart, D. & Davies, H. V. Application of metabolite and flavour volatile profiling to studies of biodiversity in solanum species. In Nikolau B. L & Wurtele, E.S. (eds) *Concepts in Plant Metabolomics*, Springer, Dordrecht, The Netherlands, 259-268 (2007). .

14. Bedair M., & Sumner L. W. Current and emerging mass-spectrometry technologies for metabolomics. *Trends in Analytical Chemistry* **27**, 238-249 (2008).

15. Fiehn O. Extending the breadth of metabolite profiling by gas chromatography coupled to mass spectrometry. *Trends in Analytical Chemistry* **27**, 261-269 (2008).

16. Halkett, J. M., Waterman, D., Przyborowska, A. M., Patel, R. K. P., Fraser, P. D. & Bramley, P.M. Chemical derivatization and mass spectral librarties in metabolic profiling by GC/MS and LC/MS/MS. *J. Exp. Bot*. **56**, 219-243 (2005).

17. Schweer, H. Gas chromatography-mass spectrometry of aldoses as o-methoxime, o-2-methyl-2-propoxime and o-n-butoxime pertrifluroacetyl derivatives on OV-225 with methylpropane as ionisation agent. *J. Chromatogr*. **236**, 355-360 (1982)

18. Katona, Zs. F., Sass, P. & Molnar-Perl, I. Simultaneous determination of sugars, sugar alcohols, acids and amino acids in apricots by gas chromatography-mass spectrometry. *J. Chromatogr. A* **847**, 91-102 (1999).

19. Adams, M. A., Chen, ZL., Landman, P. & Colmer, T. D. Simultaneous determination by capillary gas chromatography of organic acids, sugars, and sugar alcohols in plant tissue extracts as their trimethylsilyl derivatives. *Anal. Biochem*. **266**, 77-88 (1999).

20 Fiehn, O., Kopka, J., Trethewy, R.N., Willmitzer, L. Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. Anal. Chem. 72, 3373-3580 (2000).

21. Birkenmeyer, C., Kolasa, A. & Kopka, J. Comprehensive chemical derivatization for gas chromatography-mass spectrometry-based multi-targeted profiling of the major phytohormones. *J. Chrom. A* **993**, 89-102.

21b Leimer K. R., Rice, R. H. & Gehrke, C. W. Complete mass spectra of the per-trimethylsilylated amin acids. *J. Chromatogr. A* **141**, 355-375 (1977).

22. Beckmann, M., Enot, D. P., Overy, D. P. & Draper J. Representation, comparison, and interpretation of metabolome fingerprint data for total composition analysis and quality trait investigation in potato cultivars. J. Agric. Food Chem. (2007).

23. Kovats E. Gas chromatographische charakteriserung organischer verbindungen. I. Retentions indices aliphatischer halogenide, alkohole, aldehyde und ketone. *Helvetica Chimica Acta* **41**, 1915-1932 (1958).

24. Morgenthal, K., Wienkoop, S., Scholz, M., Selbig, J. & Weckwerth, W. Correlative GC-TOF-MS based metabolite profiling and LC-MS-based protein profiling reveal time-related systemic regulation of metabolite-protein networks and improve pattern recognition for multiple biomarker selection. *Metabolomics* **1**, 109-121 (2005).

25. Stein , S. E. An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *J. Am. Soc. Mass Spectrom*. **10**, 770-781 (1999)

30. Sumner, L.W. et al. Proposed minimum reporting standards for chemical analysis. *Metabolomics* **3**, 211-221 (2007).

31. Bino, R.J. et al. Potential of metabolomics as a functional genomics tool. *Trends in Plant Science* **9**, 418-425 (2004).

Appendix 3A Metabolite peak list

SCRI annotations			TOF	Quad	Aber annotations			Golm 2003 Binbase selections		
	Charac.		Apex	Apex		Charac.	Apex		Charac.	873080
Metabolite	m/z	RRI	Rt (s)	Rt (s)	Metabolite	m/z	Rt (s)	Metabolite	m/z	RI
L-alanine (TMS)2	116	1095	1.49	1.12	Alanine	116	4.31	alanine	116	244920
					Hydroxylamine	133	4.43	hydroxylamine	133	255690
like 2-Aminobutyric acid (TMS)2	130	1169	2.14	1.72	U-080	130	4.78	isobutyric acid 2-amino	130	286860
U1174	241	1174	2.19	1.75	U-081	241	4.79			
Urea(TMS)3	261	1177	2.21							
					U-082	86	4.85	y2645	86	295330
L-valine (TMS)2	144	1216	2.55	2.18	Valine	144	5.05	valine	144	313700
					U-110	278	5.17	y467	278	325317
Unknown like Ethanolamine (TMS3	174	1233	2.69	2.31						
urea (TMS)2	189	1244	2.78	2.42				urea	117, 189	327700
Ethanoamine (TMS)3	174	1266	2.99	2.62	Ethanolamine	174	5.36			
Phosphate (TMS)3	299	1270	3.02	2.67	Phosphate	299	5.35	glycerol	147, 205	343570
leucine (TMS)2	158	1272	3.03	2.68	Glycerol	205	5.36	phosphoric acid	299	344710
glycerol (TMS)3	205	1275	3.06	2.71	Leucine	158	5.37	leucine	158	345890
L-isoleucine-(TMS)2	158	1291	3.2	2.85	Isoleucine	158	5.49	isoleucine	158	358940
L-proline (TMS)2	142	1293	3.23	2.87	Proline	142	5.54	proline	142	364640
glycine (TMS)3	174	1300	3.29	2.94	Glycine	174	5.57	glycine	174	368140
succinic acid (TMS)2	247	1315	3.37	*3.02*						
2,3-dihydroxypropanoic-acid (TMS)3	189	1333	3.48	3.13				glyceric acid	147, 189	376660
fumaric acid (TMS)2	245	1359	3.63	3.30	Fumaric_acid	245	5.78	fumaric acid	147, 245	390730
L-serine (TMS)3	204	1366	3.69	3.35	Serine_main	204	5.82	serine	204	394140
2-Piperidinecarboxylic acid (TMS)2	156	1369	3.70	3.35						
								cycloleucine	156	402890
U1376	141	1376	3.75	3.45	U-083	141	5.89	y487	141	403640
dihydroxydihydrofuranone (TMS)2 (lactone of threonic acid)	247	1380	3.76	3.41						
L-threonine (TMS)3	218	1393	3.85	3.52	Threonine	218	5.95	threonine	218	408140
b-alanine (TMS)3	174	1438	4.12	*3.78*	beta-Alanine	174	6.17	beta-alanine	174	434000
Homoserine (TMS)3	218	1463	4.26	*3.93*				homoserine	117, 218	442550
Malic-acid (TMS)3	233	1500	4.48	4.15	Malic_acid	147	6.42	malic acid	147, 233	461870
U1509	243	1509	4.55	4.21	U-084	243	6.54			
L-methionine (TMS)2	176	1525	4.64	4.30	Aspartic_acid	232	6.56	aspartic acid	100, 232	478980
oxo-proline (TMS)2	156	1526	4.65	4.30	Methionine	176	6.60	methionine	128, 176	482230
L-aspartic-acid (TMS)3	232	1527	4.66	4.32						
					Oxoproline	156	6.62	oxoproline	156	485080

Appendix 3A (continued)

SCRI annotations					Aber annotations			Golm 2003 Binbase selections		
	Charac.		Apex			Charac.	Apex		Charac.	
Metabolite	m/z		Rt (s)		Metabolite	m/z	Rt (s)	Metabolite	m/z	RI
g-aminobutyric acid (TMS)3	174	1535	4.71	4.36	GABA	174	6.64	gamma-aminobutyric acid	174	487820
Threonic acid (TMS)4	292	1562	4.87	4.53				threonic acid	147, 292	495820
U1567	218	1567	4.90	4.56	U-106	218	6.85	y amine	100, 218	500820
U1586	227	1586	5.01	4.66	U-116	227	6.90	y515	154, 227	516950
					U-085	159	6.93	y516	147, 159	522140
Glutamic acid (TMS)3	246	1618	5.18	4.83	Glutamic_acid	246	6.98	glutamic acid	246	527940
					Asparagine_minor	188	7.01			
L-phenylalanine (TMS)2	192	1623	5.20	4.84	Phenylalanine	218	7.06	phenylalanine	218	536530
Asparagine (TMS)4	188	1625	5.21	4.85						
Trihydroxypentanoic acids (2,3,5; 2,4,5) (TMS)4 good fit	245	1649	5.31	4.97						
Gluconic acid O-methyloxime (TMS)5	204	1656	5.35	5.00						
Gluconic acid O-methyloxime (TMS)5	204	1663	5.38	*5.03*						
Asparagine (TMS)3	116	1670	5.41	5.07	Asparagine_main	116	7.19	asparagine	116	552960
Glutamine (TMS)4	227	1736	5.70	5.35	U-121	227	7.42	y532	147, 227	578045
putrescine (TMS)4	174	1742	5.73	5.36				putrescine	174	586510
								glycerolphosphate alpha	101, 299	590510
					U-104	217	7.54	glucose-1-phosphate degr.prod.	217	594020
Ribonic acid (TMS)5	292	1768	5.84	5.49				isoribonic acid	103, 292	598820
a-glycerophosphate (TMS)4	299	1769	5.85	*5.50*						
L-glutamine (TMS)3	156	1781	5.90	5.54	Glutamine	156	7.59	glutamine	156	599840
U1809 *likeunoximated fructose (TMS)5*	217/437	1809	6.02	5.67	U-060	217	7.68	y896	217	608980
unoximated fructose (TMS)5	204/217/437	1820	6.07	5.70	U-059	217	7.70	y cho	217	612620
citric acid (TMS)4	273	1824	6.09	5.74	Citric_acid	273	7.74	citric acid	147, 273	617200
								citrulline	157	621470
								dehydroascorbic acid	173	633010
Quinic acid (TMS)5	345	1860	6.25	5.89	Quinic_acid	345	7.89	quinic acid	147, 345	633360
Fructose O-methyloxime (TMS)5	307	1873	6.31	5.95	Fructose_1	307	7.92	fructose meox1	103, 307	637820
Fructose-O-methyloxime (TMS)5	307	1882	6.35	5.99	Fructose_2	307	7.95	fructose meox2	103, 307	641730
Allantoin (TMS)4	331	1885	6.36	6.00				mannose	157, 319	644720
Mannose-O-methyloxime (TMS)5	319	1887	6.37	*6.01*	U-014 (galactose??)	319	7.97	galactose	147, 319	646750
Galactose-O-methyloxime (TMS)5	319	1891	6.39	*6.02*				allantoin	100, 331	647660
Glucose-O-methyloxime (TMS)5	319	1896	6.41	6.05	Glucose_1	319	8.02	glucose meox1	147, 319	649210
Allantoin (TMS)5	188/428/518	1906	6.45	*6.09*						

Appendix 3A (continued)

SCRI annotations					Aber annotations			Golm 2003 Binbase selections		
	Charac.		Apex			Charac.	Apex		Charac.	
Metabolite	m/z		Rt (s)		Metabolite	m/z	Rt (s)	Metabolite	m/z	RI
Glucose-O-methyloxime (TMS)5	319	1914	6.49	6.12	Glucose_2	319	8.10	glucose meox2	147, 319	658190
Histidine (TMS)3	254	1919	6.51	6.15	Lysine	174	8.13	lysine	156, 174	662510
L-lysine (TMS)4	174	1923	6.53	6.16	Histidine	254	8.13			
Mannitol (TMS)6	319	1927	6.55	6.18				mannitol	147, 319	663260
								histidine	154, 254	663718
Sorbitol (TMS)6	319	1933	6.57	*6.21*				sorbitol	147, 319	665730
L-tyrosine (TMS)3	218	1939	6.60	6.23	Tyrosine	218	8.20	tyrosine	218	670636
U1948	232/203/449	1948	6.64	*6.28*				y amine	147, 232	672906
U1953	217/361	1953	6.66	*6.30*	U-056	217	8.23	isopropyl b-D-thiogalactopyranoside	217	674420
Glucose (TMS)5	191, 204	1973	6.75	6.39	U-042	204	8.25	y cho	204, 191	678920
Galactonic acid (TMS)6	292, 333	1984	6.8					galactonic acid	147, 292	690200
Pantothenic acid (TMS)3	103	1985	6.81					pantothenic acid	103	690660
Gluconic acid (TMS)6	292, 333	1989	6.82							
								saccharic acid	147, 292, 333	697790
U2020 (carbohydrate)	204	2020	6.94	6.58	U-122	204	8.44			
								mucic acid	147, 333, 292	709370
U2023	185	2023	6.95					y902	185	709570
					Gluconic_acid_main	333	8.53			
Glucaric (saccharic) or Galactaric (mucic) acid (TMS)6	292/333	2036	7.00	6.62						
					Hexadecanoic_acid	313	8.56	palmitic acid	117, 313	713960
Inositol (TMS)6	217	2086	7.18	6.80	Inositol	305	8.69	inositol myo-	147, 217, 305	729110
U2105 (carbohydrate)	205/319/245	2105	7.25	6.88	U-088	245	8.71			
Caffeic acid (TMS)3	396	2138	7.37	7.00				caffeic acid	219, 396	748272
U2190	218/130	2190	7.56	7.17	U-094	218	9.07			
					U-093	167	9.11	y2449	167	775730
7.648 Tryptophan (TMS)3	202	2212	7.64	7.25	Tryptophan	202	9.12	tryptophan	202	780120
Octadecanoic acid (TMS)	117/341	2250	7.78					stearic acid	117	787400
spermidine(TMS)5	144	2251	7.78	7.40				spermidine	144	791480
					Octadecanoic_acid	341	9.18			
Fructose-6-phosphate oxime-(TMS)6	315	2300	7.96	7.59				fructose-6-phosphate	315	803740
glucose (or galactose)-glycerol conjugate (TMS)6	204/337	2309	7.99	7.61	U-016	204	9.31	y cho	204	804030
Gluocse-6-phosphate oxime-(TMS)6	387	2313	8.00	7.62				glucose-6-phosphate	147, 387	808270

Appendix 3A (continued)

SCRI annotations					Aber annotations			Golm 2003 Binbase selections		
	Charac.		Apex			Charac.	Apex		Charac.	
Metabolite	m/z		Rt (s)		Metabolite	m/z	Rt (s)	Metabolite	m/z	RI
Gluose-6-phosphate oxime-(TMS)6	387	2330	8.07							
U2545 (polysaccharide)	204,217, 437	2545	8.75		U-099	217	9.47	y cho	204, 217	863310
U2559 (polysaccharide)	103/204/217	2559	8.80		U-123	204	9.79	y cho	103, 217, 361	873180
U2567 (polysaccharide)	204/217	2567	8.82		U-124	217	9.98	y cho	204, 217, 361	879340
U2589 (polysaccharide)	186, 361, 437	2589	8.88					y678	186	897690
U2615	103, 204, 217	2615	8.96		U-092	283	10.17	y amine	116, 283	904050
sucrose (TMS)8	361	2637	9.03	8.65	Sucrose	361	10.22	sucrose	147, 361, 217	912730
U2660	204	2660	9.100		U-125	217	10.32	inulobiose1	217, 361	924490
U2677	204	2677	9.150		U-126	217	10.36	inulobiose2	217, 361	929980
U2684 (mixture?)	188, 204, 361, 491	2684	9.170		U-127	217	10.41	levanbiose	204, 217, 361	937020
U2710	204	2710	9.250					y cho	217, 361	946937
U2721	204	2721	9.285					y cho	204, 217	948965
Maltose-O-methyloxime (TMS)	204, 361	2736	9.32					maltose meox2	147, 217	954620
Maltose Meox2		2766	9.415							
unknown polysaccharide	204, 217, 361	2781	9.460		U-130	217	10.54	y cho	217	956730
glycerol-1-(Inositol-1-phosphate) (TMS)8	103, 299, 318, 461	2795	9.504					y cho	204, 217	962960
unknown polysaccharide	204, 217, 361	2851	9.675							
U2864 - (polysaccharide, quite strong)	204, 217, 319, 361, 437	2864	9.704					y cho	204, 217, 361	970390
unknown polysaccharide	204, 217, 361	2885	9.771					y cho	204, 217,	972590
unknown polysaccharide	204, 296, 331, 361	2895	9.804					y cho	204, 217,	988430
Unknown polysaccharide	191, 204, 217, 418, **433**	2932	9.91					y cho	147, 204, 217	991020
								y cho	204, 217	998110
								y cho	129, 217, 204	999910
								y cho	204, 217,	1001051
								y cho	204, 217	1003816
U2974 (polysaccharide)	204	2973	10.04	9.65	U-019	204	11.22	y cho	204, 217	1015400
U2993 (polysaccharide)	204	2933	10.09	9.72						
trans-5-O-caffeoyl-D-quinic acid (TMS)6 Chlorogenic acid (TMS)6	345	3107	10.40	10.02						
U3118 (polysaccharide)	204, 217, 361, 597	3118	10.42		U-055	204	11.59	y cho	204, 217, 361	1046200
trans-4-O-caffeoyl-D-quinic acid (TMS)6 chlorogenic acid isomer	307, 255		10.56					caffeoyl quinate	147, 345	1049400
								isocaffeoyl quinate	307	1065393
trans-3-O-caffeoyl-D-quinic acid (TMS)6 chlorogenic acid isomer	345	3197	10.63							

Appendix 3A (continued)

SCRI annotations				Aber annotations			Golm 2003 Binbase selections		
	Charac.		Apex		Charac.	Apex		Charac.	
Metabolite	m/z		Rt (s)	Metabolite	m/z	Rt (s)	Metabolite	m/z	RI
							y trisaccharide	103, 361, 217	1069200
							y fatty acid	117	1086200
U3353 polysaccharide	204, 217, 361	3353	11.03	U-128	361	12.15	y trisaccharide	217, 361	1107000
Raffinose	217, 361	3375	11.1	Raffinose/1-Kestose	361	12.83	raffinose	129, 361, 217	1118000
U3378 (polysaccharide Kestose?)	217, 361	3378					1-kestose	217, 361	1119800
							inulotriose meox1	217, 361	1122000
							inulotriose meox2	217	1130000
							melissic acid (iso-C27)	117,	1136800
U3445 (polysaccharide)	204, 217, 361	3445	11.33				y trisaccharide	204, 217, 361	1140600
U3470 (polysaccharide)	204, 217, 361	3470	11.42						
polysaccharide	204, 217, 361, 597	3821	12.81						
polysaccharide	204, 217, 361, 597	4022	13.6						

🟧	"core" metabolite identified by all three institutes
🟩	metabolite identified by 2 or 3 insititutes, but differing in elution order
🟦	metabolite identified by 2 or 3 insititutes, but differing in elution order
🟨	metabolite identified by 2 of 3 institutes
🟪	unknown metabolite suspected to be present in 2 or 3 institute peak list

Appendix 3B “Core” metabolites identified from GO2 datasets independently by SCRI, UWA and Golm.

Identified metabolite derivatives are listed by compound class and described by a characteristic mass or group of masses.

Metabolite	Charac m/z	TOF	Quad
Amino acids			
Alanine (TMS)2	116	X	X
Aminobutyric acid (TMS)2	130	X	X
Valine (TMS)2	144	X	X
Leucine (TMS)2	158	X	X
Isoleucine-(TMS)2	158	X	X
Proline (TMS)2	142	X	X
Glycine (TMS)3	174	X	X
Serine (TMS)3	204	X	X
Threonine (TMS)3	218	X	X
β-Alanine (TMS)3	174	X	X
Homoserine (TMS)3	218	X	
Methionine (TMS)2	176	X	X
Oxo-proline (TMS)2	156	X	X
Aspartic-acid (TMS)3	232	X	X
γ-Aminobutyric acid (TMS)3	174	X	X
Glutamic acid (TMS)3	246	X	X
Phenyalanine (TMS)2	192	X	X
Asparagine (TMS)4	188	X	X
Asparagine (TMS)3	116	X	X
Glutamine (TMS)4	227	X	X
Glutamine (TMS)3	156	X	X
Histidine (TMS)3	254	X	X
Lysine (TMS)4	174	X	X
Tryptophan (TMS)3	202	X	X
Tyrosine (TMS)3	218	X	X
Carboxylic acids			
2,3-Dihydroxypropanoic-acid (TMS)3	292	X	
Citric acid (TMS)4	396	X	X
Fumaric acid (TMS)2	273	X	X
Malic-acid (TMS)3	345	X	X
Threonic acid (TMS)4	103	X	
Ribonic acid (TMS)5	292	X	
Galactonic acid (TMS)6	292, 333	X	
Gluconic acid (TMS)6	292, 333	X	X
Glucaric (saccharic) or Galactaric (mucic acid (TMS)6)	292/333	X	
Pantothenic acid (TMS)3	103	X	

Metabolite	Charac m/z	TOF	Quad
Carboxylic acids (continued)			
Quinic acid (TMS)5	345	X	X
Caffeic acid (TMS)3	396	X	
trans-5-O-Caffeoyl-D-quinic acid (TMS)6 Chlorogenic acid (TMS)6	345	X	
trans-4-O-Caffeoyl-D-quinic acid (TMS)6 Chlorogenic acid isomer	307, 255	X	
Carbohydrates			
Fructose (TMS)5	204/217/437	X	X
Fructose O-methyloxime (TMS)5 anomer 1	307	X	X
Fructose-O-methyloxime (TMS)5 anomer 2	307	X	X
Mannose-O-methyloxime (TMS)5 anomer 1	319	X	
Galactose-O-methyloxime (TMS)5 anomer 1	319	X	
Glucose-O-methyloxime (TMS)5 anomer 1	319	X	X
Glucose-O-methyloxime (TMS)5 anomer 2	319	X	X
Glucose (TMS)5	191, 204	X	X
Sucrose (TMS)8	361	X	X
Maltose-O-methyloxime (TMS) anomer 1	204, 361	X	
Maltose -O-methyloxime (TMS) anomer 2	204, 361	X	
Raffinose	217, 361	X	X
Gycerol (TMS)3	205	X	X
Inositol (TMS)6	217	X	X
Mannitol (TMS)6	319	X	
Sorbitol (TMS)6	319	X	
Glucose (or galactose)-glycerol conjugate (TMS)6	204/337	X	X
Phosphorylated compounds			
Phosphate (TMS)3	299	X	X
α-Glycerophosphate (TMS)4	299	X	
Fructose-6-phosphate oxime-(TMS)6	315	X	
Glucose-6-phosphate oxime-(TMS)6	387	X	
Nitrogenous compounds			
Urea (TMS)2	189	X	
Allantoin (TMS)4	331	X	
Allantoin (TMS)5	188/428/518	X	
Spermidine(TMS)5	144	X	
Putrescine (TMS)4	174	X	

G03012 Final Report - Section 4

Sue Verrall/Derek Stewart/Manfred Beckmann

Assessment of software for automated alignment and annotation of GC-MS and LC-MS data (Task 03.01)

4.1 BACKGROUND TO EXPERIMENTAL SECTION

Utilitarian, robust software are key to the use of any analytical approaches to developing unified data models and pre-processing strategies to yield meaningful and standardized statistical analyses of metabolome variability in crop plants is. Highlighted as part of this proposal application was the diversity of approaches used to monitor and report upon the plant metabolome and the potential that these may have for food safety and compositional studies. Indeed as part of the previous Agency-funded projects under the Safety Assessment of GM Foods research programme (G02) there were 12 collaborating sites in 6 projects produced metabolomics data for potato, wheat, barley, *Arabidopsis* and tomato. These projects utilised a broad range of technical (analytical) approaches to create the compositional data and further levels of complexity were created by the diversity of the onboard (technology manufacturer supplied) software to analyse this data. This is symptomatic of current metabolomics, where there are no rigorously applied and widely applied standards for experimental procedures, although efforts to address this are underway via the Metabolomics Standards Initiative (http://msi-workgroups.sourceforge.net/) and refereed reports (Fiehn et al 2006, 2008). As a result the large amount of data collected is not necessarily, therefore, comparable. This was further confounded by the fact that the diversity of the projects submitted and approved under the G02 programme meant that no prior arrangements were defined to ensure or enhance comparability of data from different projects. However this apparent problem offered an opportunity to use the data generated in disparate labs with many different technologies, extraction and preparation approaches as a resource that could provide a meaningful and durable description of food raw material composition for any future safety or quality assessments.

To achieve this aim a selection of samples /metabolic profiles derived from G02 and/or G03 (bespoke spectra run as part of the new project) were used to cross compare the utility of the different user metabolomic analytical platforms (GC-MS and LC-MS). Furthermore, several software packages were evaluated using G02/3 derived raw data for attempted automation of peak alignment and spectral deconvolution for metabolome profile data formats available to the G03 project

4.2 COMPARISON OF INSTRUMENT VENDOR SOFTWARE FUNCTIONALITY; LIMITATIONS AND PROBLEMS OF USING AUTOMATED OUTPUT

The vendor platform technology software, here was confined to those onboard the collaborating partners Thermo-Finnigan and Leco instruments, were cross compared to determine key similarities and differences and how these would impact upon metabolomic analysis and the detail and quantity of the data derived therein.

Research at both SCRI (Shepherd et al, 2007; Dobson 2007, 2008) and Aberystwyth (Beckmann et al, 2007, 2008; Enot and Draper,2007; Enot et al, 2007, 2008; Overy et al, 2008) has already developed many tuber profiles derived from the various platforms. The work in SCRI has centred on a Thermo Tempus 1 (GC-tof-MS) and DSQ (Quad) and Agilent machines whilst in Aberystwyth the work utilised a LECO GC-tof-MS Pegasus 2. All of these platforms have broadly similar technologies with respect to ionisation and ion mass measurement. Therefore, with the appropriate similar columns and chromatographic conditions a direct comparison of the on board operating and deconvolution software should be feasible. The corresponding vendor software packages have several similarities but distinct differences which are listed below.

Agilent

- The generated chromatograms are analyzed using the onboard software, Chemstation™, AMDIS and by manual examination to develop a bespoke (user) library of all peaks that are reproducibly seen.
- Each metabolite is defined by a retention window and characteristic masses (ions).

- The onboard software (Chemstation™) is used to analyse for the presence of the metabolites in the bespoke library with the manual assessment for the absence of library peaks (peak checking).
- A table is constructed compiling the relative intensity for the characteristic ions for each metabolite.

Thermo-Finnigan

- AMDIS is used to mine the total ion current chromatograms leading to the construction of a bespoke (user) library of all peaks that are reproducibly seen
- Each metabolite is defined by a retention window and characteristic masses (ions) and depending on the user setup, a predictive structure as defined by comparison with the bespoke and/or commercial MS databases.
- The on board software (Xcalibur™) generates Selected Ion Chromatograms (SICS) for each metabolite within the retention window.
- A table is constructed compiling the relative intensity of the selected ion chromatograms for the characteristic ions of each metabolite

Leco

- The chromatograms are analysed using proprietary LECO software Chromatof™ and a bespoke library is constructed of all peaks that are reproducibly seen and their spectra.
- The onboard software (Chromatof™) is again used to mine the chromatograms along with manual confirmation of peak identities, the calculation of relative peak areas leading to the construction of a metabolite table

Clearly the approach for each platform looks very similar but there are key differences with the LECO and Thermo systems with the former employing the deconvolution software PeakFind (part of ChromaTOF™) which is constructed to primarily develop *de novo* peaks lists for each run. Conversely the Thermo Tempus system, with its Xcalibur™ software utilises the NIST derived AMDIS software to deconvolute peaks leading to the development of a library of expected metabolite peaks. This list is then used in an automated format to screen subsequent data sequences (see Hargrove *et al.,* 1981, Herron *et al.,* 1996, Stein and Scott, 1994; Stein 1999; Halket et al 1999; Zhang, 2006).

Analysis showed that the approaches by the AMDIS based and Chromatof™ base software produced significant differences in the numbers of components identified and the qualitative results. These largely focus on the lesser known or unknown components with AMDIS particularly prone to generating a significant number of false positive peaks due to several issues (see below). In addition although each instrument should theoretically handle multiple data formats there was significant difficulty in interchanging data formats and this issue had led to the proliferation of 3rd party software for such analyses. The commercial and free software available are highlighted in Tables 4.1 and 4.2 respectively,

As a result these differences in peak identification and spectrum deconvolution make routine comparison of datasets produced on different machines and at different sites very difficult and therefore unacceptable for the development of routine food compositional analysis. However this may be better addressed once (or when/if) a uniform format can be adopted as noted above.

The AMDIS software is now routinely used as a metabolomic deconvolution approach and despite its obvious attractiveness the downside of false positive peak generation, as discussed earlier, does detract from its utility. Developed by NIST (www.amdis.net), AMDIS was created as a method for extracting individual component spectra from GC/MS data files and then using these spectra to identify target compounds by matching to spectra in a reference library. Based on the peak method of Dromey et al (1976) for the basis for spectrum extraction (deconvolution) both which showed this approach was shown to produce reliable results in large-scale tests (Shackleford et al, 1983) and importantly it followed an approach similar to that of an analyst. Its utility is such that it has been adapted repeatedly to become a programme that accepts multiple data formats as shown below:

- Bruker (*.msf)
- Finnigan (GCQ, INCOS, and ITDS formats) (*.ms;*.mi;*.dat)
- HP Benchtop and MS Engines (*.ms)
- HP Chemstation (*.d)
- Inficon GCMS (*.acq)
- MassLynx NT Formats (*.*)
- MicroMass (*.)

- NetCDF (*.cdf)
- Perkin-Elmer Turbo Mass (*.raw)
- Saturn SMS (*.sms)
- Shimadzu MS Files (*.R##)
- Shrader/GCMate (*.lrp)
- Varian Saturn Files (*.ms)
- Xcalibur (*.raw)

AMDIS utilisation generally takes the form of four distinct processing steps: 1. noise analysis, 2. component perception, 3. spectrum deconvolution, 4. compound identification. These all come with limitations that can impact directly upon the resultant peak tables and therefore food safety and compositional studies. The AMDIS approach of peak maximization (Dromely et al 1976) as the only means for perceiving components can cause problems particularly in the complex biological systems analysed in G02/03. The co-elution and co-maximisation of peaks can lead to them being identified as a single metabolite with a resultant (two compound-derived) single spectrum. In addition if there is close elution/peak maximisation accompanied by broad peak tops leading to several local maxima a component may be identified more than once leading to the generation of several false positives. An additional problem seen much more significantly in LC-MS is that if the peaks are broad and significantly larger than the analyses window setting then the peaks can be missed completely leading to the generation of false negatives.

Perhaps the main problem associated with AMDIS is that of the requirement for a present or absent decision be made concerning the existence of a component leading to the missing value problem down the line in subsequent data analysis. For the biological complex spectra generated as part of the G02/03 programmes there will be uncertainty for several compounds due to their inherent low level in a specific experiment leading to its elimination as noise. This drawback is significant since there is a huge difference between a sample not being present and not being detected especially where food safety and toxicology is concerned .

Table 4.1 Commercial software tools for metabolomic data processing (Katajamma and Oresic, 2007).

Name	Vendor	Features	Main Application and examples
BlueFuse.	BlueGnome, Cambridge, UK Filtering, peak detection, and alignment	Univariate and multivariate methods for data analysis	Metabolomics with MS and NMR data
Chenomx	NMR Suite Chenomx, Edmonton, Canada	Data conversion, analyzing spectra to compounds and concentrations	Metabonomics with NMR data (Suade et al 2006)
Metabolomics module	Expressionist Genedata, Basel, Switzerland	Genedata Filtering, peak extraction, *m*/*z* and retention time alignment. Metabolite identification using third-party databases. Includes also analysis and interpretation modules and integrated database	Cross-omics platform for transcriptomics, proteomics and metabolomics works with MS data
LineUp	Infometrix	Alignment of chromatographic data. Alignment can be used also for spectroscopic data	Chromatographic alignment
MarkerLynx	Waters, Milford, MA, USA	Peak detection and alignment. Principal component analysis (PCA)	Metabonomics with LC–MS data (Idborg et al 2005; Lenz et al 2005; Whitfield et al 2005
MarkerView	Applied Biosystems, Foster City, CA, USA	Peak detection and alignment. PCA and *t*-test methods for data analysis. Visualization and reporting	Metabolomics with LC–MS data (Wagner et al 2007)
MassHunter	Profiling software Agilent Technologies, Santa Clara, CA, USA	Feature extraction and alignment	Proteomics with LC–MS data
Metabolic Profiler	Bruker Daltonic & Bruker BioSpin, Billerica, MA, USA	Bucket raw data into retention time, *m*/*z* table with intensities. Identification using libraries. PCA for data analysis	Metabolomics with MS and NMR
metAlign,	PlanResearch International B.V., Wageningen, The Netherlands	Filtering, baseline correction, peak detection, alignment	Broad, LC–MS and GC–MS data (Vorst, O. et al, 2005; Tikunov et al 2005; Keurentjes et al, 2006; Bino et al 2005).
MS Resolver	Pattern Recognition Systems, Bergen, Norway	Resolve multicomponent data from multidetection instrumentation into individual contributions	Broad, LC–MS and GC–MS data (Idborg et al 2005; Eide et al 2002)
Profile	Phenomenome Discoveries, Saskatoon, Canada	File conversion, peak detection and alignment. Tools for statistical analysis and data mining	Metabolomics with MS data
Rosetta Elucidator	Rosetta Biosoftware, Seattle, WA, USA	Peak detection and alignment, statistical analysis and visualization	Proteomics with LC–MS data
Sieve	Thermo Fisher Scientific, Waltham, MA, USA	Direct comparison approach to comparing multiple LC–MS datasets.	Uses ChromAlign for chromatographic alignment Proteomics with LC–MS data (Sadygov et al 2006)

Table 4.2 Freely available software for metabolomic data processing (Katajamma and Oresic, 2007).

Name	Features	Main Application and examples	License type	Platform
Chrompare (Frenzel et al, 2003)	Comparison of chromatographic peak lists and raw chromatograms, automatic and manual normalization	Metabolomics with GC-FID	Licence type unknown, available for download	Microsoft Excel Visual Basic
COMSPARI (Katz et al, 2004)	Visualization to aid searching for differences between pair of runs	Metabolomics with LC–MS and GC–MS data (Katz et al, 2004)	GNU General Public License	Implemented in C, for any recent platform including Linux and Windows
Continuous profile models (Listgarten et al, 2007)	Alignment and normalization of time series data	Proteomics with LC–MS data	Free for educational and research use, source code available	Toolbox for Matlab
HiRes (Zhao et al, 2006)	Processing and analysis spectral data	Metabolomics with NMR data	Free for research and clinical purposes	Implemented in C++, for Windows platform
LCMSWARP (Jaitly et al, 2006)	Retention time alignment and feature clustering	Proteomics with LC–MS and LC–MS/MS data (Umar et al, 2007)	Unkown	Implemented in C++
MapQuant (Leptos et al, 2006)	Noise filtering, peak detection and visualization.	Proteomics with LC–MS	Harvard University open-source compatible license	Implemented in C, for Windows and Linux Platforms
MathDAMP (Baran et al, 2006)	Direct comparison of raw data sets without peak picking. Includes methods for preprocessing (binning, baseline subtraction, smoothing) and normalization	Metabolomics with LC–MS, GC–MS and CE–MS data	Free	Package to Mathematica
MET-IDEA (Broekling et al, 2006)	Extracts ion intensity data for listed ion/retention time values from multiple runs	Metabolomics with LC–MS, GC–MS and CE–MS data	Freely available to academic users upon request	Windows, .NET platform
MSFACTs (Duran et al, 2003)	Alignment and comparison of raw chromatograms or peak lists generated with a third-party software	Metabolomics with GC–MS and LC–MS data	Freely available upon request for academic and non-commercial use	Implemented in Java
MSight (Palagi et al 2005)	Visualization and visual analysis and comparison of multiple runs	Proteomics with LC–MS data	Free of charge	Windows
msInspect	Peak detection, alignment,	Proteomics with LC–MS data	Free software available	Implemented in Java. Requires R statistical

(Bellew et al, 2006)	normalization and visualization		under Apache 2.0 License	language
MZmine (Katajamaa et al, 2006)	Noise filtering, peak detection, alignment, normalization and visualization. Distributed computing	Metabolomics with LC–MS and GC–MS data (Kind et al 2007; Rischer et al, 2006)	GNU General Public License	Implemented in Java
SpecArray (Li et al, 2005)	Noise filtering, centroiding, peak detection, alignment and visualization	Proteomics with LC–MS data	GNU General Public License	Implemented in C, for Linux platform
SuperHirn (Mueller et al 2007)	Peak detection, alignment and normalization. Includes also analysis capabilities	Proteomics with LC–MS data	Not Yet available	Not yet available Implemented in C++, for Unix platforms
Xalign (Zhang et al, 2005)	Peak detection, alignment between samples and quality control	Proteomics with LC–MS data	Available upon request from the author	Implemented in C++
XCMS (Katajamaa & and Oresic, 2005)	Noise filtering, peak detection and alignment	Metabolomics with LC–MS and GC–MS data (Kind et al, 2007; Go et al, 2006; Want et al, 2006)	GNU General Public License	Implemented in R statistical language

4.3 REVIEW AND SELECTED TEST OF PLATFORM-INDEPENDENT SOFTWARE FOR AUTOMATED GENERATION OF PEAK TABLES FROM GC-MS AND LC-MS DATA.

The increasing interest in the application of metabolomics to biological systems has also been accompanied by an increase in instrument independent software available to (theoretically) do the same function. There are certainly many software packages freely and commercially available (Tables 4.1.& 4.2) for such approaches with many alignment programs for GC–MS, LC–MS, nuclear magnetic resonance (NMR) well reported (Duran et al, 2003; Katajamaa and Orešič, 2005; Katajamaa et al, 2006; Kind et al, 2007; Hongmei et al, 2008). Of these many packages there is a wide range of metabolomic applications and few are focussed on the really problematic area of MS deconvolution. The universality and general acceptance, albeit piecemeal, of the packages deemed it necessary to select examples of these for attempted automation of peak alignment and spectral deconvolution of the metabolome profile data formats available to the G03 project. To best address this G02 and nascent potato metabolomic data were used and the following parameters were focussed on with the assessments of selected and most commonly used software tabulated in Appendix 4.1:

- Cost, availability and conditions of use
- Maintenance and support - can we expect improvements in the future?
- Are the implemented methods explained or published?
- Operating system
- Open source
- Import/export data format
- Dependence on the analytical system
- Scalability: data set size restrictions/running time decent for medium size (100 runs) high throughput metabolomics experiments
- Level of parameterisation and user knowledge necessary to get the best out of the technique

Furthermore, it was important to note from an operational point of view the operating system required and whether or not the code was open source. Finally, a range of interoperability criteria were assessed in relation to import/export data format; whether dependent on the analytical system; scalability, including data set size restrictions/running time decent for medium size (100 runs) to high throughput metabolomics experiments; level of parameterisation and user knowledge necessary to get the best out of the technique. From all of the available software only two of the packages were successful in robustly delivering on the challenges faced from complex plant/crop metabolomic data; MetAlign and XCMS.

These seemed to offer the desired characteristics and these were assessed in greater detail for functionality using potato GC-MS and LC-MS data files generated at SCRI.

4.3.1 MetAlign

Metalign was devised at the Plant Research Institute, the Netherlands to distinguish between and filter out statistically significant differences between pre-defined classes of full scan LC or GC mass spectrometry data sets. It has been used several times (Tikunov et al, 2005; America et al, 2006; De Vos et al, 2007) and although initially launched as an expensive (€5000) commercially available platform independent package it is now freely available on request.

To asses this package G02/3 GC - & LC MS data was utilised. The LC-MS data utilised was of a cultivar experiment (G02) and this had been derived from 23 cultivars: Desiree, Record, P.Dell, Shelagh, Stirling, Torridon, Glenna, Morag, Eden, M.Piper, P.Javelin, Cara, P.Crown, Brodick, Barbara, Pink Fir Apple, G.Wonder, Inca Sun, Mayan Gold, Lumpers, Fortyfold, Anya.

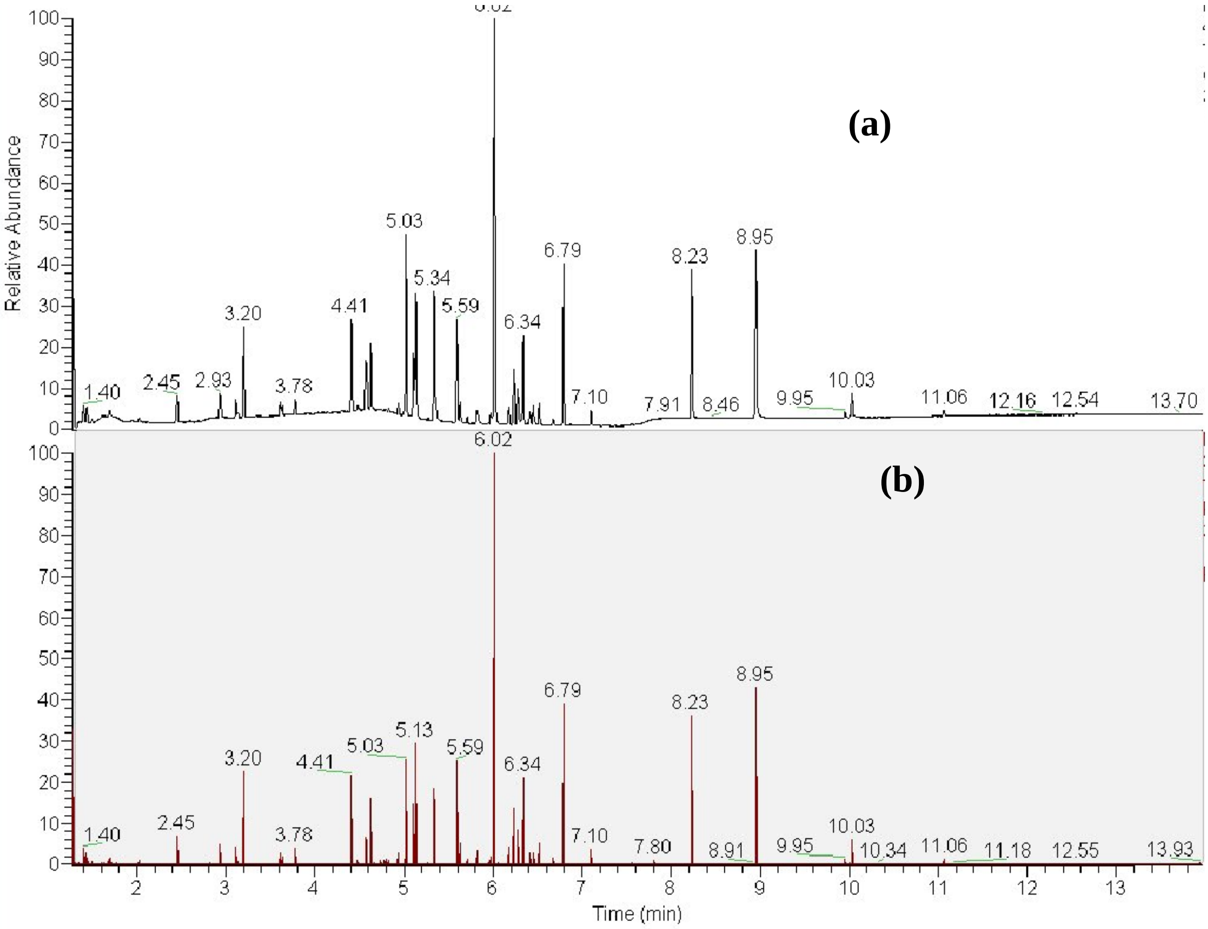


Figure 4.1 A comparative screen capture of the (a) Thermo-Xcalibur derived raw GC-MS data of a polar extract of the reference potato cultivar Cara analysed using the G02001 derived SOPs and (b) the same raw GC-MS data subject to MetAlign directed file conversion to net.cdf, background subtraction, peak finding and signal reduction to one scan width.

With respect to GC-MS data nascent chromatographic runs were used. These were derived from SCRI RERAD funded experiments into potato flavour. Briefly, the cultivars Desiree, Mayan gold and Cara were grown as described in the experimental section of SCRI's G02001 final report (Transcriptome, proteome and metabolome analysis to detect unintended effects in genetically modified potato). Following harvest all tubers were subjected to 4 weeks "skin set", stored undercover at 10°C as per normal practice. Selected tubers were then stored at 0 weeks/4°C, 6 weeks at 4°C and 17 weeks at 4°C. The tubers were then analysed using the SOPs generated in G02001 and the raw GC-MS data subject to analysis by MetAlign

The immediate possibilities and advantages of MetAlign are obvious with respect to its ability to background subtract, a dominant feature of MetAlign (Figure 4.1).

This rationalization of the data is more clearly seen if the data is focussed in on a smaller time window (Figure 4.2). It is clear that the MetAlign processed data is "simpler (Fig 4.2b), having baseline corrected and reduced each peak to a single scan event. At each of these scan events MetAlign has identified the masses which peak at this retention time. This process is similar to deconvolution, producing spectra which can be used for external library searches to tentatively identify components in the data set.

When challenged with multiple chromatograms from the potato experiments the reproducibility of MetAlign was robust when providing reproducible peaks for the same compound in different chromatograms (Figure 4.3). With respect to the output three options are available: a detailed Ascii output, Excel Compatible output and Differential Retention Display. The utility of the former is that the secondary statistical packages, such as Genstat™, Simca-P10™ etc, can be used. The output generally consists of amplitudes, noise and retention times for every mass scan event.

The significant differences between MetAlign and the other commercial platform onboard software are with respect to the processing (Table 4.3)

Table 4.3 The comparative difference between MetAlign and Xcalibur software for metabolomics chromatogram processing

	MetAlign	**Xcalibur**
Total peak found	20,000	100
Peak Indent	Scan number & mass	Specific ident from bespoke processing method preparation and standards/database comparison
Run time processing runtime (14 files)	360 minutes	30 minutes
Peak validation	None	Dependent on databases to search against (mins/hours)
Compound Identification	unknown	Seconds with a % probability or match to the databases

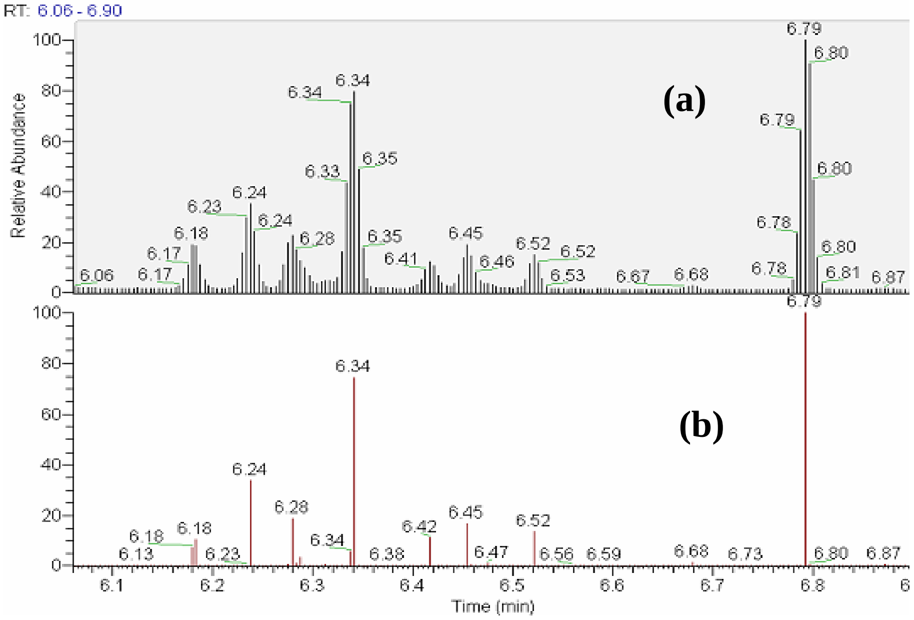


Figure 4.2 (a) A comparative screen capture of the raw Thermo-Xcalibur derived GC-MS data of a polar extract of the reference potato cultivar Cara analysed using the G02001derived SOPs; retention window 6.0-6.9 minutes. (b) The raw GC-MS data of (a) subject to MetAlign directed file conversion to net.cdf, background subtraction, peak finding and signal reduction to one scan width; retention window 6.0-6.9 minutes.

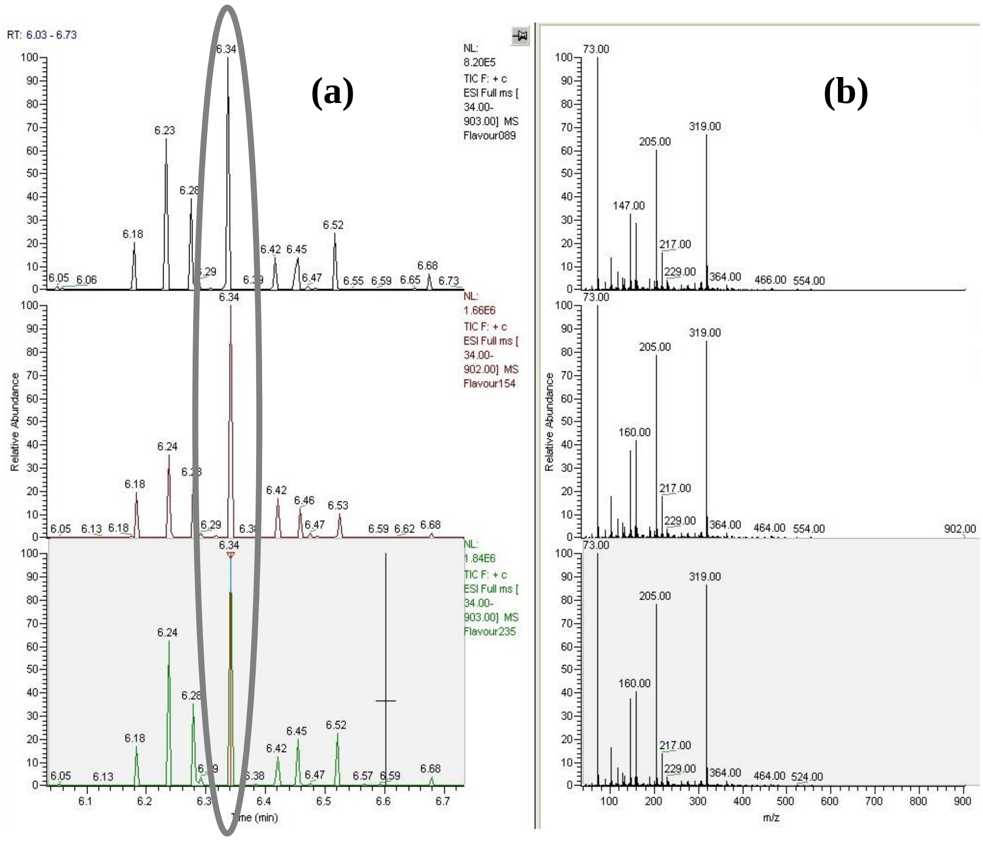


Figure 4.3 Three MetAlign baseline corrected files (a) verifying the same compound (highlighted) using extracted mass spectra (b) at retention time 6.34.

The results of principal component analysis (PCA) of the Xcalibur and MetAlign data sets generated from the potato storage experiments were remarkably similar following multivariate analysis. (Figure 4.4). It is clear from these plots that the segregations are very similar, despite the PCA generating data formats and albeit in different directions. However it is when this cumulative multivariate data are data mined that the problems with MetAlign can arise. Since the datapoints for MetAlign are described by scan number and mass, unlike Xcalibur which has a compound identification capability, the analysis of the PCA plots to define the data points responsible for the segregation becomes tortuous and in essence meaningless for MetAlign derived data. This is highlighted in Figure 4.5 where the PCA plots of Figure 4.4 are re-interpreted to show the data points causing, or driving, the segregation. For the Xcalibur processed data the processing methods is set up to annotate the datapoints and can have multiple masses describing a peak. This means that for the PCAs clear identification/annotation of the compounds is utilised to construct them allows the user to see, in this example, that the compounds glucose, fructose and sucrose were driving the segregation of the Cara (17 weeks/4°C) data away from the bulk of the other samples. This supports the previous reports of the potato cultivar Cara's susceptibility to low temperature sweetening and starch breakdown (Bournay et al, 1996; Rivero et al., 2003; Dobson et al 2007). However, with the MetAlign generated loading plots the lack of a "screened" or annotated processing method leaves many data points (scan number and mass) to describe the segregation leading to excessive over population of the plot and an inability to easily interpret the data.

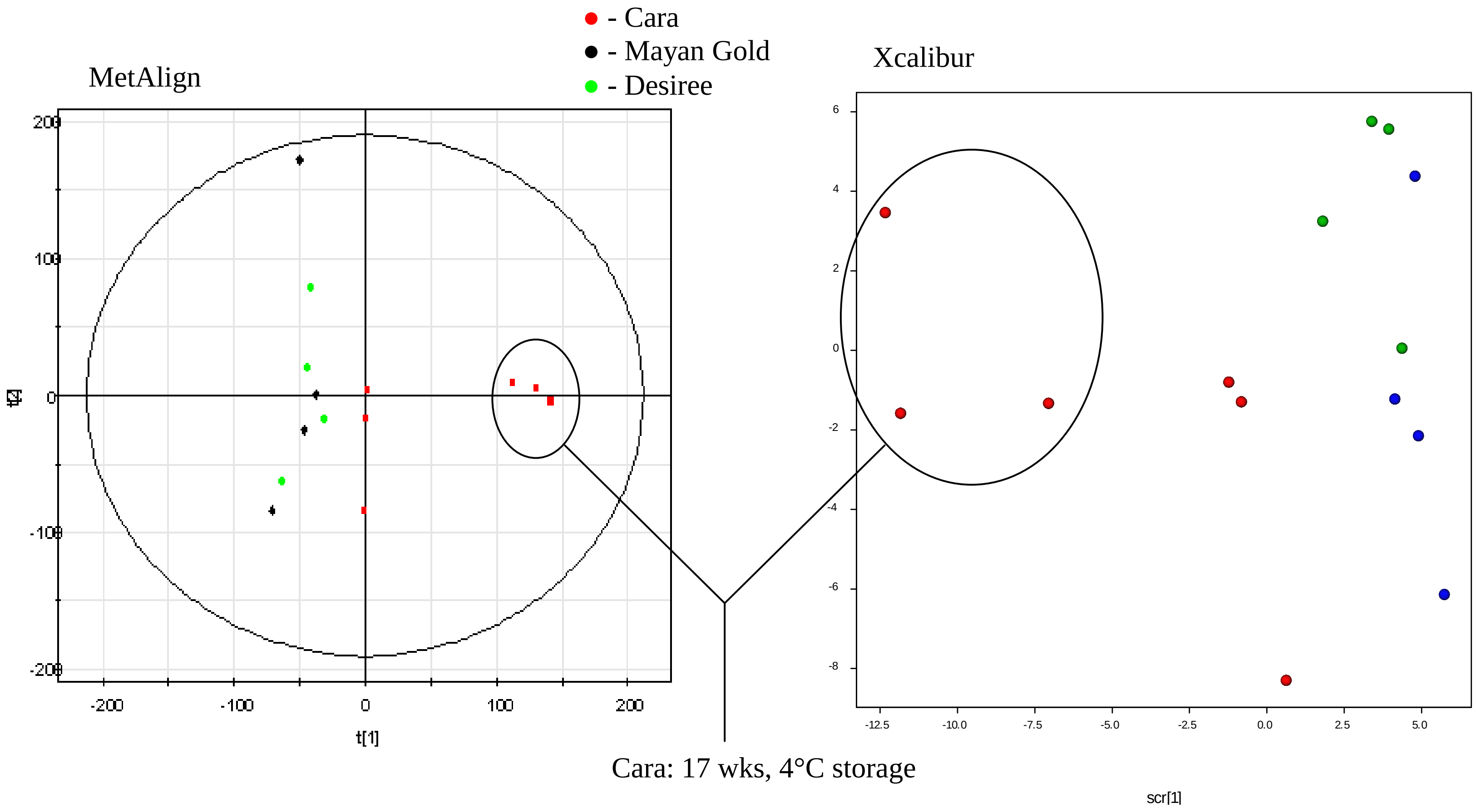


Figure 4.4 Application of the Xcalibur and MetAlign processing software to the raw chromatograms generated on Thermo Tempus GC-ToF-MS from the potato storage experiment. The outputs derived from MetAlign were analysed by Simca-P10 and those from Xcalibur by Genstat to generate principle component analysis plots

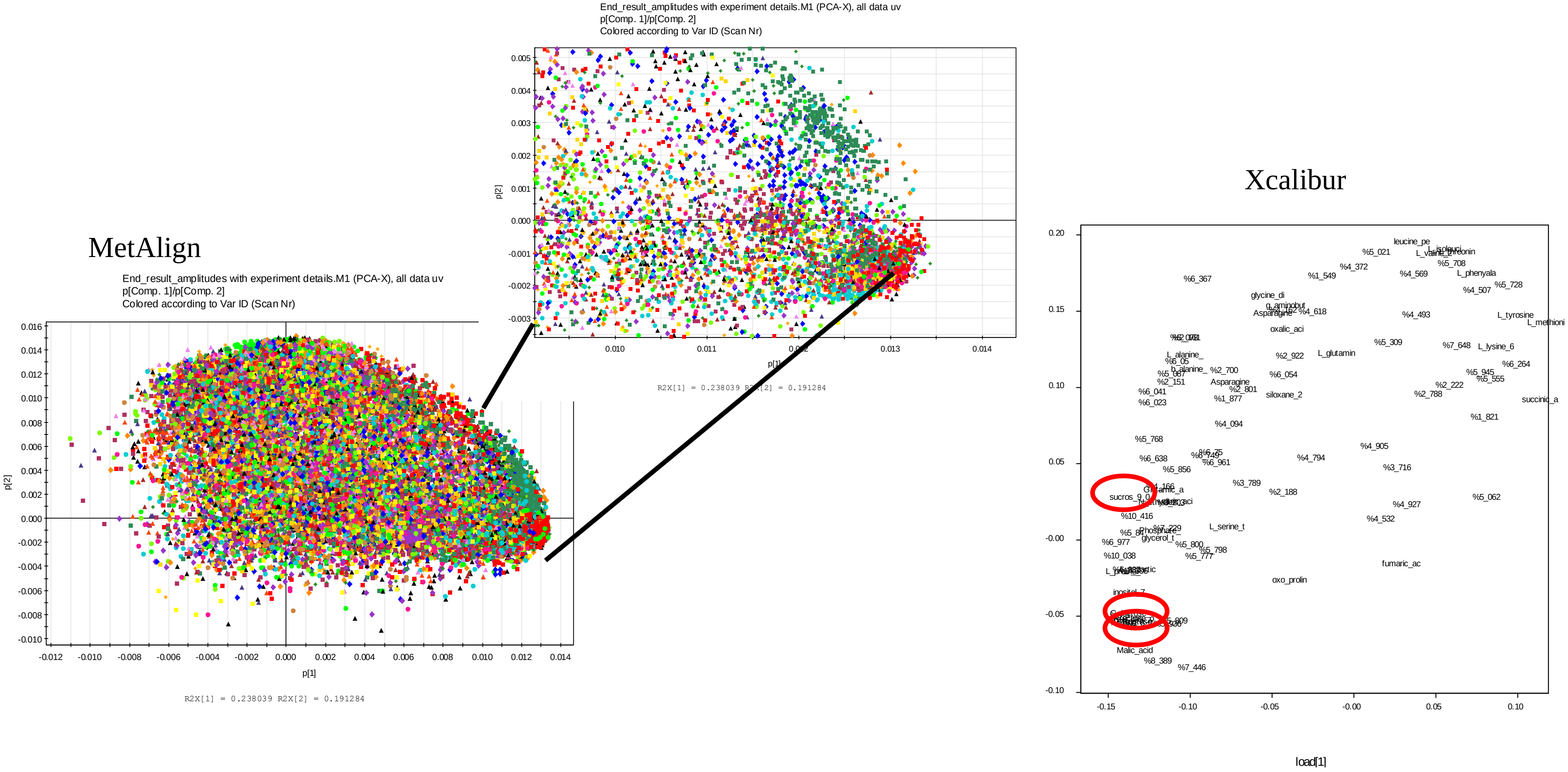


Figure 4.5 The loadings plots generated from the PCA displayed in Figure 4.4 (potato storage experiment). The loading plot on the left is generated from the MetAlign derived data with an associated exploded view of part of the plot. The loadings plot on the right is derived from the Xcalibur annotated processing method and, although showing the same segregations are much easier to analyse and interpret within a biological frame of reference.

It is clear (particularly in the zoom view; Figure 4.5) that clusters of colour (individual masses, same scan) are causing the 3 long term storage Cara samples to separate. These masses could be from a single component which then requires the user to return to the raw data file and library searching at that specific the scan number to identify and annotate the compound/peak. At present this is not automated in MetAlign making it too labour intensive for utilitarian metabolomics assessment. It does however have mean that MetAlign has utility as a process to check (relatively) quickly that a data set is suitable for further onboard (Xcalibur, ChromaTof etc) processing.

With respect to LC-MS analysis MetAlign was again shown to have comparative drawbacks. It can realistically handle well chromatographed data (Figure 4.6) but as peaks start to shift or if the scan intervals are not regular (due to user setup differences; a not uncommon occurrence when cross comparing data from numerous or different sites) then MetAlign can fail to align the chromatograms within a data set leading to unusable data. (Figure 4.7)

In summary MetAlign has a place in metabolomic studies but it is unlike to replace the technology associated onboard software used in conjunction with an experience user. It was suitable when running small data sets, but large sets require significant computing power due to the generation of extremely large temporary files whilst chromatogram aligning. These are stored where the software has been installed, so space restrictions or PC interruptions can lead to fatal program crashes. More restricting is the fact that the generated peak list contains no annotation; consist only of scan number and mass. The identification of components must be done manually by using the raw data files and an appropriate mass spectrum library.

4.3.2 XCMS

The other metabolomic analytical software package gaining favour is XCMS (Dunn et al, 2008; Nordström et al, 2008; Wikoff et al, 2007; Nordström et al, 2006; Smith et al, 2006; Want et al, 2006). The XCMS package reads and processes LC/MS data stored in NetCDF (AIA/ANDI), mzXML, and mzData files. The authors claim that XCMS includes numerous options for visualizing and interacting with that data. In addition, it is claimed to include functionality for peak picking, non-linear retention time alignment, and relative quantification. The programme comes packaged with R embedded. R is a language and environment for statistical computing and graphics and provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, etc.). The combined package should therefore be capable of simultaneously pre-processing, analyzing, and visualizing the raw data from hundreds of samples.

Once downloaded the program was trialled successfully with the packaged data set which produced successful a text report of the peaks found. However production of extracted ion chromatographs using the example commands did not appear possible or feasible using the example statements.

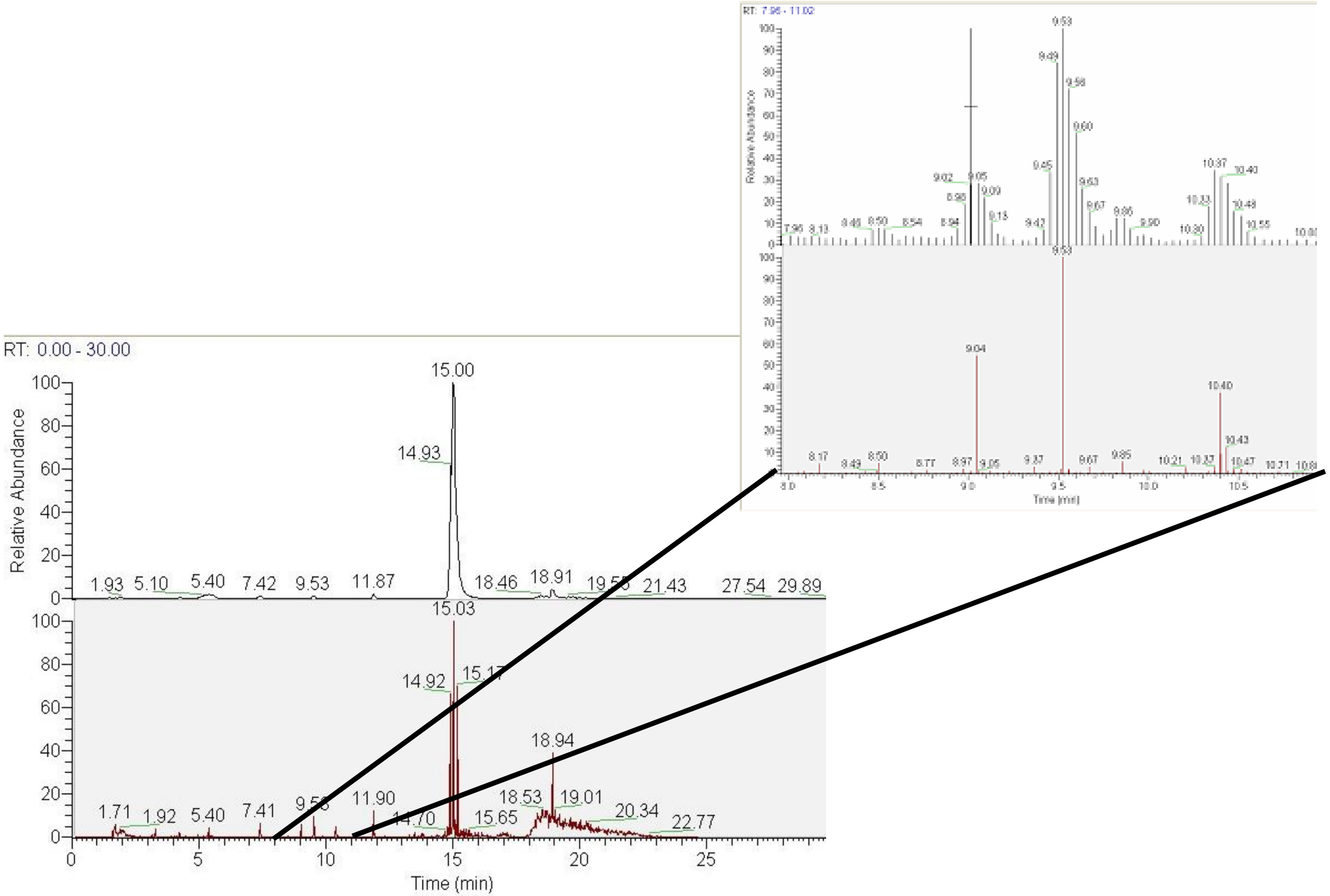


Figure 4.6 Selected LC-MS chromatogram processed by Xcalibur (upper) and MetAlign (lower) software. The exploded sections s hows the retention window 8.0-11.0 minutes and how the MetAlign software has reduced the Xcalibur peaks to stick peaks

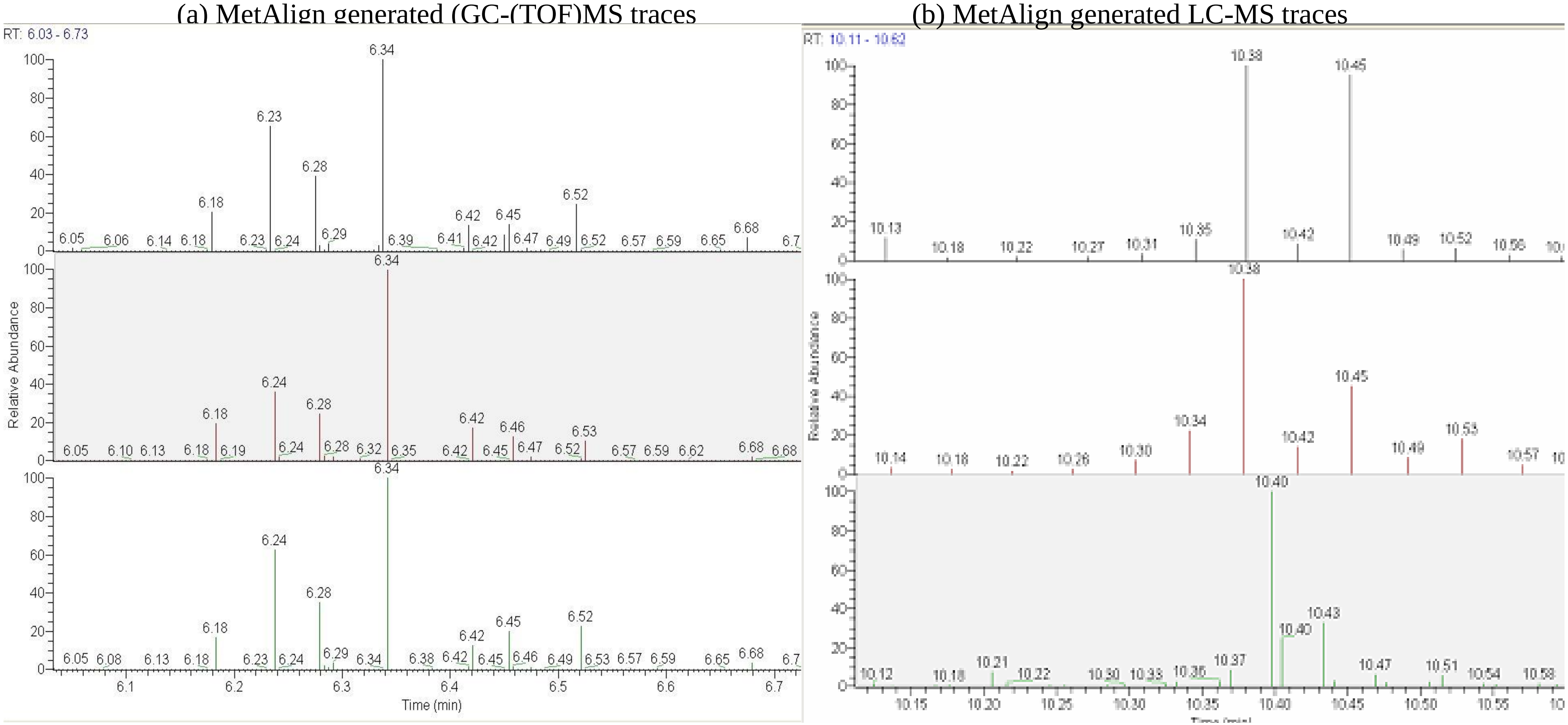


Figure 4.7 Potato (Cara cv.) polar extract GC-ToF-MS and LC-MS (iontrap) traces generated using MetAlign. The software can handle data generated with regular scan intervals, such as GC-ToF-MS derived data (a) thereby allowing alignment. However any mismatch on scan intervals renders chromatogram alignment unusable (b) such as some LC-MS derived data. The top two LC-MS MetAlign generated traces plots (b) have slightly different scan intervals whilst the bottom plot clearly has a different time intervals between the scans leading to an inability to align the peaks and subsequently process to yield robust peak tables.

It was important to note that to employ and exploit the facilities available within XCMS the data had to be full scan LC-MS, unlike the vendor software that can deal with truncated or partitioned chromatograms. As a point of routine the data sets collected at SCRI were all LC-MS/MS (LC-MS^2) which are ***incompatible*** with XCMS. A data set containing Desiree (S Tuberosum), Phureja (S Tuberosum ssp phureja, a diploid breeding line), Inca Sun (S Tuberosum ssp phureja), Mayan Gold (S Tuberosum ssp phureja), Montrose (S Tuberosum) and Pentland Dell (S Tuberosum) was extracted and re-run using full scan LC-MS in positive mode as outline in the G02001 SOPs. Within this set Desiree and Phureja were the sequence quality controls and the other four varieties of one replicate from 3 harvest dates. The initial test run used the Desiree and Phureja samples (Fig 4.8) since their files represented repeat injections and the replicates therefore contain the same major peaks.

These and the other associated LC-MS raw data were set in the XCMS program folder (as instructed by the XCMS help files) but these raw data files were not recognised. XCMS is run using syntax and structure as defined by the embedded program R, but this is non-intuitive and requires a user with specific experience in this environment which makes subsequent analysis of the data set problematic for the uninitiated. However the analysis was eventually run and the output was in the form of a retention time deviation plot (Fig 4.9) which highlighted 89 retention time groups on a first analytical pass and 236 following a second pass. This data was then worked on to identify the significant peaks. Close inspection of the data set showed that there were approximately 12 peaks between 2.5 minutes and 14 minutes. These XCMS-derived significant peaks are shown overlaid on the original LC-MS TIC trace (Fig 4.10). The XCMS programme did miss peaks, for example at 11.8 minutes (708 secs, highlighted) with a mass of 903, however these missing peaks may be “found” by passing the data though the “peakfind” routine repeatedly. This highlights the drawback of XCMS: how many repeats are required to fully describe the metabolome? Realistically this “chemistry free” analysis can be set as part of the routine and allowed to repeat until the changes in the significant peaks reduces to < 2%.

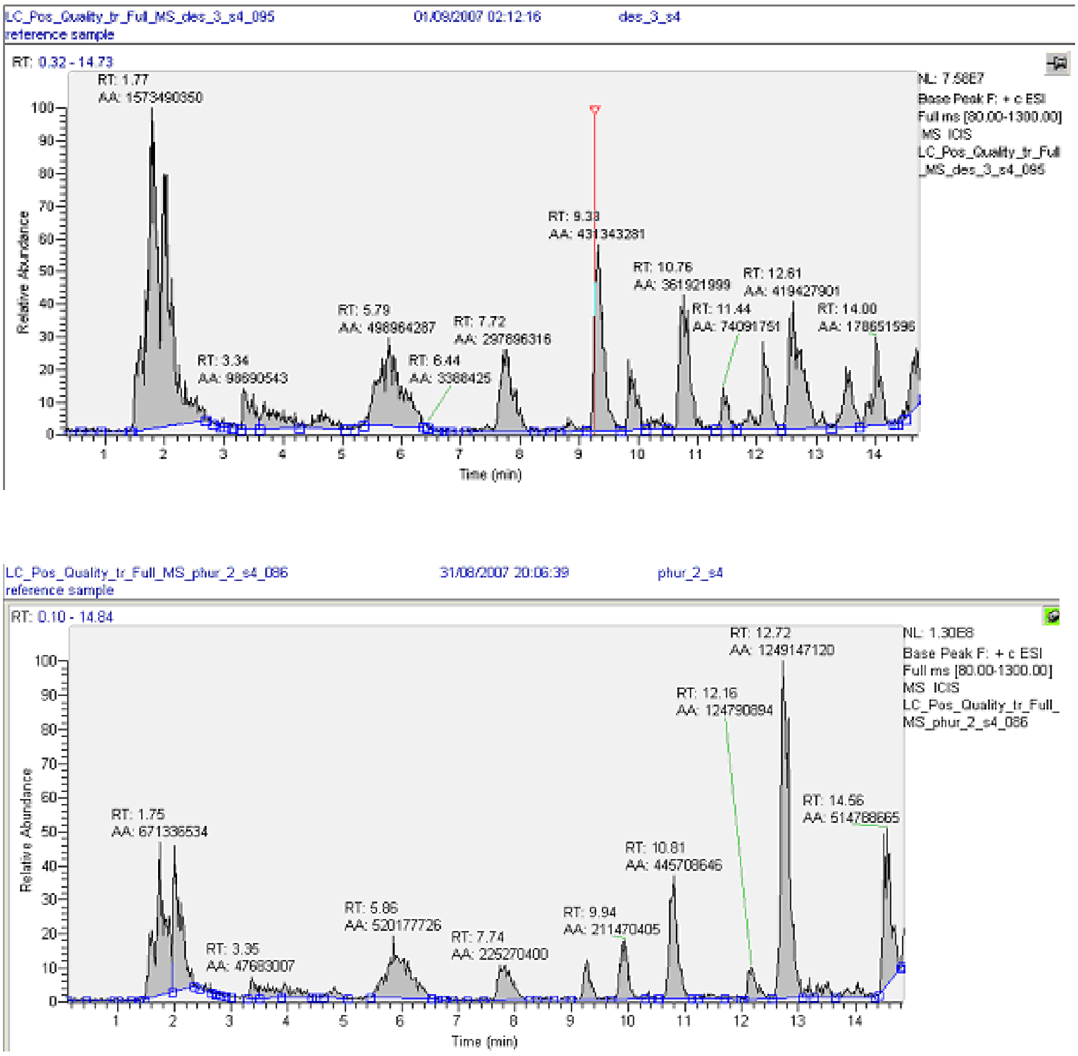


Figure 4.8 The LC-MS (Thermo LTQ systems) polar extract chromatograms of Phureja (upper) and Desiree (lower).

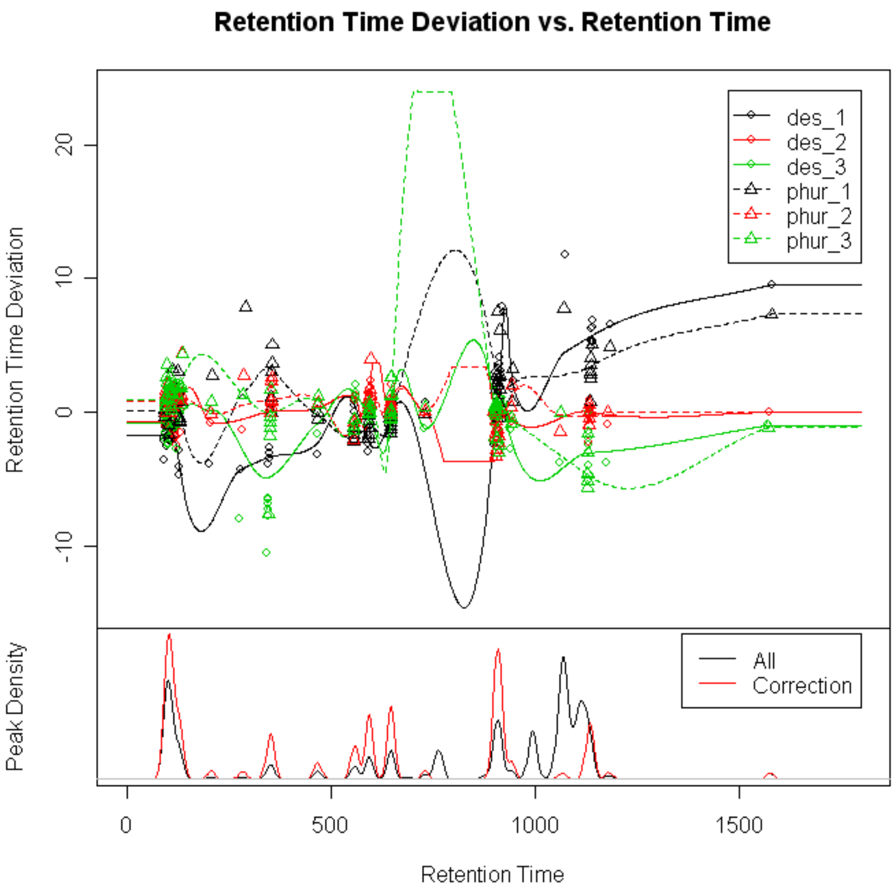


Figure 4.9. XCMS output file for analysis of LC-MS raw data derived from the polar extracts of Desiree, Phureja potato controls and the varieties Inca Sun, Mayan Gold,, Montrose and Pentland Dell. The first two varieties are diploid phureja subspecies whilst the latter two are common tetraploid tuberosum varieties. The output from XCMS is in the form of a retention time deviation vs. retention time plot.

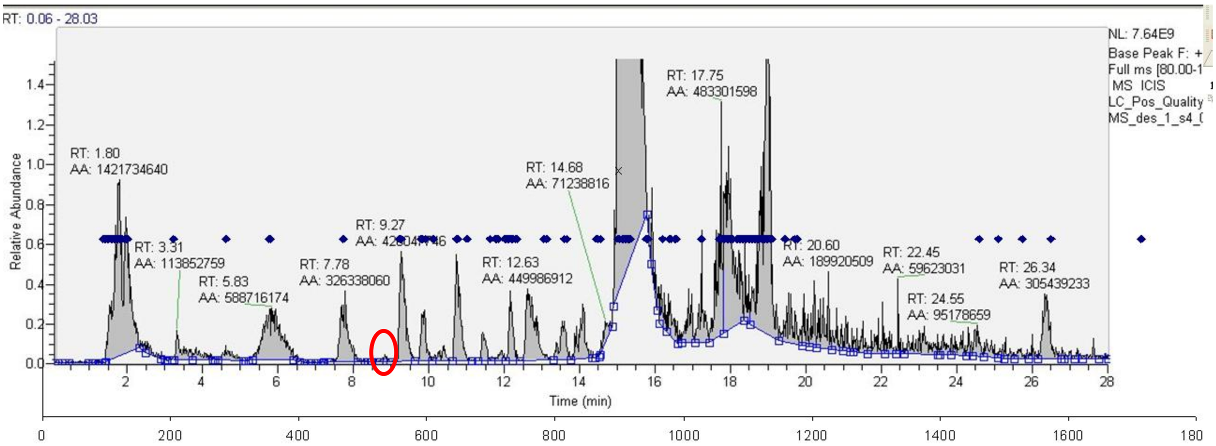


Figure 4.10 The combine original LC-MS TIC and the overlaid XCMS generated (♦) significant peaks. An example of how XCMS can miss peaks is highlighted (**O**) with this peak not present in the overlaid XCMS generated (♦) significant peaks.

Unlike many of the other packages XCMS comes with a peak identification ability via a link to the Metlin metabolite database (http://metlin.scripps.edu/) which requires a mass range as an input. However this metabolite database is limited with respect to the analytes/metabolites available since for the XCMS data reports generated from the potato experimental set the mass 852 (chaconine, a neurotoxic glycolalkaloid common at low levels in all potatoes) yielded 14 different putative structures all of which were either complex fatty acids or glyceride derivatives; Chaconine was not registered within the Metlin database. This is a fatal flaw for food safety and compositional applications

To realistically assess XCMS the full LC-MS experimental dataset was trialled. This contained the cultivars with 3 harvest points, 2 quality control samples and 3 injections were analysed. This represented 6 cultivars (4 test + 3 control) x 3 harvests/reps x 2 peak-find passes and this took a total time of 12 minutes. The data was saved to text file and analysed using SIMCA (PCA analysis). The exact same data was also subject to analysis using the Thermo Xcalibur software and similarly analysed by PCA.

Desiree and Phureja (QC data) cluster in both data sets and are segregated away from the experimental data set. Figures 4.11 highlights that close similarities between the Xcalibur and XCMS processed data as represented on PCA plots. The associated loading plots are represented in Figs 4.12 a & b and 4.13 a & b, respectively. In the Xcalibur processed data peak P205RT9.51 (Fig 4.12b) is the same peak as M205T591 in the XCMS data set (Fig 4.13b) and within the XCMS data there are other peaks at the same retention time suggesting breakdown products from a single compound (mass fragments). In general there is a broad agreement with the segregation of compounds (Xcalibur) and masses/RT (XCMS) but the latter (XCMS) is predisposed to miss peaks unless pre-set on a punishingly repetitive re-scan schedule. For example in the XCMS dataset are peaks with masses 166 and 103 with a retention time of 7.4 minutes were missing but present in the Xcalibur dataset. This was a function of the data being subject to only 2 passages through the XCMS peakfind routine. Passes through the peakfind routine in conjunction with statistical analysis assessing the validity of the additional information gained will guide the user to the number of peakfind passes required to extract all relevant data. However this will be completely dependent on the analytical setup procedure and cannot be enshrined in an SOP.

In conclusion, XCMS runs on the R platform which is open source, and becoming widely recognised as a portable platform for statistical analysis. The limited trial highlighted the utility of this approach in comparison to targeted bespoke software (in this case Xcalibur). Furthermore, XCMS is not intuitive and required a greater degree of programming skills than are routine found amongst the chemical/biochemical community. However, if used as part of a metabolomics initiative involving programmers the R-based XCMS software could form the basis of a centralised food database metabolomics repository. Indeed, proficiency within the R environment would allow for scaling using an internal standard, and further statistical procedures could be written to analyse datasets, and these routines would add further consistency when analysing multiple experiments.

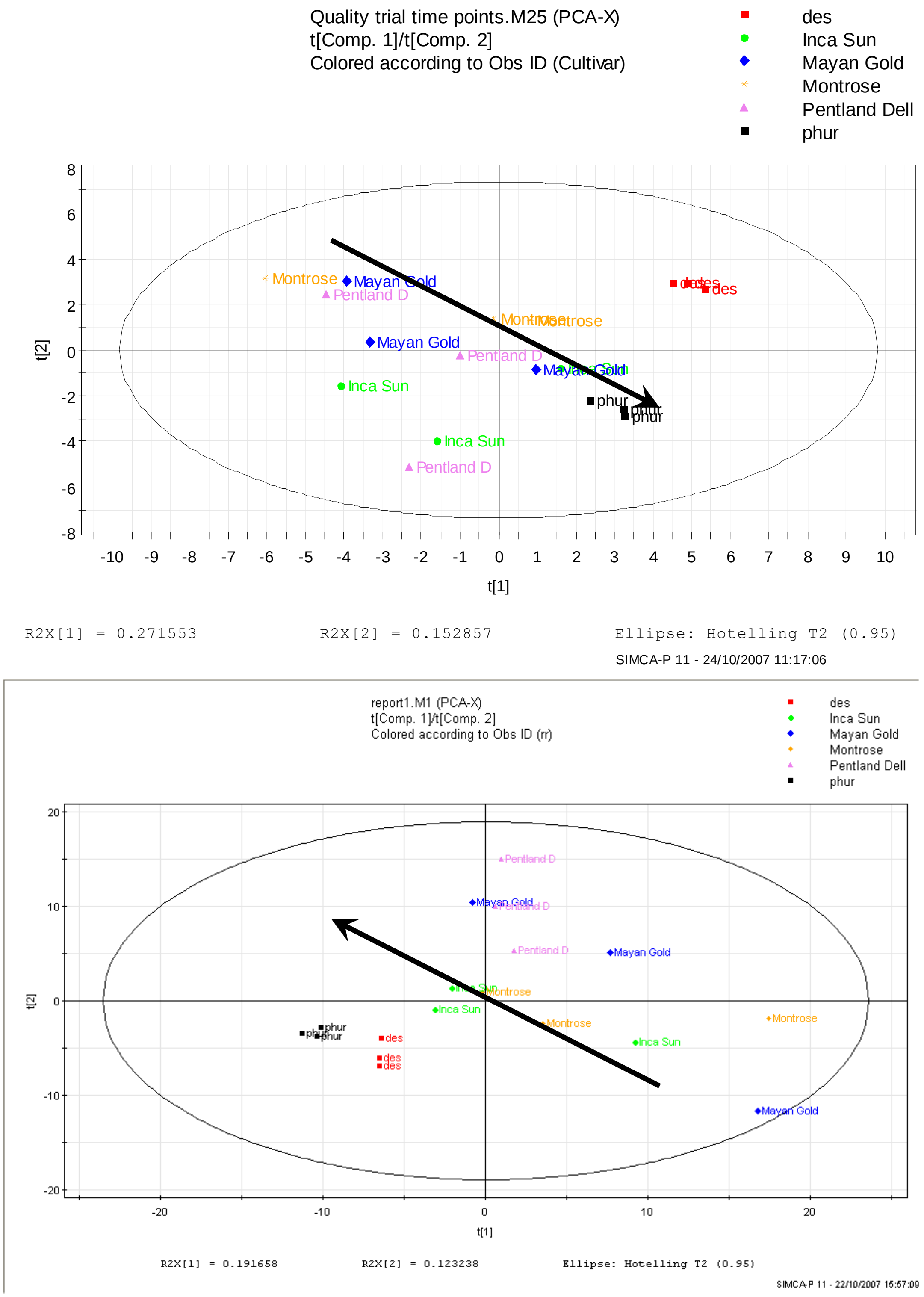


Figure 4.11. PCA plots derived from the Thermo LTC-LTQ-MS systems. The samples were Desiree - ■, Phureja - ■ (potato controls) and the varieties Inca Sun - ♦, Mayan Gold - ♦, Montrose - × and Pentland Dell - ▲. All material was harvested at 3 time points. The PCA data was derived from and Xcalibur (upper) and XCMS (lower) processed peak tables. The arrow represent the trend of harvest date/maturity.

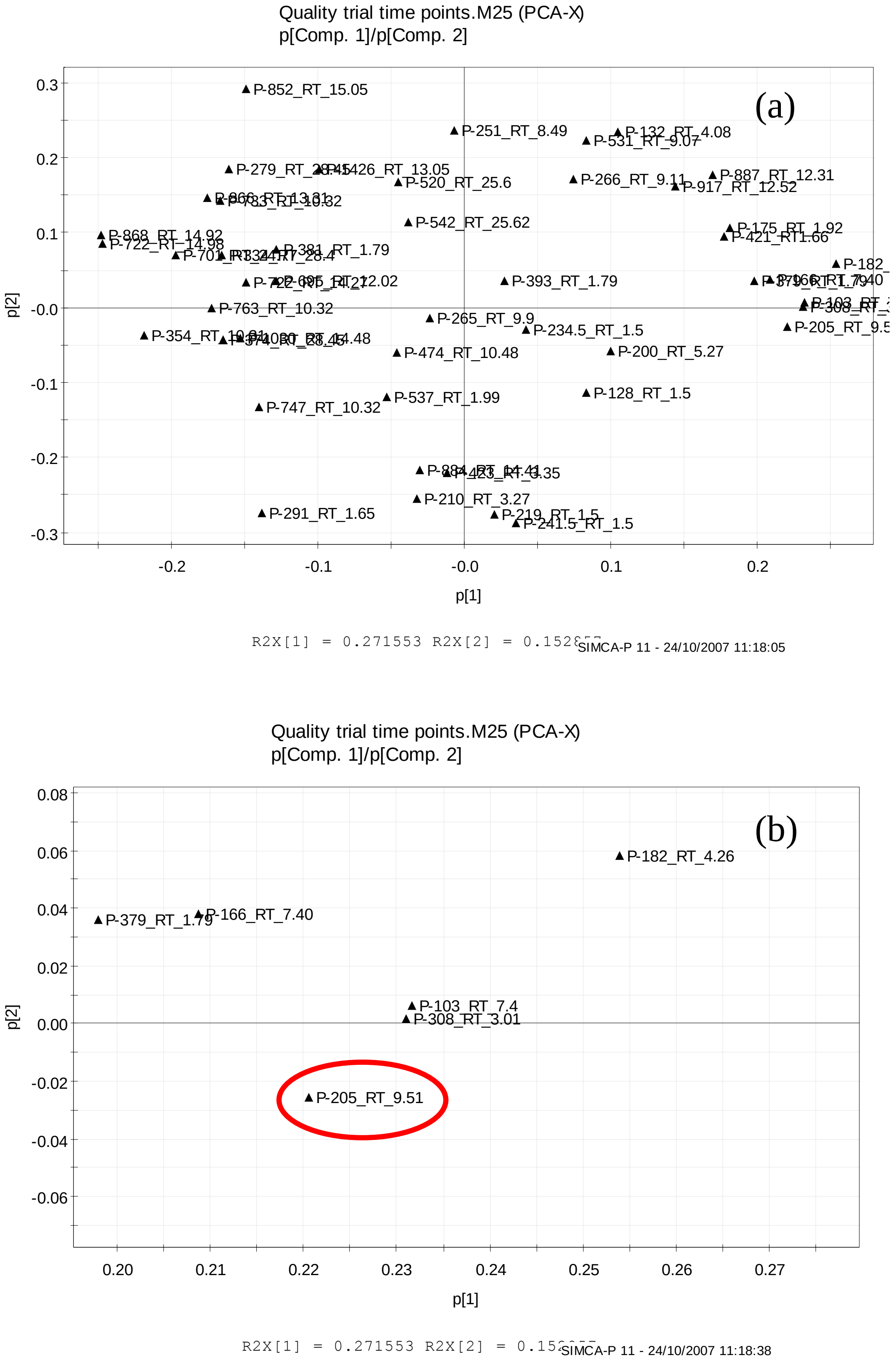


Figure 4.12. The full (a) and exploded (b) PCA loadings plot derived from the Xcalibur derived PCA scores plot Fig 4.11 upper. The outlined data point has a corresponding match in the XCMS dataset (Figure 4.13b).

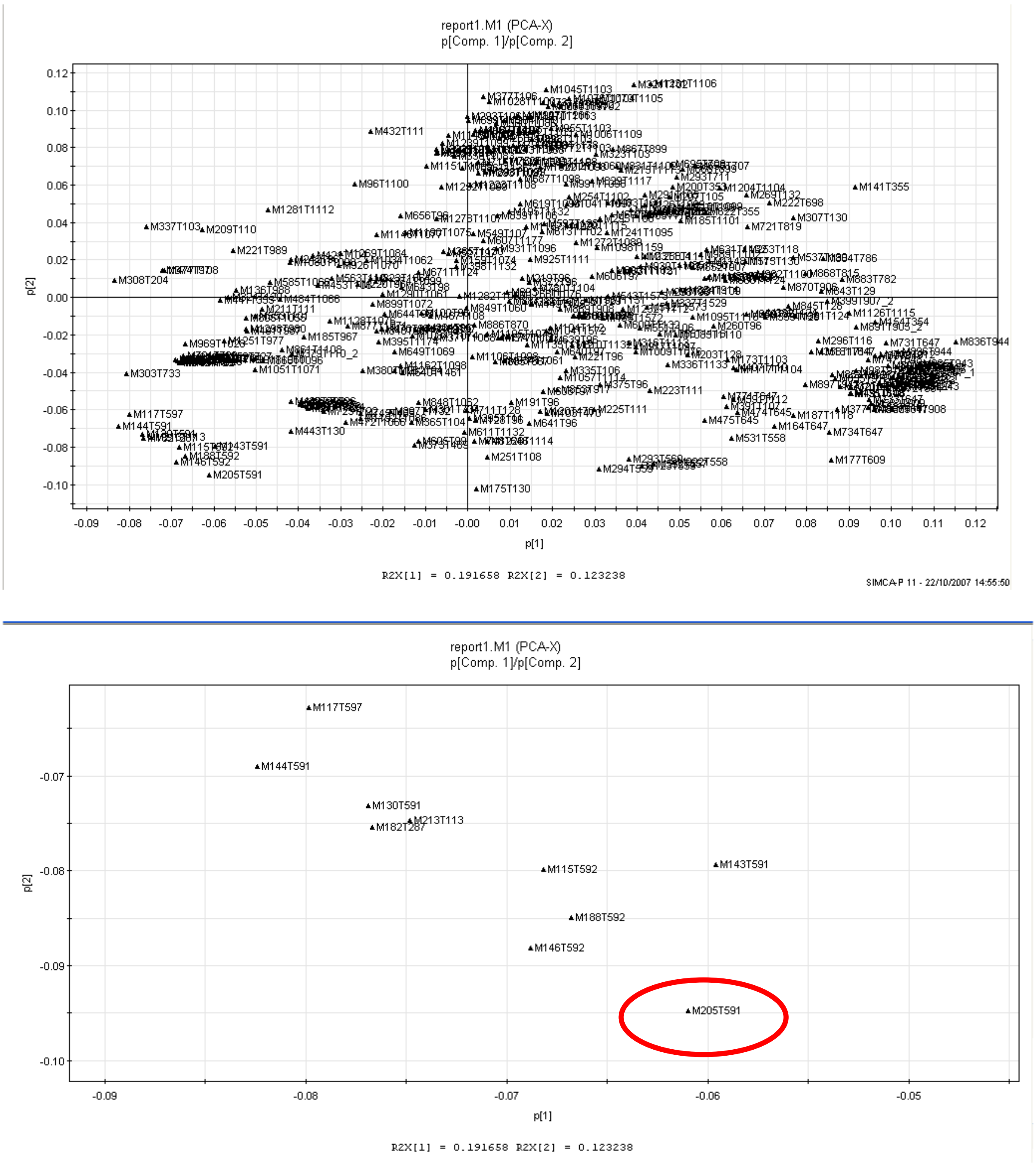


Figure 4.13. The full (a) and exploded (b) PCA loadings plot derived from the XCMS-derived PCA scores plot Fig 4.11 lower. The outlined data point has a corresponding match in the XCMS dataset (Fig 4.12b)

REFERENCES

America A.H. et al (2006) Proteomics. 6, 641-53.
Baran, R. et al (2006) BMC Bioinformatics 7, 530.
Beckmann, M. et al (2007). J. Agricultural and Food Chemistry 55, 3444-3451.
Beckmann, M. Et al (2008). Nature Protocols 3,486-504.
Bellew, M. et al (2006) Bioinformatics 22, 1902.
Bino, R.J. et al (2005) New Phytol. 166, 427.
Bournay, A.S et al, (1996) Nucleic Acids Research, 24, 2347-2351
Broeckling, C.D. et aI (2006) Anal. Chem. 78, 4334.
De Vos R.C. (2007) Nat Protoc. 2, 778-91.
Dobson, G.D. et al (2007) Application of Metabolite and Flavour Volatile Profiling to Studies of Biodiversity in Solanum Species. In Concepts in Plant Metabolomics. Eds. Nikolau, B.J. & Wurtel E.S., Springer Netherlands, 209-219
Dobson, G.D. et al (2008) J. Agric Food Chem. In Press.
Dromey, R.G; (1976) Anal. Chem. 48, 1368-1375.
Dunn WB et al. (2008) J Chromatogr B Analyt Technol Biomed Life Sci. [Epub ahead of print]
Duran AL (2003) Bionformatics 19, 17 2283–2293.
Duran, A.L. (2003) Bioinformatics 19, 2283.
E.J. Saude, C.M. Slupsky, B.D. Sykes, Metabolomics 2 (2006) 113.
Eide, I. et al (2002) Environ. Health Perspect. 110, 985.
Enot, D. et al (2008). Nature Protocols, 3, 446-70.
Enot, D.P. & Draper, J. (2007) Metabolomics 3, 349-355.
Enot, D.P. et al (2007). Metabolomics 3, 335-347.
Fiehn, O. et al (2006) OMICS.10 158-63.
Fiehn, O. et al (2008) Plant J. 53, 691-704.
Frenzel, T. (2003) Eur. Food Res. Technol. 216, 335.
Go, E.P. et al (2006) J. Proteome Res. 5, 2405.
Halket et al., (1999) Rapid. Commun. Mass Spectrom. 13, 279-284;
Hargrove, W.F. et al., (1981) Anal. Chem. 53, 538-539;
Herron, N.R., et al., (1996) Amer. Soc. Mass Spectrom. 7, 598-604;
Hongmei L et al (2008) Trends in Analytical Chemistry, 27,215-227.
Idborg, H. et al (2005) J. Chromatogr. B 828, 14.
Jaitly, N. et al (2006) Anal. Chem. 78, 7397.
Katajamaa, J and Orešič M. (2005) BMC Bioinformatics 6, 179-191.
Katajamaa, J. et al (2006) Bioinformatics 22, 634–636.
Katajamaa, M. and Oresic, M (2005) BMC Bioinformatics 6, 179.
Katajamaa, M. et al (2006) Bioinformatics 22, 634.
Katz, J.E. et al (2004) J. Am. Soc. Mass Spectrom. 15, 580.
Keurentjes, J.J.B et al (2006) Nat. Genet. 38, 842.
Kind, T. et al (2007) Anal. Biochem. 363,185.
Kind, T. Et al (2007) Analytical Biochemistry 363 (2007) 185–195
Lenz, M.E. (2005) Biomarkers 10, 173.
Leptos, K.C. et al (2006) Proteomics 6, 1770.
Li, X.J. et al (2005) Mol. Cell. Proteomics 4, 1328.
Listgarten, J. et al (2007) Bioinformatics 23, e198.
Mueller LN, et al, (2007) Proteomics. 2007 Oct;7(19):3470-80.
Nordström A etal (2006) Anal Chem. 78, 3289-95.
Nordström A. et al. (2008) Anal Chem. 80, 421-9

Overy, D.P. et al (2008). Nature Protocols, 3, 471-85.
Palagi, P.M. et al (2005) Proteomics 5, 2381.
Rischer, H. (2006) Proc. Natl. Acad. Sci. U.S.A. 103, 5614.
Rivero, R.C. et al (2003) Food Chemistry 80, 445-450
Sadygov, R.G et al (2006) Anal. Chem. 78, 8207.
Shackelford, W.M; (1983) Analytica Chim. Acta 146, 25-27.
Smith C.A. et al (2006) 2006 78, 779-87.
Stein, S.E. (1999) J. Am. Soc. Mass Spectrom. 10, 770-781
Stein, S.E. (1999) J. Am. Soc. Mass Spectrom. 10, 770-781;
Stein, S.E. and Scott, D.R. J (1994) J. Amer. Soc. Mass Spectrom. 5, 859-866;
Tikunov Y. et al (2005) Plant Physiol. 139, 1125-37.
Tikunov, Y. et al (2005) Plant Physiol. 139, 1125.
Umar, A. et al (2007) Proteomics 7, 323.
Vorst, O. et al (2005) Metabolomics 1, 169.
Wagner, S. et al (2007) Anal. Chem. 79, 2918
Want E.J. et al (2006) Anal Chem. 2006 Feb 1;78(3):743-52.
Want, E.L. (2006) Metabolomics 2, 145.
Whitfield, P.D. et al (2005) Metabolomics, 1 215.
Wikoff WR (2007) Clin Chem. 53, 2169-76.
Zhang, W. et al Rapid Commun Mass Spectrom. 2006;20(10):1563-8.
Zhang, X. et al (2005) Bioinformatics 21, 4054.
Zhao, Q. et al (2006) Bioinformatics 22, 2562.

Appendix 4.1 Evaluation of software packages

Name	MZmine
Source	MZmine - LC/MS Toolbox (http://mzmine.sourceforge.net/features.shtml
Reference	BMC Bioinformatics 2005, 6:179 (http://www.biomedcentral.com/1471-2105/6/179/abstract)
Availability	GNU GPL - Java software - multi platform - source code available
Last release	V 0.60, April 2006 (possibly still maintained)
General Use	Tool for processing of LC/MS raw data
Features	smoothing and filtering of spectra; visualisation; peak picking and automatic peak detection.
General Comments	Peak table generation seems a rather rudimentary approach as it is based by searching the closest peaks from known master peak list - A peak is defined by {mz, δmz, rt, δrt, height, area}
Input	Import NetCDF/mzXML files
Output	Peak table of dimensions: Num. of samples * Num. of peaks in the master peak list

Name	Chromatof
Source	**LECO Instruments -** http://www.leco.com/products/sep_sci/chromaTOF/chromaTOF.htm The software cannot be bought independently of the machine
Reference	
Availability	Commercial instrument software for operating Leco GC-tof-MS instruments
Last release	V3.25 optimized for Pegasus
General Use	Tool for processing of GC and LC/MS raw data
Features	Built-in automatic deconvolution; peak-list generation; peak-list alignment.
General Comments	Peak list alignment for metabolomics data not satisfying as the algorithm is based on deconvolved mass spectra, which do not always represent the actually deconvolved peak
Input	Leco raw-data only
Output	NetCDF files, CSV files, peak lists including database (NIST) search results and deconvolved mass spectra

Name	AnalyzerPro
Source	**http://www.spectralworks.com/MatrixAnalyzer.aspx**
Reference	http://www.spectralworks.com/analyzerpropublications.asp
Availability	Commercially available
Last release	
General Use	data processing application for MS data
Features	Accurate Mass Capable - handles high precision data as well as integer mass. Any Chromatographic MS data - any data that has some aspect of time resolution Intuitive User Interface - builds on web browser familiarity Qualitative Analysis - because you can expect the unexpected Quantitative Processing - using external or internal standards Combined Quan and Qual Reports - more extensive sample information in less time Custom Reports - readily generation to be application specific Automatic Spectral Enhancement - no user intervention or prior knowledge required Seamless Integration with NIST - and any NIST format library Targeted and non-Targeted Analysis - flexible and comprehensive Vendor Independent - one system for all including legacy instruments
General Comments	Peak list alignment for metabolomics data not trivial.
Input	most vendors raw data and NetCDF files
Output	Target Component Libraries; Target Quantitation Libraries

Name	BinBase
Source	http://fiehnlab.ucdavis.edu/projects/binbase_setupx
Reference	Fiehn O, Wohlgemuth G, Scholz M (2005) Automatic annotation of metabolomic mass spectra by integrating experimental metadata. Proc. Lect. Notes Bioinformatics 3615, 242439 (http://fiehnlab.ucdavis.edu/publications/publications/DILS%202005_Fiehn_Scholz_Wohlgemuth.pdf
Availability	Web interface freely accessible
Last release	19/06/2008 V3.0
General Use	web-based metabolomics LIMS, Database system for automated metabolite annotation
Features	high performance clustering: fatty acid methyl ester based retention index correction; kovatch retention index based retention index correction; automated library generation; central configuration; linking with external resources; replacing of empty values; support for GC-Tof with Leco pegasus 3.0/ 2.*; support for quadrupole; GCxGC support; import msp files as reference library; export of msp files; LGPL compatible; highly customizable matching of samples; integration with setupX; integration with external LIMS systems; possible to access using web services; multi user support; based on meta informations for a smarter matching and bin generation; fully automated; retention index correction based on internal/external standards: quantification possible.
General Comments	Not intuitive requiring a significant level of chemometric trailing
Input	UniqueMass; S/N; Purity; R.T (seconds); Quant S/N; Quant Masses; Dimension Time 2
Output	XML. XLS , TXT

Name	MET-IDEA (Metabolomics Ion-Based Data Extraction Algorithm)
Source	http://bioinfo.noble.org/download/
Reference	Broeckling CD, Reddy IR, Duran AL, Zhao X and Sumner LW. MET-IDEA: data extraction tool for mass spectrometry-based metabolomics. Anal Chem. 2006 Jul 1;78(13):4334-41. (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=16808440&query_hl=4&itool=pubmed_docsum)
Availability	Freeware
Last release	07-01-2008, V2.0
General Use	Data processing tool to quantify ion abundances from GC(LC)-MS data - General idea is that the user defines the master list of compounds in a form of "ion/retention time" pair. The software will then extract an estimation of the area of each "IRt" in every sample
Features	
General Comments	Fine as far as it goes but it has limited scope for datamining unknowns
Input	Raw NetCDF fies: ASCII file with IRt pairs - can also be manually edited in the software extracted from AMDIS output.
Output	peak table - each variable representing one IRt

Name	MetaQuant
Source	http://bioinformatics.org/metaquant
Reference	MetaQuant: a tool for the automatic quantification of GC/MS based metabolome data, Bioinformatics Advance Access 17/10/2006 (http://bioinformatics.oxfordjournals.org/cgi/content/abstract/btl526v1?papetoc)
Availability	Freeware
Last release	14/01/2008
General Use	
Features	Import of GC/MS data in a particular CSV or netCDF format Automatic, lab tested peak recognition and peak integration algorithms Automatic calibration function Different regression algorithms Usage of retention times or retention indices for compound definition, calibration and analysis

	Quantification with or without prior calibration Improved peak analysis correction functions with possibility of summarization and manual integration Export of quantification results into MS-Excel, XML , CSV and SBML file format Graphical Batch-Analysis with possibility of analysis comparison Powerful commandline-based Batch-Analysis with or without MetaQuant config file (.mcfg) Commandline analysis creates a MetaQuant result file (.mres), which can be opened with the GUI version Compound classification options (e. g. KEGG, CAS) Internal standard can be described on preferences panel leading to normalized peak areas Easier calibration with equimolar mixes
General Comments	Quantification of GC-MS metabolomics profiles - Performs automatic quantification of a predefined set of compounds in one or several chromatograms
Input	raw GC/MS profile in CSV/NetCDF format
Output	XLS, XML and SBML peak tables

Name	MSFACTs (Metabolomics Spectral Formatting, Alignment, and Conversion Tools):
Source	http://www.noble.org/PlantBio/MS/MSFACTs/MSFACTs.html
Reference	Duran, A.L., Yang, J., Wang, L. and Sumner, L.W. (2003). Metabolomics Spectral Formatting, Alignment and Conversion Tools (MSFACTs). Bioinformatics 19(17): 2283-2293. Similar approach described in "Progressive peak clustering in GC-MS Metabolomic experiments applied to Leishmania parasites", Bioinformatics 2006 22(11):1391-1396
Availability	Freeware for academia
Last release	Aug 2003 - no recent update/release - Software support/maintenance doubtful
General Use	Alignment of integrated chromatographic peak lists
Features	
General Comments	Difference of retention times between two adjacent peaks in one run must be smaller than RT differences between 2 runs. Decision regarding the window size must be taken. RT alone is used to resolve "collisions" (two peaks from one run are falling in the same cluster). Reported results are based on fairly well resolved chromatogram. The application for true metabolomics is therefore limited Completeness of RTalign has not been reported
Input	ASCII files
Output	tab delimited data

Name	GASP
Source	http://www.genedrift.org/gasp.php
Reference	
Availability	Freeware to academia
Last release	01-10-07
General Use	Aligns peaks from multiple gas chromatograms enabling the user to detect statistically significant differences between levels of compounds present in the specimen under study
Features	The program reads output files from Xcalibur, HP Chemstation and AMDIS, converts and align these files in order to make the comparisons. Format of file not known Can also align chromatographic data processed by AMDIS deconvolution software (available from NIST). R plugins to allow simple forms of statistical analysis
General Comments	Alignment of integrated chromatographic peak lists - No description of the techniques to correct RT, neither to perform multiple alignment
Input	Xcalibur, HP Chemstation and AMDIS peak list files
Output	Peak tables

Name	MetAlign
Source	http://www.pri.wur.nl/UK/products/MetAlign/
Reference	. Tikunov, A. Lommen, C.H.R. de Vos, H.A. Verhoeven, R.J. Bino, R.D. Hall, P. Lindhout, A.G. Bovy, Plant Physiol. Break Through Technologies Section 139 (2005) 1125-1137. A Novel Approach for Non-targeted Data Analysis for Metabolomics. Large-Scale Profiling of Tomato Fruit Volatiles.
Availability	Formerly commercial now freeware
Last release	08-01-07
General Use	GC and LC-MS
Features	Includes highly parameterized smoothing, baseline correction and statistical chromatographic alignment options. These options were recently extended by a mass alignment procedure for the accommodation of high mass-accuracy instruments.
General Comments	similar to that of Xcalibur, however deciphering the components of the dataset which are driving these results from the many thousands of data points produced using MetAlign becomes arduous. Also MetAlign is applicable to GC-EI-TOF-MS analysis, but appears **not to perform deconvolution**, in other words the combination of extracted mass tags into full MSTs.
Input	Universal acceptor
Output	

Name	XCMS
Source	http://metlin.scripps.edu/download
Reference	Smith CA, Want EJ, O'Maille G, et al. XCMS: Processing mass spectrometry data for metabolite profiling using Nonlinear peak alignment, matching, and identification. Analytical Chemistry, 2006, 78(3), 779-787
Availability	Freeware
Last release	V 1.11.22, 28-04-08
General Use	peak picking and nonlinear alignment software predominantly for LC-MS
Features	
General Comments	Processing tools for LC(GC)-MS data - Performs automatic peak identification and integration by fitting a Gaussian peak model on one trace - Peak are then grouped together if they close enough - Missing peak infilling by integration the expected time window - Allows iterative retention time correction based on "well behaved" peaks
Input	Reads and processes LC/MS data stored in NetCDF (AIA/ANDI), mzXML, and mzData files
Output	Peak table of selected peaks

GO3 Final Report - Section 5

Manfred Beckmann/John Draper

Development of methodology for standardized generation and pre-processing of ESI-MS metabolite fingerprint data (Task 02.02 and parts of Task 04.02 and 04.03)

5.1 BACKGROUND TO EXPERIMENTAL SECTION

A key aspect of the G03 project was to ensure that the metabolite fingerprinting techniques developed in G02 had both in-built data quality assessment and importantly will produce data that might be used for 'predictive' classification of samples in the future. An important practical consideration is that the techniques developed must be sufficiently robust to allow direct comparison of data generated on different instruments. For example there is little point in developing a technique if current data can only be compared with data generated on the same machine within the same run batch. By definition, global metabolite fingerprinting techniques are not quantifying a set number of discrete variables in comparison with authentic standards, but instead are providing 'snapshots' of the relative ratios of metabolite signals in exceedingly complex mixtures. Against this background it is clear that protocols for assessing data quality and ensuring any data 'calibration' to cope with inherent instrument drift or detection of occasional operator-related variance have to be developed. As the data are highly dimensional (1000's of variables) then fingerprint data quality cannot be checked by eye or checked by using simple univariate statistics, instead multivariate analysis techniques must be used to reduce dimensionality in order to assess reproducibility**.** Thus procedures need to be developed to ensure that robust data comparisons can be made both within and between sample batches analysed at different times over a period of several months.

At the time of writing the merged G03 project it was assumed that NMR fingerprinting data from IFR relevant to potato would be available. This proved to be the case, but on deeper investigation it was clear that the data set were only produced on a single instrument and thus there was no scope for aligning data output from different machines. In addition the data was limited to samples generally not covered by other major profiling/fingerprinting experiments and so integration of NMR results with other metabolomics data was not possible. Additionally, from discussions with NMR experts it was also clear that NMR data was absolutely quantitative and thus pre-processing was generally rigorously handled by instrument software and as such there were few instances were operator involvement was required. For these reasons the limited work on examining the IFR NMR data was dropped from the G03 project.

During the process of merging the two separate GO3 projects it was agreed with the FSA that research would use non-transgenic potato data collected during the G02 project as well as new data sets produced in external projects for validation purposes. As much work had already been carried out on the transgenic potato tuber data sets used in the G02 project it was important to have some independent samples to validate, especially the data pre-processing and data mining methodologies. It was initially hoped that further metabolomics data sets relevant to potato would become available from external laboratories in the SAEEFOODERA programme in which SCRI was a participant. In the event these data were not available in time and were replaced by more extensive and challenging sample populations produced by SCRI and Aberystwyth in different projects. These samples included a major focus on plant tissue extracts of both starchy tubers (potato) and leaves (*Brachypodium*, rice and *Arabidopsis*). In particular a complex experiment involving rice blast (a fungal pathogen) infection of *Brachypodium* leaves provided a very challenging sample set for data alignment and was used to validate both the extraction, data pre-processing and data mining SOPs. As an aside, for the purpose of developing standardized metabolomics methods for ESI-MS the extraction protocols were also tested with further commonly encountered sample types including microbial cultures (yeast), blood fractions and urine.

5.2 OVERALL OBJECTIVES OF EXPERIMENTAL SECTION

The ability to compare multivariate data from different time batches is imperative to success. In the last four years in G02006 (and other internal projects) we have explored several routines for mass alignment to validate ESI-MS fingerprint data with acquisition dates from within a week to as much as 6 months apart. Our first major intention was to examine the methodology which was used previously in the G02 project to generate ESI-MS data sets on two very different systems (a Micromass LCT and Thermo-Finnigan LTQ linear ion trap respectively). At the same time we also reviewed recent publications concerning the development of mass spectrometry fingerprinting. From this background we developed a standardised SOP describing in detail all the common steps (with alternative instructions were necessary) to allow the production of quality assured ESI-MS fingerprint data on any instrument. The methodology review covers aspects of data generation including considerations relating to overall experimental statistical design, sample extraction and approaches for quality assurance. Following this we validated a data generation and data pre-processing procedure which can be used to prepare ESI-MS data in a common format for data mining. Once shown to be working in potato the validation was undertaken using an independent data set representing disease progression in Brachypodium leaves infected by the rice blast fungus (data produced in a separate BBSRC project in Aberystwyth: see Parker et al., 2008 [49]). For convenience this validated protocol is described separately from the introduction and discussion sections as a separate section (5.4).

A key element of these procedures is the development of standardised software tools for data pre-processing. The development of these tools continued to be in conjunction with staff on the BBSRC MetRO project and this activity is described in detail in Section 8 of the report which will need to be consulted for background. The final protocols are presented in an instructional format in Section 5.4 and have been published recently in a Nature Protocol (Beckmann et al., 2008 [60]). Finally, one of the overriding requirements of FSA GO3 project is to develop a standardised fingerprinting procedure for assessment of overall potato tuber chemical composition. With this in mind we will explore the use of ESI-MS fingerprinting to classify and explore the differences between traditionally-bred potato cultivars used as control in the G02006 project developing tools to test substantial equivalence between GM and non-GM potato tubers (see G02006 report).

Ultimately in the G03 project we will explore a strategy for fingerprint data alignment and comparison to develop a database of fingerprint information that will have value for compositional comparisons many years ahead. This is described in detail in Section 9 of the report and has been published in Beckmann et al., 2007 [48].

5.2.1 Summary of objectives

- Review literature on direct infusion/flow injection metabolite fingerprinting and introduction to the major issues related to standardization of FIE-MS fingerprinting procedure
- Develop standardized protocols for generation of FIE-MS data which can used on both ion trap and tof instruments (LTQ, LCT). These should include QA procedures for ESI-MS fingerprinting.
- Develop standardized approaches for data pre-processing (base line corrections, outlier detection, normalization.

5.3 INTRODUCTION TO METABOLITE FINGERPRINTING USING NOMINAL MASS FLOW INJECTION ELECTROSPRAY MASS SPECTROMETRY (ESI-MS)

Genomic technologies are now commonplace in experimental strategies aiming to understand specific gene function, in the search for molecular 'biomarkers' linked to disease or to determine drug mode of action[1, 2]. Likewise, plant breeding programmes investigating complex traits relevant to agriculture or researchers attempting to optimise fermentation processes increasingly utilise global profiling methods to assay gene expression and metabolism under circumstances where the genetic basis of any differences are unknown [3, 4]. In contrast to transcriptomics or proteomics experiments the global metabolite content of the majority of biologically-derived samples is only partially understood [5]. Metabolite diversity estimates suggest that upwards of 200,000 natural compounds exist [6] but unfortunately pure chemical standards are widely available for only a small percentage of this number. Additionally, unlike DNA and proteins which are made of repeating units of a small number of basic components (nucleotides or amino acids, respectively), natural metabolites are represented by a very wide range of chemistries. Developing meaningful information on the total metabolite composition of a biological sample is therefore a difficult task.

Traditional approaches for assaying metabolite content involve the use of chromatography to first attempt to separate metabolites before detection, thus generating quantitative or semi quantitative information on individual metabolites. Such metabolite profiling methods require exquisite control over the chromatographic process to obtain reproducibility and demand rigorous [7] approaches to pre-process data in order to correctly deconvolve, align and annotate peaks. Importantly any chromatography column matrix will undergo gradual detoriation with repetative use, resulting in significant changes in data characteristics ('instrument drift') after a period of constant operation in larger (> 200 samples) profiling experiments. An alternative approach to capture information relating to total metabolite content is to develop a spectrometric 'fingerprint' without any chromatography [8-10]. In fingerprint data single variables do not relate to individual metabolites; depending on the approach, many chemicals can contribute to several signals in any spectrum and it is equally common for several metabolites to contribute to the same signal [4, 9].

5.3.1 Metabolite profiling and metabolite fingerprinting technology

Established technology for quantitation of targeted metabolites in complex mixtures includes gas or high pressure liquid chromatography (GC and HPLC) linked to a variety of detectors [11-15]. However, such approaches are by definition selective [5, 6, 16-18], providing information only on a sub-set of metabolites that are well resolved during chromatography and for which pure standards are available. Using more sensitive 'time of flight' detectors GC-tof-MS 'profiles' comprising the relative ratios of 2-3 hundred polar metabolites have been described [4, 6, 7, 12, 13].

By attempting to profile all metabolite peaks detected automatically by instrument software this latter approach was likely to provide a more comprehensive coverage of the metabolome. However, over 60% of peaks in GC-tof-MS peak tables generally represent unknown chemistry with poor or no matches to existing reference spectra, which makes peak annotation and alignment non-trivial [5-7, 16]. Recent advances in ultra high pressure chromatography instrumentation have stimulated improvements in metabolite profiling using UHPLC [19]. However, peak finding, peak alignment and peak annotation in large sample batches remains challenging as a result of aging of chromatography columns and instrument drift [20, 21].

A more 'global' overview of total sample metabolite composition can be obtained from 'fingerprinting' techniques which do not incorporate a chromatographic step [4, 5, 9, 16, 22]. For example, spectrometeries, such as Fourier Transform Infra Red (FT-IR) [23] and Nuclear Magnetic Resonance (NMR) [10, 24] generate global chemical 'fingerprints', however, they generally require a further level of directed analysis to link any differences in wavenumber

(FT-IR) or chemical shifts (NMR) to specific chemistry [16]. In contrast, ‘fingerprinting’ techniques based on mass spectrometry such as flow injection electrospray ionisation mass spectrometry (FIE-MS) or direct infusion mass spectrometry (DIMS) offer the advantage that the measured ‘variables’ (mass to charge [*m/z*] ratios) can be more directly linked to an individual metabolite by the additional information of atomic mass [25-28].

5.3.2 Flow Injection Electrospray Mass Spectrometry (FIE-MS)

Amongst different types of ion sources (Electrospray ionisation, ESI; Atmospheric pressure chemical ionisation, APCI; Atmospheric pressure photo ionization, APPI; multimode sources combining ESI and APCI) the ESI source can be regarded as standard and is probably the most commonly used ion source. In FIE-MS the analyte is introduced to the ion source in solution as the flow from liquid chromatography, whilst in DIMS the analyte is delivered directly either from a syringe pump, or, via a nano-volume pipette using a liquid handling robot [26, 29, 30]. During the electrospray process ionization is accomplished by the loss or gain of a proton or other adducts, and charged analyte molecules can carry either single or multiple charges. Electrospray is a very ‘soft’ method of ionisation as very little residual energy is retained by the analyte upon ionisation. As such minimal (usually no) fragmentation is produced and the ionised analyte is often referred to as the "pseudo molecular ion” [31, 32]. For studies requiring structure elucidation, this leads to an additional need for tandem mass spectrometry in which parent ions of analyte molecules can be fragmented in a second phase of analysis at higher ionisation energies. Although ideally suited to more polar chemicals, even molecules that do not have acidic or basic groups can be charged through the formation of adducts with various ions such as Cl^- ions in negative ion mode or commonly K^+ or Na^+ in positive ion mode [31]. Common adducts and neural loss fragments detected in the present procedure are described in Table 5.1. The formation of such adducts is highly dependent on the salt content of the crude sample matrix. Initially, these adduct ions may appear to confuse the spectrum, but actually they have diagnostic value in the determination of molecular weight (M) by searching for the simultaneous occurrence of adduct mass differences [4, 9, 27-29, 31-33].

Cation	Mass	Cation	Mass	Anion	Mass
[M+H-HCOOH]1+	**M - 45**	[M+NH_4]1+	M + 18	[M-3H]3-	M/3 - 1
[M+H-H_2O]1+	**M - 17**	**[M+Na]1+**	**M + 23**	[M-2H]2-	M/2 - 1
[M+H-NH_3]1+	**M - 16**	[M+H+CH_3OH]1+	M + 33	[M-H_2O-H]1-	M - 19
[2M+H]1+	**2M + 1**	**[M+K]1+**	**M + 39**	[M-H]1-	M - 1
[2M+Na]1+	**2M + 23**	**[M+2Na-H]1+**	**M + 45**	**[M+Na-2H]1-**	**M + 21**
[2M+2H+3H_2O]2+	2M + 28	[M+H+2K]1+	M + 77	**[M+Cl]1-**	**M + 35**
[2M+K]1+	**2M + 39**	**[M+2H]2+**	**M/2 + 1**	**[M+K-2H]1-**	**M + 37**
[2M+2Na-H]1+	**2M + 45**	**[M+H+Na]2+**	**M/2 + 12**	**[2M-H]1-**	**2M - 1**
[M+H]1+	**M + 1**	**[M+H+K]2+**	**M/2 + 20**	**[2M+Na-2H]1-**	**2M + 21**
[M+Li]1+	M + 7	[M+2Na]2+	M/2 + 23	[2M+Cl]1-	2M + 35
[2M+H+Na]1+	M + 20			**[2M+K-2H]1-**	**2M + 37**

Table 5.1 Expected nominal mass ion adducts commonly found in FIE-MS data. The list of nominal mass ions is restricted to adducts which are likely to form in an electrospray source. The conditions described in this protocol. Mass ions regularly detected in FIE-MS analyses are highlighted in blue. Their occurrence and abundance is however matrix dependent and metabolite specific.

Commonly available LC-MS instruments suitable for FIE-MS vary with regards to detection technique, ion source design, mass range, sensitivity and mass resolution, but most, if not all, can produce a spectrum in which ion signals within the mass range *m/z* 65 to *m/z* 1000 can be designated into *nominal mass* signals differing by 1 *m/z*. Ion trap mass spectrometers are particularly useful for FIE-MS as they allow the isolation of a large number of ions for fingerprint generation which in instruments with tandem detectors are also available for secondary ionisations when studies require further structural information. The basic process used to generate a FIE-MS fingerprint involves the injection of extract into a stream of solvent entering the ionisation chamber at a steady rate. A typical infusion profile may last only 1 min rising sharply to the peak's apex and then tailing off more gradually as analytes slightly delayed by weak interactions and diffusion processes within the solvent lines enter the ion source (Fig. 5.1).

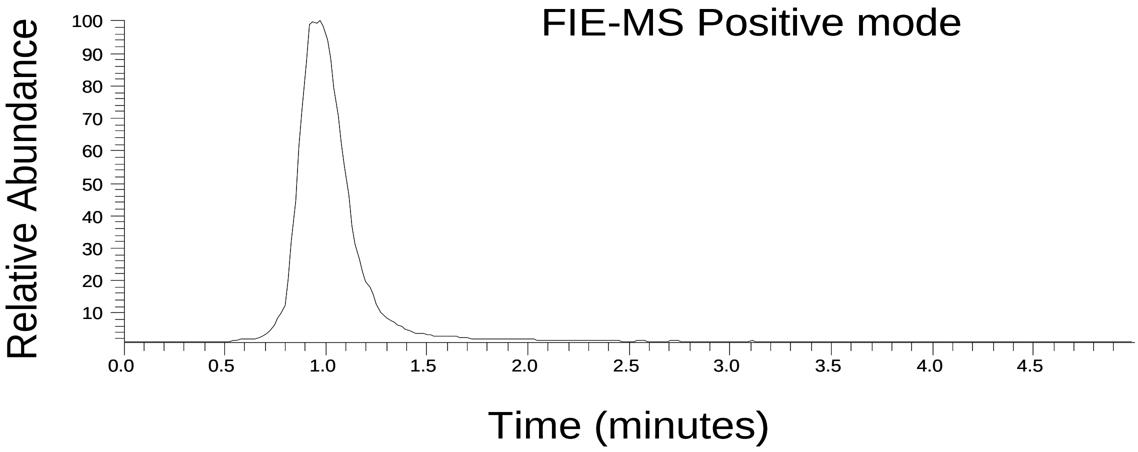


Figure 5.1 Ideal Flow-Infusion chromatogram of *Brachypodium* leaf extract acquired on the LTQ mass spectrometer.

A delay of a couple of minutes before the next sample is infused into the solvent stream can avoid any 'carry over ' between consecutive injections and provides a region of signal 'noise' that can be used for background subtraction (Fig. 5.2).

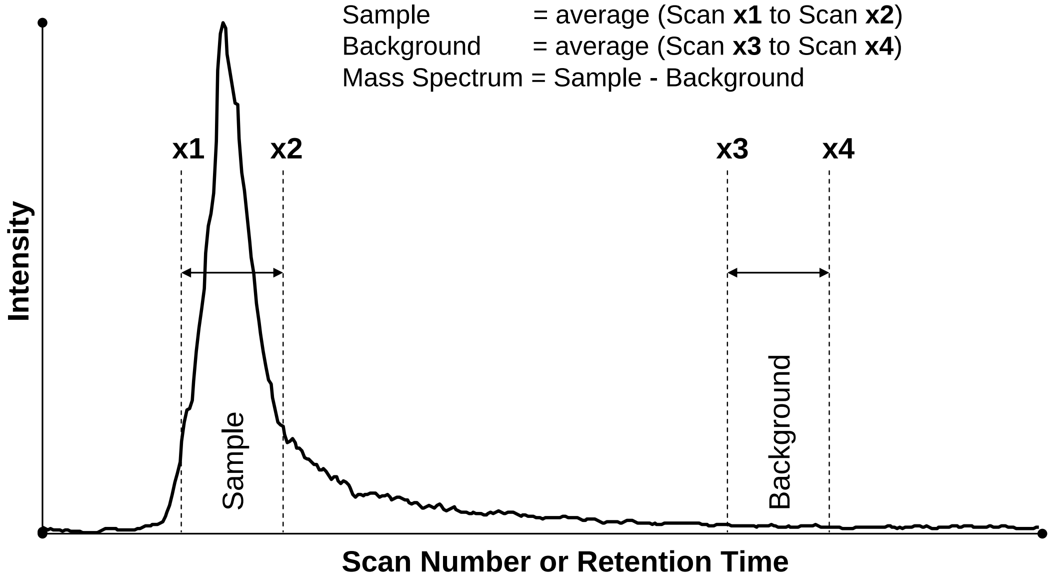

Sample = average (Scan **x1** to Scan **x2**)
Background = average (Scan **x3** to Scan **x4**)
Mass Spectrum = Sample - Background
x1
x2
x3
x4
Intensity
Sample
Background
Scan Number or Retention Time

Figure 5.2 Applying background subtraction for signal processing.

The scan range comprising approximately the central three quarters of the major infusion peak is used to generate mass fingerprint data by averaging all mass-to-charge (*m/z*) intensities (Fig. 5.3).

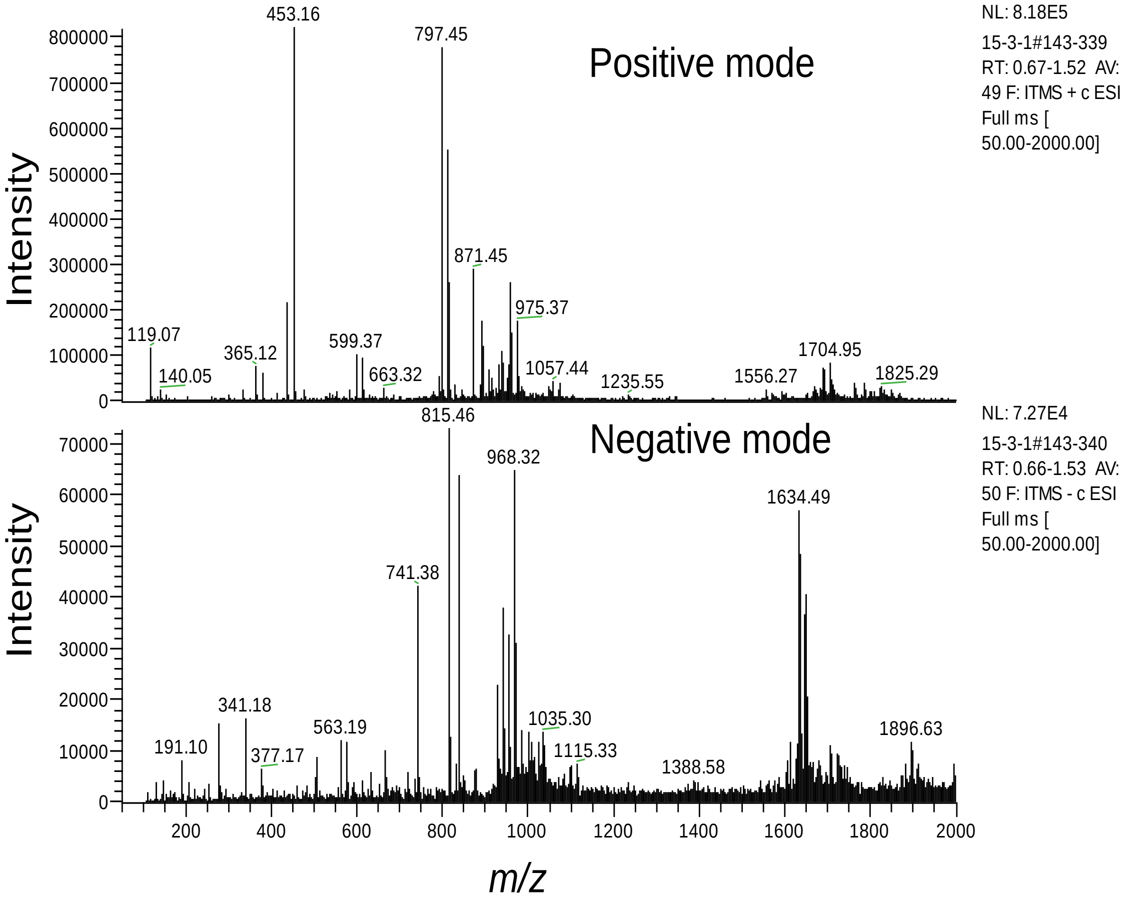


Figure 5.3 Averaged mass spectra of a typical infusion peak from potato tubers

Nominal mass spectra generated by FIE methods thus contain approximately 1000 *m/*z variables where ideally each nominal *m/z* signal represents a metabolite that is separated from different metabolites by virtue of their molecular mass (Fig. 5.4). Clearly, it is possible for several metabolites, associated adducts or other ionisation products, to have very similar or the same nominal molecular weight and thus in FIE MS fingerprints each *m/z* ‘bin’ integrates signals from all metabolite ions which share the same nominal mass [33].

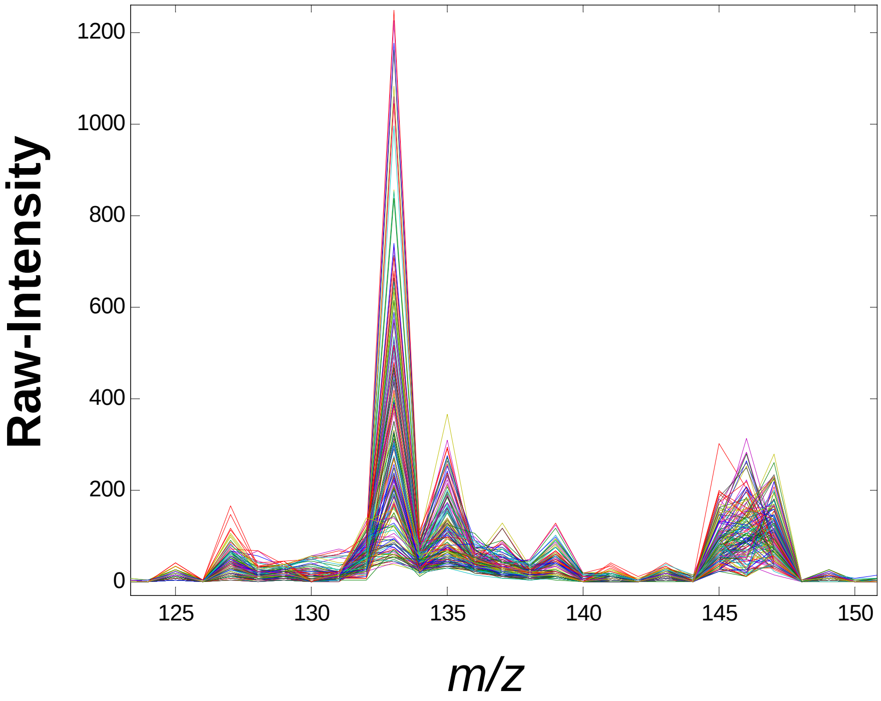


Figure 5.4 Example plot of a nominal-mass raw-intensity matrix after signal processing

MS fingerprints each *m/z* 'bin' integrates signals from all metabolite ions which share the same nominal mass [33].

5.3.3 Constraints in the use of nominal mass FIE-MS

Global compositional assessments of biological samples is made challenging by the chemical complexitity of the metabolome, coupled with huge differences between the concentration ranges of individual metabolites and large natural biological varaibility. Consequently, no single analytical approach can truly offer comprehensive coverage of metabolite diversity and provide strictly quantitative measurements. Generally, metabolite fingerprinting using FIE-MS can provide a useful 'first pass' method for compositional analysis as part of an overall metabolomics investigation (Fig. 5.5) before embarking on a deeper analysis using hyphenated mass spectrometry techniques (GC-MS and LC-MS) targeted to specific metabolite classes [11, 14, 15]. More specifically, FIE-MS is suitable in experiments demanding high sample

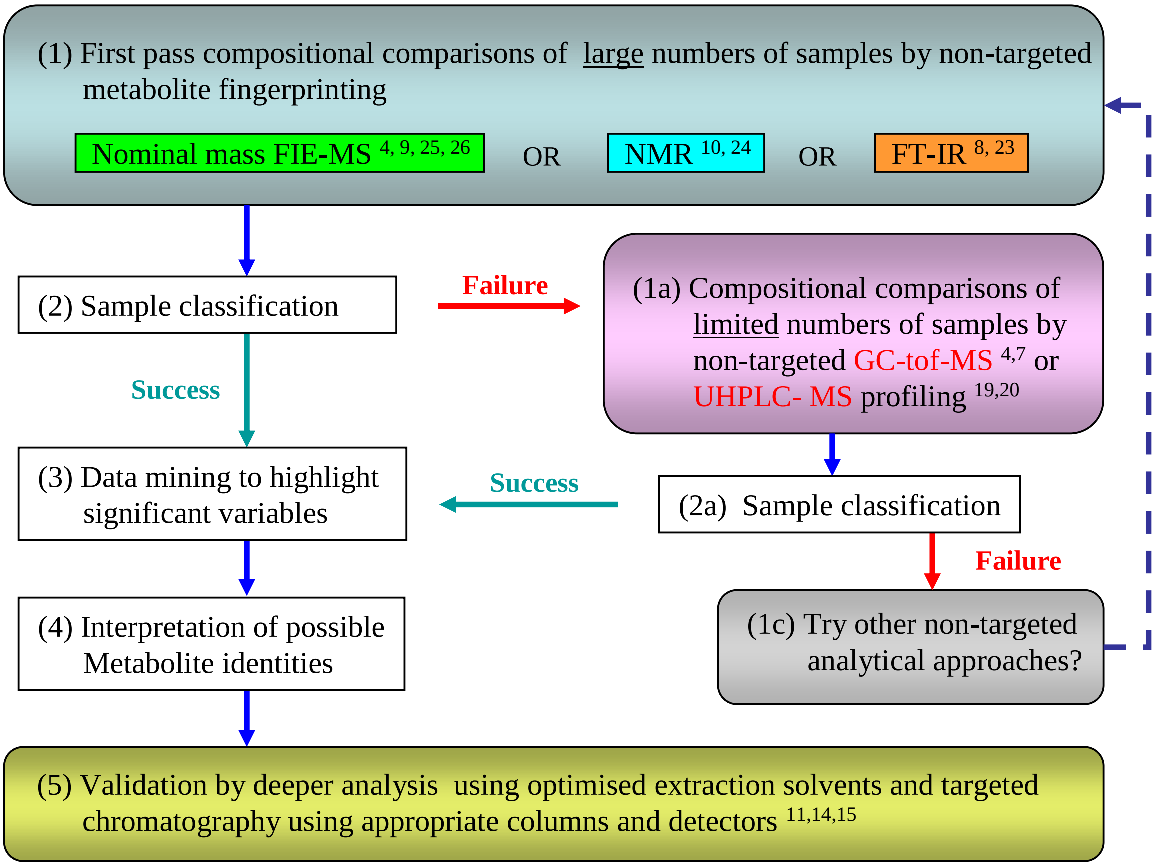

(1) First pass compositional comparisons of large numbers of samples by non-targeted metabolite fingerprinting
Nominal mass FIE-MS 4, 9, 25, 26
OR
NMR 10, 24
OR
FT-IR 8, 23
(2) Sample classification
Failure
(1a) Compositional comparisons of limited numbers of samples by non-targeted GC-tof-MS 4,7 or UHPLC- MS profiling 19,20
Success
(3) Data mining to highlight significant variables
Success
(2a) Sample classification
Failure
(4) Interpretation of possible Metabolite identities
(1c) Try other non-targeted analytical approaches?
(5) Validation by deeper analysis using optimised extraction solvents and targeted chromatography using appropriate columns and detectors 11,14,15

Figure 5.5 Nominal mass, non-targeted FIE-MS fingerprinting in the context of a metabolomics experiment

throughput, as with no chromatography requirements, little sample pre-treatment, simple data pre-processing and instrument cycle times of less than 5 minutes it is feasible to construct FIE-MS experiments capable of analysing more than 1000 samples per week without significant problems in data quality. Additionally, FIE-MS fingerprinting can be applied in a wide range of circumstances where metabolite profiling might prove unproductive as prior detailed understanding of sample chemistry is unnecessary.

The protocol developed in the G02 project and validated in the G03 project focuses on the generation of low mass resolution spectra in which *m/z* signals are binned to nominal mass and concentration measurements are presented only as relative intensity ratios. As such the procedure can be carried out on a wide range of instruments, including common, relatively cheap ion trap machines. Related variants of this protocol can be found in the literature, including methods that demand higher mass resolution instruments where the aim is to generate semi-quantitative data [22, 27-29]. Such methods using high mass accuracy differ from the present qualitative procedure as the *m/z* signals necessarily are veiwed as discrete variables linked to a specific metabolite. Particularly in DIMS experiments producing semi-quantitative data then factors such as ion supression [34-37], need to be taken into account and thus a more detailed understanding of expected sample chemistry is important [27-29.]

5.3.4 Applications of metabolite fingerprinting by FIE-MS

FIE-MS has been used to study a range of biological systems focusing to date mainly on polar extracts from microbes and plants [4, 8, 9, 22, 25-29, 38-44]. Some success also has been reported in the use of FIE-MS for more targeted metabolite fingerprinting using mammalian biofluids (blood and urine) to detect individuals with metabolic disorders [45] or investigate metabolic differences between animal breeds (*e.g.* discriminating dog breeds by urine fingerprinting; MB, unpublished results). Investigations with microbes have included the development of chemical taxonomy approaches to investigate the genetic diversity of fungal contaminants in food [28] or identity of bacterial species in mixed populations [38]. Similarly, FIE-MS has found utility also in the optimisation of fermentation processes producing valuable secondary metabolites using either plant or bacterial cells [40, 41]. Studies investigating quantitative traits in introgression lines between two species of tomato [43] or classification of siblings in *Arabidopsis* crosses [26] have demonstrated the power of metabolite fingerprinting for large scale analysis of breeding populations. Further studies have used FIE-MS to confirm that only the intended changes in metabolism had occurred in plants genetically engineered to exhibit novel enzyme activity [4, 27]. In functional genomics studies on specific plant and microbial mutants FIE-MS has been used to investigate the phenotypic effects of gene disruption or unscheduled expression [9, 25]. Similarly, metabolite fingerprinting using mass spectrometry provides a first pass tool to investigate whether there are fundamental changes in metabolism associated with plant responses to either environmental or development signals or biotic stress [27, 42, 44].

In FIE-MS data any potentially ‘explanatory’ (statistically significant and biologically relevant in context) *m/z* signals [46] highlighted by data mining can be linked directly to candidate metabolites by atomic mass, providing further advantages over other metabolite fingerprinting methods such as NMR or FT-IR [4, 9, 27-29]. The high throughput nature and reproducibility characteristics of FIE-MS should also make the technique ideal for large scale classification tasks involved in screening crude samples derived from human population studies relating to nutritional/health status and disease pre-disposition in the emerging era of personalised medicine or nutrition [47]. Similar utility is expected in typical screens used in the pharmaceutical industry to evaluate drug mode of action and in the food industry to make comparative assessment of raw food material provenance and quality [48].

5.3.5 Considerations for overall experimental design

Multidisciplinary collaborations

To utilise FIE-MS fingerprinting to its full potential demands the efficient integration of analytical chemistry expertise with biologists, biostatisticians and chemometricians to achieve the workflow outlined in Figure 5.6.

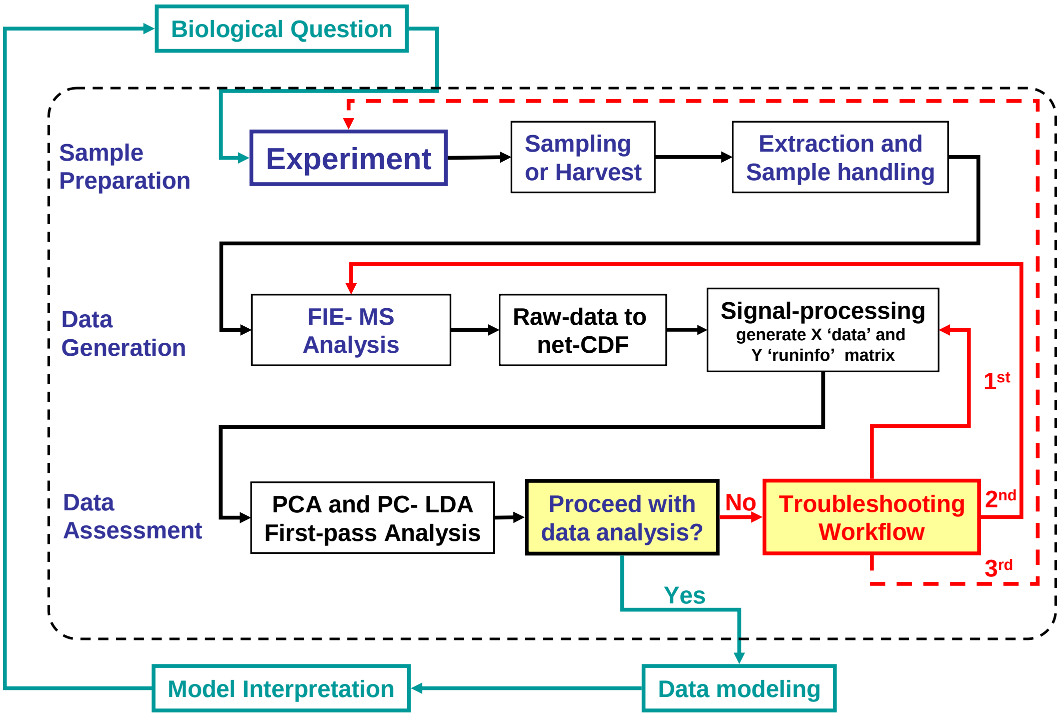


Figure 5.6 Work flow in FIE-MS high-throughput analysis

As outlined in the overview section the basic methodology was develop in the G02 project and here is validated using largely FIE-MS fingerprinting in the context of data generated for the first pass analysis of metabolome changes occurring during a plant-pathogen interaction which is described in detail in a Nature Protocol by Parker et al. [49]. Although this section of the report focuses on the generation and processing of FIE-MS data the collaborative work with statisticians is also presented which aims to confirm that fingerprints are suitable for subsequent in-depth data analysis. The data analysis approaches and software resources designed specifically for the investigation of FIE-MS data quality and in-depth data mining are described in detail in Section 8 and in a Nature Protocol by Enot et al. [50].

Sample characteristics and sampling

FIE-MS has been shown capable of using major metabolome characteristics to classify individual species [22]. It is equally powerful enough to investigate, within a single species, both subtle multivariate differences between distinct stable genotypes (ecotypes, cultivars, varieties, introgression lines) as well as determining specific metabolite signals that are explanatory of individual genetic modifications [4, 9]. Thus, it is extremely important that samples are comparable to one another if FIE-MS fingerprinting is to be used to address biological questions in an adequate way. Success in experimental design will depend on prior understanding of the sources of potential variability in any biological system allowing appropriate and pragmatic (*i.e.* realistic and cost effective) steps to be taken to minimise such variance. Care must be taken to sample tissue explants or biofluids from organisms at the same developmental phase and exposed to the same environment wherever possible. In all cases adequate meta-data must be recorded to account for any unforeseen confounding factors relating to individual samples [51]. For example plants are generally grown in specific dark-light cycles and will exhibit entrained diurnal changes in metabolism and so must be sampled at the same time of day.   Similarly, samples derived from animals should take into consideration, hormonal, feeding/drinking or behavioural patterns, particularly when homeostatic bio-fluids such as blood and urines are sampled. From an instrument perspective samples should be analysed using identical protocols and within a reasonably short, previously defined, time window to avoid the effects of 'instrument drift'. A detailed description of minimal information required for a metabolomics experiment is currently undergoing standardisation [52]. As an example Appendix 5.1 provides a summary of appropriate meta-data fields for a typical sample in the plant-pathogen interaction system used to illustrate the principle of FIE-MS metabolite fingerprinting.

Apart from controlling biological variability in samples and instrument variability it is important to stress the need to ensure that sample matrix effects on MS signal generation are minimised. One factor commonly associated with electrospray ionisation is the phenomenon of ion suppression [34-37]. The main cause of ion suppression is a change in the spray droplet solution properties caused by the presence of relatively abundant non-volatile or less volatile solute[28, 37] In the present protocol no attempts are made to compensate for these affects as it is assumed that FIE-MS fingerprinting is used most effectively for the 'first pass' quantitative comparison of similar sample matrices as a prelude to subsequently more informed analysis by targeted metabolite analysis using methods incorporating chromatographic techniques (**Fig. 5.6**). Under such circumstances by adhering rigorously to experimental protocols to maintain reproducibility it is assumed that ion suppression would affect all samples equally making any fingerprint comparisons meaningful. The core methodology focuses on a description of metabolite fingerprinting applied to plant leaf extracts as a typical sample.

Statistical experimental design

It is very important that the experimental statistical design is adequate for the problem under study. The high dimensionality of metabolome fingerprints (often approaching 2000 variables) dictates that powerful multivariate data analysis techniques are required for adequate modelling [9, 16, 46, 50, 53]. Each data analysis technique requires specific numbers of sample replicates per biological class to develop adequate models with high classification accuracy that home in on variables statistically significant to the biological problem under study rather than 'noise' in the data [9, 16, 46, 50, 53]. Supervised multivariate methods require sufficient samples to provide both training and test sets and importantly the 'generalisability' of any metabolome model can only be checked if completely independent experimental sample sets are available for validation purposes. Under conditions, particularly of sample paucity it is possible to develop 'overtrained' or uninterpretable metabolome models that unknowingly are confounded by class-specific regular variance ("noise") within the data. As a general 'rule of thumb' we suggest that a minimum of 18 independent biological replicates are normally sufficient for most studies (*i.e.* 18 leaves each from different plants), but wherever practical this should be increased depending on the degree of variance within sample classes and magnitude of difference expected between the biological classes under investigation.

Sample extraction

Basic considerations for sampling, metabolite extraction and extract storage in metabolomics experiments have been described previously in a Nature Protocol [7] with reference to higher plants. The extraction solvent used will not only dictate the range of metabolite classes eventually contributing to the FIE-MS fingerprint but will also effect ionisation behaviour.

The present extraction protocol was developed specifically to provide as comprehensive coverage of the metabolome as possible when applied to a wide range of sample classes and was validated using GC-tof-MS non-targeted profiling [7]. As individual metabolites are not measured the data is not quantitative. If there is a need to target specific chemical classes (e.g. non-polar lipids) or a requirement to generate quantitative data then the extraction methodology needs to be modified to use suitable solvents and internal standards, which is outside of the scope of the present high-throughput, non-targeted, qualitative fingerprinting procedure.

Although in principle any tissue extract (plant material like leaves or phloem exudates, single cells like yeast or fungi, bio-fluids like urine or blood) can be submitted to FIE-MS analysis, preference is given to extraction procedures which result in a predominantly aqueous methanol matrix to aid ionisation of metabolites in the ion source. In practise, most biological samples can be extracted routinely using a modified Bligh and Dyer solvent mix [54] consisting of pre-chilled (-20°C) chloroform:methanol:water at a ratio of 1:2.5:1. In the case of ostensibly aqueous samples (*e.g.* cells in culture media, fresh potato tubers) one part of water in the solvent mix can be omitted and to one part of the sample a chloroform:methanol mix in the ratio 1:2.5 is added. Therefore, a crude single phase extract is still obtained for subsequent FIE-MS analysis. The extraction protocols developed for potato tubers and *Arabidopsis* leaves in the preceding GO2 project protocol has been validated in the present GO3 project by optimized extraction of freshly harvested *Brachypodium* leaf material as a typical matrix [49].

LC-MS instrument specification, sample run batches and quality control

For reasons of clarity the described protocol for FIE-MS fingerprinting in Section 5.4 focuses on the use of a single instrument, the LTQ linear ion trap (Thermo Finnigan), mainly because of its versatility and ability to acquire data for both ionization modes in a single analysis in a mass range from *m/z* 15 to *m/z* 2000. If required, multiple MS/MS experiments are possible to follow up secondary ionisations of either the most abundant or predefined mass ions. Additionally, tuning and mass calibration for high-throughput FIE-MS routine analyses has to be performed only three to four times a year. This is of advantage if for example biological experiments involving dynamic changes in metabolome (*e.g.* time dependent host-pathogen interactions) have to be compared in replicate experimental batches several months apart [49].

During the G03 project we have shown that present procedure can be adapted for use with a wider range of LC-MS instruments which might possess various degrees of sensitivity and mass resolution. For example, this protocol has also been applied to FIE-MS analysis on LCT and Q-Tof instruments (Micromass) [48, 55]. FIE-MS using high-resolution and accurate mass instruments equipped with time-of-flight detectors may need some attention regarding mass-drifts in long analysis sequences with many samples as there is an apparent sensitivity of flight tubes to temperature fluctuations [56]. A 'Lock-Spray Exact Mass Ionization Source' is highly recommended if the acquired accurate mass information is to potentially be used in further metabolite identification. Alternatively, the most abundant mass ions can be used to initially align mass spectra to avoid potential binning errors [21]. Generally speaking, increases in mass resolution are concomitant with a proportional increase in data dimensionality, which in turn effects experimental design with regards to the numbers of replicates required to achieve statistical robustness [50, 53]. For the high throughput, first pass *m/z* fingerprinting procedure presently described we advocate integration of all signals to the nearest nominal mass to constrain data dimensionality [22]. This strategy also avoids any problems with false positive or false negative signals that can result from drifts in instrument mass accuracy over time. From a pragmatic perspective, non-targeted, nominal mass FIE-MS fingerprinting also has the advantage that any instrument in any laboratory should be able to replicate any measurements thus possibly providing extended scope for future data integration [48, 55]. To accomplish this objective the data analysis strategy is based on the measure of relative ratios of *m/z*-signals in each fingerprint generated from large numbers of samples in a single analytical batch.

In high-throughput analyses there is a possibility of subtle levels of instrument baseline drift caused by 'carryover' effects from preceding samples. This effect adds to the overall variance in the total data, which demands appropriate numbers of replicates for mining robust data models. Randomizing sample classes in the run sequence is mandatory to avoid bias due to the position of samples within the injection order and should be practised generally at all levels within the protocol to minimize systematic errors. Carryover effects are not fully compensated by auto-zeroing the detector at the beginning of an analytical run; because of the 'plug' flow of a newly injected sample, where the infused sample volume replaces the actual mobile phase, some of the deposition inside the solvent lines will be carried forward to the end of a run. Applying background subtraction has been proven advantageous for classification and data mining. Within this context raw data signal acquisition and data pre-processing is therefore performed in four rudimentary steps: 1) binning to nominal mass thus decreasing data dimensionality and minimising mass accuracy effects; 2) background subtraction of individual sample-attributed ion intensity as a simple baseline correction; 3) log10-transformation reducing data-set variance; and 4) normalization of data-set to total ion count providing relative ratios of *m/z*-signal abundance. These four steps are sufficient for the quality assessment of the FIE-MS analysis, process variability and the underlying experiment. Further in-depth data analysis only requires the background subtracted matrix (output of steps 1 and 2 above). Any decision to additionally use log10-transformation and normalisation is dependent on assessment of process variability; the robustness of classification models and overall accuracy will further dictate the workflow of statistical analyses [50].

A quality control (QC) sample of known composition is not necessarily required in FIE-MS experiments. Instead, a ‘mastermix’ (MM) sample composed of all the sample extracts to be analysed in a particular batch [48] (see REAGENT SETUP in **Section 5.4**) can be used to monitor instrument response throughout the analysis of a large batch of samples. It is recommended to infuse at least four MM-samples at the beginning of a batch of samples to create a constant system background in acquired mass spectra of subsequent infusions. This prevents the detection of outliers of early samples when investigating injection order effects. Specifically unsupervised data modelling techniques such as Principal Components Analysis (PCA) are generally useful to check on data quality in large experiments by visualising distinct sample outliers or irregular instrument drifts in PCA score plots [50, 53, 57-59].

5.4 PROTOCOL DESCRIBING STANDARDISED PROCEDURE FOR FIE-MS FINGERPRINTING

5.4.1 REAGENTS

- Water: Milli-Q quality (≤ 18 MΩ*cm)
- Chloroform: GLC – Pesticide residue grade (Fisher C/4963/15)
- Methanol for extraction: trace analysis grade (Fisher M/4020/17)
- Methanol for mobile phase: HPLC grade (Fisher M/4056/PB17)
- Acetonitrile: Far UV, HPLC gradient grade (Fisher A/0627/17)
- Acetone: Analytical reagent grade (Fisher A/0600/17))
- Liquid nitrogen (BOC)
- Argon 5.0 (BOC)
- Helium 5.0 (BOC)

5.4.2 EQUIPMENT

- LTQ linear ion-trap detector (Thermo Finnigan, http://www.thermo.com/com/cda/product/detail/1,,22534,00.html)
- Surveyor (Thermo Finnigan) liquid chromatography (LC) system consisting of Autosampler and MS-pump
- HPLC-glass vials, 2 mL crimp top (2-CV, Chromacol, http://www.chromacol.com/)
- Micro-vial insert, 200 µL (Chromacol, 02-NV)
- Aluminium crimp cap, rubber/PTFE seal (11-AC7, Chromacol)
- Cap crimper and decrimper
- Dewar (thermally insulated flask) for liquid nitrogen, 2 x 1L (Dilvac)
- Personal protective equipment (PPE) for lab work
- Appropriate PPE for handling liquid nitrogen
- Long forceps (to retrieve sample tubes from liquid nitrogen)
- Single hole puncher (5 mm ID, Fisher, http://www.fisher.co.uk/)
- Pair of scissors, single hole puncher , cork borer (harvesting)
- 1000 µL pipette and tips (Gilson, http://www.gilson.com/Products/product.asp?pID=67)
- 100 µL pipette and pipette tips (*e.g.* Gilson)
- Balance (*e.g.* Sartorius BP 211 D)

- Glass measuring cylinders 500, 250, 100 mL
- Screw-cap glass bottles between 50 mL and 2L (ensure solvent resistant *e.g.* Duran)
- Eppendorf 'safe-lock' tubes 2mL (Eppendorf AG, http://www.eppendorf.com/int/index.php)
- Refrigerated centrifuge (Hettich, EBA 12R)
- Orbital shaker (IKA Vibrax VXR) in cold room at 4 °C
- Mixer mill, MM200 or MM301 (Retsch, http://www.retsch.com/44.0.html?&L=0)
- Teflon adaptor for 1.5–2.0 ml micro-vials (Retsch)
- Stainless steel cone balls (Retsch)
- Stainless steel balls of ball bearing quality (5 mm diameter) and acetone washed
- SpeedVac; Centrifugal vacuum concentrator Univapo 150H with Unijet II Refrigerated Aspirator (UniEquip, http://www.uniequip.com)
- Xcalibur mass spectrometry data management software (Thermo Finnigan)
- Matlab (The MathWorks, Inc., http://www.mathworks.com/)
- R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, http://www.R-project.org)
- Computer: PC (Microsoft Windows 2000 or XP) or Mac (e.g. G5)

5.4.3 REAGENT SETUP

Bligh and Dyer extraction mix

Ensure enough extraction mix is prepared for all samples of the experiment to minimize variability. Use a pre-cleaned (thoroughly rinsed with water, methanol and chloroform) screw cap glass bottle of appropriate size. Add solvents in the order water-methanol-chloroform in the ratio 1:2.5:1. **C**lose bottle, shake to mix thoroughly and store at -20°C. For sample extraction aliquot extraction mix into 50 ml screw cap glass bottle for convenience and avoid cross-contamination of extraction mix with samples during pipetting.

Master-Mix samples (MM)

MM is ideally a mix of aliquots of extracts from every sample in the experiment that can be used to monitor instrument performance throughout a long run series. For practical purposes it has been proven advantageous to first prepare master mix samples for each biological class separately. For each sample class pipette a minimum of 50 µl solvent extract of each biological replicate into a fresh 2 mL Eppendorf tube and mix thoroughly. In a second step pipette a minimum of 100 µl of the single class master mixes into a fresh 2 mL Eppendorf tube to obtain the MM for the experiment. Mix thoroughly and store all master mix samples at -80 °C.

Mobile phase for HPLC

The mobile phase used in FIE-MS analysis of soft leaf tissue sample extracts (*Brachypodium*, *Arabidopsis*, Tobacco) is methanol:water (50:50, v:v). Always use freshly prepared and sufficient solvent mix for an analytical batch to minimize sources of variability due to mobile phase changes. In practice, different methanol-water ratios have been used as a mobile phase for FIE-MS analysis ranging from 30 % to 70% methanol. Because of the 'plug' flow in FIE-MS experiments the mobile phase does not directly affect the ionisation of sample metabolites, but does help to prevent build-up of contamination in transfer lines, possible carry-over effects, the increase of background signals in large batches and, therefore, minimises sample related machine variability. As a rule for any LC-MS system to be used, the run time, mobile phase composition and occasionally the flow rate (up to 100 µl/min for *e.g.* plasma, urine) should be adjusted to insure that signal intensities at the end of an infusion profile equal those intensities acquired at the beginning of a run.

5.4.4 PROCEDURE

The general assumption is that the majority of users reading the current protocol will have mass spectrometry experience and an experienced operator for their LC-MS system when applying FIE-MS analysis. It is not possible to provide detailed information on essential maintenance and calibration of all mass spectrometry instruments, but the following information emphasises important equipment quality checks, which should be undertaken prior to and during any batch of FIE-MS analyses (see also Quality Control below). Recommended analytical settings for an LTQ instrument (Thermo-Finnigan) are part of the protocol (PROCEDURE, steps 14 to 16.

Before starting a run sequence ensure analytical instrument is fully operational and specifically that: (A) the system has been appropriately calibrated and tuned (follow vendor's specifications); (B) Syringe, injection line and injector are flushed; (C) Tip of spray capillary is in good condition and properly adjusted; (D) Ion source is clean; (E) Solvent reservoirs contain fresh solvent sufficient for the batch of samples, and (F) A stable spray and a constant mass spectrum is achieved for each ionisation mode and mass range using the mobile phase only.

As this protocol for high throughput analysis of plant leaf extracts is based on the injection of samples in the extraction mix we have to expect precipitation of sample constituents within the mobile phase transfer lines. Tubing contaminated with plant leaf extracts can be easily cleaned with 100 % acetonitrile after a batch of samples. On a day to day basis however preceding samples with different matrix characteristics might have contaminated the analytical instrument, which requires more drastic manual cleaning. To remove precipitation in solvent lines apply a sequence of solvents with different polarity and acidity (see also vendor's recommendation. Depending on the mobile phase delivery system, presence of restrictors, tubing, ion source geometry and detector characteristics of different LCMS-systems sample volumes and concentrations have to be adjusted and for higher flow rates a flow splitter has to be considered to maintain a mobile phase flow rate of around 50 µl into the ion source. Check vendor's recommendations and optimize sample delivery system in test runs.

FIE-MS analysis optimisation and quality control (QC)

It is generally advisable to undertake test analyses/pilot experiments to optimise extraction, sample concentration, injection volumes, flow rates and mobile phases when working with new sample matrices or different instruments. The examination of a dilution series of extracts or master mixes of samples under investigation is of particular importance to identify potential ion suppression effects, which might be caused by high salt concentrations or highly abundant metabolites. In such instances the dilution of samples could increase the information content of mass spectra by preventing detector overload, or by increasing injection time in ion traps. Related instrument characteristics (*e.g.* dead time or automatic gain control; check vendor's specifications) generally dictate the maximum number of ions or ion intensities to be detected. General guidelines to assessing fingerprint quality include empirical measures such as total ion count of infusion profiles and the number and intensity of mass ions. It is good practice in high throughput FIE-MS analyses to routinely check the performance of the analytical system and the quality of acquired data visually. Raw data of solvent extracts generally have a file size of 5 to 7 Mega Bytes (MB) in centroid mode and vary within an experiment of 200 samples by ±0.5 MB. Depending on the matrix the total ion count (TIC) at the apex of an infusion peak is commonly between $1*10^6$ and $1*10^7$. A rapid drop in the file size of raw data files by more than 1 MB as well as a distorted or missing infusion peak in the chromatogram indicate instrument-related problems, which are most likely caused by a blocked needle or injector transfer lines, a blocked ion transfer line; or contaminated or not appropriately adjusted spray needle in the ESI source.

Sampling and extraction

1 Label two sets of 2ml Eppendorf 'safe-lock' tubes uniquely at the side of the tube; one set for harvest/extraction and the second set for storage. [**Troubleshooting**: Milling [steps 1 and 6]

2 For harvest/extraction (set 1 tubes) add a stainless steel ball to each Eppendorf tube and close caps loosely. New stainless steel balls are protected by an oil film and should be thoroughly cleaned using acetone before use (wash at least three times and air-dry in fume hood).

3 Quickly sample 1 x 40 mm leaf segment per plant (in this case the model grass species *Brachypodium distachyon*) into a 2-ml Eppendorf tube containing a stainless steel ball [49]. **[**Troubleshooting**:** Harvest Location [step 3]]. Close tube tightly and rapidly flash freeze sample-containing tube in liquid nitrogen. For *Arabidopsis* leaves, use 4 x 5 mm diameter leaf cores [7]. For tobacco leaves, use 2 x 8 mm diameter leaf lamina cores, avoiding major veins and midrib. For other sample types see **Box 5.1.**

CAUTION Work with liquid nitrogen only in well ventilated areas; use PPE

4 Transport harvested material to lab in a low-temperature liquid within a suitable Dewar.

5 Insert sample tubes into pre-cooled (low-temperature liquid or freezer) Mixer-mill teflon adaptors suitable for 1.5–2.0 ml micro-vials. Homogenize in Mixer-mill for 30 sec at 30 Hz.

6 Quickly open tube and add 1 ml of **extraction mix** (chloroform:methanol:water, 1:2.5:1, pre-cooled at -20 °C; see 5.4.3 REAGENT SETUP) using a calibrated pipette and close tube tightly. For LTQ-FIE-MS fingerprinting no internal standard is required; however, internal standards can now be added if the same sample is also required for other analytical techniques (*e.g.* GC-MS analysis [7]). [**Troubleshooting:** Milling]

CAUTION Chloroform may be a potential carcinogen; use PPE.

CRITICAL STEPS 4 TO 6 Avoid thawing of samples

Box 5.1 Other sample types

During extraction keep samples on ice and use appropriately pre-chilled solvents.

Potato tuber

- Prepare 'Pot-Extraction Mix': $CHCl_3/MeOH/H_2O$ (2:5:1) and store at -20°C;
- Start with Step 1;
- In Step 2 use stainless steel cone balls instead of stainless steel balls;
- In Step 3 use one 10 mm diameter x 3 mm thickness core; take sample of inner medulla at least 1 cm below skin or take a core with outer skin;
- In Step 5 defrost potato tuber disk on ice and add 150 µl pre-chilled 'Pot-Extraction Mix' prior to milling; after milling a slurry should be obtained;
- In Step 6 add only 850 µl 'Pot-Extraction Mix'; check for phase separation because of high water and sugar content in potatoes;

Yeast (chemostat)

- Start with Step 1 and ignore Steps 2 to 7;
- Use 500 µl culture medium + 500 µl MeOH + 500µl glass beads;
- 3 freeze/thaw/vortex cycles; (beat-beating 30 sec);
- Proceed with Step 8;

Urine

- Start with Step 1 and ignore Steps 2 to 10;
- Use 50µL sample + 100µL H_2O + 350µL MeOH ;
- Vortex and centrifuge for 6 min at 14,000*g;*
- Proceed with Step 11;

Plasma/Serum

- Prepare 'P/S-Extraction Mix': $CHCl_3/MeOH/H_2O$ (2:8:1) and store Duran bottle labels as 'P/S-Extraction Mix' at -20°C;
- Start with Step 1 and ignore Steps 2 to 7;
- For extraction add a micro spoon full of glass beads to one set of Eppendorf tubes;
- Add 100 µl serum/plasma*;*
- Add 820 µl 'P/S-Extraction Mix'
- Insert sample tubes into pre-cooled (low-temperature liquid) PTFE adaptors. Homogenize in mixer-mill for 30 sec at 30 Hz
- Proceed with Step 8

7 Vortex for 10 s and place on crushed ice for the next step [**Troubleshooting** Phase Separation]

8 Shake tubes by hand to ensure that all sample is washed into the extraction mix if any milled plant material has deposited onto the inside of the cap.

9 Agitate for 15 min in an orbital shaker (950 rpm) at 4 °C in the dark.

10 Centrifuge for 3 min at 14,000*g* and 4 °C.

11 Use 1 mL Gilson pipette to transfer supernatant to second set of labelled tubes, and store in a -80 °C freezer.

PAUSE POINT Samples can be stored at -80 °C for up to 3 months. For short periods (up to a month) sample degradation by air is negligible as there is sufficient solvent vapour in the tube. It is however advisable to store samples for longer periods under inert gas to prevent extracts from oxidization and degradation by reactants in atmospheric air. Gently replace atmospheric air in tubes with inert gas. Argon gas is preferred because it is heavier than other air constituents and will isolate the liquid from atmospheric oxygen. Use nitrogen gas if argon is not at hand. If you as a user are concerned about the stability of metabolites (e.g. when samples have a high content of fatty acids) you are advised to store extracts in solvents free of dissolved oxygen and/or add antioxidants (e.g. butylated hydroytoluene, BHT).

12 Retrieve stainless steel ball and discard precipitate-containing tubes appropriately.

CAUTION Halogenic reagents and solutions should be disposed with halogenic waste.

LTQ instrument set up and sample analysis

13 Before starting the analytical run sequence ensure the LTQ instrument is fully operational and the sample concentrations are optimised for FIE-MS analysis.

14 Set ion source parameters and instrumental conditions as follows: sheath gas (nitrogen) pressure at 40 arbitrary units; Auxiliary gas (nitrogen) pressure at 5 arbitrary units; helium as collision gas (incoming pressure: 40 psi); 4.5 kV (ESI+) and 4.0 kV (ESI-) spray voltage and +15 V (ESI+) and -13 V (ESI-) capillary voltage; temperature of the heated transfer capillary is set to 380 °C.

15 Set data acquisition method to acquire data in positive and negative ionisation mode as follows: 5 min MS run time (duration); 1 segment; 4 scan events with scan rate 'normal', scan type 'full' and data type centroid; scan event 1: positive polarity, normal mass range from *m/z* 50 to 2000; scan event 2: positive polarity, low mass range from *m/z* 15 to 200; scan event 3: negative polarity, normal mass range from *m/z* 50 to 2000; scan event 4: negative polarity, low mass range from *m/z* 15 to 200. **[Troubleshooting** Acquisition Time and MS Detector]

16 Set autosampler injection and pump parameters as follows: autosampler tray temperature to 15 °C and sample injection mode to ‘partial loop’. Set mobile phase flow to 60 µl*min^{-1} of premixed methanol:water (50:50).

17 For FIE-MS analysis transfer 60 µl of supernatant to HPLC crimp cap glass vials containing a 200 µl micro glass insert.

18 Load HPLC vials, including biological extracts, mastermix (MM) samples and extraction solvent blanks, into autosampler in a randomised run order and edit sequence in Xcalibur (Thermo-Finnigan) files describing loading order. Set sample injection volume to 20 µl. **[Troubleshooting** Sample Batch [step 18]**]**

19 Start FIE-MS analysis run (see EQUIPMENT SETUP: FIE-MS analysis optimisation and quality control (QC)).

Signal-processing

All software resources derived from the G03 project are maintained on an Aberystwyth University web site describing the FIEmspro data pre-processing and data mining package. This package has resulted from efforts to convert all software code used for data mining into a common platform language ‘R’ and was developed jointly with David Enot on the BBSRC MetRO project and has been published in a Nature Protocol by Enot et al., 2008 [50] . It is assumed that all individuals using FIEmspro will have some training in command line programming and as part of the GO3 project the whole package was tested independently by a seconded computer scientist Dr Chuan Li. The final format of the data pre-processing and data mining package and instruction manual was validated and refine during interactions with biologists/analytical chemists with no experience of computer programming.

The basic FIEmspro package contains the following resources all in an ‘R’ format:

- original code in the FIEmspro package
- tutorials of how to use FIEmspro
- an example data set
- an example work flow covering data extraction, pre-processing, quality checking and data mining

An overview of the development of the FIEmspro package and how to use it are presented in **Sections 8 and 9** of the GO3 report. In the current section reference is made only to scripts used for data extraction from instruments. A range of examples of potential QA problems with FIE-MS data sets are described as a series of anticipated results which explain the sources of the problems and give advice on how to correct them. The next stage of the generating of FIE-MS fingerprinting data is to extract the raw data from the instrument and develop a matrix of binned mass spectra as shown previously in Figure 5.4.

20 Export data by converting raw data files to ANDI NetCDF files using XConvert (Thermo Finnigan).

21 Download all software required for raw data binning as outlined in **Box 5.2.** Use R-package 'FIEmspro' [50] and the script provided in **Box 5.3** to generate a matrix of binned mass spectra (see Fig. 5.4).

22 Identify in '*scan event 1*' (positive polarity, normal mass range from *m/z* 50 to 2000) of a representative infusion profile (see Fig. 5.1) scans '**x1**, **x2**' and '**x3**, **x4**' for sample and background, respectively (see illustration in Fig. 3). For further details see R-package *FIEmspro* manual [50] and ANTICIPATED RESULTS.

Box 5.2 Software required for raw data binning

- R: download and install the latest release at http://cran.r-project.org/
- R-Packages: follow the link 'Software'/'Packages' at http://cran.r-project.org/ download by clicking the appropriate 'Available Bundles and Packages'. Required are: Package *e107*1; Package *randomForest*; Package *MASS*; Package *ncdf*; Package *impute*; Package *KernSmooth*.
- At (http://www.aber.ac.uk/...) download R-Package *FIEmspro* and example NETCDF files for test purposes.
- Download and install an editor for Windows (*e.g.* Tinn-R, http://www.sciviews.org/Tinn-R/ or MacOS X (*e.g.* SubEthaEdit 2.2, http://www.codingmonkeys.de/subethaedit/).
- Use a spread-sheet program (*e.g.* MS Excel) to generate the comma separated values file '`runinfo.csv`' as provided for test purposes for LTQ data in **Box 7** or for LCT/Q-Tof data in **Box 8**
- After successful installation of the R-program, required R-packages and an appropriate editor use the scripts provided in **Box 7** and **Box 8** to check the functionality of functions *fiems_lct_main* and function *fiems_lct_main*, respectively.

23 Calculate scan range values '**y1**' to '**y4**' for vector '*scrng'* [see R-package *FIEmspro* manual [50]] required for function '*fiems_ltq_main*' as follows: subtract '1' (the Scan Event) and then divide by '4' (total of 4 scan events). Therefore, the scan range values for the sample are **y1**=(**x1**-1)/4 (begin scan sample) and **y2**=(**x2**-1)/4 (end scan sample), and for the background **y3**=(**x3**-1)/4 (begin scan background) and **y4**=(**x4**-1)/4 (end scan background), with equal scan differences for sample and background **y2-y1** = **y4-y3**.

Box 5.3 Example for processing NetCDF files of 'LTQ' FIE-MS metabolite fingerprint data using FIE-MSpro

1. Factor list 'runinfo.csv' for LTQ-binning of NetCDF files

Example factor list to test functionality of function *fiems_ltq_main* in R. A comprehensive list is imbedded in the list *data(abr1)* of R-package *FIEmspro*. Copy content of table into a *e.g.* spreadsheet editor and save as comma separated values file. Change content of column 'pathcdf' to the path of the folder containing LTQ NetCDF files.

injorder	pathcdf	filecdf	remark	name	rep	class
1	C:/Xcalibur/ANDI-LTQ/050509-Abr1	01.cdf	ok	12_2	2	2
2	C:/Xcalibur/ANDI-LTQ/050509-Abr1	02.cdf	ok	13_3	3	3
3	C:/Xcalibur/ANDI-LTQ/050509-Abr1	03.cdf	ok	15_4	5	4

2. Script for LTQ-binning of NetCDF files

For a full description of arguments and functionality of function `fiems_ltq_main` see R-Package *FIEmspro* manual. The functionality of function `fiems_ltq_main` is currently limited to the use of 4 scan events, where it is assumed that respectively two scan events differ only in the value for the ionisation mode.

```
##  Workflow to process LTQ ANDI NetCDF files, which are
##  generated by XConvert (Thermo-Finnigan), into a M x N matrix of
##  'M' mass spectra and 'N' nominal mass-to-charge ratios using
##  fiems_ltq_main(my_path,runinfo,y1,y2,y3,y4,limit)

## START SCRIPT ==================================================

library(FIEmspro)

my_path <- "H:/050509-Abr1"
runinfo <- "runinfo.csv"
tmp <- fiems_ltq_main(my_path,runinfo,35,95,190,250,0.7)

## END SCRIPT ====================================================

For a full description of arguments and functionality of function
'fiems ltq main' see Package FIEmspro manual.
```

24 Define further input arguments for function '*fiems_ltq_main*' [see R-package *FIEmspro* manual [50]]. Argument '*my_path*' is the location of the folder where results are to be saved and where to find the file is specified in the argument '*runinfo*'. Argument '*runinfo*' is the name of a CSV-file (*e.g.* runinfo.csv) created in a spreadsheet editor (*e.g.* MS-Excel) containing run and sample information (list of factors) of the

experiment including the path and the name of saved NetCDF files (**Box 5.3**). As a good starting point set argument '*limit*' to '0.7'.

25 Save modified script (**Box 5.3**) as 'LTQcdf2mat.r' and run script in R. The returned values are saved by default as LTQ-mean.RData in folder '*my_path*'. Additionally, single items are saved by default as TEXT files: posh.txt, posl.txt, negh.txt negl.txt, myparam.txt (containing '*scrng*' and '*limit*' for reference purposes

First-pass data analysis to evaluate overall data quality

26 Assess data quality by visual inspection. Check size of raw data file: about ±0.5 MB of the mean file size is expected. A sudden drop or gentle decrease of the file size of samples by injection order by more than 1 MB, a missing or smaller than an accepted infusion peak (see Fig. 5.1) indicates instrument related problems (e.g. empty solvent reservoir; blockage in sample injection or solvent lines; blockage of ion source).

27 Assess data quality by multivariate analysis as outlined in **Box 5.4** by performing Principle Components Analysis (PCA) and additionally Principle Component Linear Discriminant Analysis (PC-LDA). Examples of known problems and solutions are given in ANTICIPATED RESULTS and details on how to further assess data quality (*e.g.* function '*ticstats'* in R-package *FIEmspro*) are provided in Section 8 and 9 of the present report and published in a Nature Protocol by Enot et al. [50].

28 Make a decision to continue analysis based on previous step: Either (A) submit data set for further in-depth analysis including data mining, feature selection and model validation if evaluation of PCA and/or PC-LDA reveal that data set is of suitable quality, (B) repeat data pre-processing if binning errors are detected, (C) thoroughly maintain instrument and repeat instrumental analysis if machine variability is greater than the expected biological variability, or (D) repeat and/or redesign biological experiment if troubleshooting has failed to improve data quality sufficiently to produce robust models. Bear in mind that steps 2 to 19 are potential sources of experimental variability adding to overall process variability.

CRITICAL STEPS 27 AND 28 It is generally assumed that an individual with experience in multivariate data analysis will be involved in signal processing and assessment of data quality. A Nature Protocol by Enot et al. [50] provides detailed protocols and suitable algorithms.

Box 5.4 Use of Principal Components Analysis and Linear Discriminant Analysis to preliminarily assess FIE-MS data quality

1) Perform Principal Component Analysis (PCA) to examine both consistency of samples from each class (time points) and reproducibility between different batches of infected plants.
2) Check that all sample classes shows clear grouping when analysed individually by PCA and are reasonably well discriminated from others.
3) Examine PCA scores plots to identify any potential sample outliers for removal in order to improve data quality (*e.g.* see Figure 5a).
4) Use PCA loadings plots (*e.g.* see Figure 8) to identify possible binning errors and reprocess the data using adjusted binning limits (*e.g.* see Figure 9).
5) Examine Linear Discriminant Analysis (PC-LDA) scores plots for any overlap between sample classes (*e.g.* Fig. 9b). Good discrimination between classes could indicate potentially good prospects for data modelling even if biological variance is relatively high [50] .
6) Examine confusion matrix for misclassification. Columns of the matrix represent instances in a predicted class, while rows represent the instances in the actual class. Therefore, high values (the number of biological replicates per class) in the diagonal and low values (ideally zeros) in all other fields indicate low misclassification rates between actual and predicted classes.
7) Calculate associated class separation metrics (*e.g.* eigenvalues) and classification accuracy and make a decision whether to continue with the experimental material for further in depth studies. As a general rule eigenvalues of less than 2.0 in models comparing two sample classes indicate poor class discrimination.

5.4.5 TROUBLESHOOTING

Milling [steps 1 and 6]

As caps of sample tubes might crack if not properly fixed in the milling device sample information written on the cap could be lost. If lids show signs of stress replace with a new one to avoid spillage of solvents during extraction.

Harvest Location [step 3]

Ideally, samples should be harvested in the growth environment to minimize disturbances and potential stress responses of the plants (other than excisions/puncturing) due to transport of material other locations.

Phase Separation [step 7]

At this step phase separation might occur. Possible reasons are: aged extraction mix, higher than expected tissue water content, too much leaf material resulting in too much additional cell-water, too much leaf material resulting in high metabolite and/or salt concentration.
Use only freshly prepared extraction mix. If phase separation is noticed in a relatively small number of samples add stepwise 50 µl methanol (up to four times) until a single phase has been formed. In general, phase separation can be prevented by either increasing the methanol proportion or decreasing the water content in the extraction mix (preferred). In contrast, phase separation can be forced by adding 100 to 200 µl water. If it is necessary to extract a larger number of leaf explants (*e.g.* 4 in the case of *Arabidopsis*) then the volume of extraction mix needed should be determined in test extractions. The extraction capacity of the solvent mix is limited by a conserved solvent-to-tissue ratio [53].

MS Detector [step 15]

Depending on individual LC-MS instrument detector characteristics sample sets may have to be analysed for positive and negative ionisation mode separately which might require regular tuning. It is advisable to prepare two separate sets of samples as the effect of an already pierced septum after the first injection in combination with the dwell time in the autosampler might alter the composition of the samples. Time-of-flight detectors could show an unwanted mass drift, which is however of no importance when binning *m/z* values to nominal masses.
If it is anticipated to use high resolution/accurate mass data the inclusion of an internal standard, an external standard (quality control, QC-sample; mastermix, MM-sample) or a

lock-mass system may be considered. A mass range from *m/z* 50 to 1200 is sufficient for detecting most metabolites.

Acquisition Time [step 15]

The acquisition time depends on parameters like length and diameter of tubing, flow rate and sample matrix composition. Test runs should give information on when infusion profiles show a constant baseline to prevent carryover. This can normally be achieved after between 1 and 6 min. However, a system set up resulting in peak widths of at least 30 sec are anticipated as a chromatographic effect can be observed in the sample peak, which helps to minimize ion suppression effects.

Sample Batch [step 18]

Depending on the characteristics of the analytical instrument up to 240 leaf samples can generally be analysed successfully in a single day. It is however recommended to work in smaller batches of for example 4 batches of 5 biological replicates comprised of 12 classes where samples are injected in a different randomized order in each batch. Batches can then be combined in a single run sequence sequentially. In the case of technical problems, one has to repeat only the affected batches in the sequence. Master-Mix samples run with each batch are of diagnostic value to investigate potential problems.

5.5 ANTICIPATED RESULTS

The major objective of the present protocol is to describe how to generate a non-targeted, FIE-MS metabolite fingerprint of complex biological samples that would be suitable for in-depth data analysis. The procedure has been used to analyse many sample types and the specific biological samples we use to describe anticipated results relate to the study of plant-pathogen interactions as described by Parker et al. [49]. The experiment contains a time series of leaf tissues sampled from *Brachypodium* leaves following infection with the fungal pathogen *Magnaporthe grisea.* This biological system is expected to display moderate levels of biological variability at each time point associated with the efficiency of fungal spore germination and the level of synchrony of disease development between different infected leaves. At the same time as the disease progresses there will be substantial changes in the metabolome of pathogen challenged leaves, thus providing biological sample classes with distinctly different metabolome characteristics. These sample features represent a challenging biological system which will require a large number of replicate samples to be analysed for adequate data modelling [9, 53] and provide an excellent context to describe high throughput, non-targeted, metabolite fingerprinting.

5.5.1 Signal processing

In order to reduce data dimensionality *m/z*-values are binned to nominal mass in steps of 1 atomic mass unit (1u) by summation of mass intensities of consecutive *m/z* values between e.g. u-0.3 and u+0.7 (binning limit = 0.7) of each nominal mass. The value for the binning limit has to be defined by checking *m/z*-values in mass spectra of vendor’s software and is generally set between 0.7 and 0.8 (see also: First-pass data analysis). To generate a single mass spectrum for an FIE-MS analysis the averaged *m/z* intensities of the mass spectrum of the background between scans **x3** and **x4** (see Fig. 2) are subtracted from the averaged *m/z* intensities covering three quarters of the mass spectrum of the infused sample (scan range **x1** to **x2**). Scan ranges for sample and background should be of equal length. Mass spectra are then combined and aligned in an *m x n*-matrix in sample injection order with *m* rows of analyses and *n* columns of nominal *m/z*-values for each ionisation mode and mass range. The values in this matrix can be plotted as shown previously in Fig. 4. FIE-MS experiments on an LTQ instrument (Thermo-Finnigan) as described in this protocol results in four data matrices: positive low mass range [*m/z* 15-200], positive normal mass range [*m/z* 50-200], negative low mass range [*m/z* 15-200], and negative normal mass range [*m/z* 50-200]), which are analysed separately. In low mass range the multipole RF voltage is lowered from 400 V (normal) to approximately 120 V to allow the detection of mass ions below *m/z* 110.

5.5.2 First pass data analysis

A detailed overview of concepts and procedures for sample classification and data pre-processing using FIE-MS data is presented in Section 8 and 9 of the present report. The multivariate data analysis methods use sample clustering or sample discrimination approaches to determine whether FIE-MS fingerprints derived from samples of the same class are reproducible (**Box 5.4**). Such methods allow for the detection of sample outliers (caused by unexpected experimental variance) which when removed from the total data set improves its quality and increases the prospects of informative data mining. Suitable software for raw data pre-processing and quality assessment are available as part of the R-package 'FIEmspro' [50] on the Aberystwyth University, Institute of Biological Sciences web site (http://users.aber.ac.uk/jhd).

Figures 5.6 to 5.10 and Tables 5.2 and 5.3 illustrate process related problems, which are commonly encountered, using the FIE-MS negative ion low mass range (*m/z* 15 - *m/z* 200) of infected *Brachypodium distachyon* leaf material [49]. The raw intensity data matrix obtained after signal processing was first log10-transformed and then normalized to total ion count (TIC). Outliers are obvious in PCA (Fig. 5.6a) but difficult to detect in PC-LDA (Fig. 5.6b). Here, samples 17 and 20 appear well separated from the rest of the samples in the PCA score plot indicating a potential problem with these two samples related to biological or overall process variability. Two useful output metrics from Linear Discriminant Analysis (PC-LDA) are classification accuracies and Eigenvalues. Classification accuracy is best visualised as a 'confusion matrix' which displays the

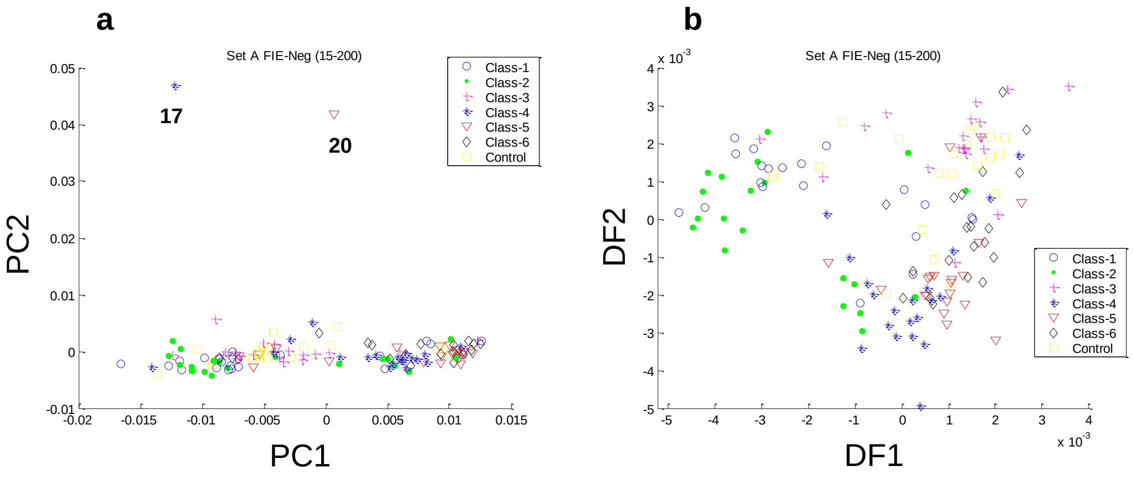


Figure 5.6 Use of PCA and PC-LDA for initial investigation of data quality

(a) Classification with initial data

		True sample class						
		1	**2**	**3**	**4**	**5**	**6**	**C**
Predicted Class	**1**	**10**	3	1	1	1	4	0
	2	3	**10**	1	4	1	1	0
	3	2	2	**13**	0	1	1	1
	4	1	0	1	**14**	2	1	1
	5	0	1	2	0	**15**	1	1
	6	1	0	1	2	2	**11**	3
	C	3	0	1	1	1	2	**12**

(b) Classification with data after outlier removal

		True sample class						
		1	**2**	**3**	**4**	**5**	**6**	**C**
Predicted Class	**1**	**10**	3	1	1	1	4	0
	2	2	**11**	1	4	1	1	0
	3	1	2	**14**	0	1	1	1
	4	1	0	1	**13**	2	1	1
	5	0	1	2	0	**14**	1	1
	6	1	0	1	2	2	**11**	3
	C	1	0	1	1	1	2	**14**

(c) Classification with new data after instrument cleaning

		True sample class						
		1	**2**	**3**	**4**	**5**	**6**	**C**
Predicted Class	**1**	**20**	0	0	0	0	0	0
	2	0	**20**	0	0	0	0	0
	3	0	0	**20**	0	0	0	0
	4	0	0	0	**20**	0	0	0
	5	0	0	0	0	**20**	0	0
	6	0	0	0	0	3	**17**	0
	C	0	0	0	0	0	0	20

(d) Classification after correcting binning problem

		True sample class						
		1	**2**	**3**	**4**	**5**	**6**	**C**
Predicted Class	**1**	**20**	0	0	0	0	0	0
	2	0	**20**	0	0	0	0	0
	3	0	0	**20**	0	0	0	0
	4	0	0	0	**20**	0	0	0
	5	0	0	0	0	**20**	0	0
	6	0	0	0	0	2	**18**	0
	C	0	0	0	0	0	0	**20**

Table 5.2 Confusion matrices from Linear Discriminant Analyses (PC-LDA) illustrated in Figures 5, 6, 7 and 9 after sample data quality assessment and problem troubleshooting. The negative ionization, low mass range data from twenty samples of each biological class were included in each PC-LDA. The numbers in red on the diagonal highlight the number of samples correctly predicted in each class.

number of times a sample is correctly classified or designated as a different class. In the present experiment each biological class is represented by 20 samples. Eigenvalues are a measure of the degree of class separation in each discriminant function (DF) vector. Confusion matrices and Eigenvalue data for the data models displayed in each of Figures 5.6, 5.7, 5.8 and 5.9 are summarised in Table 5.2 and Table 5.3 respectively. The original data shows Eigenvalues smaller than 2.0 which is indicative of a poor data set [50] (Table 5.2), suggesting that metabolome differences between compared sample classes are barely significant. Low eigenvalues are reflected in poor classification accuracies with up to 50% of samples per class confused with another class (Table 5.2a).

Outlying samples are common in complex biological experiments and 20 biological replicates allow the removal of up to two samples per class from the data matrix. Unfortunately, removal of outliers in this example did not improve the quality of score plots (Fig. 5.7).

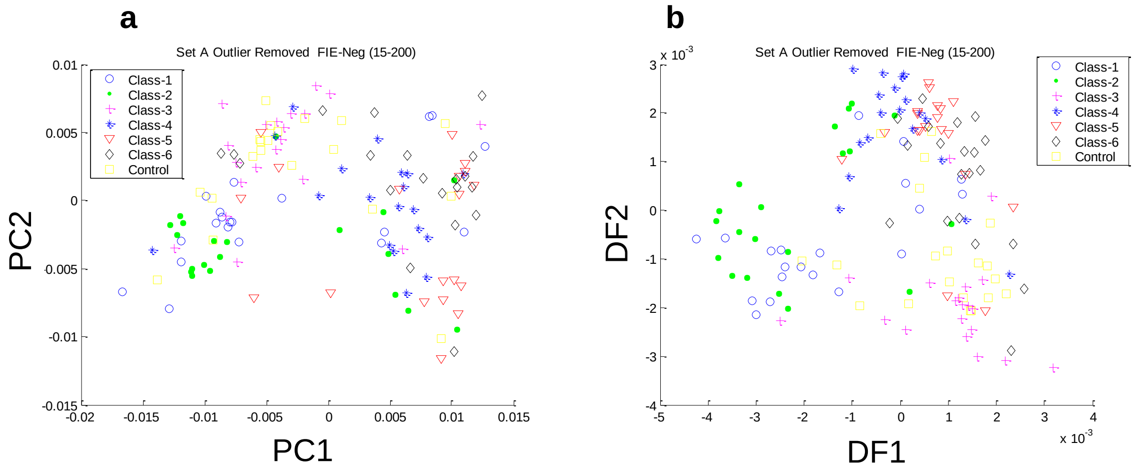


Figure 5.7 Reanalysis of data after removal of sample outliers.

Eigenvalues were still below an acceptable value of 2.0 (Table 5.3) and too many samples remained misclassified (Table 5.2b). In the present example the source of variability was identified, after troubleshooting the analytical instrument, as a small particle in the sample transfer tube between injection port and injection valve of the autosampler. Following reanalysis of the sample set using a thoroughly cleaned analytical system (PROCEDURE Steps 13 to 25) first pass data analysis using PCA and PC-LDA produced the scores plots shown in Fig. 5.8. PC-LDA models had good classification rates and high Eigenvalues for the first four discriminant functions (Table 5.2c and Table 5.3).

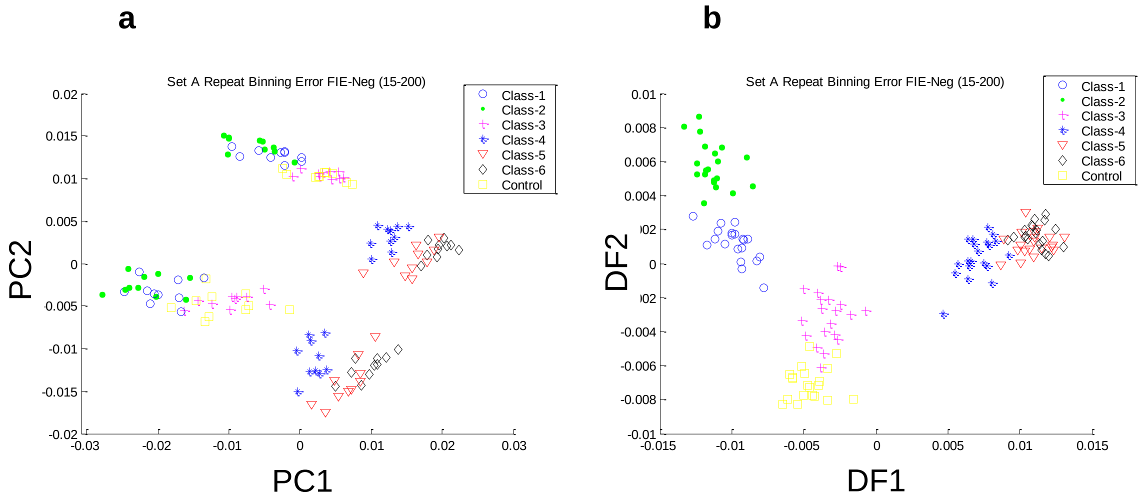


Figure 5.8 Scores plots obtained after instrument troubleshooting and sample reanalysis

However, clustering in PC2 dimension (Fig. 5.8) showed that samples split into two distinctive groups which is often indicative for a mass binning error. The PC2-loading plot (Fig. 5.9) revealed that raw *m/z*-values for nominal mass 97 and/or 98 were not correctly binned.

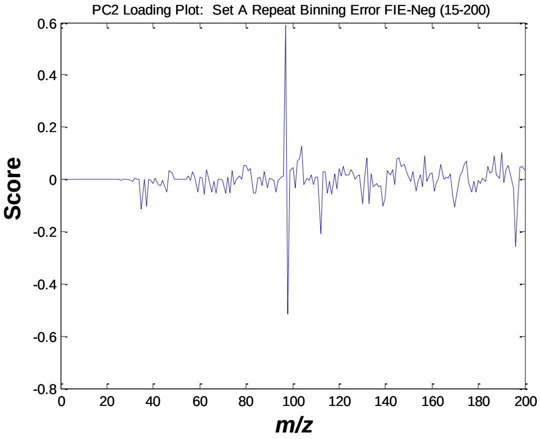


Figure 5.9 PC2 Loading Plot revealing a mass binning error.

Repeating signal processing using a correct binning limit in function '*fiems_ltq_main*' produced a data matrix of very good quality: samples cluster into classes in PCA (Fig. 5.10a), show good separation in PC-LDA (Fig. 5.10b), have high Eigenvalues in the top five DFs (Table 5.3) and show hardly any confusion between classes (Table 5.2d). The data matrices can now be subjected to in-depth statistical analysis including data mining, feature selection and model validation.

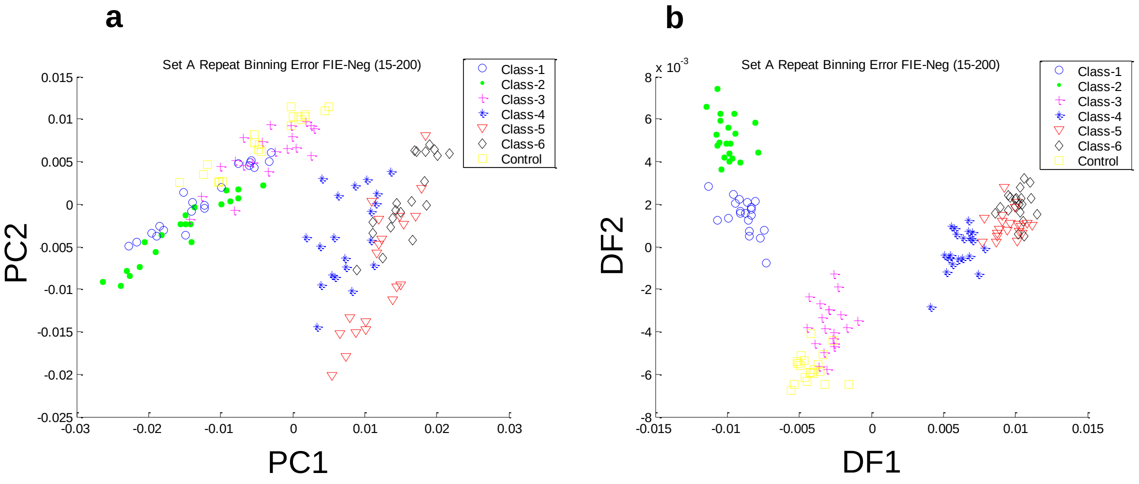


Figure 5.10 Scores plots obtained after repeating data pre-processing using an adjusted binning limit.

		Eigenvalues after addressing sources of variability			
	DF	Initial data	Outlier removal	Instrument cleaning	Binning limit adjusted
Discriminant Function	1	0.81	0.87	36.71	54.39
	2	0.70	0.71	6.92	10.22
	3	0.56	0.58	4.66	6.53
	4	0.45	0.51	2.44	4.04
	5	0.31	0.33	1.04	1.83
	6	0.19	0.24	0.63	0.92

Table 5.3 Eigenvalues for first six discriminant functions (DF) from PC-LDA illustrated in Figure 5.6, 5.7, 5.8 and 5.10 after sample data quality assessment and problem troubleshooting.

REFERENCES

1. Martin, D.B. and Nelson, P.S. From genomics to proteomics: techniques and applications in cancer research. *Trends in Cell Biology* **11,** S60-S65 (2001).
2. Saghatelian, A. & Cravatt, B.F. Global strategies to integrate the proteome and metabolome. *Current Opinion in Chemical Biology* **9**, 62-68 (2005).
3. Murray, D.B., Beckmann, M. & Kitano, K. Regulation of yeast oscillatory dynamics. *Proc. Natl. Acad. Sci. USA* **104**, 2241-2246 (2007).
4. Catchpole, G.S. et al. Hierarchical metabolomics demonstrates substantial compositional similarity between genetically modified and conventional potato crops. *Proc. Natl. Acad. Sci. U S A* **102**, 14458-14462 (2005).
5. Bino, R.J. et al. Potential of metabolomics as a functional genomics tool. *Trends in Plant Science* **9**, 418-425 (2004).
6. Fiehn, O. Metabolomics - the link between genotypes and phenotypes. *Plant Molecular Biology* **48**, 155-171 (2002).
7. Lisec, J., Schauer, N., Kopka, J., Willmitzer, L. & Fernie, A.R. Gas chromatography mass spectrometry–based metabolite profiling in plants. *Nature Protocols* **1**, 387-396 (2006).
8. Kaderbhai, N.N., Broadhurst, D.I., Ellis, D.I., Goodacre, R. & Kell, D.B. Functional genomics via metabolic footprinting: monitoring metabolite secretion by Escherichia coli tryptophan metabolism mutants using FT-IR and direct injection electrospray mass spectrometry. *Comparative and Functional Genomics* **4**, 376-391 (2003).
9. Enot, D.P., Beckmann, M., Overy, D. & Draper, J. Predicting interpretability of metabolome models based on behavior, putative identity, and biological relevance of explanatory signals. *Proc Natl Acad Sci U S A* **103**, 14865-14870 (2006).
10. Ward, J.L., Harris, C., Lewis, J. & Beale, M.H. Assessment of H-1 NMR spectroscopy and multivariate analysis as a technique for metabolite fingerprinting of Arabidopsis thaliana. *Phytochemistry* **62**, 949-957 (2003).
11. Roessner, U., Wagner, C., Kopka, J., Trethewey, R.N. & Willmitzer, L. Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant Journal* **23**, 131-142 (2000).
12. Taylor, J., King, R.D., Altmann, T. & Fiehn, O. Application of metabolomics to plant genotype discrimination using statistics and machine learning. *Bioinformatics* **18**, S241-S248 (2002).
13. Wagner, C., Sefkow, M. & Kopka, J. Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. *Phytochemistry* **62**, 887-900 (2003).
14. Tolstikov, V.V., Lommen, A., Nakanishi, K., Tanaka, N. & Fiehn, O. Monolithic silica-based capillary reversed-phase liquid chromatography/electrospray mass spectrometry for plant metabolomics. *Analytical Chemistry* **75**, 6737-6740 (2003).
15. Tolstikov, V.V. & Fiehn, O. Analysis of highly polar compounds of plant origin: Combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. *Analytical Biochemistry* **301**, 298-307 (2002).
16. Goodacre, R., Vaidyanathan, S., Dunn, W.B., Harrigan, G.G. & Kell, D.B. Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends in Biotechnology* **22**, 245-252 (2004).
17. Sumner, L.W., Mendes, P. & Dixon, R.A. Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry* **62**, 817-836 (2003).
18. Fiehn, O. et al. Metabolite profiling for plant functional genomics. *Nature Biotechnology* **18**, 1157-1161 (2000).
19. Dear, G.J., James, A.D. & Sarda, S. Ultra-performance liquid chromatography coupled to linear ion trap mass spectrometry for the identification of drug metabolites in biological samples. *Rapid Communications in Mass Spectrometry* **20**, 1351-1360 (2006).
20. Startin, J.R., Hird, S.J. & Sykes, M.D. Determination of ethylenethiourea (ETU) and propylenethiourea (PTU) in foods by high performance liquid chromatography-atmospheric pressure chemical ionisation-medium-resolution mass spectrometry. *Food Addit Contam* **22**, 245-250 (2005).

21. Wang, W. et al. Quantification of proteins and metabolites by mass spectrometry without isotopic labelling or spiked standards. *Anal Chem* **75**, 4818-4826 (2003).

22. Smedsgaard, J., Hansen, M.E. & Frisvad, J.C. Classification of terverticillate Penicillia by electrospray mass spectrometric profiling. *Studies in Mycology*, **49**, 243-251 (2004).

23. Goodacre, R. et al. Detection of the dipicolinic acid biomarker in Bacillus spores using Curie-point pyrolysis mass spectrometry and fourier transform infrared spectroscopy. *Analytical Chemistry* **72**, 119-127 (2000).

24. Nicholson, J.K. & Wilson, I.D. Understanding 'global' systems biology: Metabonomics and the continuum of metabolism. *Nature Reviews Drug Discovery* **2**, 668-676 (2003).

25. Allen, J. et al. High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nature Biotechnology* **21**, 692 - 696 (2003).

26. Scholz, M., Gatzek, S., Sterling, A., Fiehn, O. & Selbig, J. Metabolite fingerprinting: detecting biological features by independent component analysis. *Bioinformatics* **20**, 2447-2454 (2004).

27. Aharoni, A. et al. Nontargeted metabolome analysis by use of Fourier Transform Ion Cyclotron Mass Spectrometry. *Omics* **6**, 217-234 (2002).

28. Smedsgaard, J. & Frisvad, J.C. Using direct electrospray mass spectrometry in taxonomy and secondary metabolite profiling of crude fungal extracts. *Journal of Microbiological Methods* **25**, 5-17 (1996).

29. Hopfgartner, G. et al. Triple quadrupole linear ion trap mass spectrometer for the analysis of small molecules and macromolecules. *Journal of Mass Spectrometry* **39**, 845-855 (2004).

30. Zamfir, A.D. et al. Thin chip microsprayer system coupled to quadrupole time-of-flight mass spectrometer for glycoconjugate analysis. *Lab on a Chip* **5**, 298-307 (2005).

31. Nielsen, K.F. & Smedsgaard, J. Fungal metabolite screening: database of 474 mycotoxins and fungal metabolites for dereplication by standardised liquid chromatography-UV-mass spectrometry methodology. *J. of Chromatography A*, 1002, 111-136 (2003).

32. Gorlach, E. & Richmond, R. Discovery of quasi-molecular ions in electrospray spectra by automated searching for simultaneous adduct mass differences. *Analytical Chemistry* **71**, 5557-5562 (1999).

33. Overy, D.P., Enot, D.P., Tailliart, K., Jenkins, H., Parker, D., Beckmann, M. & Draper, J. Explanatory signal interpretation and metabolite identification strategies for nominal mass FIE-MS metabolite fingerprints. *Nature Protocols* **3,** 471-485 (2008).

34. Antignac, J.P. et al. The ion suppression phenomenon in liquid chromatography-mass spectrometry and its consequences in the field of residue. *Analytica Chimica Acta* **529**, 129-136 (2005).

35. van Hout, M.W.J., Niederlander, H.A.G., de Zeeuw, R.A. & de Jong, G.J. Ion suppression in the determination of clenbuterol in urine by solid-phase extraction atmospheric pressure chemical ionisation ion-trap mass spectrometry. *Rapid Communications in Mass Spectrometry* **17**, 245-250 (2003).

36. Muller, C., Schafer, P., Stortzel, M., Vogt, S. & Weinmann, W. Ion suppression effects in liquid chromatography-electrospray-ionisation transport-region collision induced dissociation mass spectrometry with different serum extraction methods for systematic toxicological analysis with mass spectra libraries. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences* **773**, 47-52 (2002).

37. Annesley, T.M. Ion suppression in mass spectrometry. *Clinical Chemistry* **49**, 1041-1044 (2003).

38. Vaidyanathan, S., Kell, D.B. & Goodacre, R. Flow-injection electrospray ionization mass spectrometry of crude cell extracts for high-throughput bacterial identification. *J Am Soc Mass Spectrom* **13**, 118-128 (2002).

39. Mazzella, N. et al. Use of electrospray ionization mass spectrometry for profiling of crude oil effects on the phospholipid molecular species of two marine bacteria. *Rapid Communications in Mass Spectrometry* **19**, 3579-3588 (2005).

40. Favretto, D., Piovan, A., Filippini, R. & Caniato, R. Monitoring the production yields of vincristine and vinblastine in Catharanthus roseus from somatic embryogenesis. Semiquantitative determination by flow-injection electrospray ionization mass spectrometry. *Rapid Communications in Mass Spectrometry* **15**, 364-369 (2001).

41. Zahn, J.A., Higgs, R.E. & Hilton, M.D. Use of direct-infusion electrospray mass spectrometry to guide empirical development of improved conditions for expression of secondary metabolites from actinomycetes. *Applied and Environmental Microbiology* **67**, 377-386 (2001).

42. Goodacre, R., York, E.V., Heald, J.K. & Scott, I.M. Chemometric discrimination of unfractionated plant extracts analyzed by electrospray mass spectrometry. *Phytochemistry* **62**, 859-863 (2003).

43. Overy, S.A. et al. Application of metabolite profiling to the identification of traits in a population of tomato introgression lines. *Journal of Experimental Botany* **56**, 287-296 (2005).

44. Koulman, A. et al. High-throughput direct-infusion ion trap mass spectrometry: a new method for metabolomics. *Rapid Communications in Mass Spectrometry* **21**, 421 - 428 (2007).

45. Rashed, M.S., Al-Ahaidib, L.Y., Aboul-Enein, H.Y., Al-Amoudi, M. & Jacob, M. Determination of L-pipecolic acid in plasma using chiral liquid chromatography-electrospray tandem mass spectrometry. *Clinical Chemistry* **47**, 2124-2130 (2001).

46. Kell, D.B., Darby, R.M. & Draper, J. Genomic computing. Explanatory analysis of plant expression profiling data using machine learning. *Plant Physiology* **126**, 943-951 (2001).

47. Kay, C.D., Mazza, G., Holub, B.J. & Wang, J. Anthocyanin metabolites in human urine and serum. *Br J Nutr* **91**, 933-942 (2004).

48. Beckmann, M., Enot, D. P., Overy, D & Draper, J. Representation, comparison and interpretation of metabolome fingerprint data for total composition analysis and quality trait investigation in potato plants. *J. Agric. Food Chem.,* **55**, 3444-3451 (2007).

49. Parker, D., Beckmann, M., Enot, D.P., Overy, D.P., Rios, Z.C., Gilbert, M., Talbot, N. & Draper, J. Rice blast infection of *Brachypodium distachyon* as a model system to study dynamic host/pathogen interactions. *Nature Protocols* 3, 435-445 (2008).

50. Enot, D., Lin, W., Beckmann, M., Parker, D. Overy, D., & Draper, J. Press (2008). Pre-processing, classification modelling and feature selection using Flow Injection Electrospray Mass Spectrometry (FIE-MS) metabolite fingerprint data. *Nature Protocols,* **3,** 446-470.

51. Jenkins, H. et al. A proposed framework for the description of plant metabolomics experiments and their results. *Nature Biotechnology* **22**, 1601-1606 (2004).

52. Castle, A.L., Fiehn, O., Kaddurah-Daouk, R. & Lindon, J.C. Metabolomics Standards Workshop and the development of international standards for reporting metabolomics experimental results. *Briefings in Bioinformatics* **7**, 159-165 (2006).

53. Broadhurst, D.I. & Kell, D.B. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* **2**, 171-196 (2006).

54. Bligh, E.G. & Dyer, W.J. A rapid method of total lipid extraction and purification. *Can J Biochem Physiol* **37**, 911-917 (1959).

55. Enot, D.P., Beckmann, M. & Draper, J. Detecting a difference - assessing generalisability when modelling metabolome fingerprint data in longer term studies of genetically modified plants. *Metabolomics* **3**, 335-347 (2007).

56. Chernushevich, I.V., Loboda, A.V. & Thomson, B.A. An introduction to quadrupole-time-of-flight mass spectrometry. *Journal of Mass Spectrometry* **36**, 849-865 (2001).

57. Nieuwoudt, H.H., Prior, B.A., Pretorius, I.S., Manley, M. & Bauer, F.F. Principal component analysis applied to Fourier transform infrared spectroscopy for the design of calibration sets for glycerol prediction models in wine and for the detection and classification of outlier samples. *Journal of Agricultural and Food Chemistry* **52**, 3726-3735 (2004).

58. Gullberg, J., Jonsson, P., Nordstrom, A., Sjostrom, M. & Moritz, T. 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. *Analytical Biochemistry* **331**, 283-295 (2004).

59. R_Development_Core_Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-900007-900050, URL http://www.R-project.org (2006).

60. Beckmann, M., Parker, D., Enot, D.P., Duval, E. & Draper, J. (2008) High throughput, non-targeted metabolite fingerprinting using nominal mass Flow Injection Electrospray Mass Spectrometry**.** *Nature Protocols*, **3**, 486-504.

Appendix 5.1

Meta-data describing experimental factors that could be sources of variability prior to tissue extraction in samples derived from rice blast infected Brachypodium distachyon leaves

1. Protocol Reference

Parker, D., Beckmann, M., Enot, D.P., Overy, D.P., Rios, Z.C., Gilbert, M., Talbot, N. & Draper, J. Rice blast infection of *Brachypodium distachyon* as a model system to study dynamic host/pathogen interactions. *Nature Protocols* 3, 435-445 (2008)[49]

2. Plant Genotypes and Fungal Strains

- **Plant species:** *Brachypodium distachyon*
- **Genotype:** ABR 1
- **Seed source:** Aberystwyth University, IBS
- **Seed batch:** 2003
- **Pathogen strain:** *Magnaporthe grisea* GUY11
- **Source:** Aberystwyth University, IBS
- **Fungal stock batch:** September 2004

3. Fungal inoculum preparation

- **Institute:** Aberystwyth University, IBS
- **Location of trial:** Fisons Cabinet No. 3
- **Fungal culture medium:** 10 plates of Complete Medium agar
- **Fungal culture inoculation date:** 13.04.2005
- **Fungal culture conditions:** 26 ^{o}C, 16h light/dark cycle
- **Conidia germination check date:** 23.04.2005
- **Conidia germination frequency:** 85%
- **Conidia harvesting date:** 25.04.2005
- **Filtered conidia concentration:** 1 x 10^{5} conidia mL^{-1} in gelatine water

4. Plant Growth Conditions

- **Institute:** Aberystwyth University, IBS
- **Location of trial:** IBS growth room Polysec 5
- **Layout:** 6 Trays with 90-100 plants each
- **Sowing date:** 05.04.2005
- **Tray size [mm]:** 15” x 19” trays (Vacapots, H. Smith Plastic Ltd.)
- **Soil:** Levington Universal Extra (Scotts UK Professional) supplemented with gravel (50:50; v/v) to improve drainage
- **Day/Night cycle setting [hours]:** 16/8 [day from 3:00 – 19:00]
- **Day/Night temperature setting [^{o}C]:** 23/21
- **Average Day/Night humidity:** 68/68 +/- 10%
- **PAR value [mmol s-1 m-2]:** 140 +/- 10

- **Light source:** 55W (Osram, Sylvania, Mrnich) high-frequency lighting tubes (4,580 lumen output) supplemented with 2* 30W clear tube cooled lighting (Osram)

5. Pathogen Challenge Conditions

- **Plant age:** 5 leaf stage (21-28d post sowing)
- **Plant infection date:** 25.04.2005
- **Inoculum volume per plant tray:** 10ml
- **Inoculum application:** artist's airbrush
- **Post inoculation treatment:** water well and bag for 2 days
- **Disease progression assessment date:** 29.04.2005
- No. Infection sites per cm leaf: 35, type 3-4 on Valent scale

6. Harvest and Storage

- **Harvest date(s):** Harvest days 1-6 from 26.04.2005 to 30.04.2005
- **Time of day:** 11.00 – 12.00
- **Hours of light:** 8 – 9
- **Tissue:** Central 4 cm of leaf
- **No of samples per plant:** 1, stored separately in Eppendorf tubes containing a stainless steel ball for milling
- **Fresh weight:** 40-60 mg per sample (estimate)
- **Storage**: snap freeze in liquid nitrogen within approx. 10 sec and store at –80^{o}C until extraction
- **Sample IDs:** DP.12.001 - DP.12.240

GO3 Final Report - Section 6

David Overy/John Draper/Nigel Hardy

Development of a pilot database for annotation of nominal mass ESI-MS data (part of Task 03.02)

6.1 BACKGROUND TO EXPERIMENTAL SECTION

Before the project started it was agreed with the FSA that the development of standardised data generation, pre-processing and data mining for metabolomics would be based largely on the potato tuber data generated in Aberystwyth, IFR, Golm and SCRI. One of the first objectives of the G03 programme in Section 1 was to make contact with all laboratories that had produced metabolomics data sets for the G02 project. As outlined earlier, the subsequent development and assessment of a G02 data archive indicated that there was only one small NMR data set and only a small amount of LC-MS data containing a relatively (compared to GC-tof-MS) small number of measured peaks. A decision was thus made early on to base most of the development of standardised metabolomics methods using GC-MS to represent profiling data and FIE-MS to represent fingerprinting data.

An overall aim of the G03 project was to develop a list of the metabolites commonly found in potato tubers that might be usefully measured in any future analyses of composition. Those chemicals might be represented as peaks in GC-MS or LC-MS data or *m/z* signals in FIE-MS fingerprints. At the end of the G02 programme the FSA were already aware that many of the consistently measurable metabolite peaks seen in both GC-tof-MS and LC-MS data were of unknown chemistry. In Section 3 of the present programme we integrated GC peak lists developed on 4 different instruments, however, as expected, this did not lead to the annotation of many more previously unknown peaks using major GC-MS spectral database such as NIST. A particular problem with the LC-MS data was that most peaks are not represent by a spectrum (electro spray ionisation [ESI] rarely fragments molecules), but only by a retention time and a typical mass signal.

We had initially intended to start generating LC fractions of potato tuber extracts for more detailed analysis of unknown metabolite peaks. However, as the G03 project was getting underway there was a major shift internationally to stop trying to deconvolve good quality LC peaks and instead use automated chromatographic data alignment software (e.g. MetAlign or XC-MS: see Section 4) to process the LC-MS raw data to develop a retention time versus *m/z* signal matrix, which often contained more than 3000 features. Thus in such LC-MS data a 'feature' is simply a combination of a retention time and an *m/z* value, with the latter corresponding to the ion signals integrated in nominal mass bins in FIE-MS fingerprints of the same sample. Thus, whether *m/z* signals are separated in time (LC-MS), or integrated into nominal mass bins (FIE-MS) the annotation of ions is of paramount importance to help describe the composition of any food raw material.

6.2 OVERALL OBJECTIVES OF EXPERIMENTAL SECTION

Unlike GC-tof-MS (in which molecules are derivatized and fragmented by electron impact ionisation) the 'soft' ionisation afforded by ESI in both LC-MS and FIE-MS approaches means that there is a good chance that the parent metabolites can be putatively identified directly based on their mass. A major effort was thus focused on the development of a nominal mass ESI-MS signal annotation database (ArMEC) strategy within the G03 project. The details of the actual software engineering required to develop the ArMEC database is presented separately in Section 2.5.

6.2.1 Summary of objectives

- Review the literature on the problems of annotation of ESI-MS signals
- Develop a pilot database architecture to hold information on metabolites predicted to be present in potato tubers (described in Section 2.5)
- Develop tables of predicted ESI-MS signals for adducts and fragments of each expected metabolite.
- Develop examples of measured ESI-MS fingerprints using chemical standards
- Develop query and output tools to interpret fingerprinting data (development described in Section 2.5).

6.3 INTRODUCTION TO DEVELOPMENT OF A PILOT ANNOTATION DATABASE FOR LC-MS DATA

Metabolite fingerprinting techniques using soft ionization mass spectrometry are applicable to a variety of research fields: from screening for metabolic changes, to the assessment of food standards in crops and food products [1-4]; from microbial taxonomy and physiology [5,6] to analysis of transgenic plants and breeding population genotyping [3,7]. Fingerprinting techniques such as flow injection electrospray (FIE) ionization or atmospheric pressure chemical ionization (APCI) mass spectrometry strive to provide a high-throughput, non-targeted representation of a suspended moment in the metabolism of an organ, organism or interaction between organisms[8]. Applied as a "first pass" screening protocol, discrimination between samples can be achieved using fingerprinting techniques without prior knowledge of expected differences [8-10]. FIE-MS methodologies are quite rapid as they do not involve a chromatographic or derivatization step, with average analytical run times reported between 1-5 minutes per sample [2-9]. Additionally, the use of soft ionization mass spectrometry, in comparison to electron impact ionization, results in a predominance towards molecular ions rather than molecular ion fragments, which in turn, enhances the ability of the user to identify metabolites in complex matrices [9].

The standardised procedure used the generate FIE-MS data in the G02 project is described in detail in Section 5 of this report. The protocol focuses on the use of datasets in which *m/z* signals have been binned to nominal masses (*i.e.* rounded up or down to whole numbers) as described in the Nature Protocol by Beckmann *et al.* [8]. Nominal mass data is particularly useful as many laboratories do not have routine access to high mass resolution instruments capable of high throughput analyses. From a practical perspective the use of nominal mass data not only reduces data dimensionality but also obviates potential problems associated with mass drift in high accuracy instruments and offers a much to simpler strategy for future data alignment and dataset integration with other laboratories.

6.3.1 Ionisation products generated in electrospray ion source

The largest constraint regarding FIE-MS fingerprinting is that analytes must dissolve in a solvent exhibiting moderate conductivity and, if not already in an ionic form when dissolved, must inductively ionize while passing through the conductive electrospray capillary. Molecules expected to be detected using FIE-MS range from non-polar species that undergo electrochemical oxidation/reduction, to polar and neutral species undergoing dynamic proton association/dissociation as well as association with ion of organic or inorganic salts [11]. A FIE-MS fingerprint of an organism's metabolome, covering a mass range from 50-2000 Da, can be quite complex. With a representative estimate of the metabolome of any given organism ranging between 4000-25000 metabolites and an estimate of the chemical diversity in the plant kingdom alone in excess of 200,000 metabolites [12-14], signal interpretation and metabolite annotation of FIE-MS fingerprints may seem a daunting task. However metabolite diversity within a specific organ/tissue type is obviously substantially lower and, as concentrations of individual metabolites may differ from each other by several orders of magnitude, only a small proportion (perhaps up to 10%) of these huge numbers will be routinely detected by FIE-MS in any typical crude extract made using a single extraction solvent.

Compounding the complexities of a wide chemical diversity, metabolites are often represented by several signals in FIE-MS fingerprints. Along with (de)protonated molecular ions and their associated isotopes, salt adducts of alkali metals, neutral losses (*i.e.* of water, ammonium or formate), homogeneous dimer and/or dimer ion pair adducts are routinely observed in FIE-MS profiles. The formation of dimer complexes occurs in solution through hydrogen bonding interactions with their respective molecules and the energy imparted by the electrospray interface is insufficient to break these complexes prior to analysis in the mass spectrometer [15]. In the gas of the electrospray, strongly bonded Na^+ and K^+ complexes form with molecules which do not readily ionize in solution, such as sugars and glycols, but which have polar groups that are electrochemically suitable for producing ionized analytes [16]. Metabolites possessing an acidic moiety (*i.e.* carboxylate, sulfonate and phosphonate groups), are more susceptible to deprotonation and efficient ionization in negative mode electrospray and are therefore more prominent in negative FIE-MS fingerprints [15]. Chloride ion adducts, readily identified by the natural isotope distribution pattern of chlorine (Cl^{35} and Cl^{37}), have been reported for a variety of compounds, including triacylglycerols, aromatic and aliphatic carboxylic acids, amides, amino acids, aromatic amines, phenols and non-ionic surfactants, neutral glycolipids and glycosphingolipids [17].

6.3.2 Identification of explanatory signals in FIE-MS data

A major output of metabolomic experiments involving FIE-MS fingerprinting, is the identification of highly significant signals (*i.e.* explanatory variables) that account for compositional differences between sample classes. Data processing and feature selection in metabolomics data is described in Sections 8 and 9 of the present report which should be consulted for details and in the Nature Protocol by Enot *et al.* [18]. Of the various supervised data mining options available, Decision Trees have proved in the past to have high classification power as well as the ability to identify explanatory variables in the complex tree structures that they generate [3, 7, 18]. One of the best ways to improve the performance of Decision Tree based algorithms is to grow 'ensembles' of trees in order to identify variables that are consistently highlighted when several thousand different trees are developed from the same data. In the G02006 project we showed previously that one such 'ensemble' method provided by the supervised classification tree algorithm Random Forest (RF) efficiently ranked metabolite signals for significance in their ability to explain the difference between different sample classes [7, 18].

RF models provide insight into the underlying data structure and cope exceptionally well with high dimensional data sets [18, 19]. While various statistical approaches may be employed to identify discriminatory signals, the use of RF was justified as it provided a measure of both model robustness and variable importance without a requirement for prior data dimensionality reduction or data transformation to fulfil statistical assumptions. Importantly, RF retains all variables in the final model, including those exhibiting covariance because they are related to a single metabolite. This property of RF models allows the user to intuitively determine the relationships between signals derived potentially from the same metabolite. Particularly a correlation analysis of variables within the list of selected signals can prove very useful to find signals linked to the sample metabolite and thus help in the annotation process [7]. For example, correlation analysis of the explanatory signals (*m/z*) highlighted by RF in FSA G02006 data successfully confirmed signals having consecutive mass to charge ratios (*m/*z) to represent isotopes and in some cases highlighted additional salt adduct signals relating to the same metabolite [7]. By understanding the relationship between *m/z* signals and predicting those likely to be derived from the same metabolite it was clear that the top 30-40 signals in RF lists of variables ranked with respect to discriminatory value in FIE-MS fingerprint comparisons were populated with multiple signals potentially relating to only a small number of metabolites. We reasoned that the mathematical relationships between such signals could be calculated mathematically and decide to develop a specific database (http://www.armec.org/) to help in annotation of ESI signals[3, 7, 19].

6.3.3 ESI-MS signal annotation using metabolite database resources

The process of annotating explanatory molecules is greatly helped by the construction of a species specific FIE-MS and FIE-MS/MS^n database. This resource can be populated by compiling a list of metabolites, that have been reported either in the literature or in other databases, for the species of interest (see Section 2); and then subsequently predicting possible adduct formation, neutral losses and dimeric associations in FIE-MS fingerprints, in addition to making in-house measurements when standards are available. When trying to annotate a nominal mass signal using commercial databases, searching on species specific databases rather than all purpose, general metabolite databases, allows for a more reasonable initial annotation of signals strictly pertaining to the metabolome of the species under investigation. Subsequent to this, signals without annotations can then be queried using more wide-ranging metabolite databases. Publicly available metabolome databases vary in relation to the overall number of metabolites they contain and the range of metabolites covered (*i.e.* natural product to synthetic or combinatorial libraries). Databases such as KEGG [20] are general to a range of organisms as they are based upon biochemical pathways; whereas others such as TAIR [21], KNApSAcK [22] and HMDB [23] are based on a species specific archive, and in the case of fully genome sequenced species such as *Arabidopsis thaliana* (TAIR), a link is made from metabolite to pathway to genome. Large libraries such as CAS (accessible using SciFinder) [24] can often be misleading for the annotation of a biological metabolome as, although it is the largest small molecule database publicly available, the content is composed of many artificially synthesized chemicals [24]. PubChem was found by Kind and Fiehn [25] to be the most reliable database for use in wide-ranging metabolite annotations as this database contains over 5 million entries and is a compilation of several databases such as KEGG, ChemIDplus and NCBI. Such databases contain a plethora of information about individual metabolites which can be considered fixed and of value in metabolite annotation (*e.g.* molecular mass, structure, synonyms) and thus transferable between databases. However, information relating to real ionisation behaviour (*i.e.* common adducts and fragments formed during soft ionisation in ESI, and spectra in MS/MS^n fragmentation experiments) of a specific metabolite has to be generated in the context of an individual sample matrix within a specific instrument following a specific analytical procedure. Similarly, literature references on the presence or absence of a metabolite, the pathways it is actually involved in and its measured concentration levels are not provided in many general databases and such information can be extremely helpful when interpreting metabolite signal identities by focusing on previously measured chemistry.

6.3.4 Consideration of Overall Experimental Design

Due to the limitation of low levels of mass accuracy, a certain level of ambiguity is present when nominal mass MS approaches are used, which makes the confirmation of putative signal annotation representing complex matrices of unknown composition difficult without further chemical investigation. This is exceptionally the case with FIE-MS fingerprinting as the chromatographic step is missing. The use of high resolution instrumentation can help solve this issue, as molecular formula can be predicted from MS measurements: however in addition to issues of maintaining correct calibration, purchasing and operating such a machine is expensive and therefore not feasible for use in all laboratories. As it is likely in most complex biological samples that more than 40% metabolites are still structurally uncharacterised within any instrument platform, then a major consideration in FIE-MS fingerprinting is the reproducible annotation of *m/z* signals highlighted by data modelling. Fragmentation of nominal mass signals by FIE-MS/MS^n technologies and comparison of the fragmentation trees with either commercially available packages (such as Mass Frontiers from Thermo Finingan) or in-house MS^n databases generated using commercial standards have proven to be successful for the annotation of ESI-MS/MS^n spectra [26-29]. As ion trees for particular adducts are unique to a given molecular structure, confirmation of the putative metabolite identifications and associated adduct/neutral loss/dimeric states can be supported by FIE-MS/MS^n experiments. However, in the case where putative identifications are shared by molecules exhibiting the same molecular weight and structural similarity (*e.g.* monosaccharides), FIE-MS/MS^n experiments will only be able to prove the identification of the class of compound and its adduct/neutral loss/dimeric association. Confirmation of putative signal annotations for these metabolites must therefore be carried out by protocols employing chromatography.

Signal annotation following gas chromatography mass spectrometry (GCMS) employing databases searching (*e.g.* NIST) and comparison to commercial standards, is an accepted method for metabolite confirmation; however, due to the derivatization step necessary for chromatography, a direct correlation to observed *m*/z signals in the FIE-MS fingerprint, although possible, remains tentative. Application of liquid chromatography coupled with electrospray ionization-MS avoids the need for derivatization, allowing for a direct relation between signal annotation and observed nominal mass; however there are some limitations in the representation of the total FIE-MS fingerprint due to issues of analyte polarity as well as the possibility of the formation of additional solvent adducts. Against this background we decided to base our studies on the use of a linear ion trap instrument (such as the Thermo LTQ) with MS/MS capability linked to a high quality HPLC system to provide an ideal platform for FIE-MS fingerprinting and the subsequent investigation of important signals.

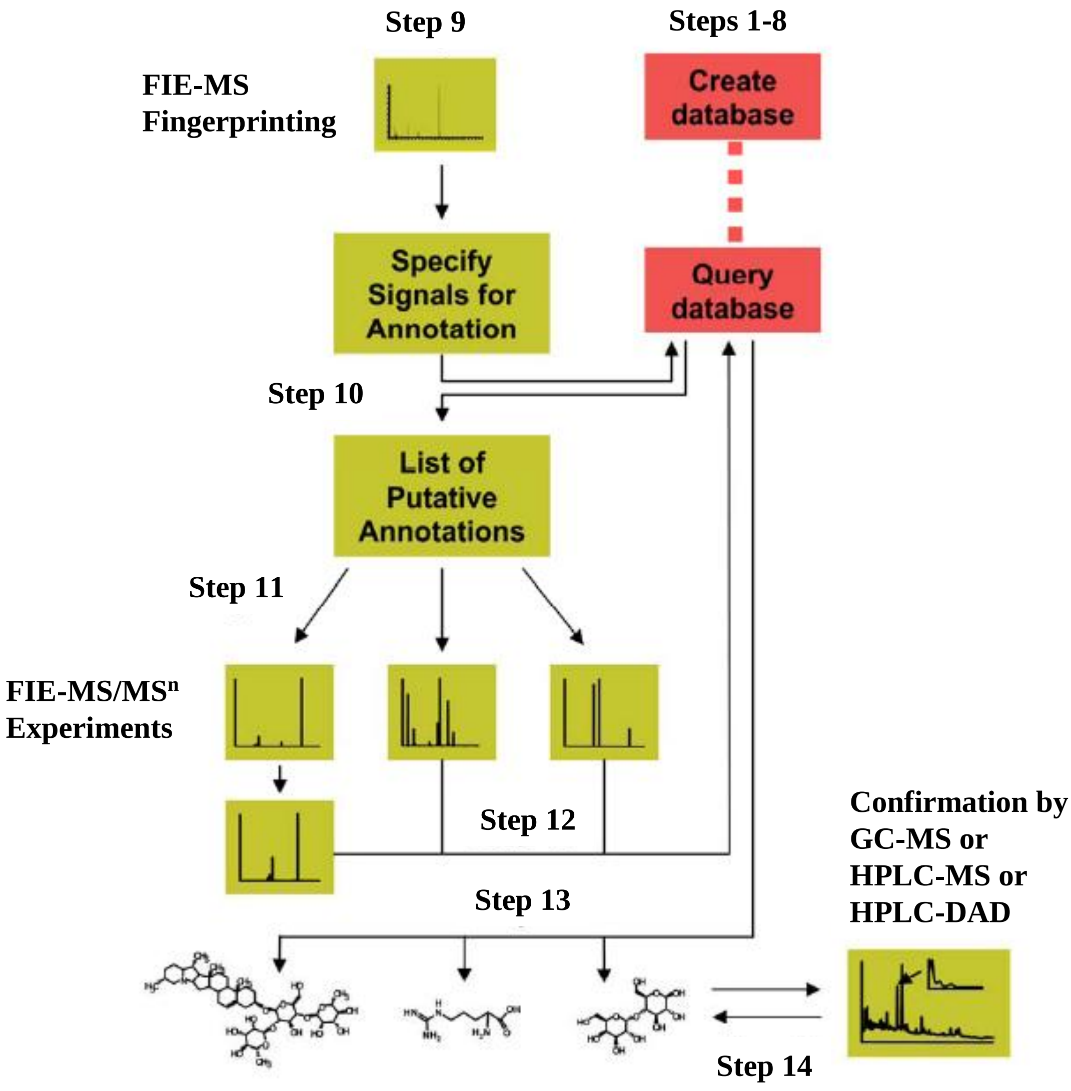

Step 9
Steps 1-8
FIE-MS Fingerprinting
Create database
Specify Signals for Annotation
Query database
Step 10
List of Putative Annotations
Step 11
FIE-MS/MS^n Experiments
Step 12
Confirmation by GC-MS or HPLC-MS or HPLC-DAD
Step 13
Step 14

Figure 6.1: **Flow diagram of steps used to annotate metabolite signals from a FIE-MS fingerprint representing a specific sample matrix.** A detailed description of each step is provided in the PROCEDURE (Section 6.4).

The first part of this section of the project was to develop a standardised way of creating a species specific ESI-MS database which culminated in the development of an on-line resource (ARMeC). This research was initiated in a BBSRC ISIS project by a visiting research Dr David Overy, who subsequently joined the G03 project. The second part of the protocol provides instruction on how to putatively identify metabolites based on database queries of signals of interest from individual FIE-MS fingerprints (see flow diagram, Figure 6.1) and from explanatory FIE-MS fingerprint signals derived from metabolome comparisons as determined by data mining (using for example RF classification trees) and correlation analyses (see flow diagram, Figure 6.2). Collaboration with data mining experts is a key factor to ensure that only good quality data models with high interpretability potential are submitted for interrogation [7].

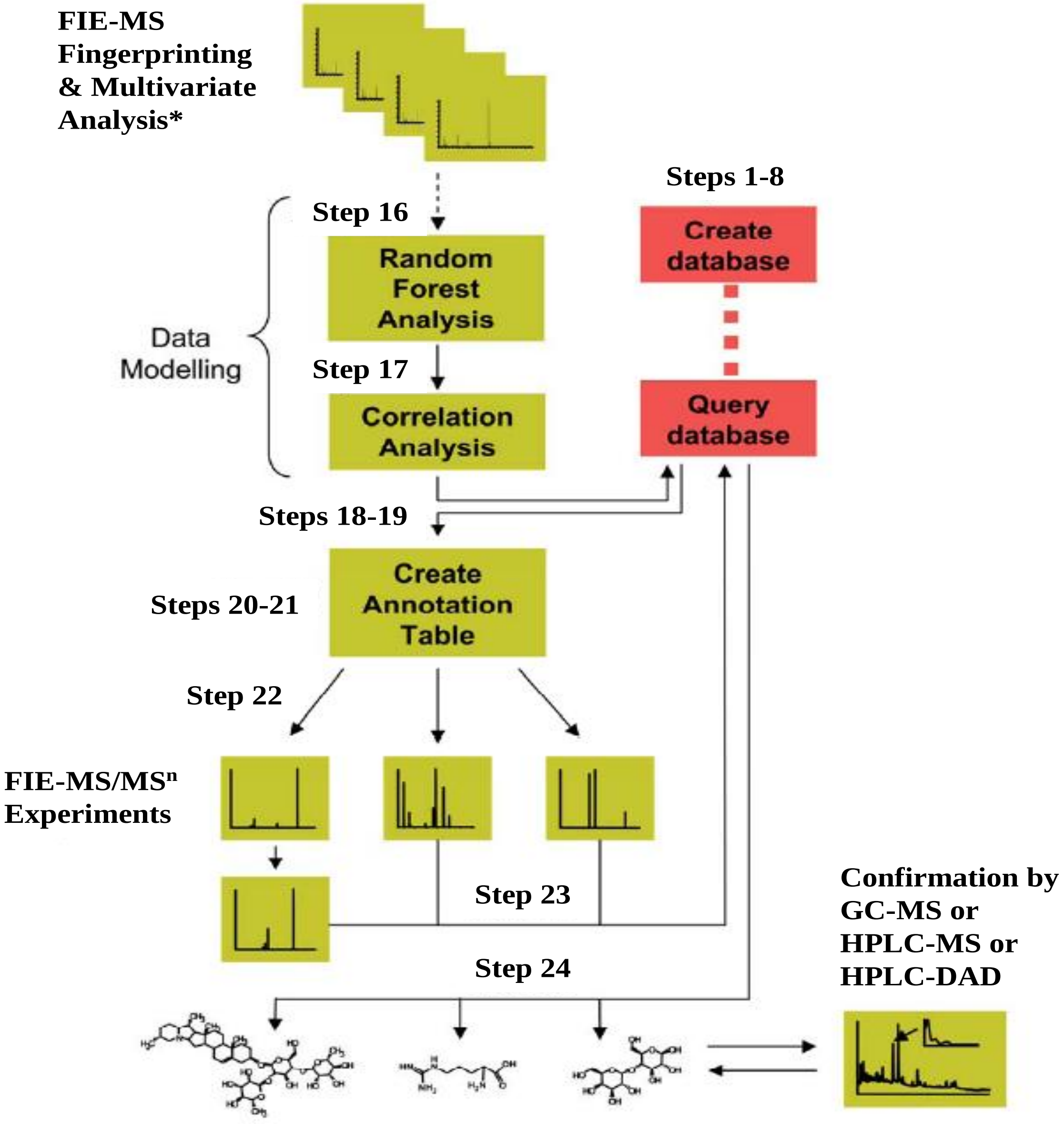


Figure 6.2: **Flow diagram of steps for the annotation of explanatory signals from FIE-MS fingerprint data modelling experiments.** A detailed description of each step is provided in the PROCEDURE (Section 6.4).

Interpreting the output of FIE-MS data mining experiments is a major use of the protocol described below. The data mining procedures are used to generate a list (often with variables ranked) of metabolite signals (*m/z*) that have statistically significant power to explain metabolome differences between two or more sample classes (explanatory variables). Because of the attributes described above, Random Forest was focused upon here for the elucidation of a ranked list of explanatory variables which is used as a starting point for the second part of the protocol. However, the approach taken in this protocol can also be applied to annotate explanatory variables derived from other data modelling algorithms such as Support Vector Machine analysis or Partial Least Squares Regression analysis [31]. Further statistical analysis expertise is also required to evaluate the behaviour of highlighted signals; strong correlations between specific ions could be indicative of a common origin from a specific metabolite or closely linked elements of a metabolic pathway. Data analysis software resources, instructions for their use and specially tailored scripts for RF data mining and correlation analysis are provided within the protocol described below and have been published in the Nature Protocol by Enot *et al.* [18].

6.4 PROTOCOL DESCRIBING METABOLITE IDENTIFICATION STRATEGIES FOR NOMINAL MASS FIE-MS METABOLITE FINGERPRINTS

6.4.1 REAGENTS

- Chloroform ($CHCl_3$), for residue analysis (Fischer Scientific, cat. no. C/4963/15)
  - **CAUTION**: Chloroform is toxic and should be handled under a fume hood using appropriate safety clothing
- Methanol (MeOH), HPLC grade (Fischer Scientific, cat. no. M/4056/17)
- Water, double distilled
- Sodium hydrogen carbonate, anhydrous ($NaHCO_3$), analytical reagent (Fischer Scientific, cat. no. S/2920/53)
- Potassium hydrogen carbonate, anhydrous ($KHCO_3$), analytical reagent (Fischer Scientific, cat. no. P/4120/50)
- All standards used were at least of 97 % purity. Organic acids were used as free acids where possible. Standards appeared generally to contain traces of sodium or potassium salts. Standards were purchased from different sources: Sigma/Aldrich, Fluka, Acros, Merck.

Extraction Solvents

- A detailed discussion of extraction solvents for use in flow injection electrospray ionisation mass spectrometry is presented in the accompanying Nature Protocol by Beckmann *et al.* [8]. Although specific solvent combinations can be used to maximize the yield of extracted metabolites, in the general case of high throughput FIE-MS fingerprinting a consensus has to be made as to a particular solvent combination employed: one that maximizes metabolite extraction. The solvent combination of $CHCl_3$, MeOH and H_2O has been recommended for plant metabolomics [32]. The solvent ratios used in the following protocol were $CHCl_3$: MeOH:H_2O (2:5:2; v:v:v).

Flow Infusion Solvents

- Depending on the solvent type selected, additional adduct ions could be generated (*i.e.* if CH_3CN is used then CH_3CN adducts are also seen). Whichever standard operating procedure has been selected for use in the lab, these conditions should be maintained as deviation from these running parameters might have direct influence over signals present in the in-house database. Use of acid in the solvent will reduce the number of salt adducts (drive the formation of $[M+H]^+$), but inhibits effective fingerprinting in negative ion mode. The flow solvent used in the following protocol was MeOH:H_2O (50:50; v:v).

○ **CRITICAL** Addition of acid to the flow solvent will hinder simultaneous monitoring in both ionisation modes.

- **Metabolite standards in extraction solvent:** Dissolve 1 mg of the metabolite standard into 1 mL of selected extraction solvent (metabolite stock). Prepare a 1:10 dilution by adding a 100 μL aliquot to 900 μL of extraction solvent for use in FIE-MS and FIE-MS/MS^n analysis.

- **Metabolite standards in $NaHCO_3$**: Prepare a 1 M stock solution of $NaHCO_3$: H_2O (84:1; mg:mL; w:v). Perform a serial dilution to obtain a 1:1000 dilution of the original $NaHCO_3$:H_2O solution by removing a 100 μL aliquot into 900 μL of extraction solvent and then repeating this with the resulting solution two more times.

To get a 1:10 dilution of the metabolite standard add 100 μL aliquot of the metabolite stock to 900 μL of the $NaHCO_3$:H_2O (1:1000) solution and vortex prior to injection.

- **Metabolite standards in $KHCO_3$**: Prepare a 1 M stock solution of $KHCO_3$:H_2O (100:1; mg:mL; w:v). A serial dilution is then prepared by removing a 100 μL aliquot into 900 μL of extraction solvent and then repeating this step with the resulting solution two more times to obtain a 1:1000 dilution of the original $KHCO_3$:H_2O solution. To obtain a 1:10 dilution of the metabolite standard add 100 μL aliquot of the metabolite stock to 900 μL of the $KHCO_3$:H_2O (1:1000) solution and vortex prior to injection.

6.4.2 EQUIPMENT & SOFTWARE

Injection parameters

- Aliquots of various standard solutions were injected at a rate of 3 μl/min using a Hamilton syringe (250 μl) into a 3 μl/min solvent flow of water/methanol (50:50) from a Surveyor liquid chromatography system (ThermoFinnigan). As data collection was interchanged between positive and negative ion mode, the solvents were not acidified.

Mass spectrometer parameters

- **Mass Spectrometer parameters for FIE-MS and FIE-MS/MS^n analysis:** A standard procedure for use of a LTQ mass spectrometer (ThermoFinnigan) is provided in Section 5 of the G03 report and in related Nature Protocol by Beckmann *et al.* [8]. To summarise, data collection is carried out in positive and negative mode and the ionization conditions are as follows: spray voltage 4.5 kV (ESI+) and 4.0 kV (ESI-), +15 V (ESI+) and -13 V (ESI-), transfer capillary temperature set to 80 °C. Calibration and tuning of linear ion trap are carried out to vendor specifications.

- o **CRITICAL** Differences in applied cone voltage will have a direct impact on the ionization (and potential fragmentation) of the analyte. Equivalent equipment parameters should be used when comparing FIE-MS spectra against previous data generated from standards.

- **Mass spectrometer parameters or MS/MS^n experiments**: Scan window set for 20 scans, isolation width set to 1 m/z, normalized collision energy 40 V, activation Q 0.250, activation time 30 ms, wideband activation turned on, source fragmentation at 20 V. Mass range settings are dependant upon molecular weight of target ion.

Database Software

To enable efficient and reliable annotation of metabolites, the database of species specific FIE-MS and FIE-MS/MS^n data must exhibit a number of characteristics. First, the data should be consistent and ideally it should be as complete as possible. However, in the case of metabolomics of food raw materials where large numbers of metabolites remain to be structurally identified it is clear that data collections will expand and undergo regular update which should incrementally improve the accuracy of annotation. Second, the data must be structured and organised in such a way as to enable efficient querying. Finally, the data must be accessible in a controlled way to multiple users at the same time. Spreadsheets, that may be produced using software packages such as Microsoft® Excel, provide a simple approach to the storage of such a data collection. However, such spreadsheets typically provide little inherent support for actions such as constraint checking and query optimisation and are not designed for simultaneous access. While this is not necessarily a problem for a standalone database of limited size that contain static data; in situations where the data collection may expand or undergo regular update, or where it is desirable for the data collection to be publicly available, a more robust solution is required.

Database Management Systems [33] (DBMS) are software tools that are designed to support development of and access to databases. They perform constraint checking to maintain data integrity and support different data organisations to optimise querying. These benefits do not come without the associated costs of design, development and population of the database, but the added utility that they give to a data collection will often outweigh those costs. We have built a database for the Oracle® 9*i* DBMS to house our collection of species specific FIE-MS and FIE-MS/MS^n data. A publicly accessible web-based interface to this database has also been developed using Java servlet JavaServer Pages and can be publicly accessed at http://www.armec.org/ .

Random Forest data mining software

It is assumed that users will generally collaborate with experienced data analysts to generate lists of 30-40 highly ranked *m/z;* appropriate data mining methods are described in the Sections 8 an d9 of the present report and have been published in a Nature Protocol by Enot *et al.* [18]. The present protocol utilises the classification tree algorithm Random Forest which has been shown to perform well for analysing FIE-MS fingerprint data [7, 18]. Random Forest analysis can be performed in the freely available and multi-platform R environment (downloadable at http://www.r-project.org/) using the additional package *randomForest* (http://cran.r-project.org/src/contrib/Descriptions/randomForest.html). Variable importance score computation requires the argument importance to be set to TRUE. One of the interesting properties of RF, is that a model does not overfit: the generalization error will reach a certain value no matter how many trees are built. Selection of an initial value for *ntree*=2000 (the number of trees calculated) is a good starting point for most analyses. All the other parameters of *randomForest* function can be used with their default value. It is advisable to perform several simulations with much larger numbers of trees and check whenever the variable importance scores are not affected by *ntree*.

Correlation analysis software

The output of any feature selection process (e.g. Random Forest analysis) applied to comparisons of FIE-MS fingerprints data is a (often ranked) list of perhaps 30-40 *m/z* signals with potential power to explain the difference between two or more biological sample classes. With each ionisation mode it cannot be assumed that all metabolites are represented simply as protonated or deprotonated masses in FIE-MS fingerprints. Depending on their chemistry metabolites could be present predominantly as specific adducts or fragments; therefore in order to predict the identity of a parent metabolite, it is important to identify signals that might be derived from the same molecule. Correlation analysis between a selected number of signals of higher explanatory value can be performed with any clustering software application. The R package Rarmec has been developed to both provide correlation calculations and correlation plots as well as automatic queries to the ARMeC website (http://www.armec.org/) to retrieve and sort all possible ionisation products corresponding to each cluster. Rarmec is freely available at http://users.aber.ac.uk/jhd.

In order, to handle both positive and negative correlations that could occur in metabolomics data, the dissimilarity measure between each variable pair is defined as: one minus the absolute value of the Pearson correlation coefficient between two signals (function *cor*). Clustering is performed with the function *hclust* with its default settings for complete agglomeration. Isotopic relationships usually exhibit the strongest correlation with an expected Pearson coefficient greater than 0.8. Please note that amongst a number of different agglomeration strategies, we advise the use of complete linkage or average linkage (UMPGA) methods to find similar clusters. Both methods are competitive and should give the same results regarding the clusters with the most correlation. Because of the nature of the similarity measure, one must check the direction of the correlation (*i.e.* positive or negative correlation) before presuming the origin of the relationships between two signals.

6.4.3 PROCEDURE

Creation of species specific ESI-MS prediction database

1) Using a spreadsheet program (*e.g.* Microsoft Excel) compile a list of reported/expected metabolites as well as information on synonyms, molecular weight, molecular structure and formula for a given species based upon literature reports and available databases (see Section 2 of the present report). For example the general metabolic pathways in KEGG [20] can be visualised to identify subsets of metabolic steps for which enzymes have been described in individual species. Databases such as TAIR [21], KNApSAcK [22] and HMDB [23] are already based on a species specific archive. A typical example of useful information can be viewed in the "Metabolite Summary" section of ARMeC (http://www.armec.org/) which, based on the G02 programme data, currently has metabolite information for potato tubers and *Arabidopsis* leaves and can easily be extended and edited to accommodate new species-specific information.

2) Calculate potential FIE-MS *m/z* signals (regarding (de)protonated ion, adduct, neutral loss and dimeric combinations) for each metabolite (see Table 1 for calculations). This can be done conveniently using a spreadsheet such as the "RF Mining Spreadsheet" found on the ARMeC website   (http://www.armec.org/).

3) Evaluate potential neutral loss predictions against structural information and eliminate predictions with structural impossibilities (for example, a molecule without a carbonyl functional group is unable to undergo a neutral loss of formate).

4) Acquire representative pure standards of the listed metabolites where possible. [see REAGENTS]

Table 6.1: Calculation table for possible ion states observed in positive and negative ion mode.

Positive Ion Mode		Negative Ion Mode	
Ion State	Formula	Ion State	Formula
$[M+H]^+$	M+1	$[M-H]^-$	M-1
$[M+Na]^+$	M+23	$[M+Na-2H]^-$	M-21
$[M+K]^+$	M+39	$[M+K-2H]^-$	M-37
$[M+H-H_2O]^+$	M-17	$[M+Cl]^-$	M+35
$[M+H-HCOOH]^+$	M-45	$[M-COOH]^-$	M-45
$[M+H-NH_3]^+$	M-16	$[2M-H]^-$	2M-1
$[2M+H]^+$	2M+1	$[2M+Na-2H]^-$	2M-21
$[2M+Na]^+$	2M+23	$[2M+K-2H]^-$	2M-37
$[2M+K]^+$	2M+39		

5) Individually profile available standards by FIE-MS, initially in the chosen extraction solvent and then in $NaHCO_3$ and $KHCO_3$ solution, to determine the pattern of actual (de)protonated ion, adduct, neutral loss and dimeric combinations. Detailed instructions for FIE-MS are provided in Section 5 of the report and published in the Nature Protocol by Beckmann *et al.* [8]; in the present protocol the extraction solvent is chloroform:methanol:water (2:5:2; v:v:v) and the flow injection solvent is composed of methanol:water (50:50; v:v).

6) Use the ion prediction list to annotate expected ion signals. Record and update the database with fragment ions not present in the prediction list.

7) At this point several trends in the formation of adducts or neutral losses may be observed. For example, many disaccharides commonly only produce $[M+Na]^+$ and $[M+K]^+$ ions in positive ion mode as illustrated by the behaviour of sucrose in Figure 6.3a-c.

8) Use FIE-MS/MS^n to create individual fragmentation trees of recorded (de)protonated nominal mass ions as well as adducts, neutral losses, dimeric combinations and associated fragments. For example Figure 3d and 3e show fragmentation spectra of the major salt adducts of sucrose standards.

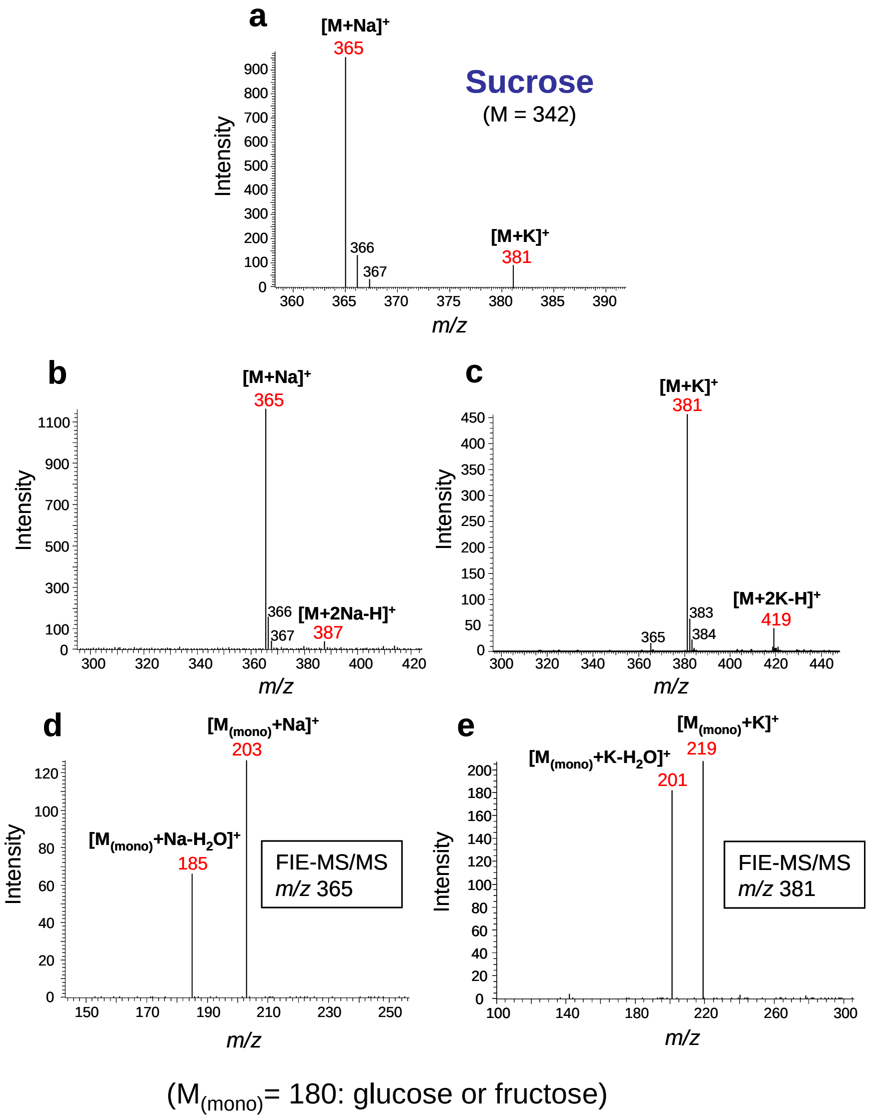


Figure 6.3: **FIE-MS fingerprinting of sucrose solutions showing the effect of salts in the sample matrix.** FIE-MS fingerprints of (a) sucrose solution, (b) sucrose in NaCOOH solution and (c) sucrose in KCOOH solution. FIE-MS/MS spectra of parent ions, (d) *m/z* 365 $[M+Na]^+$ and (e) *m/z* 381 $[M+K]^+$. In both cases the sucrose salt adducts fragment to release the corresponding monosaccharide salt adduct, either without (219 or 203) or with (201 and 185) a neutral water loss. The structural identity of prominent ion products (in red) are shown in square brackets

Developing ion trees of major m/z signals within a specific sample matrix under FIE-MS conditions to guide signal annotation

9) Prepare sample extracts and carry out FIE-MS fingerprinting according to a standardized in-house protocol (for a recommended procedure see Section 5). At this point you may wish to remove an aliquot of the extract and add to an equivalent volume of metabolite standard solution to create a "spiked" extract and carry out steps 11 to 13 for confirmation of putative annotations where *m/z* signals in the original extract fingerprint resulted from more than one predominant metabolite (see Step 13).

10) Specify *m/z* signals of interest for annotation. Query database (*e.g.* ARMeC) for all potential (de)protonated ions, adducts, neutral losses, dimeric combinations or metabolite fragments potentially corresponding to the *m/z* signal. Please note that steps 16 – 24 below describe how to interrogate a ranked list of significant *m/z* signals in order to determine which co-varying signals are potentially derived from the same progenitor metabolite.

11) Sequentially perform FIE-MS/MS^n experiments for all targeted ions. First collects a mass spectrum of a standard. In a second step increase the energy for collision induced dissociation (CID) of a all selected target (parent) mass ions and respectively collect a mass spectrum of the fragmentation product(s) (daughter ion). Thirdly, perform CID for all selected daughter mass ions detected in the second step and repeat where possible this step in further MS/MS experiments until no mass ion is detected. Examples of fragmentation trees for a major unknown metabolites signal in FIE-MS spectra derived from *Brachypodium* extracts is presented in Figure 6.4

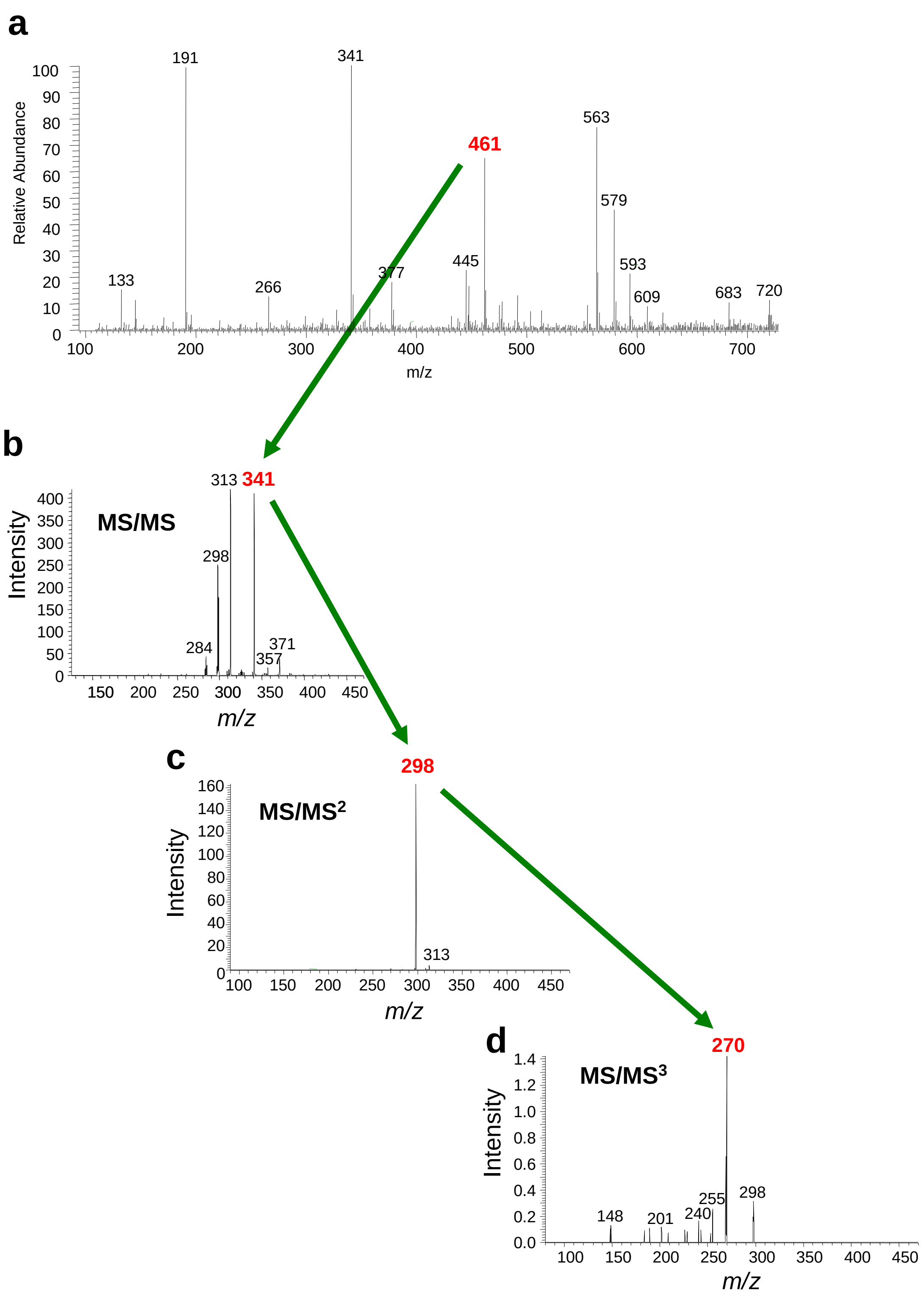


Figure 6.4: **FIE-MS/MS^3 fragmentation tree record of an unknown *Brachypodium* metabolite (*m/z* 461) in negative ion mode.** (a) FIE-MS fingerprint of a *Brachypodium* leaf section harvested 3 days after infection with the rice blast pathogen. Labelled in red is a signal (*m/z* 461) representing a highly explanatory metabolites with no known matches in the ARMeC database of MS/MS spectra. Panels (b) to (d) show a typical MS-3 fragmentation tree record for this unknown metabolite.

12) Query MS/MS^n database with parent *m /z* signal and resulting predominant daughter ion fragments (*m/z* signals with a relative abundance of 60% or greater).

13) Use daughter ion fragments to confirm identity where multiple possibilities of parent *m/z* signal identity exist. If predominant daughter ion fragments appear to be originating from more than one contributing metabolite you may wish to carry out Step 9 using aliquots of sample extracts spiked with metabolite standards corresponding to the multiple predicted signal annotations.

14) If identification of the target *m/z* signal is not possible by database query of the MS/MS^n fragmentation tree (*i.e.* if a MS/MS^n tree has not been recorded in the database for the metabolite annotations under query or multiple hits are recorded having the same MS/MS^n tree), then further profiling of the extract by an additional chromatographical method (*e.g.* HPLC-ESI-MS/MS^n or GC-ToF-MS) is required [32, 34, 35] .

15) In cases where fragmentation patterns do not match database archives, further queries should be attempted using additional, more generalized metabolite database resources.

Annotation de novo of explanatory signals from data modelling

16) Obtain list of statistically significant (explanatory) metabolites generated by multivariate analysis of FIE-MS fingerprints representing different biological sample classes. A detailed description of this process is described in Section 8 and 9 of the present report. Appropriate software resources to utilise Random Forest decision tree analysis are indicated in the Equipment section. In the present protocol the results of data mining experiments from analysis of either a time course of pathogen-infected *Brachypodium* leaves [30] or potato tuber representative of several varieties [19] provide example data models.

17) Perform correlation analysis using top ranked signals from both FIE-MS positive and negative ion modes separately. Generally all variables with a Random Forest Importance Score > 0.002 or a *P*-value < 0.001 may be considered significant [7, 18, 31].

18) Identify isotopic relationships from the strongest correlating signals of consecutive mass values having an expected Pearson coefficient greater than 0.8. Isotopes should not only be strongly correlated but should also exhibit distinctive ratios, with the *m/z* signal decreasing as the mass increases (see examples in Figure 6.10 and Figure 6.11). Additionally, isotopes should also exhibit the same intensity trends in each sample class (see examples in Figure 6.10 and Figure 6.11 in Anticipated Results).

19) Query each explanatory *m/z* signal against the ESI-MS prediction database (*e.g.* ARMeC) and annotate all possible putative identifications, corresponding to either the (de)protonated nominal mass ion, salt adduct, neutral loss, dimeric combinations or associated fragments. Salt adducts of major signals are also expected to be strongly correlated (see Figures 6.10 and 6.11 for examples).

20) Enter the *m/z* signal into a calculation spreadsheet (for a model spreadsheet and more detailed instructions refer to RF mining instructions at http://www.armec.org/) and determine all possible associated ion signals in a situation where the *m/z* signal were to represent the (de)protonated molecular ion, a salt adduct of either Na or K, a neutral loss of water, ammonium or formate, the homogeneous dimer ion or a dimer ion pair adduct of Na or K. Use the resulting list of possible associated ion signals to search against the RF ranked list and highlight all associated ion signals.

21) Record into an annotation table all of the (de)protonated nominal mass ion, salt adduct, neutral loss or dimeric combination corresponding to all putative identifications made in steps 19 and 20. Note that in our experience [3,7,17] it is not uncommon to find that 3 or 4 (or even more) explanatory *m/z* are in fact all ionisation products of a single molecule (*e.g.* see the sucrose cluster in Figure 6.10 in Anticopated Results). This intuitive analysis provides a first pass approach to predict the likely mass of the parent metabolite and allows targeted analysis of a sub-set of *m/z* signals by MS/MS^n and more logical interpretation of MS/MS^n data in instances where metabolite structure is unknown.

22) Subject each *m/z* signal representative of a potentially highly discriminatory metabolite to FIE-MS/MS^n fragmentation as described in steps 10-14 above. Note it may be difficult to generate good quality MS/MS^n spectra for low abundance signals. In such circumstances more success might be achieved by use of longer infusion times using a robotic direct

injection system, such as the Nanomate (as described Section 5 of the present report) or following separation of *m/z* signals in the sample matrix following Ultra High Pressure Liquid Chromatography (UHPLC) [35].

23) Compare the FIE-MS/MS^n fragmentation trees against the database as previously described in step 12 to reduce the number of putative identifications in the peak table.

24) If identification of the target *m/z* signal is not possible by database query of the MS/MS^n fragmentation tree (*i.e.* if a MS/MS^n tree has not been recorded in the database for the metabolite annotations under query or multiple hits are recorded having the same MS/MS^n tree), then further profiling of the extract by an additional chromatographical method (*i.e.* HPLC-ESI-MS/MS^n or GC-ToF-MS) is required [32, 34, 35]

25) In cases where fragmentation patterns do not match database archives, further queries should be attempted using additional, more generalized metabolite database resources.

6.5 ANTICIPATED RESULTS

Profiling of Standards

Several trends regarding the relationship between adduct formation, neutral losses and metabolite class should become visible after fingerprinting metabolite standards by FIE-MS. Understanding the structure of the most prevalent ionisation products of any particular metabolite is of major importance when trying to predict the molecular origin of any *m/z* signal in an FIE-MS fingerprint. When non-buffered flow solvents are used, Na^+ and K^+ ions are the most prominent adducts observed in positive mode FIE-MS. FIE-MS/MS^n fragmentation patterns of $[M+H]^+$, $[M+Na]^+$ and $[M+K]^+$ ions can differ [25]; therefore it is crucial that standards are profiled in both Na^+ and K^+ buffered solutions (see Figure 6.3 in previous section). It is important to note here that if buffering salts are used to acidify or alkalize the FIE-MS carrier solvent, the appropriate salt ion adducts should be compensated for. For example, if ammonium acetate is added to the carrier solvent, adduct predictions should include the occurrence of an ammonium ion adduct ($[M+NH_4]^+$) in positive ion mode. The use of either free acid or salt forms of metabolite standards can affect ionisation behaviour as illustrated for sodium benzoate which preferentially forms ion clusters in the presence of sodium (Figure 6.5). Additionally, depending on the carrier solvent used for flow injection, the formation of solvent ion clusters might have to be accounted for. No solvent/molecular ion clusters were observed using the flow solvent combination advised in this protocol; however solvent/protonated molecular ion clusters (*i.e.* $[M+H+CH_3CN]^+$) and solvent/alkali metal/molecular ion clusters (*i.e.* $[M+Na+CH_3CN]^+$) have been observed when CH_3CN has been used for ESI-MS [37, 38]. Depending on their chemical structure, many metabolites are commonly encountered as signals relating to neutral loss fragments as illustrated for 2-phenylglycine and 5-chlorosalicylate in Figure 6.6; *m/z* representing the loss of formate (both metabolites) or loss of an ammonia group (2-phenylglycine) are stronger signals than molecular ions. Figure 6.6b illustrates the typical appearance of chlorine isotopes differing by 2 mass units, reflecting the natural isotopic ratio of Cl_{35} and Cl_{37}.

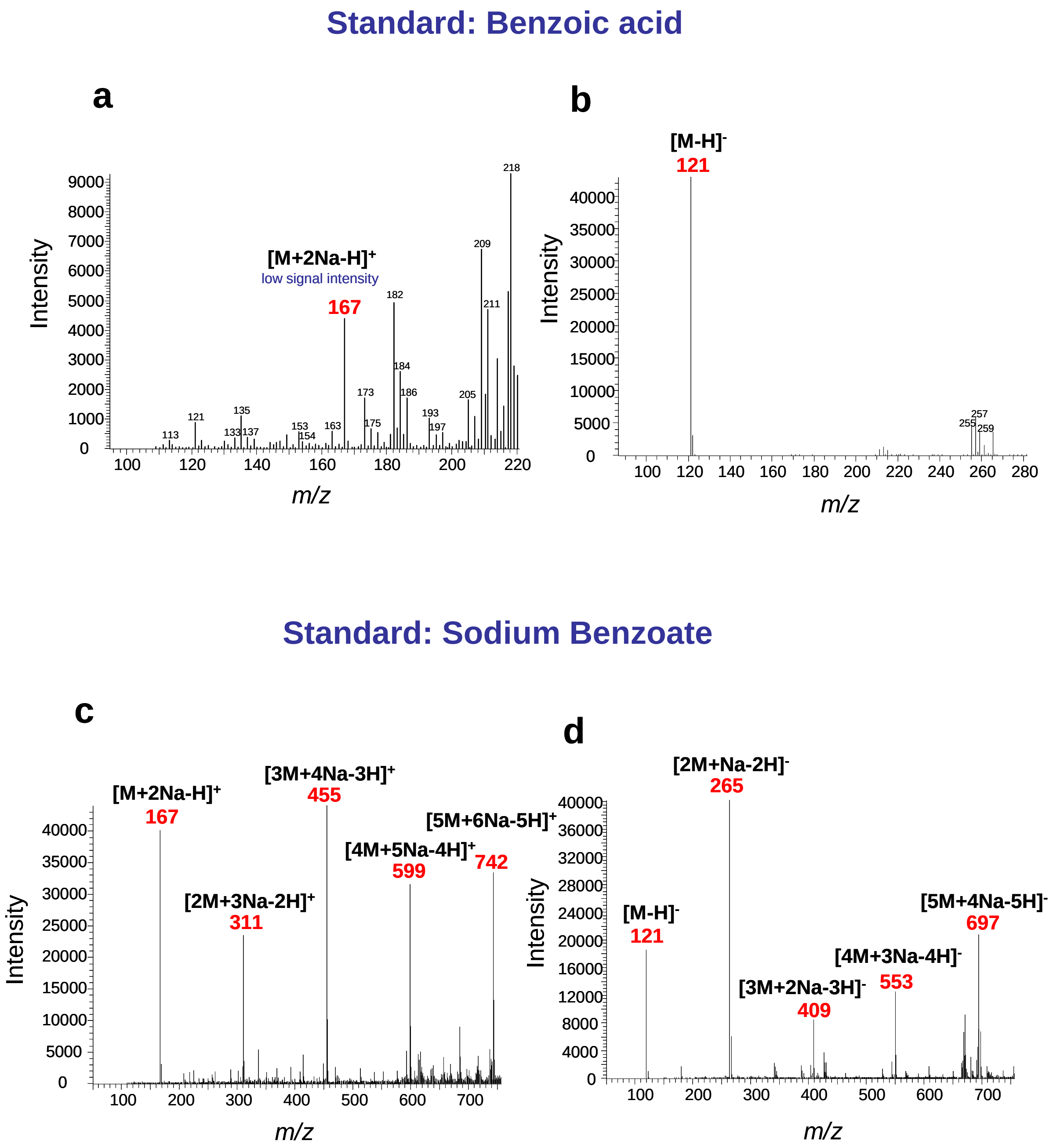


Figure 6.5: The effect of sodium on the formation of ion clusters. A standard solution of benzoic acid was analyzed by FIE-MS fingerprinting in positive (a and c) and negative (b and d) ionization modes, as either a free acid (a and b) or salt solution (c and d). The structural identity of prominent ion products (in red) are shown in square brackets.

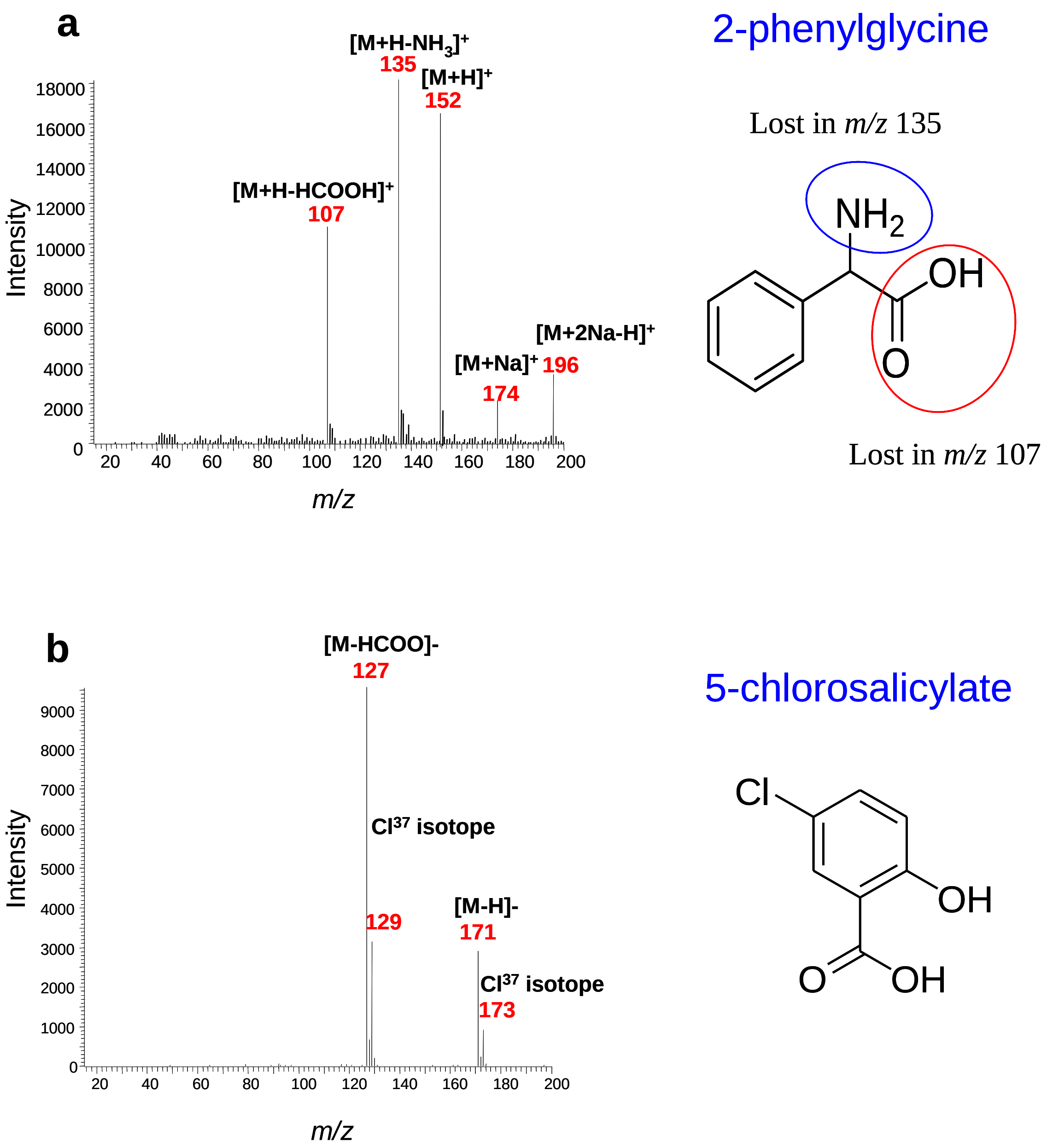


Figure 6.6: Examples of neutral losses and chlorine isotopes in metabolite spectra. FIE-MS spectra of (a) 2-phyenlglycine in positive ionisation mode and (b) 5-chlorosalicylate in negative ionisation mode illustrate the presence of metabolite fragments as major ions. The chemical moieties (amino and carboxyl groups) lost after ionisation of 2-phenylglycine are circled in (b).

Identification of Metabolites using FIE-MS/MS^n

Depending on the molecular structure and relative quantity, FIE-MS/MS^n fragmentation of metabolites within the mass range of 150-1000 Da typically yield fragmentation trees with components ranging from MS/MS to MS/MS^3 fragments; however for larger molecules such as short chain polymers (1500-2000 Da) up to MS/MS$^{6-9}$ fragments can be expected. In the case of low molecular weight molecules (approx. 50-150 Da), FIE-MS/MS experiments result in molecular fragments of a molecular weight below the mass range of the detector (<50 Da). Figure 6.7 is an example of the identification of 2 predominant nominal mass signals in the FIE-MS fingerprint of a potato tuber using FIE-MS/MS^n experiments. In Figure 6.7, the prominent nominal mass ion 852 *m/z* represents the protonated $[M+H]^+$ ion of chaconine, a glycoalkaloid having 3 glycoside units (namely one glucose attached to 2 rhamnose units), whilst *m/z* 868 represents the lower abundance glycoalkaloid solanine which similarly has 3 glycoside units (namely one galactose linked to a glucose and a rhamnose unit). Subsequent fragmentation experiments of the 852 *m/z* and 868 *m/z* ions results in a series of fragmentation trees, uniquely representative of the $[M+H]^+$ ion of the alkaloid moiety solanidine with sequential loss of a single monosaccharide unit following each round of ionisation.

Putative Annotations by Database Queries

Querying metabolome databases with the generated FIE-MS signal list will result in a multiple number of putative identifications for any one given nominal *m/z* signal. The application of a species specific metabolome database for queries initially reduces the number of probable "hits" to create a shortlist of putative *m/z* signal identifications as shown for the signal *m/z* 175 in Figure 6.8a & 6.b.

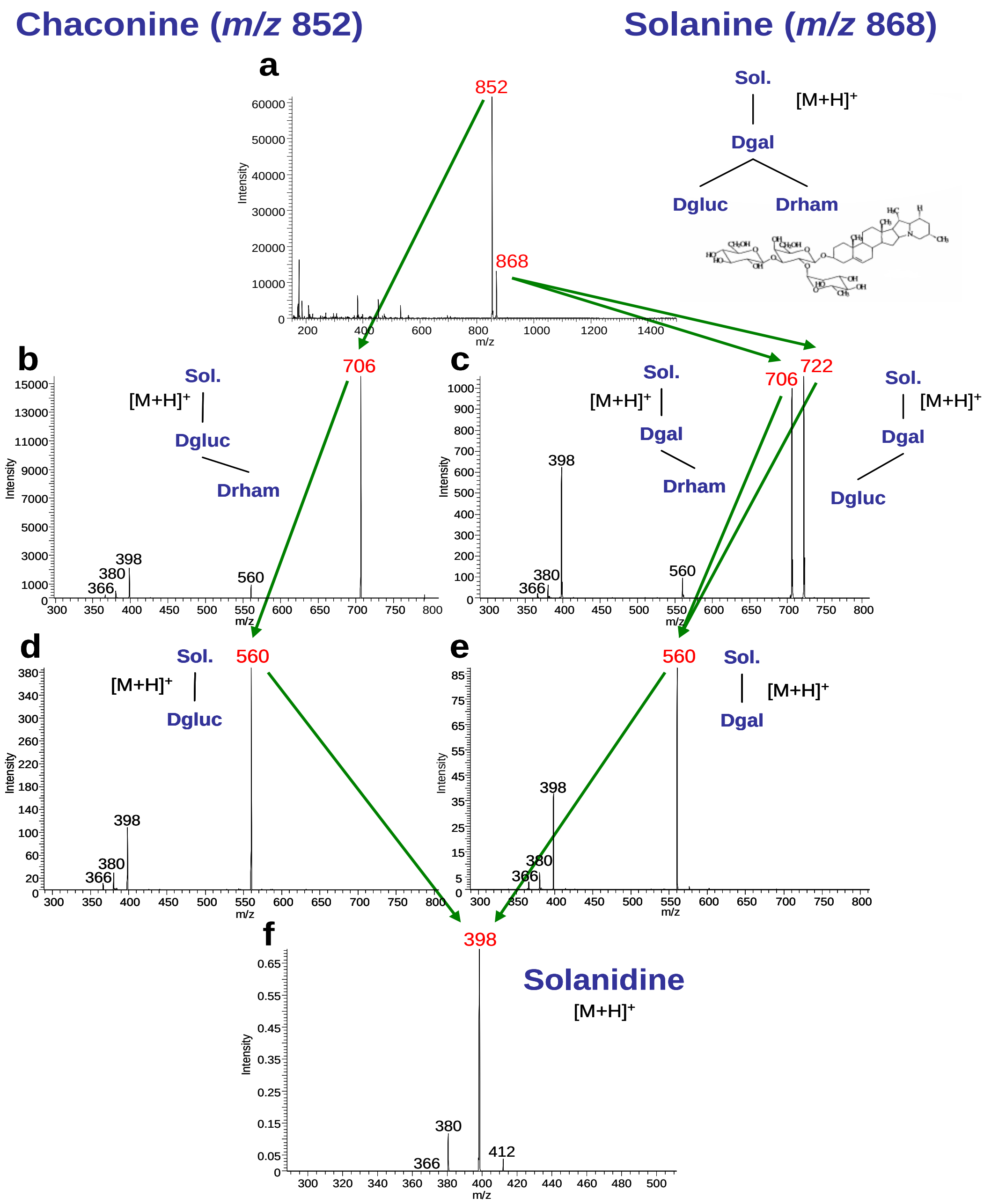
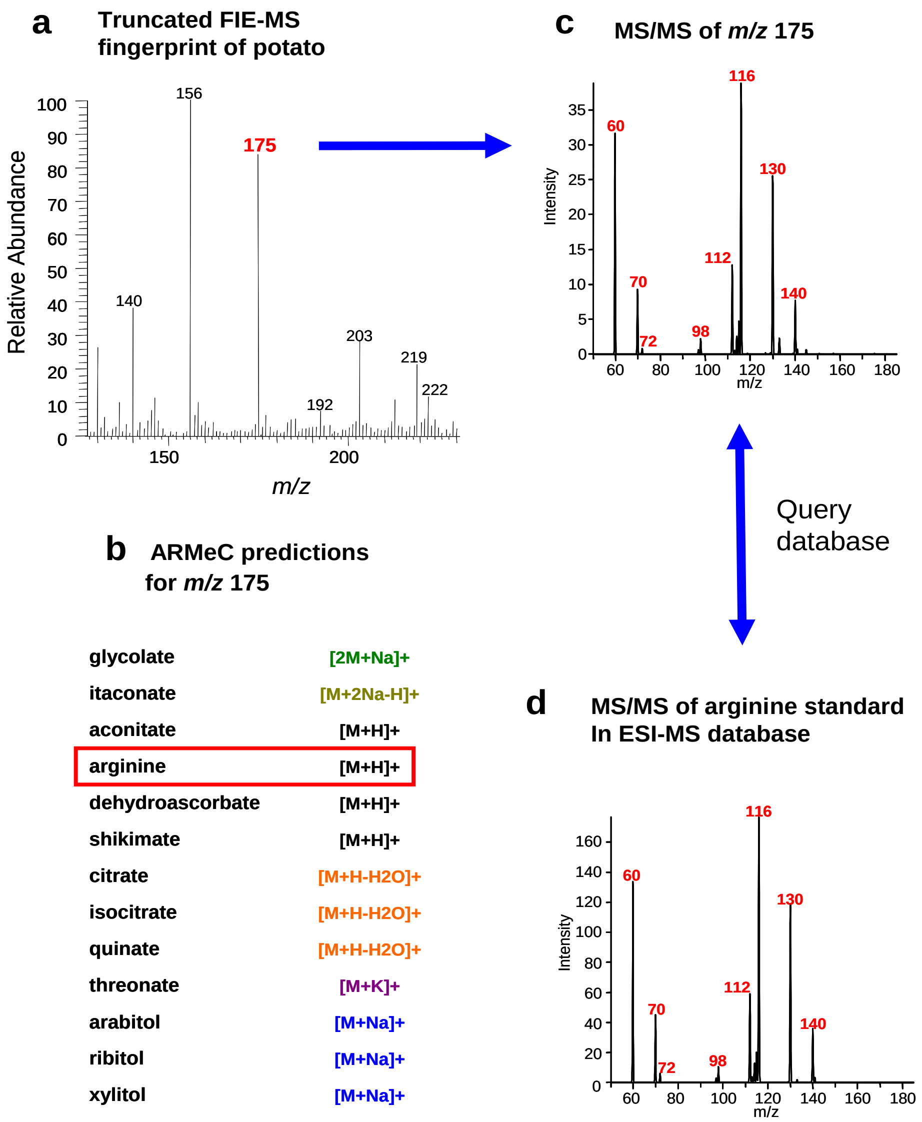
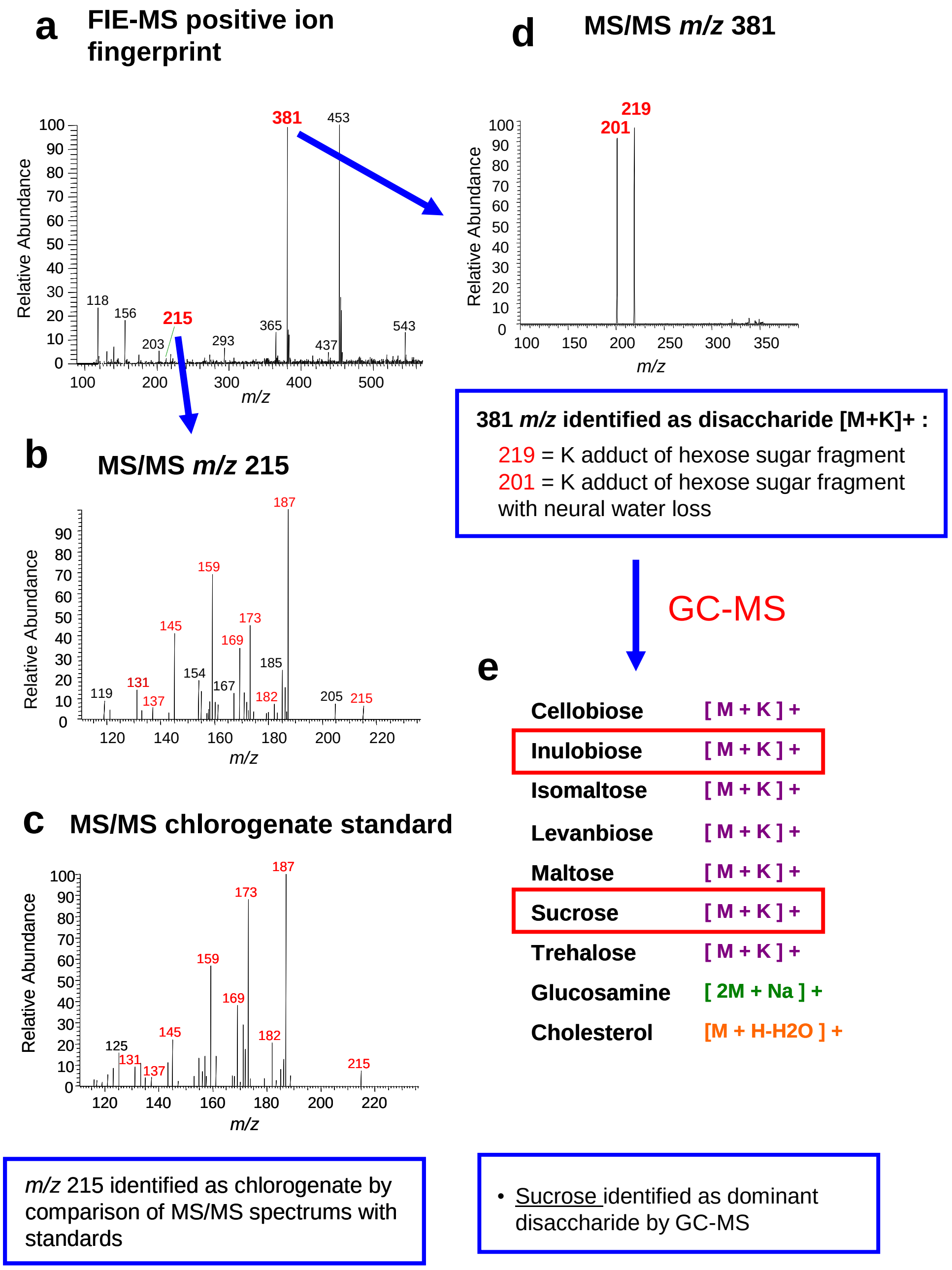


Figure 6.7: FIE-MS/MS^3 fragmentation trees of the potato tuber glycoalkaloids chaconine and solanine in positive ion mode showing sequential loss of sugar units. The chemical structures of both potato glycoalkaloids are presented and sub-unit structure summarized diagrammatically at the top of the figure (Sol. = Solanidine; Dgluc = D-glucose; Drham = D- rhamnose; Dgal = D-galactose). The loss of specific sugar units is illustrated at each stage of the fragmentation process. (a) FIE-MS fingerprint of crude potato extract with major ions representing chaconine (*m/z* 852) and solanine (*m/z* 868) labelled in red. FIE-MS/MS of parent ions, (b) *m/z* 852 and (c) m/z 868. (d and e) FIE-MS/MS^2 of parent ions *m/z* 706, and 722. f) FIE-MS/MS^3 of parent ion *m/z* 560 showing that both of the original metabolites had solanidine as the core alkaloid unit.



glycolate	[2M+Na]+
itaconate	[M+2Na-H]+
aconitate	[M+H]+
arginine	[M+H]+
dehydroascorbate	[M+H]+
shikimate	[M+H]+
citrate	[M+H-H2O]+
isocitrate	[M+H-H2O]+
quinate	[M+H-H2O]+
threonate	[M+K]+
arabitol	[M+Na]+
ribitol	[M+Na]+
xylitol	[M+Na]+

Figure 6.8: Putative annotation of an arginine [M+H]+ ion from an FIE-MS fingerprint of a crude potato extract by spectrum matching with known standards in the ARMeC database. (a) FIE-MS fingerprint of a crude potato extract with a signal at *m/z* 175 labelled in red. b) ARMeC predictions for the putative identity of *m/z* 175 in a potato tuber extract. FIE-MS/MS experiment of *m/z* 175 yielded spectrum c) which matched by database query to the ARMeC FIE-MS/MS spectrum d) of the arginine standard *m/z* 175 ([M+H]$^+$ ion).



381 *m/z* identified as disaccharide [M+K]+ :

219 = K adduct of hexose sugar fragment
201 = K adduct of hexose sugar fragment with neural water loss

Cellobiose	[M + K] +
Inulobiose	[M + K] +
Isomaltose	[M + K] +
Levanbiose	[M + K] +
Maltose	[M + K] +
Sucrose	[M + K] +
Trehalose	[M + K] +
Glucosamine	[2M + Na] +
Cholesterol	[M + H-H2O] +

m/z 215 identified as chlorogenate by comparison of MS/MS spectrums with standards

- Sucrose identified as dominant disaccharide by GC-MS

Figure 6.9: **Annotation of *m/z* signals in FIE-MS fingerprints after fragmentation and further chromatography to identify isomers.** (a) A high abundance signal (*m/z* 381 and a lower abundance signal (*m/z* 215) are shown labelled in red in a FIE-MS fingerprint of a crude extract from a potato tuber. FIE-MS/MS experimentation (b) and ARMeC database query (c) identified *m/z* 215 as chlorognic acid. FIE-MS/MS experimentation (d) and ARMeC database query (e) identified *m/z* 381 as a disaccharide $[M+K]^+$ ion. Subsequent GC-MS analysis revealed that *m/z* 381 was a combined representation of sucrose and inulobiose $[M+K]^+$.

After putative annotations have been assigned, further investigation by FIE-MS/MS^n queries is required to reduce the number of possible putative annotations (Figure 6.8c & d). As ion trees for particular adducts are unique to the particular structure of a molecule, confirmation of the putative metabolite identifications and associated adduct/neutral loss/dimeric states can be supported by FIE-MS/MS^n experiments (Figure 6.8). If the case arises where more than one dominant metabolite is contributing to the observed *m/z* signal, confirmation of metabolite identity can be achieved by subsequent FIE-MS/MS^n experiments on the extract spiked with the appropriate metabolite standard and monitoring for increased *m/z* ion intensity of targeted daughter ion fragments.

In the case where putative identifications are shared by molecules of different isomeric states, sharing the same molecular weight and structural similarity (*i.e.* saccharides), FIE-MS/MS^n experiments will only be able to prove the identification of the class of compound and its adduct/neutral loss/dimeric association. Figure 6.9 provides such an example. Here the annotation short list for *m/z* signal 381 Da is composed of several compounds, the majority of which are $[M+K]^+$ disaccharide ions. As the FIE-MS/MS^n fragmentation tree for each of these ions are identical due to similarity in structure and in turn the molecular weight of the resulting fragments, chromatographic profiling of this extract is required to confirm the identity of the disaccharide ion. In the case of Figure 6.9, GCMS profiling demonstrated sucrose to be the predominant disaccharide in the potato sample. However, GCMS results also confirmed that inulobiose was present as a minor component and therefore was also contributing to the observed 381 *m/z* signal.

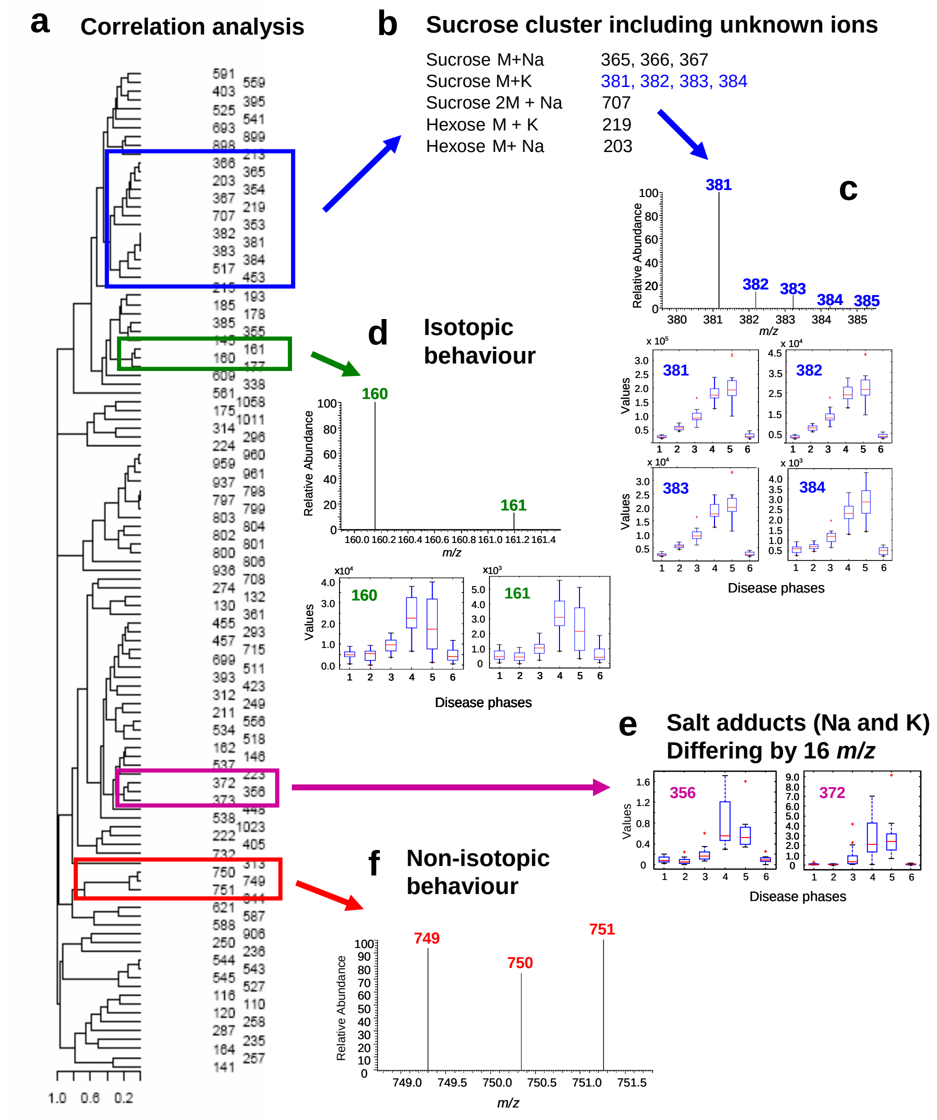


Figure 6.10: **Examination of correlation analysis results, isotopic ratios and intensity behavior to discover linked FIE-MS signals**. (a) Shows the results of a correlation analysis of *m/z* signals from positive ion mode FIE-MS with high explanatory power to discriminate between *Brachypodium* leaf tissues at different days (1-5) after infection with the rice blast fungus. The dendrogram on the left illustrates the hierarchical clustering of *m/z* signals (shown on the right) using the absolute value of signal Pearson correlation coefficient between signals as similarity matrix (scale at the base). Boxed in red are clusters of signals that display interpretable patterns. (b) Grouping of signals representing putative salt adducts isotopes, clusters and possible fragmentation products derived from sucrose. (c) The relative intensity of signals 381>382>383>384 suggest that all are isotopes of the same molecule which is reflected in the signal intensity levels at each day post-infection. Similar isotopic behaviour is shown for *m/z* 160 and 161 (d) and *m/z* 372 and 373 (e), among several other examples. The correlation of *m/z* signals (372 and 365) differing by 16 mass units (e) is indicative of Na and K adducts of the same molecule. (f) Non-isotopic behaviour of *m/z* 751 which is of greater intensity than *m/z* 750.

Identification of Isotopes by Correlation Analysis

Although *m/z* signals differing by one nominal mass unit can be present in top ranking, discriminatory RF lists, these signals are not necessarily isotopes of the same molecule. Correlation analysis performed upon RF lists is a rapid way of distinguishing possible isotopic relationships between explanatory *m/*z signals. Figure 6. 10a shows a correlation analysis of all *m/z* signals with explanatory power to distinguish between samples representing different phases of disease progression in rice blast infected *Brachypodium* leaves. Many signals clusters appear to contain an isotope series (for example in red boxes see 365, 366, 367; 381, 382, 383, 384; 160, 161; 372, 373; 749, 750, 751). With the exception of the last group, an examination of the relative abundance of ions in each series shows a large drop in signal intensity as the mass increases, reflecting the natural abundance of heavy isotopes (Figure 6.10c, d). In Figure 6.10f the *m/z* 751 has a higher intensity than both 749 and 750 and is thus unlikely to represent an isotope. In addition to isotope ratios the box plots in Figure 10c and 10d show that, as expected, each isotope has an identical signal abundance pattern in each sample class. Within the correlation analysis it is common to observe pairs of correlated signals in positive ion data differing by 16 mass units which represent Na and K adductions of the same metabolite (for example 365 and 381; 366 and 382, 367 and 383; 356 and 372; 543 and 527). Similarly, clusters can contain signals differing by 22 mass units that may represent protonated and Na addicts of the same metabolite (for example 534 and 556; 937 and 959). Similar isotope clusters are evident in negative ion data (Figure 11a and 11b). Due to natural variation in the isotopic pattern of various metal adducts, results from the correlation analysis can be used to distinguish adducts or metabolites containing Cl and S atoms. Figure 11c and Figure 11d show the correlation behaviour and isotopic relationships respectively of Cl adducts derived from chlorogenate in a potato extract.

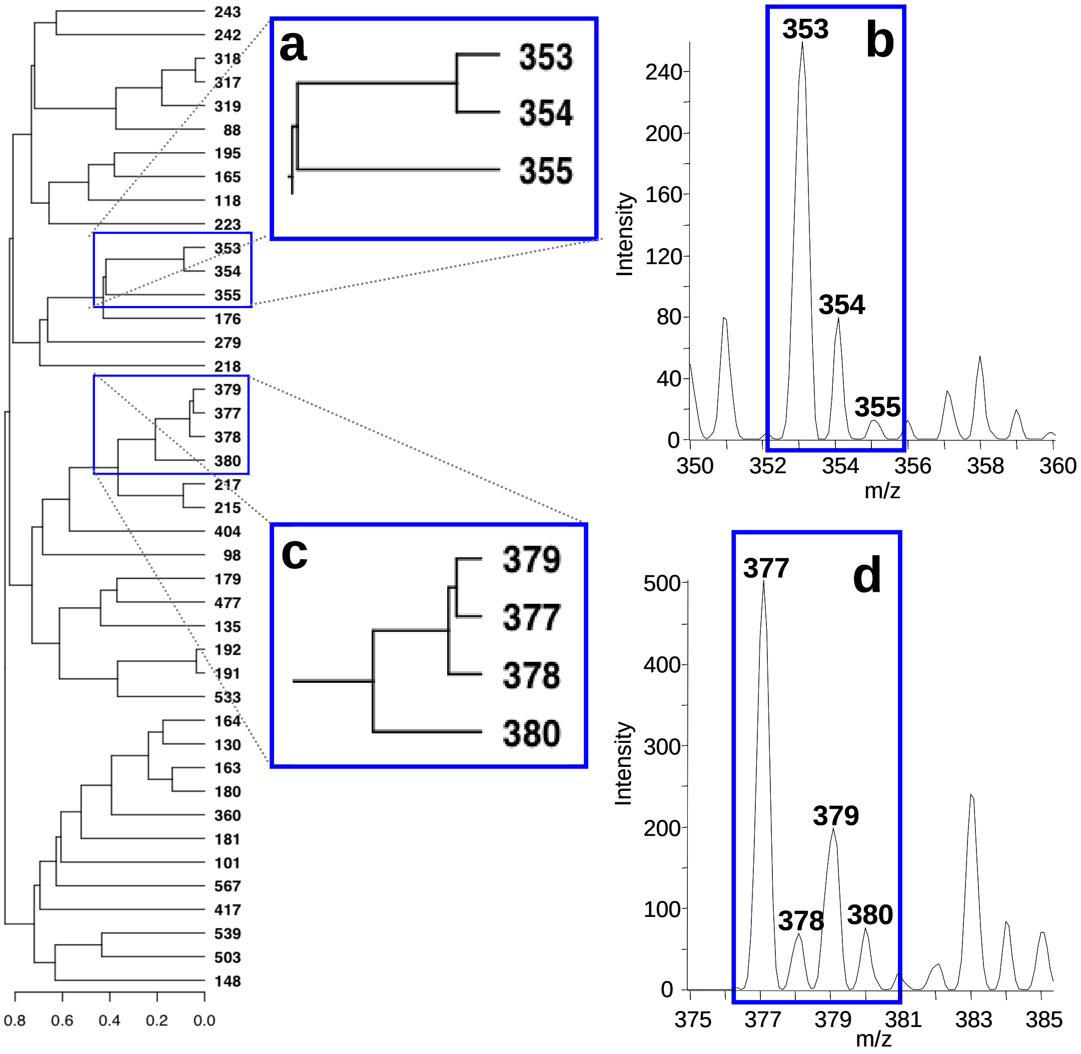


Figure 11: **Correlation analysis of explanatory variables discriminatory for quality traits in a comparison of five different potato cultivars in negative ion mode FIE-MS.** The dendrogram on the left illustrates the hierarchical clustering of *m/z* signals (on the right) using the absolute value of signal Pearson correlation coefficient between signals as similarity matrix (scale at the base). Highlighted in box (a) is an example of a typical isotopic *m/z* signal correlation as evidenced by the isotopic distribution pattern for $[M-H]^-$ ions in box (b) (*m/z* 353 = chlorogenate $[M-H]^-$). The ion cluster shown in box (c) represents salt adducts of sucrose with the isotopic distribution pattern for $[M+Cl]^-$ ions (*m/z* 377 = sucrose $[M+Cl]^-$) evident in box (d).

Putative Annotation of RF Discriminatory Ions

An important reality that should be recognized regarding nominal mass FIE-MS fingerprinting is that the fingerprinting technique is not directly measuring the level of any individual metabolite; instead it is providing a summation of ‘identical’ signals in the analytical system. When annotating FIE-MS fingerprints following multivariate data mining, especially when complex matrices are being evaluated, any particular *m/z* has the potential to relate to more than one metabolite as signals are binned (typically within a 1 Da window (see Section 5) when populating data matrices. This can have a direct impact upon explanatory signal interpretation and for this reason we query the database for all potential mathematical relationships (*i.e.* (de)protonated molecular ion, a salt adduct of either Na or K, a neutral loss of water, ammonium or formate, the homogeneous dimer ion or a dimer ion pair adduct of Na or K) within the list of explanatory variables (*m/*z) representative for a given signal interpretation. This process is repeated where the previous ion signal assumption is substituted for a subsequent interpretation representing the (de)protonated molecular ion, salt adduct, neutral loss or dimeric complex so that all possible combinations representative of the interpreted signal are queried against the RF list of explanatory variables. Output of the database queries and the mathematical determination of potential signal relationships can then be used to generate tables containing putative identifications of explanatory variables. Further investigation of signals highlighted in these tables by FIE-MS/MS^n and additional chromatographic experiments can be used to confirm the putative database annotations. Validation of these signal annotation protocols has lead to the determination of metabolome changes in transgenic experiments involving *Solanum tuberosum* and *Arabidopsis thaliana* mutants [3, 7]. In the context of the G03 programme, this protocol has also been used successfully to putatively annotate potato cultivar traits associated with food quality characteristics as described in Section 9 and published in Beckmann *et al*, 2007[19].

REFERENCES

1. Rashed, M. S. Clinical applications of tandem mass spectrometry: ten years of diagnosis and screening for inherited metabolic diseases. *J. Chromatogr. B* **758**, 27-48 (2001).

2. Wilcken, B., Wiley, V., Hammond, J. & Carpenter, K. Screening newborns for inborn errors of metabolism by tandem mass spectrometry. *N. Engl. J. Med.* **348**, 2304-2312 (2003).

3. Catchpole, G. S. *et al.* Hierarchical metabolomics demonstrates substantial compositional similarity between genetically modified and conventional potato crops. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 14458-14462 (2005).

4. Ramos Catharino, R. *et al.* Characterization of vegetable oils by electrospray ionization mass spectrometry fingerprinting: classification, quality, adulteration, and aging. *Anal. Chem.* **77**, 7429-7433 (2005).

5. Allen, J. *et al.* High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nat. Biotechnol.* **21**, 692-696 (2003).

6. Smedsgaard, J. & Frisvad, J. C. Terverticillate Penicillia studied by direct electrospray mass spectrometric profiling of crude extracts. I. Chemosystematics. *Biochem. Syst. Ecol.* **25**, 51-64 (1997).

7. Enot, D. P., Beckmann, M., Overy, D. P. & Draper, J. Predicting interpretability of metabolome models based on behavior, putative identity, and biological relevance of explanatory signals. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 14865-14870 (2006).

8. Beckmann, M., Parker, D., Enot, D.P., Duval, E. and Draper, J. High throughput, non-targeted metabolite fingerprinting using nominal mass Flow Injection Electrospray Mass Spectrometry. *Nature Protocols* **3,** 486-504 (2008)

9. Dunn, W. B., Overy, S. & Quick, W. P. Evaluation of automated electrospray-TOF mass spectrometry for metabolic fingerprinting of the plant metabolome. *Metabolomics* **1**, 137-148 (2005).

10. Gray, G. R. & Heath, D. A global reorganization of the metabolome in *Arabidopsis* during cold acclimation is revealed by metabolic fingerprinting. *Physiol. Plant.* **124**, 236-248 (2005).

11. Cole, R. B. Some tenets pertaining to electrospray ionization mass spectrometry. *J. Mass Spectrom.* **35**, 763-772 (2000).

12. Fernie, A. R., Trethewey, R. N., Krotzky, A. J. & Willmitzer, L. Metabolite profiling: from diagnostics to systems biology. *Nat. Rev. Mol. Cell Biol.* **5**, 763-769 (2004).

13. Fiehn, O. Metabolomics - the link between genotypes and phenotypes. *Plant Mol. Biol.* **48**, 155-171 (2002).

14. Trethewey, R. N. Metabolite profiling as an aid to metabolic engineering in plants. *Curr. Opin. Plant Biol.* **7**, 196-201 (2004).

15. Schug, K. & McNair, H. M. Adduct formation in electrospray ionization. Part 1: Common acidic pharmaceuticals. *J. Sep. Sci.* **25**, 760-766 (2002).

16. Kebarle, P. A brief overview of the present status of the mechanisms involved in electrospray mass spectrometry. *J. Mass Spectrom.* **35**, 804-817 (2000).

17. Zhu, J. & Cole, R. B. Formation and decompositions of chloride adduct ions, $[M+Cl]^-$, in negative ion electrospray ionization mass spectrometry. *J. Am. Soc. Mass Spectrom.* **11**, 932-941 (2000).

18. Enot, D., Lin, W., Beckmann, M., Parker, D., Overy, D. & Draper, J. Pre-processing, classification modelling and feature selection using Flow Injection Electrospray Mass Spectrometry (FIE-MS) metabolite fingerprint data. *Nature Protocols* **3,** 446-470 (2008).

19. Beckmann, M., Enot, D. P., Overy, D & Draper, J. (2007). Representation, comparison and interpretation of metabolome fingerprint data for total composition analysis and quality trait investigation in potato plants. *J. Agric. Food Chem.* **55**, 3444-3451 (2007).

20. Kaneshisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27-30 (2000).

21. Garcia-Hernandez, M. *et al.* TAIR: a resource for integrated *Arabidopsis* data. *Funct. Integr. Genomics* **2**, 239-253 (2002).

22. Shinbo, Y. *et al.* KNApSAcK: A comprehensive species-metabolite relationship database. *Biotechnol. Agric. Forest.* **57**, 165-184 (2006).

23. Wishart, D. *et al.* HMDB: the human metabolome database. *Nucleic Acids Res.* 35, D521-D526 (2007).

24. Wagner, A. B. SciFinder Scholar 2006: an empirical analysis of research topic query processing. *J. Chem. Inf. Model.* **46**, 767-774 (2006).

25. Kind, T. & Fiehn, O. Metabolomic database annotations *via* query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics* **7**, 234-243 (2006).

26. Baumann, C. *et al.* A library of atmospheric pressure ionization daughter ion mass spectra based on wideband excitation in an ion trap mass spectrometer. *Rapid Commun. Mass Spectrom.* **14**, 349-356 (2000).

27. Fredenhagen, A., Derrien, C. & Gassmann, E. An MS/MS library on an ion-trap instrument for efficient dereplication of natural products. Different fragmentation patterns for $[M+H]^+$ and $[M+Na]^+$ ions. *J. Nat. Prod.* **68**, 385-391 (2005).

28. Milman, B. L. Towards a full reference library of MS^n spectra. Testing of a library containing 3126 MS^2 spectra of 1743 compounds. *Rapid Comm. Mass Spectrom.* **19**, 2833-2839 (2005).

29. Pavlic, M., Libiseller, K. & Oberacher, H. Combined use of ESI-QqTOF-MS and ESI-QqTOF-MS/MS with mass-spectral library search for qualitative analysis of drugs. *Anal. Bioanal. Chem.* **386**, 69-82 (2006).

30. Parker, D., Beckmann, M., Enot, D., Overy, D., Rios, Z.C., Gilbert, M., Talbot, N. & Draper, J. Rice blast infection of *Brachypodium distachyon* as a model system to study dynamic host/pathogen interactions. *Nature Protocols* **3,** 435- 445 (2008).

31. Broadhurst, D.I. & Kell, D.B. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* **2**, 171-196 (2006).

32. Lisec, J., Schauer, N., Kopka, J., Willmitzer, L. & Fernie, A. R. Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nat. Protocols* **1**, 387-396 (2006).

33. Connolly, T. & Begg, C. *Database Systems: A practical approach to design, implementation and management.* Addison-Wesley, (2002).

34. Tolstikov, V.V., Lommen, A., Nakanishi, K., Tanaka, N. & Fiehn, O. Monolithic silica-based capillary reversed-phase liquid chromatography/electrospray mass spectrometry for plant metabolomics. *Analytical Chemistry* **75**, 6737-6740 (2003).

35. Dear, G.J., James, A.D. & Sarda, S. Ultra-performance liquid chromatography coupled to linear ion trap mass spectrometry for the identification of drug metabolites in biological samples. *Rapid Communications in Mass Spectrometry* **20**, 1351-1360 (2006).

36. R_Development_Core_Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-900007-900050, URL http://www.R-project.org (2006).

37. Gorlach, E. & Richmond, R. Discovery of quasi-molecular ions in electrospray spectra by automated searching for simultaneous adduct mass differences. *Anal. Chem.* **71**, 5557-5562 (1999).

38. Nielsen, K. F. & Smedsgaard, J. Fungal metabolite screening: database of 474 mycotoxins and fungal metabolites for dereplication by standardised liquid chromatography-UV-mass spectrometry methodology. *J. Chromatogr. A* **1002**, 111-136 (2003).

G03012 Final Report - Section 7

Matthieu Vignes/Jim McNicol

Dealing with missing values in metabolomics data and analysis of ranges of metabolite values and development of statistical models of metabolite distributions in conventional crops (Tasks 03.03 & 04.01)

7.1 BACKGROUND TO EXPERIMENTAL SECTION

A major issue when trying to compare the composition of any food raw material is that information from diverse external sources is rarely directly comparable. A major conclusion of Section 2 of the GO3012 Report was that few databases or publications provided the same coverage of metabolites and, where there is overlap, the concentration measures vary; in most instances the data could not be transformed easily into a common format, thus any comparative univariate analysis or correction of missing values was more or less meaningless. Partly for this reason it was not deemed appropriate to spend time developing a new implementation of ArMet to hold this external data and instead research focused on the further development of the pilot database describing information on potato tuber chemical content (ARMeC: see Section 6 of the GO3012 Report) to now include provision of links to external data describing measurements on specific metabolites. Univariate analysis, including the exploration of data ‘infilling’ procedures where there are missing values was therefore limited to the 4 large G02 GC-MS data sets (see Section 1 of G03012 Report) whose structure are summarised below.

7.2 OVERALL OBJECTIVES

In the present section we introduce the univariate and non-supervised multivariate analyses focused on, including specifically a discussion on methods to deal with missing values and then go on to describe the different sort of conclusions we were able to reach. Univariate analysis of large metabolomics datasets had clearly been an element in all of the G02 project laboratories contributing data to be analysed in the current G03012 project. Rather than repeat any of these studies, a major objective was to determine what kind of univariate approaches had value in the description of the overall trends of metabolite signal behaviour in the four major GC-MS datasets. From this study we can start to make some suggestions of what kind of distribution characteristics generally measurable metabolites overall or indeed specific metabolites might be expected to show in potato tuber extracts. From this description any unusual deviations in the future might be highlighted in new cultivars or GM varieties. As part of this study an integral task was to develop statistically robust peak table in-filling strategies to deal with missing values and data transformation methods to normalise data produced on instruments with different sensitivities. Following this study a validated peak table infilling routine was selected and implemented in the R software package developed in the G03012 project (see Section 8).

7.2.1 Summary of objectives:

- Analyse feature signal behaviour in GC-MS data and determine best fit distribution pattern.
- Compare peak table feature content, intensity distributions and effects of experimental factors on non-supervised multivariate behaviour.
- Analysis of effects of data transformation on univariate behaviour.
- Analyse missing data distributions.
- Determine effect on data infilling strategies on peak intensity distributions.

7.3 INTRODUCTION AND DESCRIPTION OF DATA SETS

The basic aim of the project was to develop procedures to describe and compare the natural variability of composition encountered across raw food materials derived both within and between individual genotypes of the same species. By examining the univariate distribution of the variables (metabolites) we hope to determine whether an appropriate distribution can be found, i.e. a distribution which describes well the variation in metabolite response. This approach will provide, via the distribution parameters, a succinct method of characterising all aspects of the metabolite behaviour for a single environment. Another goal was to detect any change in response to different environmental or experimental conditions where these are evident as ‘factors’ (e.g. field plot or harvest year or instrument operator) in the meta-data.

Using distributions, such changes are indicated by differences in the distribution parameter values which may or may not effect all metabolites in the same way. This being said it is apparent from Section 1 of the G03012 Report that none of the G02 datasets represented the same cultivars grown in different geographical regions and so this element of the project will be limited. Conventional statistical methods for the study of metabolite content define ranges of values e.g. mean+/-standard error or mean+/-3*standard error. As a simpler option, we might also consider minimum to maximum intervals; fitted distributions provide a comprehensive description rather than a summary. Missing values in metabolomics, just like in many high-dimensional datasets regardless of the field, is a common feature in the high-throughput methods, not to mention yet the merging of data from different sources. As a consequence, applying most of statistical tools designed for complete datasets is far from straightforward.

Missing values can have a strong effect on the outcome of the analysis. Simply discarding samples – *case deletion* in the statistical jargon – with (too many) missing values could exclude a too high percentage of the samples, depriving the analysis of a substantial amount of meaningful data and lowering the power of any test carried out. Accounting for 'missingness' specifically increases the level of complexity of the analysis. It is probably a necessary burden as it is now recognized that over-simple approaches can lead to spurious results. On the other hand, patterns of missing values are difficult to identify reliably and computational aspects can be an issue (possibly why this feature is not available yet in many statistical analysis packages!). As recommended by several authors (e.g. Heitjan 1997), we will first focus on describing missingness in our data: its distribution and its origin. Theoretically, when the dependency between variables and the missingness mechanism is known, including them in a model allows an easier and more reliable analysis. In practice, we will see that it is extremely difficult to decipher these links and to correctly integrate such knowledge.

7.3.1 Data Sets

Four data sets were used to address the questions in this section: ABER 2001, SCRI 2002, GOLM2001 and GOLM2003. Briefly, all four data sets are based on GC-MS spectra and hence describe volatile/semi-volatile compounds. Entering a vacuum region after the separation column, compounds are bombarded by high energy electrons that cause metabolites to fragment. Fragmentation patterns are dependent on the structure of the compound. Fits with a library of known compounds allow the identification of the actual detected compounds. Many compounds remain unknown because no match in the library was found; either the actual compound is not in the library or the experimental pattern does not correspond to the one(s) stored in the library (see G03012 Sections 3 & 4 for description of peak detection and annotation).

ABER 2001

These data are from a field trial, originally with 6 cultivars and 6 GM lines laid out in 4 blocks with a randomized complete block design. Only data pertaining to the 6 cultivars were used here. From each of the 6x4 target plots 48 tubers were selected and analyzed (note that the number of repeated measurements can vary). Day of injection was recorded. Each spectrum had 85 measured peaks of which 37 (43.5%) were 'known' (*i.e.* are assigned names that can match a precise chemical compounds *e.g.* "Alanine", "L-lysine_6.533" or "serine-major" whereas "34U-106", "5.777", "cho35" or "y_amine" aren't). Of the 6x4x48x85 data values less than 1% were missing.

SCRI 2002

These data are from a field trial with 23 cultivars with a randomized design. From each of the 23 target plots, 4 tubers were selected and analyzed (7 of them have no corresponding measures because of instrument malfunction and were consequently deleted from the analysis). Day of injection was recorded. Each spectrum has 160 measured peaks of which 36 (23%) are 'known'. 3.5% (10.7% if empty rows are not omitted) of the data values were missing. Other information (e.g. day samples frozen) are also available for quality control.

GOLM2001 and GOLM2003

GOLM2001 data and GOLM2003 data are two repeats of an experiment which was derived from the same series of field trials providing the material for the ABER 2001 project. Theses data are derived from very similar experimental structures originally from a field trial, with 6 cultivars and 6 GM lines laid out in 4 blocks with a randomized complete block design. 48

tubers were selected and analyzed by 2 operators (randomized towards block and line only in the GOLM2003 data set). Again, we only used data describing the 6 cultivars. Day of injection was recorded, only being randomized towards other factors in the GOLM2003 data set. Each spectrum has respectively 252 and 245 measured peaks in GOLM2001 and GOLM2003 data sets. Among these, respectively 81 (just above 31%) and 109 (44.5%) are ‘known’. The proportion of missing values is higher in these data sets and reaches 17 and 20% respectively.

7.4 UNIVARIATE ANALYSIS OF G02 DATA

7.4.1 Data characteristics, structure and transformation

A first step in data analysis is to produce, from the gross machine measurements a matrix expressing relative metabolite levels for the sample under consideration. This subject has been largely debated in the data pre-processing section (Section 8) of the present research project. It is assumed that a standard quality control and data-pre-processing strategy will have already been implemented for example to highlight regular patterns of missing data, sample outliers and any lack of measurement consistency between similar sample classes or detection limit issues. Hence we end up with a matrix that provides for each sample a profile over the different variables (intensity measures of detected peaks in our case). Samples are represented by rows and variables (metabolites) in columns. As determined in Section 9 of the G03012 Report, all response values were subject to a log-transformation before any other transformation or normalisation procedure. This removes skewness and follows the common convention (see van den Berg et al. 2006 and Steinfath et al. 2008 for discussions on data transformation but also on data scaling). As we have demonstrated this treatment achieves optimum classification results and more stable feature selection.

Figure 7.1 (left) gives an example profile (Asparagine in ABER2001 data set) across samples which have been sorted by order of processing. It is clear that comparison between samples can be an issue. There is clearly an effect due to day of injection and ideally this should not be present as an observed trend in the data. One way of removing such *sample effects* is to express each compound as a fraction of the sum of all compounds for that sample. Although not perfect, this normalisation created better data in terms of trends (see Figure 7.1 right; however one could argue that visible trends have only been diminished) and in terms of shape of the distribution (see next section). The effect on the same metabolite distribution across cultivars can be seen in Figure 7.2. While raw data could have led to the conclusion that cultivar Linda might be richer in Asparagine, the transformed and normalised data make it clear that no such effect exists.

Greater difficulty arises when the comparison is between measures made on different instruments. We decided to use the same transformation/scaling procedure approach. Standard procedures to provide a principled comparison of raw food material are needed. We focus here on untargeted experiments (i.e. no chemical standard is included in sample to analyse). A drawback of this process is that absolute values are not obtainable. An advantage however is that a greater variety of compounds (all resolvable peaks whose annotation is far from trivial) that need to be characterized can be measured. Since our aim is to observe the biological differences in raw food materials, we want to eliminate instrument, laboratory, operator and any other effects that can interfere with differences in raw data patterns.

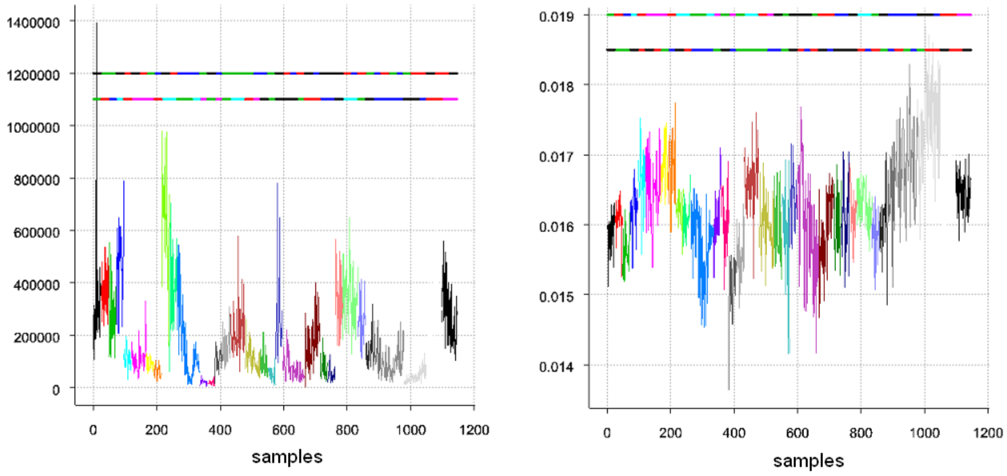


Figure 7.1: Raw asparagine peak across samples (left) and log-transformed and normalized data (right) in ABER2001 data set. The colours represent day of injection and the 2 lines at the top of the graphs indicate cultivar and field block of the sample.

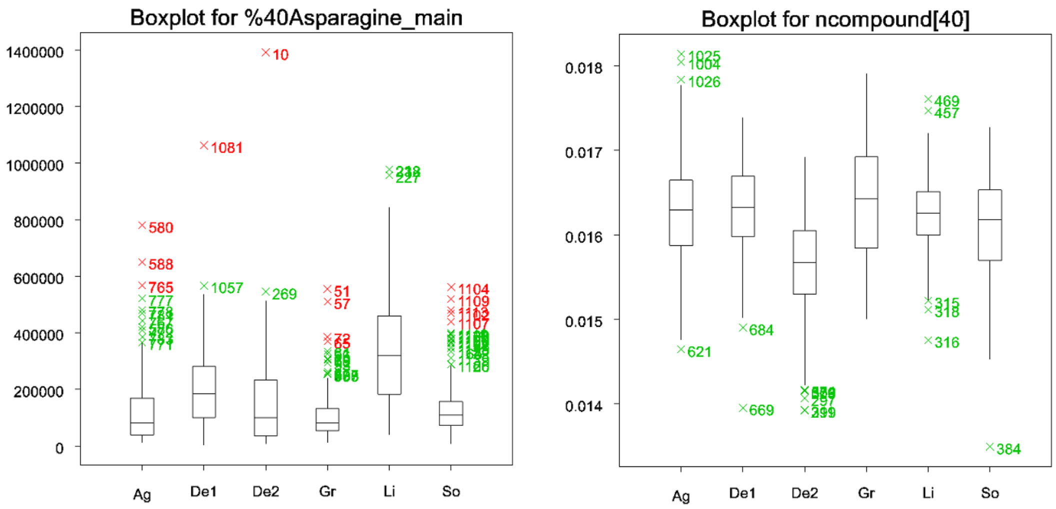


Figure 7.2: Box plot of asparagine value distribution according to the cultivar factor in raw data (left) and log-transformed and normalized data (right) in ABER2001 data set.

Several other features of the data need to be taken into account: high dimensionality, biological treatments and instrument variations. The FSA GO2006 project led to the production of powerful machine learning data analysis tools for fingerprinting or profile classification/discrimination that can lead to standardised similarity criteria for crop raw material comparison based on metabolomics data. Before going onto multivariate analysis, we focused on studying individual metabolite distributions in the 4 data sets described above.

7.4.2 Fitting a distribution to individual metabolites

The main subject of Milestone 4.1 of the original proposal was to assess the univariate distribution of common metabolites. First, we quickly scanned through histograms and smoothed histograms (kernel density estimates) of compounds to get an initial assessment of how well a ‘Normal’ distribution describes metabolite variation. To quantitatively assess how well the data approximate a particular distribution, we used a plot of the sorted values against the expected values of the ordered statistics for the theoretical distribution. A perfect fit would result in a straight line. Traditionally, such representations include Quantile-Quantile (Q-Q) plot or Probability-Probability (P-P) plots. We preferred to use a more refined stabilized probability plot (Michael 1983) that stabilizes the variance of the plotted points by an arcsine transformation. In combination with a powerful goodness-of-fit statistic (maximum deviation hence analogous to the statistic used in a Kolmogorov-Smirnov test), we can construct acceptance regions for the compound distributions (Figures 7.3, 7.4 and 7.5 show examples with 99% confidence bands). If every point lies within the defined band, the null hypothesis that data points are distributed according to the theoretical distribution is accepted. We investigated other distributions (log-normal, gamma and exponential) but none performed better than the Normal distribution.

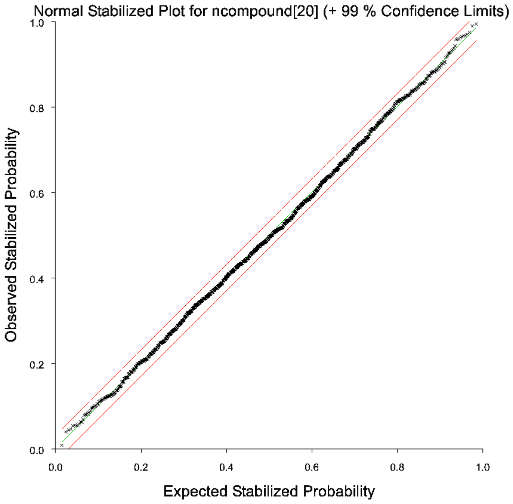


Figure 7.3: Example of metabolite (Fumaric Acid) for which the test output concludes as being Gaussian (ABER2001 data).

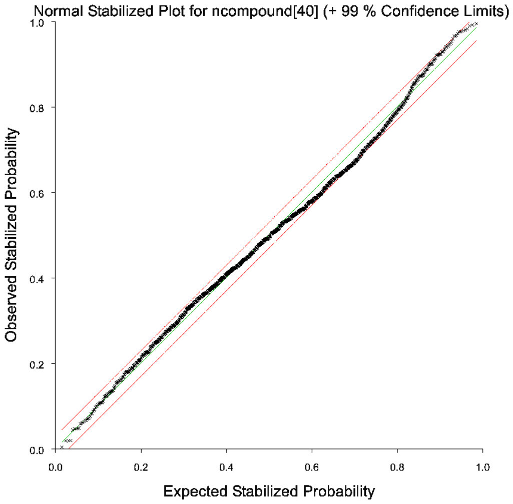


Figure 7.4: Limit example where normality can be accepted for metabolite (Asparagine) distribution, even though fit is not perfect (ABER2001 data).

We then tested all metabolite distributions for normality. Three different scenarios were encountered. Firstly, an excellent fit of the data to a Gaussian distribution (see Figure 7.3). For example, we can conclude that log-normalised fumaric acid has a normal distribution at a 99% significance level. The second situation arises in the case where data are clearly not Gaussian. This is the case for the sucrose peak of Figure 7.5; this very abundant metabolite is an example of when under-estimates of peak concentration are possible due to detector saturation. The intermediate situation also occurred. For example, from Figure 7.4, we should conclude that asparagine distribution is not Normal as a slight effect of skweness is observed. Outliers might also be observed in distributions' tail(s). When these effects are not too important (as it is the case in Figure 7.4), we will accept the approximate Normal distribution for such variables.

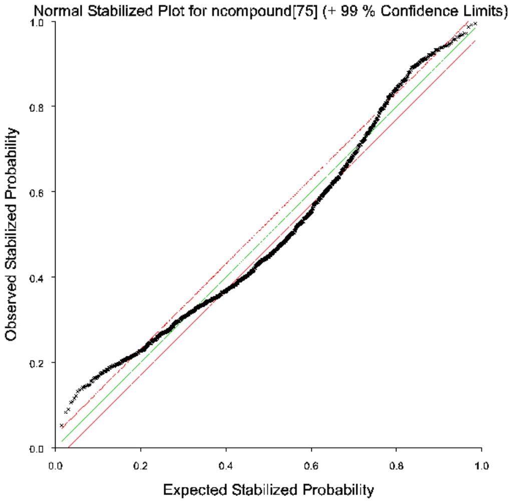


Figure 7.5: Example of metabolite (Sucrose) distribution that is not Gaussian (ABER2001 data).

Tables 7.1 and 7.2 summarize our assessment of normality of all metabolites.

	Normal	Non normal
SCRI2002	129	31
GOLM2001	75	102
GOLM2003	78	80
ABER2001	52	33

Table 7.1: Number of metabolites whose distribution is identified as normal (after log-transformation and normalization) in the 4 data sets, using a level of significance of 99%.

Table 7.2 shows that for the ABER2001 data there is not a strong dependency between a compound being known and whether it is normally distributed (the p-value for positive association is 0.35 according to a Fisher test on proportions). This observation would tend to suggest that peaks have been carefully manually aligned with little evidence of false positive or false negative values. We scanned every metabolite distribution in the 4 data sets and essentially (with one or two exceptions such as sucrose) there is not much more that can be said about metabolites that do not have a Gaussian distribution. Based on this observation we suggest that measures for deciding whether a sample (e.g. of a given cultivar) has a typical or extreme content in one particular compound is only really valid when it is reported to be normally distributed. In fact, we can use the central and dispersion values reported in our in-house compositional database (see report on crop compositional data collected from the literature or available food databases in Section 2) to give good estimates of the distribution parameters (a central value and a value reflecting the spread of the distribution). In Section 2 we reported the difficulty of aligning quantitative data relevant to food composition data and chose to input derived qualitative concentration estimates in the ARMeC database which was only intended to be developed as a pilot and cannot be used to assess metabolite distributions in new data. In the future we hope that ARMeC (Aberystwyth Repository of Metabolite Characteristics) can be made comprehensive: values in it need to be (i) filled in more comprehensively (nearly 300 compounds were reported as present at least once in literature describing potato tuber but only about 100 of them have reliable numerical information available); (ii) the literature sources quoted in ARMeC need to be checked more thoroughly to make sure that the data reflect the average crop and the data converted (if at all possible) to represent the same sample extraction and metabolite quantitation descriptors. We have already reported that there is room here for further development but that this cannot be achieved in this project given the considerable needed amount of time needed from expert potato biochemists. The implemented ARMeC will permit the update and future development of the database. In the interim we need to assume in the following sections that we would have such values at our disposal if metabolomics data is used to compare new potato genotypes to the composition of previously described potato tuber raw materials.

	Normal	Non normal
Identified	24	13
Not identified	28	20

Table 7.2: Number of metabolites whose distribution is identified as normal (level of significance 99%, after log-transformation and normalization) according to the metabolite chemical identification in ABER2001 data set.

At the present time we would be able to tell whether the value of a given compound in a new sample (from a cultivar to test or a GM variety) is 'within range' or seems to be extreme. It is a different question whether a value is typical or not for a given distribution. An important underlying assumption here is that the data we use reflects correctly the biological usual background of the crop under study. This could be checked by plotting values from metabolomics data vs. values from compositional data in order to confirm that metabolomics data are correlated to the actual content of the crop. Again, this is not feasible yet because of complexity of the external literature. The major drawback of the solution proposed here is that it focuses on metabolites individually. Considering distributions of several variables and their joint distribution in deciding whether a given sample differs from another or from predefined baseline behaviour can lead to quite different conclusions. We propose to go more deeply into this question in the section to follow in which we use a non-supervised multivariate method to visualise the similarity of samples representing the same or different cultivars.

7.4.3 Representing sample differences and similarities by non-supervised multivariate analysis

The challenge here is to make the best use of all compound distributions simultaneously. We propose Principal Coordinate Analysis (PCO) to perform this assessment, an approach based on similarities between samples. PCO is a simple case of multidimensional scaling. The basic idea of multidimensional scaling is to take the matrix of all pairwise samples distances and to seek the best representation of the samples in *d* dimensions. Ideally, *d* would be 2 allowing simple scatter plot graphs of the samples. An important point is to determine how many PCO dimensions should be used. If *d* is chosen too small, there may be some important information in the data not included in the results. On the other hand, if it is chosen too large relative to the number of samples, misleading interpretations can arise (Anderson and Willis 2003). We started with *d* =4 and looked at the results for *d* =7 without noticing additional interesting features.

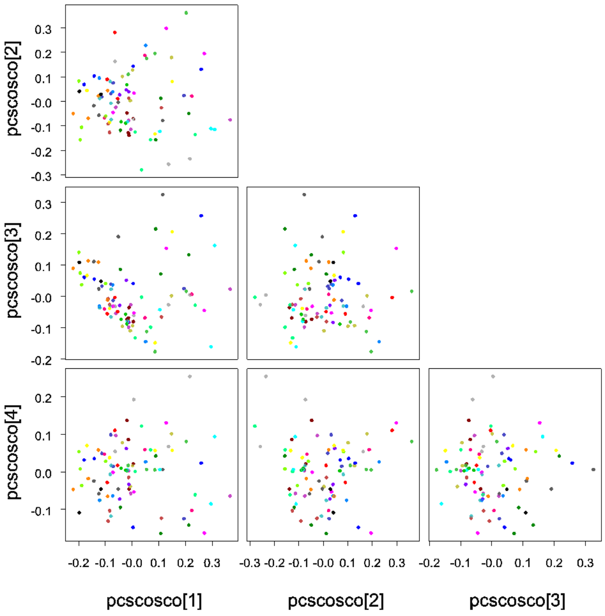


Figure 7.6: Plot of PCO scores for SCRI2002 samples. Colours stand for the 23 different cultivars. Percentage of distance accounted for by the first 4 dimensions are 21.5, 18.8, 11.6 and 8.

Any new sample (e.g. new cultivar, GM line) can be added to the PCO maps of points created from a set of samples considered to represent appropriate genetic background variation. This is a more technical procedure than adding a new variable in PCA but procedures have been developed to solve the problem (Wilkinson 1970).

We present in Figures 7.6 to 7.9 results obtained from our 4 target data sets. One goal here was to identify the most important environmental factor(s) affecting the distribution of the data. The situation in the SCRI 2002 data shown in Figure 7.6 is somewhat peculiar. It includes only 92 samples spread across 23 levels for the factor “line” (4 replicates per cultivar only) and so it would be very difficult to see any groupings. We are indeed interested in groupings because PCO score plots reflect a distance between samples. Firstly we want to check that samples that are believed to be similar can be distinguished from less similar ones on the plots. Secondly, we would want to plot a new sample in the initial PCO plot to check whether it is similar according to the distance of known samples or if there is some distinctive feature in it. Such assessment provides a useful visualization of compositional differences between samples of previously analysed or novel genotypes. To use effectively in the future appropriate genotypes to represent background variation in a gene pool would need to be selected using information on UK potatoes (http://varieties.potato.org.uk) or in the European database (http://europotato.org).

In Figure 7.7, there is a well defined cloud of points covering all samples. PCO did not clearly separate samples according to the line. However, one can spot some separation among some sets gathering different cultivars (see PCO4 vs PCO1, PCO2 and PCO3 for instance).

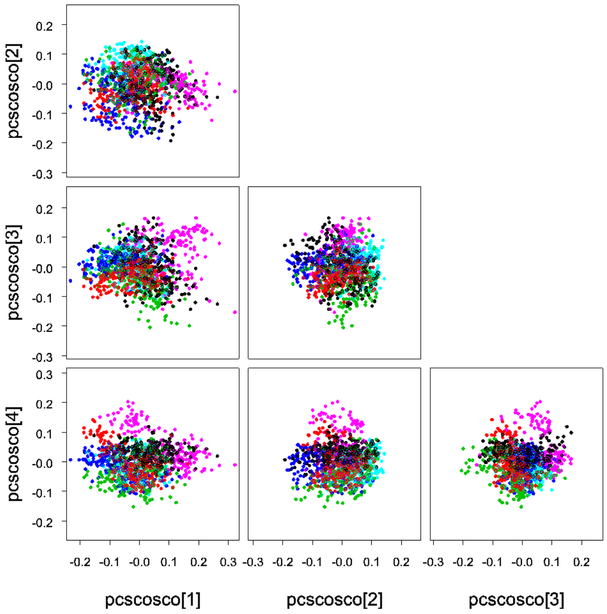


Figure 7.7: PCO scores for ABER2001 sample. Colours represent different cultivars. Percentage distances accounted for by dimensions 1, 2, 3 and 4 are: 16.1, 9.8, 8.7 and 6.8.

The GOLM2001 PCO analysis (Figure 7.8) is surprising. Two clearly separated groups, unrelated to field block or cultivar are evident in PCO. Further inquiry (Figure 7.8 (C)) revealed that this separation was mainly due to the factor relating to the operator handling the sample. A problem of randomization of the sample processing is also present is this data set in which no randomization among blocks, lines, operators and days of injection was incorporated. This example shows that PCO plots can reveal the influence of factors which are part of the experimental procedure as apposed to factors whose effect we wish to estimate (‘treatments’). In contrast the PCO analysis of GOLM2003 data (Figure 7.9) reveals only cultivar differences, demonstrating clearly the improvement of experimental procedures during the G02006 project.

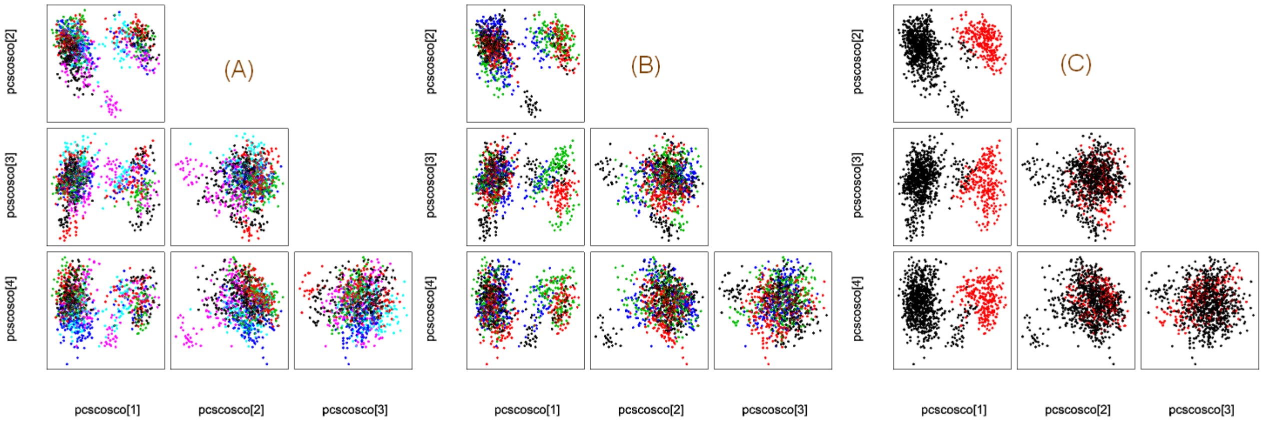


Figure 7.8: PCO scores for GOLM2001 samples. Colours stand for (A) field block (4), (B) cultivar (6) and (C) operator (2). Percentage distance accounted for by dimensions 1, 2, 3 and 4: 19.5, 9.4, 5.8 and 5.4.

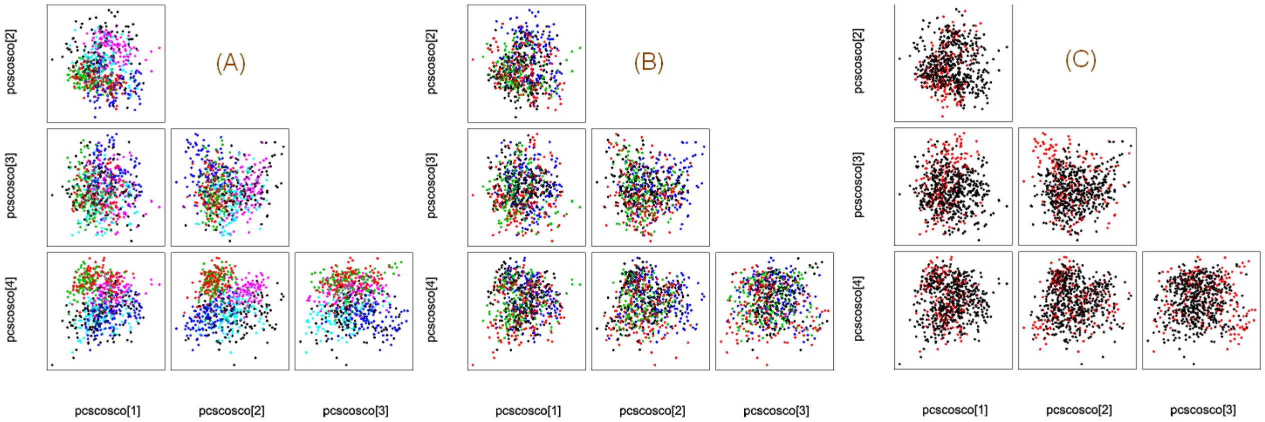


Figure 7.9: PCO scores for GOLM2003 samples. Colours stand for (A) cultivar (6), (B) field block (4) and (C) operator (2); percentage distance accounted for by dimension 1, 2, 3 and 4: 12.6, 11.1, 8.4 and 7.2.

The conclusion so far is somehow mitigated; apart from the GOLM2001 data (see discussion above), PCO score plots consist of a single cloud of points. Samples expected to be different are not really differentiated from the overall background variation typical of the potato tube matrix, although *cultivar* seems here to be the determining factor that structures the clouds of points. PCO currently seems to be a promising tool for checking similarities between data from samples run through GC-MS but the results presented here on real data might be seen as a bit disappointing. However, it is important to realise that the two extremes of experimental replication typified by the SCRI 2002 data (4 biological replicates) and Golm/Aber data (192 replicates from 4 field blocks) have different effects on variance characteristics of the data. The former does not offer sufficient replicates for accurate PCO visualisation whereas the latter suffers from experimental and environmental factors that compound variance, making it difficult to see distinctive clusters of signals representing a single cultivar. In the future it is expected that adoption of a pilot study to optimise a replication strategy (e.g. in Section 9 of the G03012 Report it was concluded that around 20 replicates will be optimal to classify potato cultivars) for a new sample matrix will overcome this problem and allow the use of PCO as a quick visualisation tool to compare metabolite distributions in new samples.

7.4.4 Conclusions on feature variance in G02 data

We conclude that the intrinsic dependencies in the data hinder their deciphering and conceal real sources of variation in these data sets produced several years ago in the G02 programme which aimed to develop procedures for metabolomics.   In fact since G02 reported the standard of data produced either at SCRI, Golm or Aberystwyth University has been incrementally improved by the production of standard operating procedures, such as those presented in the series of Nature Protocol articles produced during the G03 project. Thus, experimental factors should no longer be major sources of unwanted variation. As far as the single compounds are concerned (after log-standardisation), the first remark is that they are often normally distributed, though this could vary from data set to data set. We could not find any indicator of compounds for which a Normal distribution is not appropriate (nor are other classical distributions). The finding of a statistical distribution that fits individual metabolite distribution is a useful tool since it can detect whether a new sample is within expected range in a new sample or not. In order to analyse all compounds simultaneously, we used PCO plots to represent overall sample differences/similarities. New individual samples can be added to the plot to see where, in relation to overall variability, they will be placed. This assessment is based on visual inspection. PCO plots should also be used to understand partitioning of overall variation by colouring sample points according to levels of each known factor (meta-data).

Suggestions for further work would be to design experiments that expressly account for the expected overall variability in a potato (need to define a set of cultivars to use as well as predefined environmental/processing factors, *etc.*). Moreover, the fusion of different data sets from different laboratories, possibly using different techniques to cover different families of important metabolites (e.g. Amino Acids and Sugars for GC-MS and Glycoalkaloids for LCMS) should be used. This is a prerequisite in order to decipher which area of metabolism is covered by metabolomics data sets for example. This would offer a complementary look at raw food material and aid in safety and/or quality assessment.

7.5 DEALING WITH MISSING VALUES IN METABOLOMICS DATA TO ALLOW COMPARISON OF METABOLITE COMPOSITION.

7.5.1 Distribution of missing values in datasets: Pattern of missing values

Figure 7.10 represents missing entries (in blue) in four different metabolomics datasets available to us. Each row represent a sample that has its own features (line, environmental condition…) while each column represents one peak that has been detected by the analysis technique (here gas chromatography in all cases) and quantified (using mass spectrometry). Samples are considered as our 'individuals' and peaks as 'variables' under study from a statistical point of view. SCRI 2002 and ABER 2001 datasets do not have a high number of missing values, respectively 3.1% (if we exclude 7 samples on which no measurements have been made, otherwise 10.7%) and 1.0%. The SCRI dataset is somewhat different since the number of sample is fewer (one order of magnitude fewer, ~90 against nearly ~1000 for each other dataset). Note that its number of replicates for each cultivar is reduced (only 4 against ~100 for other datasets) but it is concerned with 22 different cultivars (6 only for the three other datasets). Particular care was taken with these data sets at the data-pre-processing stage to keep missing entries to a minimum. On the contrary, datasets GOLM 2001 and 2003 can be considered to have a high percentage of missing values, respectively at 16.9% and 20.0%. Moreover, a structure in this missing pattern is clearly visible (batch/line/column effect). Both population and variable structures influence the missing pattern. In this case, data are said to be Missing Not At Random[2]. It is recognised that the MNAR framework is difficult to tackle

[2] The key question for analysis with missing data lies in the circumstances that can alter the validity (e.g in terms of valid inferences) of answers arising from complete data sets analysis. It depends on the data absence mechanism. Data can be Missing Completely At Random (MCAR), Missing At Random (MAR) or Missing Not At Random (MNAR), see Little and Rubin 2002. Under MCAR, the probability of an observation being missing does not depend on observed or unobserved measurements. An example would be the random drop out of measurements without considering measurement values. Under MAR the probability of a value being missing will generally depend on observed values -so it does not correspond to the intuitive notion of 'random'. Examples include trial subjects removal because not enough predefined information criteria are available or the decision to

whatever analysis approach is used, even for low rates (i.e. above 5% of total data absence) of missing values. However, no test exists at the moment to determine whether a dataset is actually MNAR.

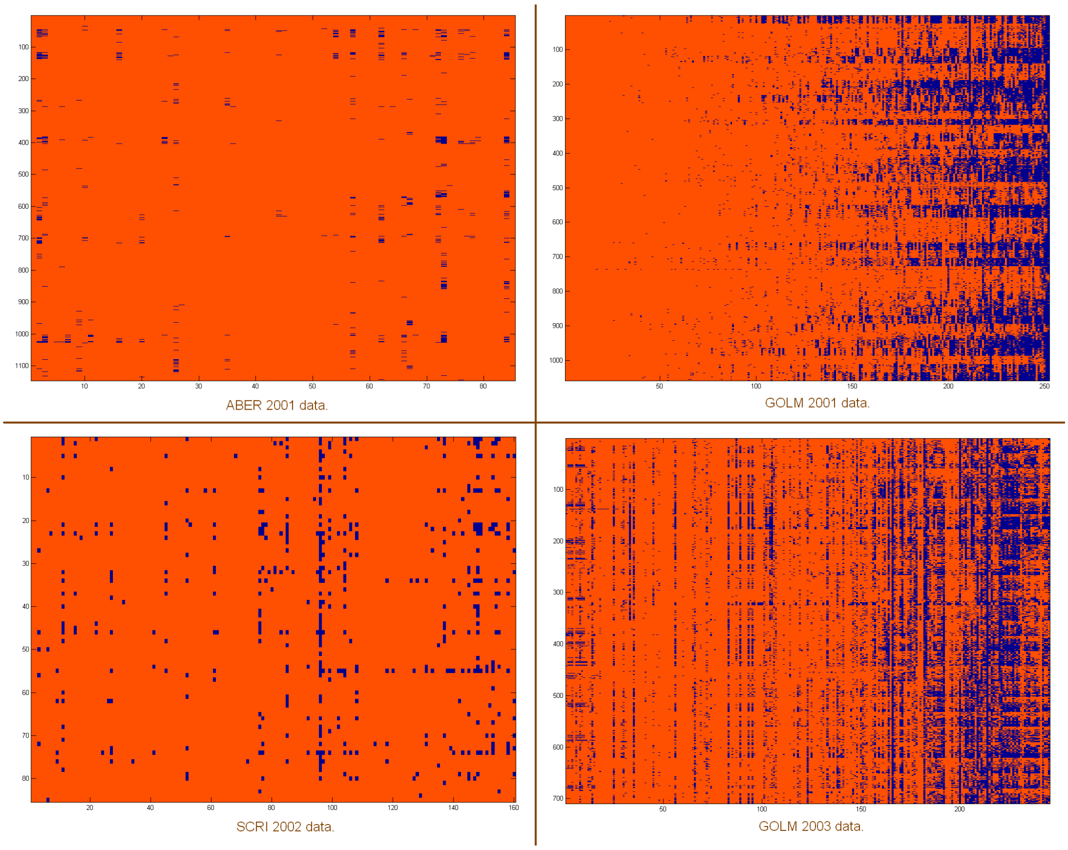


Figure 7.10: Missing value distribution in different datasets (rows are samples and columns are peaks or variables).

7.5.2 Influence of factors on missingness

We decided in the present report to focus on the relationship between missingness and other factors/variables of interest. In other words, how are “holes” in one dataset distributed according to some metadata? For example, Figure 7.11 represents the dependencies between the number of missing values in one sample and the field block where it was grown in the three datasets for which this factor was available. No particular relationship is to be noticed. In the same way, Figure 7.12 is a box plot of the number of missing values in the four datasets

make another measurement depending on the recorded value for a variable as regards a predefined criteria. In other cases, MNAR framework (probably true mechanism in most cases) needs to be considered; to obtain valid inference, the (generally unknown) absence mechanism must be included in the model. Censorship induces MNAR data.

according to the cultivar of the sample. Again, no dependency is to be noticed. Note that no real conclusion can be given for the SCRI 2002 dataset given there are only 4 replicates at most for each line. So there is no influence on the sample origin or of its nature on the value absence.

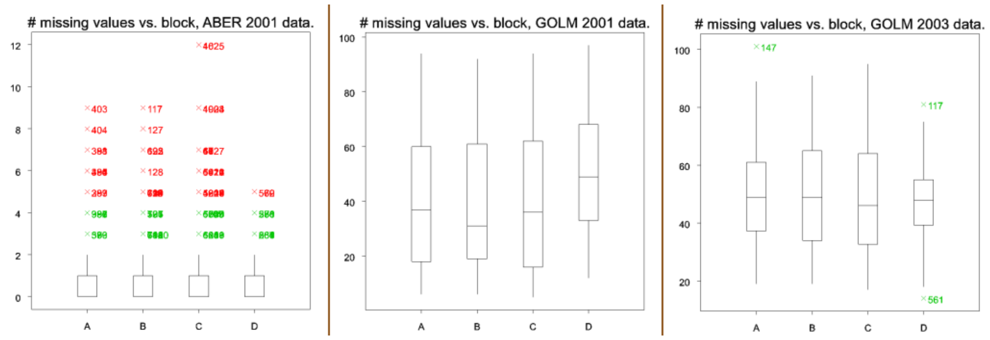


Figure 7.11: Number of missing values in different datasets according to the field block.

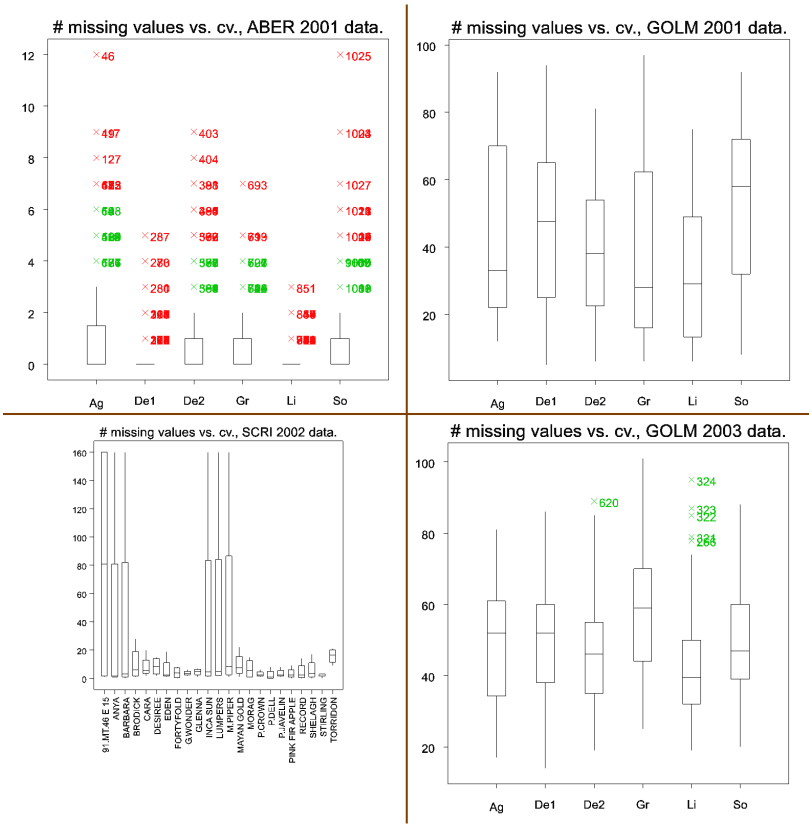


Figure 7.12: Number of missing values in different datasets according to the cultivar.

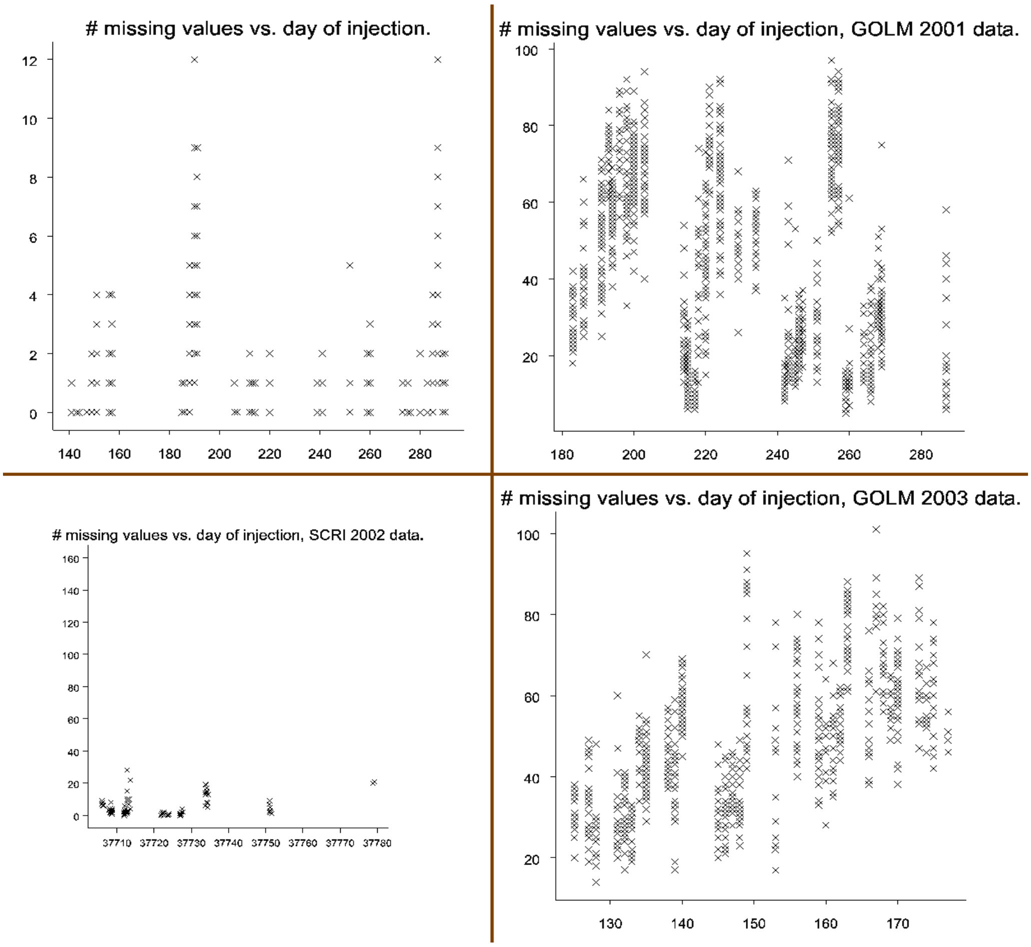


Figure 7.13: Number of missing values in different datasets according to the day of injection of the sample.

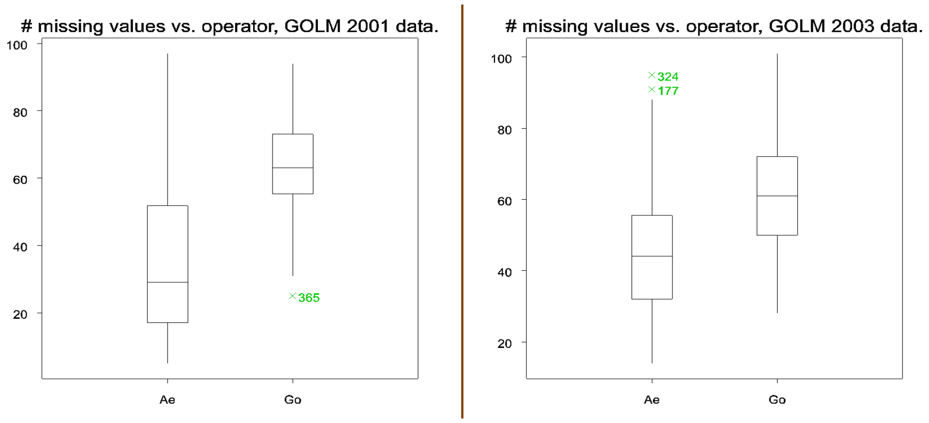


Figure 7.14: Number of missing values in two datasets depending on the operator carrying the analysis.

Figure 7.13 plots the number of missing values versus the day of injection of the sample into the machine. Again no clear structure between these two factors can be retrieved. So the order in which measurements were made does not seem to affect the missingness. We also report in

Figure 7.14 the distribution of missing values according to the operator when we had this information. Again, there does not seem to be a significant dependency between these two variables (though this can be discussed for the operator factor). The only positive dependency we were able to retrieve is shown in Figure 7.15. The number of missing values is notably larger for unidentified peaks in three datasets (SCRI 2002, GOLM 2001 & 2003), the last one (ABER 2001) having been carefully pre-processed as regards missing entries does not show this relationship. On one hand, this is rather good news since we will actually be focusing on identified variables for the univariate analysis. Relatively few missing values will have to be imputed for identified peaks. The choice of the imputation method in this case will not influence too much the results of most statistical analysis (especially if we can assume MCAR data). On the other hand, imputing missing values for unidentified peaks will become necessary for any multivariate analysis of the data.

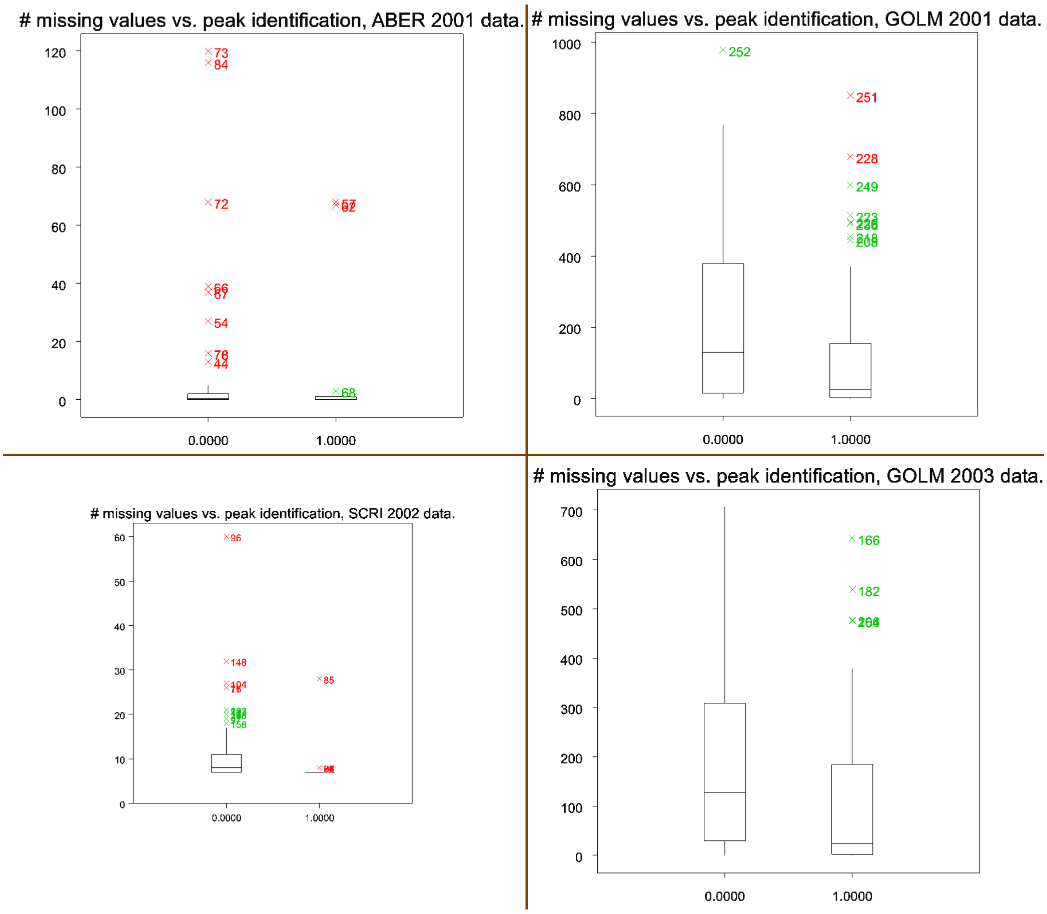


Figure 7.15: Number of missing values in different datasets according to the identification of the peak on which measures are carried out. 0 indicates that the peak isn't identified whereas 1 indicates that it is. Identification refers to a relatively accurate link to a known chemical but does not include partial chemical differentiation (e.g. cho-oligo14, amine30 or y_aliphate).

7.5.3 Causes for data absence in metabolomics data

We now briefly discuss possible causes of missing values in a metabolic datasets. Metabolites are present in plant tissue extracts in a dynamic concentration range that can differ over several orders of magnitude between individual metabolites. Thus many low concentration metabolites may fall below the threshold for accurate detection in some runs, depending on system sensitivity at the time and absolute concentration in replicate sample (Draper et al. 2004). We can encounter the case of a false negative: assay values below detection limit (true 'low' value) can point out to a specific class of metabolites. Hence the missingness is informative. The absence of an observation might also result in an inconsistency or a mismatch, for example when the analytical instrument experienced a malfunction or the metabolite was treated as an outlier. A false positive peak could have been generated by the peak alignment software in the same retention time window and thus the expected peak is not found. Quite often the automatic deconvolution or the peak detection software can be blamed (true negative) when two metabolites cannot be separated (Steinfach et al. 2007). A third possibility is that a value can be missing without us actually knowing it is missing (the missing value is hidden by another measured peak at the same mass ratio because of artificial effects: see the glossary entry for 'matrix effect' in Birkemeyer et al. 2005). We here refer to the (ionization) suppression effect well known in proteomics.

Given the typical volumes of metabolomic data the semi-manual checks performed on the ABER 2001 data are obviously too prohibitive in terms of time and money to be carried out routinely. Moreover, individual machine set ups might also imply that a different data set intensity values cannot be directly compared. All these reasons support the inclusion of missing data treatment in the analysis phase. We cannot ignore missing values as soon as the dimension of the problem at hand becomes significant. In fact, a simple calculation shows that for only 1% absent measurement in an 100*100 array (100 samples and 100 variables), 63% of the rows (samples) are affected by a missing value in the case of random value removal. In the case of pairwise correlation, the situation is not so adverse since only one or two values will be omitted. We will here only consider simple tools to modify data sets with missing values, discuss their advantages and drawbacks and comment on the work that can be considered to avoid known missing data pitfalls.

7.5.4 Missing data in combined data set: Description of the fusion of data sets

When data from several datasets are combined, the result is not only the addition of missing patterns. Another form of missingness which we describe as *structurally missing values* appears. It occurs because of the highly non-overlapping lists of variables (chemical compounds) that are detected in our varied metabolomics datasets. Figure 7.16 illustrates the complex pattern that can occur from a combination of such data. Moreover, the percentage of missing values is much greater than in individual data sets because the sets of compounds in each data set are not identical (see Appendix 3A of G03012 Report). From a purely biochemical perspective, a comparison of the metabolite list from each data set is in itself of interest. According to the diversity of the considered datasets, it will provide an overview of detected compounds in the crop of interest in different or similar conditions by a panel of analytical techniques. To this extent, the comparison with an 'external' source of the crop metabolite composition should be fruitful (see Section 2 of the G03012 Report which comments on published literature and accessible databases describing compositional analysis of crop plants).

The priority is not to estimate the full data matrix. However, comparison across datasets could lead to more robust way for estimating missing values *of interest*. For example, Little 1992 reviews the possibility of using generalized regression models for this purpose. Multiple regression models allow standard errors for estimates on the assumption of normally distributed peaks. Building up a list of reference data sets has already been successfully applied in transcriptomics (Hu et al. 2006) for missing value imputation. Additionally, this would help to determine whether an experiment reflects the actual amount of a chemical species in the crop under study and under which conditions. Whatever the approach is, the method can only be semi-automatic from our point of view. The knowledge of a biological expert is thus needed to decide whether a set of variables found to be correlated to a set of variables with missing entries will be relevant, or, to decide which reference datasets to use. Hence this is a difficult data analysis area because expertise is needed in the biological content of the data as well as in data handling and analysis.

7.5.5 Identifying a gold standard complete data set for comparison to aid imputation of missing values.

An additional difficulty is the general issue of assessing a method when no gold standard dataset exists. Trying to create it would involve estimating missing values as no metabolomic data set is complete. There are two general approaches to this problem; either missing values can be inferred using other data from databases or statistical procedures can be used to generate estimate based on the appropriate set of data to be corrected. A combination of procedures is also possible. The choice of the estimating procedure itself is a rather difficult issue that is been debated a lot in the transcriptomics literature (see for example Brock et al. 2008). A first step in this direction was our attempt in Section 2 to compare different GCMS metabolomics data sets (ABER 2001, GOLM 2001 & 2003 and SCRI 2002) with external compositional data picked from the literature or from databases. The idea is that, in light of important information coming from an ‘expected’ list of common metabolites in a given crop, we can quantify relationships between these common compounds often detected at significant levels and less frequent compounds.

7.6 EXISTING MISSING VALUE ESTIMATION APPROACHES

7.6.1 Classical in-filling procedures applied previously to transcriptomics data

There is an abundant literature on the subject of estimating missing values for transcriptomics data. Troyanskaya et al. (2001) can be considered a pioneering piece of work which used the K-Nearest Neighbours approach. Oba et al. (2003) described a singular value matrix decomposition followed by a Bayesian estimation of the parameters whilst Bo et al. (2004) advocated a least square imputation procedure and Ouyang et al. (2004) made estimations by modelling of mixture of distributions. There is no one single preferred method though it is acknowledged that different prior imputations can lead to very different results depending on data features (Scheel et al. 2005). This being said single imputation methods have two general drawbacks: lack of adjustment to errors of the statistics of interest to account for uncertainty on the imputed value and systematic bias. The latter is a result of non-random patterns in the occurrence of missing values.

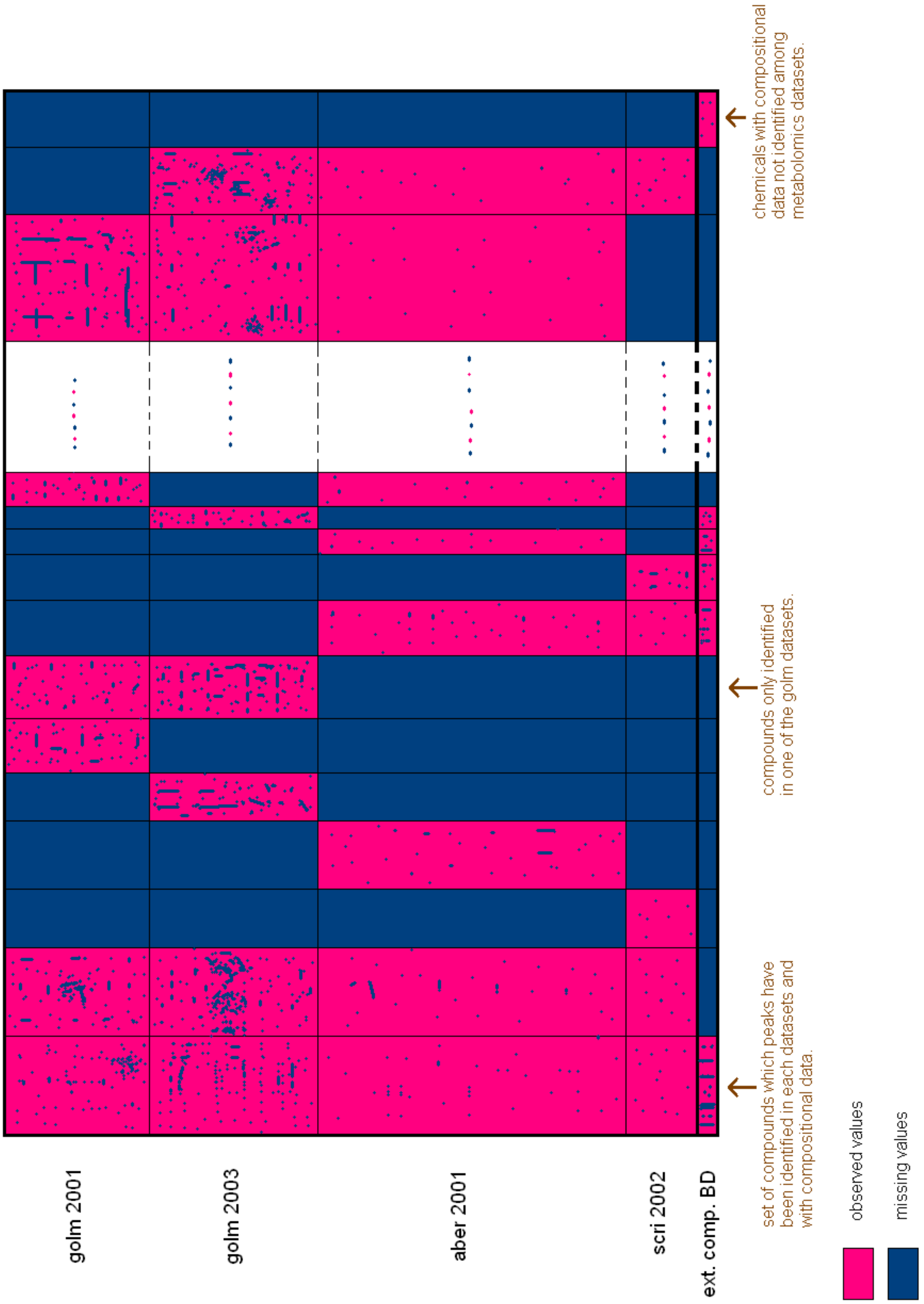


Figure 7.16 Merging metabolomics datasets causing a curse on missing blocks of data.

7.6.2 In-filling procedures in the metabolomic context

Literature in the metabolomics field concerning dealing with missing values is comparatively sparse (see Tong et al. 2007). However there is a strong need for approaches that account for any particularities of the data at hand. The situation for general principled methods implemented in statistical software (see Horton and Kleinman 2007 for the case of regression) is not really satisfactory: complex approaches don't seem to bring more reliable results than simpler naïve or heuristic approaches. The Multiple Imputation[3] (Sinharay et al. 2001 or see http://www.multipleimputation.com) approach of the Amelia II software can only deal with multivariate time-series. Multiple imputation and likelihood based[4] approaches give good estimates under 'reasonable" assumptions (typically MAR). When MNAR is encountered, modelling the absence mechanism can be useful (it theoretically cannot be ignored!) but, as we have already pointed out, can be very sensitive to model selection.

7.6.3 Taking a pragmatic approach to dealing with missing values

Given the difficulties outline above, we chose to use a simple approach. For individual metabolomics datasets, we will infer missing values using a K-Nearest Neighbours approach (intermediary between weighted mean imputation and hot-deck imputation taking an observed value from a randomly chosen similar sample). The situation for ABER 2001 and SCRI 2002 shouldn't be too problematic since the total number of missing values is low and there is no apparent missingness pattern. GOLM 2001 and 2003 data are more difficult. Hence we decided to remove variables with more than 25% of missing values. Many references agree that this relatively simple approach leads to acceptable results on simulated data or data where the absence mechanism is too simple for our case. Ideally we would have preferred to test different approaches and to assess their relative performances but we were not able to find a fair scheme and meaningful measures of error to compare them on metabolomics data. This

[3] "Multiple imputation methods randomly draw observations from a fitted distribution for the covariates and the outcome variable. For each imputed data set, the missing data are filled in with values drawn randomly [with replacement] from the distribution. Analyses are performed on each data set as though the data had been completely observed. The results of these analyses are then pooled to provide point and variance estimates for the effect of interest (Faris et al. 2002).

[4] Likelihood approaches consist in the integration of the probability distribution over all possible values for missing variables giving more weight to more plausible ones whereas multiple imputation only consider few plausible values.

would be possible if we had any complete datasets. Simulating a realistic dataset to test/compare different approaches to deal with missing values is not an option as it demands too great an understanding of the correlation structures among the compounds.

7.6.4 Missing values in the G03 external literature database

As far as the fusion of different sets of data is concerned, we chose to leave the database of external metabolite literature presented in Section 2 as it is; the complex differences between these lists of metabolites mean that attempting to calculate missing values cannot be done in any principled way as discussed above. However, although the previous compositional and the G02 metabolomics data do shed light in a complementary manner on the biological content of a sample. We suggest that in the future it will be valuable to create a reference profile data set by quantitative measurements of a list of metabolites that are deemed to be measurable in the majority of potato cultivars (see Appendix 3B). If this is achieved then such a database on the chosen crop (potato) could be accessible through ARMeC.   Any future strategy to impute missing values will need to be treated with care. Moreover this work needs to be carried forward both on the external compositional value retrieval and on the metabolomics data integration. A biological expert will be able to add data and to decide how to impute values according to references or with a best use of relationships between existing metabolomics data and compositional data.

The ultimate goal of the G03 project is to decipher natural variability as well as protocol effects on the metabolomics response of the crop. The present report clarifies some aspects of the missing value issue in metabolomics datasets and will help for later work. The KNN imputation method for dealing with missing values has been added to the suite of data pre-processing software routine developed in the G03 project (see Section 8 of the G03012 Report).

REFERENCES

Regression with missing X's: a review, Little R.J.A., *Journal of the American Statistical Association*, 1992.

Annotation: what can be done about missing data? Approaches to imputation, Heitjan D.F., *American Journal of Public Health*, 1997.

The use of multiple imputation for the analysis of missing data, Sinharay et al., *Psychological Methods*, 2001.

Missing values estimation methods for DNA microarrays, Troyanskaya et al., *Bioinformatics*, 2001.

Statistical Analysis with Missing Data, Little and Rubin, Wiley, 2nd edition, 2002.

Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analysis, Faris et al., *Journal of Clinical Epidemiology*, 2002.

A Bayesian missing value estimation method for gene expression profile data, Oba et al., *Bioinformatics*, 2003.

LSimpute: accurate estimation of missing values in microarray data with least square methods, Bo et al., *Nucleic Acids Research*, 2004.

Metabolite Peak Identification and Data Structure in a Multi-Site, Large Scale Metabolomics Experiment, Draper et al., *Proceedings Pittcon*, 2004.

Gaussian mixture clustering and imputation of microarray data, Ouyang et al., *Bioinformatics*, 2004.

A Q-technique for the calculation of canonical variates, Gower J.C., *Biometrika*, 1966.

Adding a point to a Principal Coordinates Analysis, Wilkinson C., *Systematic Zoology*, 1970.

The stabilized probability plot, Michael J.R., *Biometrika*, 1983.

Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology, Anderson M.J. and Willis T.J., *Ecology*, 2003.
Centering, scaling and transformations: improving the biological information content of metabolomics data, van den Berg R.A. et al, *BMC Genomics*, 2006.

Metabolite profile analysis: from raw data to regression and classification, Steinfath M. et al, *Physiologia Plantarum*, 2008

Metabolome analysis: the potential of in vivo labeling with stable isotopes for metabolite profiling, Birkemeyer et al., *Trends in Biotechnology*, 2005.

The influcence of missing value imputation on detection of differentially expressed genes from microarray data, Scheel et al., *Bioinformatics*, 2005.

Integrative missing value estimation for microarray data, Hu et al., *BMC Bioinformatics*, 2006.

Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models, Horton and Kleinman, *The American Statistician,* 2007.

Metabolite profile analysis: from raw data to regression and classification, Steinfach et al., *Physiologia Plantarum,* 2007.

Investigation of processing and imputation in metabolomics datasets, Tong et al., *Proceedings of AIChE,* 2007.

Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes, Brock et al., *BMC Bioinformatics*, 2008.

GO3012 Final Report - Section 8

David Enot/Wanchang Lin/John Draper

Refinement of standardised multivariate statistical methods for both describing and comparing metabolite composition of food raw materials (Tasks 04.02 and 04.03)

8.1 BACKGROUND TO EXPERIMENTAL SECTION

It is imperative that if the data processing is standardised (as described in Sections 4 & 5) then the post-processing (statistics) is also formatted. Preliminary conclusions from the G02 initiative and observations elsewhere in other GMO related projects strongly emphasised the lack of coherence and consistency of what is reported in terms of statistical and modelling results relevant to safety assessments. Thus the main emphasis in the present project was to achieve high level quality in terms of relevance, accuracy and coherence of the statistical analysis by enforcing statistical best practice using robust data mining tools, properly validated models and cross comparison between statistical models from different algorithms, techniques and commonly used software. Accessibility (including cost of software), interpretability and consistency of the statistical information are also important. **One important point to make is that the identification of data mining routines for suitability to analyse metabolomics data was undertaken previously as part of the G02006 project; the aim of the present project was get these routines in a standardised software environment and develop a standardised workflow.**

Metabolite profiles and fingerprints provide intensities over a large number of compounds. This provides a challenge – how to make use of all of these compounds simultaneously to provide a baseline against which to assess any novel product, and an opportunity – to provide a baseline which itself consists of several criteria, each based on a different combination of the compounds, thus offering a range of characterisations against which any new product must be assessed. Inherent high dimensionality coupled with both biological and instrument variance characteristics in FIE-MS and GC-tof-MS data require use of powerful, multivariate data analysis tools for comparison of raw plant material chemical compositional. To further ‘mine’ such data for variables responsible for sample class discrimination demands specific supervised ‘machine learning’ software routines incorporating some form of data dimensionality reduction and requiring support from experienced data analysts. At the inception of this project, to the best of our knowledge, nobody has reviewed or proposed a systematic way to analyze and extract meaningful information out of metabolomics data for global assessment of food composition. Through the combined experiences of SCRI and Aberystwyth in developing models of metabolome profile distributions and on clustering, we have developed standardized statistical analyses of metabolome variability in plants. A major deliverable of this part of the GO3 project is the identification of a set of statistical tests which we believe offer robust comparative descriptions of the metabolome.

8.2 OVERALL OBJECTIVES OF EXPERIMENTAL SECTION

The metabolomics technology development programme funded by the FSA in the G02006 project headed by JHD resulted in the identification of several validated, powerful machine learning data analysis tools for fingerprint or profile classification/discrimination that can both cope with the intrinsic biological and instrument variance in large scale experiments and also highlight specific metabolite signals responsible for discrimination. Such methods continue to be refined as part of the core activity of Aberystwyth researchers as part of the BBSRC Plant and Microbial Metabolomics Initiative (MetRO). This section of the present GO3012 project had several major objectives which were carried out in collaboration with the MetRO advanced data mining team (Dr David Enot and Dr Chuan Lu). The software routines reported on previously in the G02006 project were supported by a combination of commercial statistical packages, web-accessible free ware and in-house algorithms written in several different computer languages (e.g. C++ , fortran, ‘R’). These routines encompassed both tools used for raw data pre-processing (e.g. data integrity detection [batch problems, false positives/binning issues, base line correction], data normalisation, outlier detection) and data mining tools for sample classification and feature selection. A key part of the G03012 project has been to re-write all the software code for a single platform (‘R’) which is designed specifically to support statistical tools. Once this had been achieved the data analysis tools were tested by biologists and analytical chemist who had little experience in data mining and who helped to write software tutorials to describe how to use the software. The tutorials are accessible at the following URL (http://users.aber.ac.uk/jhd).

As part of the validation process the parameterisation of the software routines and a refinement of the modelling visualisations and reporting formats was undertaken. As outlined in report Section 5 the example data used to test the software came from a variety of experiments and a challenging data set (a series of rice blast challenged *Brachypodium* leaves) provided by Dr David Parker on a BBSRC project was selected for illustration purposes. This part of the project was largely the responsibility of Dr Wanchang Lin and is described as an example workflow accessible at the following URL (http://users.aber.ac.uk/jhd;) based on the pre-processing and mining of FIE-MS data described in Section 5 of the present GO3 report. This example workflow and associated software has been published as a Nature Protocol by Enot et al., 2008 [70]. These software and tutorials are now freely-available on an Aberystwyth University web site (http://users.aber.ac.uk/jhd) and are updated regularly as any reported bugs are resulted or new routines developed.

The present section of the G03 report is additionally concerned with assessing the output characteristics of the main modelling tools and describing how the data mining procedures were used to provide standardised similarity criteria for making compositional comparisons between plant genotypes. For these purposes we used largely the FIE-MS fingerprint data and GC-MS data representing the chemical composition of 5 non-transgenic potato varieties. Further 'in house' data generated on the FSA G02 project describing the chemical composition of a range of *Arabidopsis* genotypes provided a larger genotype collection with which to evaluate the reproducibility of the methodology. These experiments have been published in several recent papers by Beckmann et al., 2007 and Enot et al., 2007a, b.

8.2.1 Summary of objectives

- Introduce multivariate data analysis tools and review recent literature concerned with comparing the overall metabolite content of raw food materials based on metabolomics fingerprint and profile data.

- Develop a software package containing all appropriate data pre-processing and data mining routines in a single environment designed for statistical analysis ('R').

- Validate software by training of non-expert users and write a user manual and standardized training workflow. Develop web resources with software, training manual and example data/workflow.

8.3 INTRODUCTION TO CLASSIFICATION MODELLING AND FEATURE SELECTION USING MASS SPECTROMETRY FINGERPRINT/PROFILE DATA

By their very nature 'omics' level datasets are high dimensional and as such can often contain too few samples per class replicates to allow adherence to an experimental statistical design that can cope easily with the degree of biological and instrument-derived variance encountered when data modelling is undertaken [1-8]. Although it is possible that 'omics' level experimentation may suddenly become much cheaper or more high throughput in the near future, it is exceedingly optimistic to suppose that it will become more straightforward to access much larger numbers of biological replicates. Any data analysis strategy must therefore cope with these data characteristics and well validated approaches are required both to initially 'test' and improve the overall raw data structure and subsequently mine the data for signals that are sufficient to explain adequately the differences between individual biological samples classes ("explanatory" variables) [7-15].

Established technology for measurement of metabolites in complex biological extracts includes high pressure liquid or gas chromatography (HPLC and GC) linked to a range of detectors [14, 16-23]. Many data tables contain measurements of peaks (sometimes in excess of 60%) attributed to commonly encountered, but structurally uncharacterised, metabolite signals for which chemical standards are not available to allow positive identification [14, 20, 24-26]. Thus peak finding, peak alignment and peak annotation in complex chromatograms is challenging, operator driven, and dependent on a combination of individual metabolite concentration and the levels of resolution and sensitivity of the instrumentation employed [16, 23-26]. A more 'global' overview of total sample metabolite composition can be obtained from metabolite 'fingerprinting' techniques which do not incorporate a chromatographic step [12, 14, 27-31]. However, once data pre-processing has been achieved both metabolite fingerprint data and metabolite profile data can be analysed using identical classification and feature selection tools.

The major features of metabolite profile data are described in Sections 3 and 4 of the present GO3 report, while metabolite fingerprint data characteristics are described in Section 5. Methodology for data pre-procesing of GC-MS and LC-MS are still heavilly dependent on interactions between instrument software and operators as described in Sections 3 and 4 of the present report. In contrast FIE-MS fingerprint data is suitable for automated pre-processing and data mining by third party software and so is used as an example to demonstrate the main principles of standardised metabolome data modelling in the present section of ther G03 report. As mass accuracy capabilities differ between instruments and rigorous mass cabibration regimes are required to avoid the effect of instrument 'measurement drift', it is common place in first pass FIE-MS modelling to generate 'nominal mass' data for all signals in the detectable range [12, 14, 28-29]. Depending on data capture parameterisation nominal mass spectra generated by FIE methods can contain up to 2000 variables, where each ionisation product is integrated into a nominal mass 'bin' together with other ionisation products that have a similar molecular weight [12, 14, 28-29]. The scope of this section of the GO3 report is to outline data quality checking methods, classification modelling techniques and metrics used to compare metabolomics data sets derived from mass spectrometry experiments (for reviews [5, 13, 15 -19, 32]). In recent years such methods have been applied extensively to analyse metabolite fingerprints of samples derived from plants in order to investigate complex quality traits, population genetic structures and in investigation of gene function [12-14, 22, 29, 33-36].

8.3.1 Metabolomics data model characteristics and classification metrics

Typical outputs of any modelling process are metrics describing the accuracy of sample classification together with 'distance' measures of relatedness between biological classes and ranked lists of potentially explanatory features (i.e. metabolite signals) contributing significantly to the model. As part of the software evaluation, using commonly encountered model output metrics, we aimed to decide pragmatic statistical significance thresholds in relation to both discrimination/classification characteristics and feature ranking in good quality models [1-9, 12, 13, 15, 36, 37]; specifically, we wished to demonstrate general features and model metrics in instances where data mining has proved inadequate and alternatively show anticipated results in which robust models have been generated that that are suitable for deeper interpretation.

For the purpose of illustration the data analysis routine used in the present section will be demonstrated using FIE-MS fingerprinting data from Section 5. A series of quality checks are usually applied to the raw FIE-MS data which are linked to predicable characteristics of both the analytical approach and standardized experimental statistical design (see Section 5). Following these quality assurance procedures a range of data analysis tools may then be applied in a logical order to answer specific questions relating to the biological phenomenon under study (Figure 8.1).

A range of data analysis algorithms are available to undertake various elements of these common tasks, many of which are adequate for the job in hand and no particular series of routines will always provide an optimal approach. The aim of the present section of the GO3 report is to describe a series of validated software routines which provide easily interpretable models with mutually complementary statistical measures relating to classification power and feature ranking.

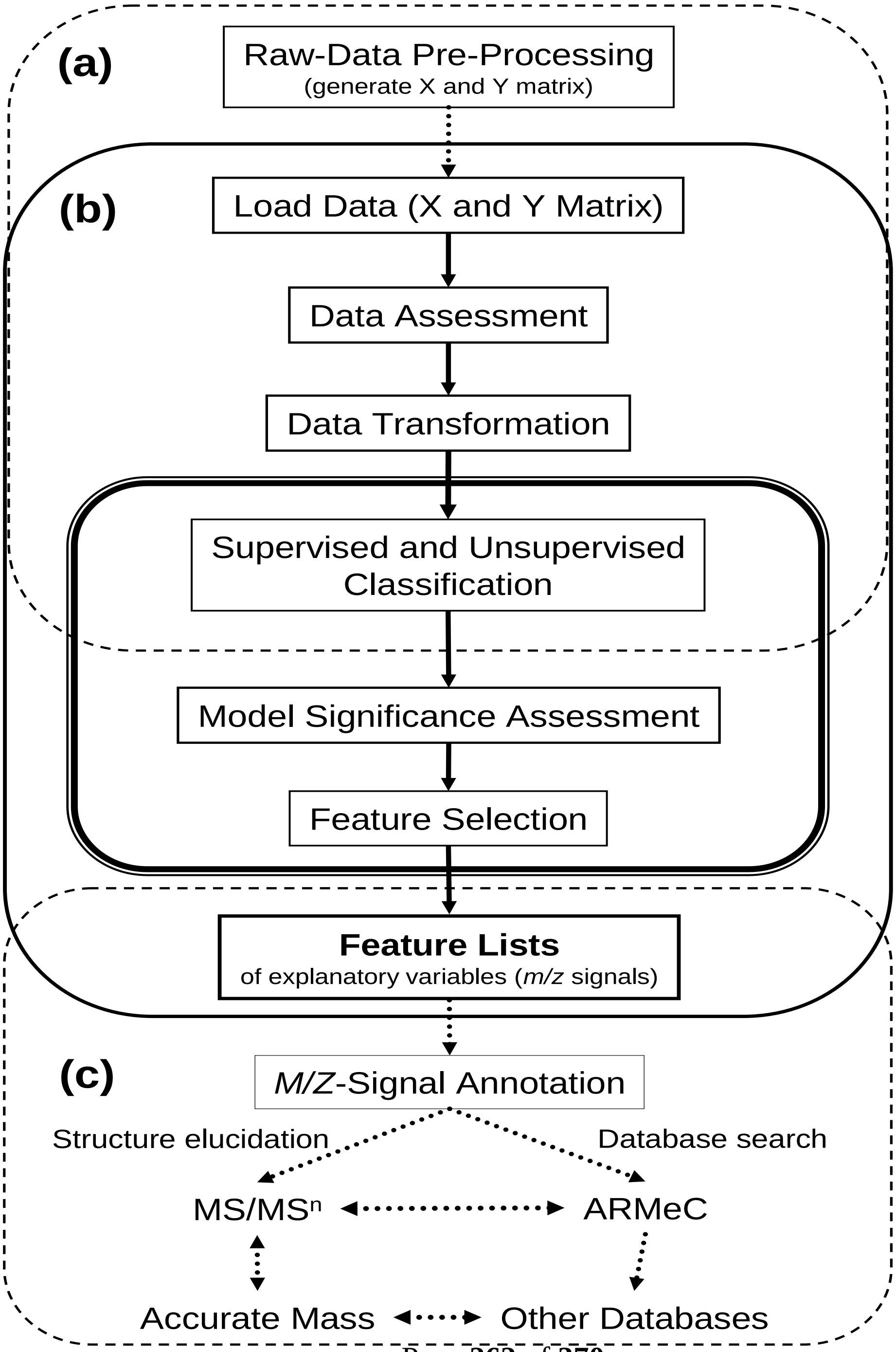


Figure 8.1 Typical work flow in FIE-MS data analysis

Data pre-processing

The ultimate goal for any metabolomics experiment is to find patterns in the fingerprint data that can describe the biological outcome [5, 12, 13, 15]. The initial step is to ensure that the quality of the data generated is meeting criteria required for further statistical analysis. This is a crucial activity to provide a sound basis for measuring and interpreting biological phenomena. Raw data pre-processing (Figure 8.1a) describes the application of one or several routines performed on binned data (see Section 5) to prepare them for further statistical analysis. Despite the fact that the raw fingerprint data may be directly utilised in a statistical treatment, data pre-processing transforms the data into a format that improves data quality and data mining accuracy. However, care should be taken so that new wrong knowledge is not created or important information lost during the "data polishing" process. Within this protocol we describe the use of various tools that are used to prepare the data including: checking for data integrity, linear transformation of the signals, baseline correction, outlier removal and fingerprint normalisation.

Data classification

Class comparison and prediction encompass a majority of the biological questions addressed by a metabolomics approach. Clustering or unsupervised modelling is useful for class discovery and provides information on data similarity: groups of observations relate to groups of variables and metabolome samples grouping together can be objectively considered to be similar [1-5, 12-15, 38, 39]. Principal Components Analysis (PCA) is probably the most widely used unsupervised approach to metabolomics fingerprints data modeling [39]. PCA is a multivariate technique that transforms the sample data matrix into a coordinate system where each new projection (also called Principal Components, PC) is a linear combination of the original variables. PCs are orthogonal so that each dimension is related to different data characteristics and source of variability in a mathematical sense. Secondly, projections are ordered according to the variance explained: the first PC encapsulates the greatest amount of variability, PC2 the second greatest, and so on. The success of PCA as a first pass technique relies on these two properties to provide initial sample clustering information and subsequent clues regarding sample behavior alongside the main directions of variance. However, due to the inherent complexity of FIE-MS fingerprints (50-2000 *m/z* range), unsupervised techniques, more often than not, tend to produce overlapping class groups, which hamper the extraction of explanatory signals responsible for class differences. Although such techniques allow a quick inspection of the underlying data structure they fail to provide decision rules for classification by objective and unbiased classification metrics [12, 15, 37]. In contrast, supervised techniques are using essential class information to generate between sample class models [4, 5, 8, 13-15, 36-47] and provide an ability to discriminate classes linked to the underlying signal behaviour within the fingerprints under comparison. For discrimination tasks, a sample per feature ratio (SFR: number of variables measured in each fingerprint compared to the number of samples available for each biological class) of around 5-10 is generally recognised as the minimum requirement to generate robust classification models. Unfortunately, SFR in FIE-MS fingerprints is always less than 1 which demands extra care during model construction to avoid producing over-optimistic models utilizing variance that is unrelated to the problem under consideration [1-4, 8, 11-15].

Model interpretation and validation

In very small SFR contexts, it is almost always possible to fit data and derive an input-output relation that will adapt to some peculiar structure in the data [1-5, 8, 11, 36, 37, 43]. As a consequence, conclusions regarding model performance cannot be reached from the same observations that were used to delineate the model. The classic solution relies on forming an independent set (validation or test set) consisting of observations that were not used for the training phase [1-5, 8, 15]. Under proper practices, test data are only employed to assess model goodness and not to decide which features to use in the final model or the number of components to include for further modeling. However, in many situations the sample size rarely allows formation of a large enough independent test to achieve adequate statistical significance estimates [1-5, 7, 8, 10-12] whilst at the same time keeping the training set big enough to construct a robust classifier (model). Such data paucity can lead to discrepancies in both interpretation and validation of the results. Thus resampling strategies [48], based on a repetitive partitioning of the original sample space, must be applied for most metabolomics modeling situations. Amongst these strategies, cross validation and bootstrap approaches [1-4, 49-51] are widely used in the machine learning communities to estimate classifier performance. Preference towards one approach is mainly driven by a variance-bias trade-off. Although cross validation is known to be fairly unbiased, this strategy suffers from a large variance and can potentially lead to what seems to be significant but which are in fact irreproducible results [2, 3, 7, 8, 43]. Contrastingly, bootstrapping based techniques have a better control of the performance variance but are biased toward overestimation of errors [1-4, 47, 49-51].

Receiver operating characteristic (ROC) curves are increasingly becoming a systematic alternative approach to investigate the predictive behavior of a binary classifier [5, 9, 37, 50]. ROC curves display the relationship between sensitivity (true-positive rate) and specificity (false-positive rate) across all possible threshold values that define the decision boundary. As a performance measure, the Area Under the ROC Curve (AUC) specifies the probability that the decision boundary assigns a higher value to a positive sample than a negative one, both chosen randomly. AUC takes a value of 0.5 when the samples from both classes are uniformly distributed across the decision boundary and a value of 1 when both classes are perfectly separated. In addition to accuracy/error estimates or ROC analysis, other statistical measures may also be derived from the actual classifier characteristics to achieve an alternative measure of model performance. For example in the G02 project we showed that complementary measures calculated from the Linear Discriminant Analysis Eigen system [39, 52, 53] and the *average sample margin* (model strength) in Random Forest [54] model generalizability and interpretability potentials extremely well [12, 36].

8.3.2 Feature ranking and identification of 'explanatory' variables

Ultimately the goal of any metabolomics experiment is to identify metabolites responsible for differences between experimental treatments (normally sample classes). It is rarely the case that one unique signal (or a combination of very few uncorrelated variables) will fully describe the property under study. Nominally redundant signals (often showing collinear variance) with FIE-MS data may reflect specific aspects of metabolic mechanisms or are different charged forms (e.g. adducts or fragments) of a unique compound. The following protocol describes an approach to determine an optimal set of features that can adequately describe the problem at hand. Unless we are interested in comparing very different sample matrices like a strawberry with an egg, a strong assumption associated with most metabolomics fingerprinting techniques (not only FIE-MS) is that the number of irrelevant signals is always much larger than the number of signals that actually contain meaningful information. Despite this statement, finding the optimal number of such features is not a trivial task in machine learning [2, 11, 34, 43], as considerable effort is directed towards delineating a model that uses a small number of variables (parsimonious models are more likely than complex ones). In many experiments it is possible that different sub-sets of variables are sufficient to explain a particular biological problem with an acceptable degree of mathematical statistical significance, but whether these multiple solutions have any biological meaning is difficult to ascertain [1, 2, 8, 10-12, 15].

Classic parametric or non-parametric univariate strategies can be employed to test the null hypothesis that treatments are not different [4, 48, 55]. The outcome is the computation of a p-value describing the likelihood that signal intensity measurements are different under some assumptions of the statistical method used [4, 11, 48, 55]. Feature ranking using univariate measures are not always necessarily applicable in a metabolomics context as an assumption is made that variables are independent, where the assessment of the discriminatory power of the individual variables exclude potentially interesting interactions with other variables. On the other hand, multivariate techniques use interactions between variables during feature ranking, either explicitly as in decision tree techniques [12, 55], or implicitly in the case of non-parametric methods like nearest neighbour. Commonly used methods such as Support Vector Machine [37, 44] or Random Forest [36, 37, 47, 54, 56] can be used to rank fingerprint signals based on an importance score or the frequency of variable selection. Because such methods involve an optimization process, selection of a sub group of signals is decided by a heuristic method usually requiring extra validation steps. Obtaining a final ranked list remains a complex and time consuming task (up to several hours of cpu time for a simple two class problem) and still the definition of relevant and irrelevant features is not trivial as it is difficult to prescribe *a priori* a threshold metric for model selection [1-4, 10-12]. In order to balance computing cost and the robustness of the explanatory variables list prior to deeper chemical analysis or database searching, we approach the validation of feature ranking by looking at the stability of individual features across multiple ranked lists by resampling the original training data [38, 54, 56]. Intuitively, ranking of relevant features should be consistent despite perturbing the data, whereas the ranking of irrelevant signals fluctuates substantially.

8.3.3 Applications of data mining

Metabolite fingerprinting coupled with multivariate data mining has been used to address a range of conceptual problems using mainly unfractionated polar extracts of samples derived largely from microbes and plants [14, 28-31, 57-64]. More structurally targeted metabolite fingerprinting approaches have also been reported using mammalian biofluids such as blood derivatives and urine to classify individual with metabolic disorders [19, 61]. Investigations with microbes have included the development of phylogenetic tools (‘chemical taxonomy’) to assess genetic diversity using unsupervised clustering techniques [31]. In other instances supervised classification approaches have been shown to have utility to actually identify the microbial species present in mixed populations [59]. Studies investigating quantitative traits in crosses between two species of tomato have used PCA to detect introgression lines that display metabolic phenotypes that differ both from each other and from the parent line contributing the majority of the genome [62]. Metabolite fingerprinting using mass spectrometry coupled with a range of data modelling techniques has provided a first pass tool to investigate whether there are fundamental changes in metabolism associated with plant responses to either environmental or development signals or biotic stress [30, 58, 59]. Further studies have used FIE-MS combined with a range of machine learning data analysis approaches to not only classify or discriminate samples, but also to identify which variables were responsible for such differences. For example metabolite fingerprinting has been used to confirm that only the intended biochemical changes in metabolism had occurred in plants genetically engineered to exhibit novel enzyme activities [14, 30]. Functional genomics studies using *Arabidopsis* mutants has utilised FIE-MS coupled with Random Forest and Linear Discriminant Analysis to investigate the phenotypic effects of gene disruption or unscheduled expression [12, 37].

8.3.4 Considerations regarding choice of the modeling techniques

In a high throughput context, the generation of robust classification models is first required before identifying model features related to underlying biological based characteristics within the data structure. Although many classification methods can achieve high predictive accuracy, the chosen method must produce directly interpretable models rather than exceedingly complex models that are opaque to further interpretability [1, 5, 12, 13, 15]. From the results of the previous G02 project it was decided that the classification methods to be employed as a default are Linear Discriminant Analysis (LDA) [5, 12, 36, 45, 52, 53] and Random Forest (RF) [12, 47, 54]. Both of these methods fulfil the previous objectives and can be applied on multi-class problems without special reformulation of the problem.

Linear Discriminant Analysis

Linear Discriminant Analysis (also known as Discriminant Function Analysis in the metabolomics literature), is a supervised classification method that provides discriminant functions (or canonical variates) to describe separation between pre-defined groups (classes) of samples in 'n-1' (n = number of classes) dimensional space [47, 52, 53]. Discriminant function (DF) scores can be plotted to visualize the 'amount of separation' between samples and to predict group membership of samples of an unknown class (DFs are orthogonal so that the discrimination among groups doesn't overlap). The objective of LDA is to obtain a projection that maximises between-class separability (S_B) while minimising within-class variability (S_W) [39]. However, due to issues associated with sample paucity (as is often a problem in metabolomics experiments), S_W is always singular and/or unstable and the inversion of S_W cannot be possible without a prior reduction of data dimensionality, usually obtained by a preceding principal components analysis (PCA) [52, 53, 65]. To avoid the introduction of bias in the estimation of the separability measure used in selecting the number of components, we chose to inverse S_W as proposed by Thomaz *et al.* [52] Eigenvalues derived from solving the eigensystem of ($S^{-1}{}_W\ S_B$) can be used as a measure of similarity of the discriminate function where the greater the distance between classes (large between-class scatter, S_B) and the more compact the classes are (small within-class scatter, S_W), the larger the eigenvalue and the more resolvable class separation becomes.

Random Forest

One of the best ways to improve the performance of single Decision Tree based algorithms is to grow ensembles of trees [54, 56]. Random Forest (RF) is a randomized and aggregated version of the standard decision tree [54]. Contrary to the standard decision tree, in RF each tree of the forest is built on bootstrapping the initial training dataset; at each node, the best split is chosen among a random subset of the initial pool of variables; and each tree is grown large, probably overfitting the bootstrap sample. The main advantage of using RF is that the final model will not overfit as the error rate will only reach a certain value no matter how many trees are built. Additionally, RF can deal with highly dimensional and correlated datasets without an initial reduction of dimension of the dataset. In the scope of this protocol, two elements can be derived from a RF model that assists deeper interpretation:

- As opposed to single decision trees, Random Forest does not produce an explicit model of the relationships between variables and the outcome. Instead, Breiman described the variable importance in a forest of trees as the measure of how each feature contributes to the prediction accuracy in this population. For an individual tree, the *importance score* of a variable is determined by comparing the performance of the classifier on the original data and that obtained when the variable is randomized. The procedure is repeated for every variable used in a tree and results are aggregated over all the trees.

- Class membership assignment in RF corresponds to the class with the largest score and a large score means that we can be confident in the class attribution and be certain that the example belongs to the actual class. From this property, *sample margin* (confidence in sample class attribution) is defined as the difference between the score for the true class and the largest score of the rest of the classes (i.e. the most probable misclassification) [12, 36, 54, 66]. Hence, a positive margin means that the example is correctly classified.

8.3.5 Considerations for overall experimental design

A structure for generating FIE-MS fingerprint raw data is presented in Section 5 of the present report. Typically FIE-MS datasets measure from 200 to 2000 variables and describe a considerable range of biological variance, which is especially prevalent in real world situations, for example when using field grown plants as apposed to organisms developed under controlled environment conditions. Given this feature, coupled with the non-targeted nature of the signals measured and the lack of any *a priori* guarantee that any specific biological question can be addressed by the proposed experimental approach, an adequate sample size is a rather important prerequisite to unequivocally derive relevant knowledge from the experiment. Defining adequate sample replication regarding the biological question is still under debate in most 'omics' communities with metabolomics being no exception [1, 5, 7, 10, 11]. The choice of sample numbers representing each class is mainly driven by cost, sample availability, technical facilities, machine reliability before instrument measurement drift and the need to overcome predictable levels of biological variance rather than strictly mathematical considerations. Because the technical variability in FIE-MS fingerprinting is usually smaller by comparison with the biological and experimental variability, biological replicates should be preferred to technical replicates as multiple measurements from the same sample will not make the experiment more effective. The present protocol is designed for pilot to medium size experiments. It is assumed that the data has a minimum of a thousand variables, and in most instances that between 10 to 50 sample replicates per class have been analyzed. We also assume that adequate sample collection and reliable good laboratory practices have been implemented not only to reduce variability but also to limit erroneous effects confounding biological treatments with irrelevant sampling and technical issues. It is assumed that samples have also been properly randomized during collection, storage and chemical analysis. Particularly in larger experiments it is crucial to confirm that the user has analyzed samples in a randomized order to avoid the embedding of structured variance in the data relating to injection order or any other experimental factors as described in metadata in Section 5. In situations where fewer samples are available, methods for model validation and feature mining can be rather limited (e.g. 'leave-one-out' cross validation) and criteria for definitive conclusions become highly severe. When larger numbers of samples are available, alternative validation strategies can be employed to assess model predictive abilities.

8.3.6 Outline of model data used to illustrate data pre-processing and data mining

Although the data pre-processing routines are described in the context of FIE-MS data, some data pre-processing routines, classification, feature ranking and validation tools have all been used successfully with NMR fingerprints [19, 27] and with metabolome profiling data generated using hyphenated mass spectrometry, for example GC-MS [12, 14, 17, 22, 25, 59] or LC-MS [23, 26]. The core elements of methodology develop in GO3 are illustrated using a FIE-MS data matrix developed from analysis of samples representing a time course of pathogen attack in a model plant species (*Brachypodium distachyon*) as outlined in a Nature Protocol by Parker *et al.* [35]. The data was developed in a single batch with all samples randomised using a Thermo LTQ linear ion trap as outlined in Section 5 and published as a Nature Protocol by Beckmann *et al.* [33]. As synchronous development of disease symptoms is difficult to control in any host-pathogen interaction this data provides a relatively challenging level of biological variance in a sample matrix that is changing rapidly as the fungus colonises the plant. Anticipated Results illustrating some of the optional steps are demonstrated in the context of other large data sets as indicated.

The protocol is designed to work with the R [67] package FIEmspro that includes specific routines for analysing FIE-MS fingerprints (http://users.aber.ac.uk/jhd). It is aimed at supporting data analysis for researchers who already have some basic experience of R command-line usage. The package developed in conjunction with researchers on the BBSRC MetRo programme, is updated regularly with regard to bug fixing, code improvement and new functions based on current research findings. For Windows and MacOS X platforms, straightforward installing can be realised via the binary packages available from the website. For any other platforms, source code can be downloaded and compiled appropriately. The reference manual supplies details on how to use the functions provided by the package, including a simple example for each function. A typical workflow that produces the results illustrating this protocol is also available on the package webpage as well as an extended document based on this workflow which is provided for both R beginners and advanced users. In the protocol below, *italic text* indicates the appropriate function name in R or in the FIEmspro package when specified. Data can be imported in R using the *read.table* command if they are in ASCII format such as those produced by Excel. Once loaded into R, it is more convenient to save them all as *.RData or *.rda format. For later usage, the function *load* allows to import them back. In the rest of the protocol, **X** designates the fingerprint matrix of dimensions (num. of samples x num. of signals), each data point corresponding to measured intensity. **Y** corresponds to the matrix of metadata including experimental factors namely class information, injection order, fingerprint file names and date of analysis for example (see Table 8.1b).

Table 8.1: FIE-MS data matrix organisation for analysis in the R environment

Sample number	Measured Variables (positive ion *m/z*)											
	P110	**P111**	**P112**	**P113**	**P114**	**P115**	**P116**	**P117**	**P118**	**P119**	**"**	**"**
1	1.52	11.53	0.51	29.62	6.08	2.57	6.30	8.55	3672.78	156.03		
2	4.91	8.00	0.00	19.05	8.92	9.83	7.10	0.00	5660.17	215.98		
3	7.50	5.91	0.00	194.02	0.00	115.49	5.23	8.78	2742.41	476.44		
4	18.38	5.68	0.51	19.90	4.42	22.26	82.92	13.38	10098.91	672.24		
5	0.50	3.68	0.41	46.69	3.71	5.34	4.37	0.00	4932.41	165.17		
6	15.90	4.78	4.06	31.63	3.30	12.73	107.77	4.51	5713.53	268.47		
7	6.02	1.06	0.00	26.45	0.89	6.94	2.88	0.35	5992.55	241.26		
8	6.83	1.62	16.97	7.14	3.42	22.53	11.18	2.68	6904.46	365.81		
9	10.32	4.63	5.10	24.41	3.46	5.09	35.14	0.73	7361.00	383.49		
10	14.97	4.77	4.14	17.20	9.07	2.13	130.65	0.66	6770.98	342.83		
11	2.83	6.36	3.74	26.30	10.30	15.99	20.99	7.19	8641.96	434.61		
12	11.11	5.65	22.70	32.74	12.51	49.34	62.88	4.58	9040.92	454.75		
13	6.40	2.33	12.26	47.66	7.27	11.62	35.71	0.76	11241.93	621.49		
14	6.99	3.87	0.00	26.97	10.97	2.00	9.52	0.16	6840.72	297.68		
15	39.97	2.61	0.00	13.25	1.99	6.26	84.59	1.05	7685.99	404.12		
16	54.31	4.07	29.70	17.90	3.45	32.33	1525.89	18.20	10032.82	603.81		
17	4.31	1.98	7.78	15.93	4.49	4.21	57.96	12.84	5543.85	266.34		
18	2.89	2.53	0.17	13.11	4.70	1.38	32.19	4.37	5461.45	259.08		
19	1.74	6.67	0.44	11.90	0.42	4.00	7.27	1.01	5400.78	217.89		
20	10.21	6.04	0.92	18.58	0.97	4.54	82.46	14.80	8412.83	418.31		
21	0.73	2.45	1.79	26.47	0.47	10.17	41.19	5.70	4517.02	217.32		
22	0.36	1.86	1.14	14.50	0.55	16.98	33.42	0.00	6049.36	235.20		
23	26.72	8.64	21.76	51.70	12.84	16.82	177.57	1.63	10211.82	550.66		
24	7.08	2.68	10.91	14.17	0.62	16.42	26.66	1.54	7489.76	403.95		
25	57.46	3.34	0.00	48.70	1.00	18.47	165.01	21.25	7236.41	378.80		
"												
"												
"												

(a) X-Data file: consisting of binned signal intensities (positive ionization mode) from nominal mass *m/z* 110 to 119 for 25 samples

(b) Y-Data file: containing columns of meta-data attached to the same 25 samples in (A) relating to experimental factors

	Experimental Factors*						
Sample Number	injorder	pathcdf	filecdf	remark	name.org	day	class
1	1	C:/Xcalibur/ANDI-LTQ/050509-Abr1/cdf	01.cdf	ok	12_2	2	2
2	2	C:/Xcalibur/ANDI-LTQ/050509-Abr1/cdf	02.cdf	ok	13_3	3	3
3	3	C:/Xcalibur/ANDI-LTQ/050509-Abr1/cdf	03.cdf	ok	15_4	4	4
4	4	C:/Xcalibur/ANDI-LTQ/050509-Abr1/cdf	04.cdf	ok	12_1	1	1
5	5	C:/Xcalibur/ANDI-LTQ/050509-Abr1/cdf	05.cdf	ok	12_2	2	2
6	6	C:/Xcalibur/ANDI-LTQ/050509-Abr1/cdf	06.cdf	ok	11_1	1	1
7	7	C:/Xcalibur/ANDI-LTQ/050509-Abr1/cdf	07.cdf	ok	14_2	2	2
8	8	C:/Xcalibur/ANDI-LTQ/050509-Abr1/cdf	08.cdf	ok	11_4	4	4
9	9	C:/Xcalibur/ANDI-LTQ/050509-Abr1/cdf	09.cdf	ok	13_H	H	6
10	10	C:/Xcalibur/ANDI-LTQ/050509-Abr1/cdf	10.cdf	ok	15_H	H	6
11	11	C:/Xcalibur/ANDI-LTQ/050509-Abr1/cdf	11.cdf	ok	14_4	4	4
12	12	C:/Xcalibur/ANDI-LTQ/050509-Abr1/cdf	12.cdf	ok	14_5	5	5
13	13	C:/Xcalibur/ANDI-LTQ/050509-Abr1/cdf	13.cdf	ok	12_1	1	1
14	14	C:/Xcalibur/ANDI-LTQ/050509-Abr1/cdf	14.cdf	ok	15_2	2	2
15	15	C:/Xcalibur/ANDI-LTQ/050509-Abr1/cdf	15.cdf	ok	13_H	H	6
16	16	C:/Xcalibur/ANDI-LTQ/050509-Abr1/cdf	16.cdf	ok	11_5	5	5
17	17	C:/Xcalibur/ANDI-LTQ/050509-Abr1/cdf	17.cdf	ok	14_3	3	3
18	18	C:/Xcalibur/ANDI-LTQ/050509-Abr1/cdf	18.cdf	ok	11_3	3	3
19	19	C:/Xcalibur/ANDI-LTQ/050509-Abr1/cdf	19.cdf	ok	13_2	2	2
20	20	C:/Xcalibur/ANDI-LTQ/050509-Abr1/cdf	20.cdf	ok	12_H	H	6
21	21	C:/Xcalibur/ANDI-LTQ/050509-Abr1/cdf	21.cdf	ok	11_4	4	4
22	22	C:/Xcalibur/ANDI-LTQ/050509-Abr1/cdf	22.cdf	ok	15_3	3	3
23	23	C:/Xcalibur/ANDI-LTQ/050509-Abr1/cdf	23.cdf	ok	13_5	5	5
24	24	C:/Xcalibur/ANDI-LTQ/050509-Abr1/cdf	24.cdf	ok	12_4	4	4
25	25	C:/Xcalibur/ANDI-LTQ/050509-Abr1/cdf	25.cdf	ok	12_H	H	6
	"						
	"						
	"						

* Experimental factors: '**injorder**', sample injection order; '**pathcdf**', reference to instrument raw data file [including processing software, instrument, date stamp, experiment label, file type]; **'filecdf',** unique label for each converted fingerprint file; **'remark',** general observation of individual fingerprint quality; '**name.org**', original name with experimental details [in this case the first digit is a reference to plant genotype (ABR1), the second digit is a reference to growth location (tray number) and the finall digit refers to treatment]; '**day',** treatment detail (used for plots as often visually more informative than class labels); '**class**', class label.

8.3.7 Sources of variability in metabolomics data and importance of exploring metedata

FIE-MS data structure will always reflect its origin (*e.g.* instrumentation and its local parameterisation, operator preferences and experience) and it is essential to obtain as much meta-data (ie. Y-data in Table 8.1b) as possible related to the generation of the dataset before initiating pre-processing. The analysis of high throughput FIE-MS data has an underlying assumption that the only source of variability is reflected in the class label and thus reflects the biological question under investigation. However, metabolomics experiments involve many steps, some of which may by the cause of unwanted variability. As a standard practice, a data analyst does not seek to obtain as much meta-data as possible in order to check that the experimental design is robust; however, it might help to better understand statistical output. Common assumptions made when examining large data sets include:

- the experimental protocol has been followed accurately for each sample and the efficiency of metabolite extraction is identical for each sample;
- all samples have been labelled correctly;
- mass calibration, instrument ionisation, mass resolution and detector sensitivity is identical for each sample.

A simple check that samples have been adequately randomised whenever possible (*i.e.* auto-sampler loading order from Y-data matrix) can help to counteract any variability resulting from gradual instrument drift. Apart from this a preliminary data analysis (*e.g.* PCA) may reveal the presence of potential outliers, or shows sub-clustering of samples which really should be in the same biological class. From a data analysis point of view, the benefits of checking meta-data can be summarised as follows:

- ➔ Relate outlying samples to experimental conditions.
- ➔ Relate group of samples behaviour to experimental conditions (including batch or sub-population).
- ➔ Moderate the use of some signals if they could be affected by changes during the whole process leading to data generation.
- ➔ Moderate the interpretation of the models if unplanned factors have been noticed.
- ➔ Adjust the experimental design including sample removal or creation of a new experimental factor.

The list in italics below provides some examples of metadata that are commonly associated with sources of experimental variability and worth investigating prior to deeper analysis.

Injection order and date of analysis:

On the basis of the common assumptions outlined above, machine performance represents a main factor over which we have the most control. In fact re-analysing a batch of samples is often more convenient than reproducing the whole experiment. Implicitly given by the sample order in the **X** data matrix (e.g. Table 8.1a) or explicitly as a separate **Y** factor, injection order and date of analysis related effects (Table 8.1b)can be outlined during the data analysis process by several means:

- effect of analysis date on average TIC of samples – *e.g. alterations in machine settings by a new operator or a maintenance event;*
- presence of isolated outliers in PCA analysis – *e.g. insufficiently or otherwise unusual extracted sample;*
- injection order related confounding effects – *e.g. lack of proper randomisation combined with unanticipated instrument drift;*
- typical 'horse shoe' shaped grouping of samples in PCA often reflects continuous gradual changes occurring in samples – *e.g. samples underwent gradual compositional (chemical) changes in an autosampler at ambient temperature.*

Sample tag and analytical sample file name:

As opposed to the sample ID in the data matrix, sample tag and/or instrument file name relating to each sample have normally detailed information attached to it which can be investigated for variants in the protocol. In addition to obvious traceability issues, these sources of information are primarily related to:

- machine calibration and tuning as well as maintenance events - *e.g. a different experiment has been analysed between two batches of samples resulting in alterations of the machine performance;*
- changes in extraction process – *e.g. use of different batch of solvent, glass ware,*
- identification of potential environmental contamination – *e.g. some m/z signals may characterize changes in laboratory environment like malfunction of air conditioning or serviced water purification system.*

Technical staff and research facilities involved:

Individuals may inadvertently introduce variability even when adhering to a standard protocol:

- examine extraction date/time – *e.g. confounding effects like sampling or extraction related to different researchers;*
- metadata related to provenance of group of samples – *e.g. block effects, where growth or sample collection site could reveal unforeseen sample behaviour or clustering requiring new experimental factors;*
- individual sample characteristic in a single location – *e.g. mislabelling, outliers*

Observational information regarding specific samples/batches:

Experimentalists probably know more about the question addressed than any other person. Their comments, even if subjective, are valuable to help explain unexpected sample variability. The list of observational information is context dependant and includes, non-exhaustively, features such as:

- "This batch of plants looks discoloured."
- "The samples were lost in post. "
- "The tubers got chilled."

8.4 PROTOCOL DESCRIBING PRE-PROCESSING, CLASSIFICATION MODELLING AND FEATURE SELECTION USING FIE-MS)METABOLITE FINGERPRINT DATA

8.4.1 Equipment

- Alternative specifications of computers used:
    - PC (Microsoft Windows 2000 or XP)
        - Pentium 4 1.70 GHz, 1.50 Gb Ram
        - Pentium 4 Dual Core 3.60 GHz, 1.00 Gb Ram
    - Mac G5 (Dual 1.8 GHz PowerPC G5, 2.00 Gb Ram)

- R, the latest release (2007-07-03) R-2.6.0.; download at http://cran.r-project.org/
- R-Packages required:
    - download following the link 'Software'/'Packages' at http://cran.r-project.org/ by clicking the appropriate 'Available Bundles and Packages':
        - Package *e1071*
        - Package *randomForest*
        - Package *MASS*
        - Package *ncdf*
        - Package *impute*
        - Package *KernSmooth*
- R-package FIEmspro and related documents:
    - download at http://users.aber.ac.uk/jhd

- Alternative Editors for computer platforms:
  - Windows (*e.g.* 'Tinn-R'; download at http://www.sciviews.org/Tinn-R/
  - MacOS X (*e.g.* 'TextWrangler'; download at http://www.barebones.com/products/textwrangler/)
- Spread-sheet program (*e.g.* MS Excel)

Choice of software packages

A wide choice of commercial data analysis software packages are available, most of which can perform adequately the majority of the analytical procedures outlined in the present protocol. The present protocol specifically provides guidance on the use of software tools that we have found particularly useful for the analysis of high dimensional metabolite fingerprint and profile data. A metabolomics data analysis package FIEmspro (http://users.aber.ac.uk/jhd) programmed in the R environment [67] has been developed to assist in the pre-processing of the raw FIE-MS fingerprints from the mass spectrometry instrument, data checking, model robustness evaluation and explanatory features mining. An web-based resource (http://www.armec.org/MetaboliteLibrary/index.html) described in Section 6 of the present G03 report and published as a Nature Protocol by Overy *et al*. 2008 [34] is available for the interrogation of signals of interest against a database of chemical standards. R is a free, open source, well documented and multi-platform environment for statistical computing. It is therefore possible to read and understand the code provided and individual users can adapt, evolve or improve the current implementation to their needs. The set of proposed methods can easily be extended with already available statistical and machine learning routines (http://cran.r-project.org/src/contrib/Views) available on the R project website (http://cran.r-Project.org/src/contrib/PACKAGES.html) and its genomic specific sister project Bioconductor (http://www.bioconductor.org).

The functions for metabolomics data processing in the FIEmspro package specifically include data matrix structure checking, baseline correction, imputation of low or missing values, outlier detection, sample classification/discrimination, feature ranking and validation of multiple classifiers. The current package is targeted to a general user who can benefit from already implemented routines for ranking features or existing classification techniques as long as their implementation satisfies general R standards for predictive modelling (readers are referred to consult the help pages of *fs.techniques* and *accest*). Several feature ranking methods are provided (e.g. t-test, Random Forest, AUC, mutual information) and different classifiers (e.g. RF, SVM, LDA) can be applied directly. FIEmspro can also be utilised by more advanced R users wanting to customise both feature ranking and discrimination methods or integrate unconventional code. Incorporation of parallelised code for few functions can relieve long processing times if access to supercomputing facilities is available. The validation mechanism has been designed in a way that classifiers can be strictly compared to each other and that resampling-based feature ranking can be performed with an identical data partitioning as for discrimination studies. Several feature ranking methods are provided such as ANOVA, Welch test, random forest, area under receiver operating curve (AUC) and mutual information. To obtain a stable feature ranking list, these methods can be validated by cross-validation or bootstrap. Users additionally can use their own feature ranking methods and validate them using the methods implemented in the package. The validation mechanism can also applied to discrimination using different classifiers, for example, RF, SVM and LDA.

8.4.2 PROCEDURE

The workflow described below represents the minimal procedure that is normally applied to a good quality data set which contains little influence from sample or method-related variance. Optional procedures (e.g. baseline correction, data infilling, normalisation) to improve data pre-processing to cope with specific commonly encountered sources of variability in larger data sets are indicated through the procedure and are addressed in section 8.5. Anticipated results with a sample workflow are shown in section 8.6.

Steps 1 to 10 relate to the checking of data integrity prior to pre-processing

CAUTION FIE-MS data structure will always reflect its origin (*e.g.* instrumentation and its local parameterisation, operator preferences and experience) and it is essential to obtain as much meta-data as possible related to the generation of the dataset before initiating pre-processing.

1. Load data into R either by reading data directly using R function *read.table* or load the saved R data format (*.RData*, *.rda*) using R function *load*. Ideally metabolomics data should be supplied via a curated database such as ArMet (http://www.armet.org/) which adheres to internationally agreed standards (*e.g.* http://www.smrsgroup.org/) to ensure data integrity and meta-data completeness.

Steps 2 to 6: Preliminary data assessment

2. Check that both FIE-MS (**X**) and experimental factor (**Y**) matrices (see Table 8.1) contain the same number of rows and correct with user if they do not. This can be performed easily using R function *dim*.
3. Check that variable names in **X** have labels corresponding to the actual nominal mass.
4. Check the number of zero values in each variable in X and remove the variable if exceeding a predefined limit. In the present example, if 80 or more values of an individual variable are missing across the set of 120 samples then the variable might be considered for removal.

CAUTION Before removing any variables from the data used for analysis it is important to check whether the patterns of missing values are random or potentially related to specific classes. If the latter seems apparent then it may be prudent to retain such variables or important data could be removed from the matrix.

5. Check that rows in both the **X** and **Y** matrices are in the same sample order.
6. Examine meta-data and assess if there might be unplanned interactions between experimental factors, injection order and analytical batches. Common examples may be poor (or no) randomisation of samples in terms of injection order or alterations to instrument set up within a batch. If potential problems are suspected then the impact of such factors should be investigated to determine whether they might prove confounding while carrying out steps 7-17 below.

Steps 7-10: Check data intensity measurements by calculating the sample Total Ion Count

7. **7.** For each sample, calculate the Total Ion Count (TIC) by summing up all signal intensities. Plot sample TIC (y-axis) against sample identifier in order of injection (x-axis).
8. **8.** Build a robust regression to model the effect of the injection order on the sample TIC by using the function *ticstats* in package FIEmspro. Investigate samples for which the residuals is 3 times higher that the median absolute deviation (MAD) of all residuals. If these samples prove to show abnormal behaviour, remove the corresponding row in both **X** and **Y** datasets (see Figure 8.2 in Anticipated Results).
9. **9.** Assess for underlying structure within the data relating to injection order (see Figure 8.3 in Anticipated Results).
10. **10.** Create a new experimental factor column in Y if an injection order related structure (*e.g.* batch effect) is identified.

CAUTION Unless samples indicated as outlying at this stage present an abnormal profile, their exceptionally high/low intensities may be blurred after normalization.

The following OPTIONS and Steps 11 to 17 relate to the pre-processing of data to counteract any experimental-related regular variance both within individual fingerprints and across the whole data set. See Section 8.5 for details.

11. Optional: Perform baseline correction and/or imputation of low/zero values following the instructions in Section 8.5.1 and 8.5.2. Please note that baseline correction of binned raw data (see Figure 8.4 in Anticipated Results) does not necessarily improve significantly data structure and, therefore, classification and feature ranking results. Baseline correction requires option ‘imputation of low/zero values’. Also note that zero or low values (see Table 8.2 and Figure 8.5 in Anticipated Results) are mainly a problem when data undergo log transformation as it leads to missing data that, in turn, is not well handled by most statistical/machine learning algorithms.

12. Perform log transformation using R function *log* directly (see Figure 8.6 for anticipated results) or set *method=”log”* in function *preproc* in FIEmspro.

13. Optional: Perform sample normalisation as described in Section 8.5.3 (and see Figure 8.7 for Anticipated Results). Please note that while normalisation of each fingerprint to their TIC separately is the preferred option, it does not necessarily improve data structure significantly in every case.

CAUTION Normalisation to sample TIC may lead to dramatic effects when the TIC intensity is correlated to the class of interest.

Steps 14 to 17: Outlier detection.

14. Perform Principal Components Analysis (PCA) for each sample class individually using function *prcomp* with the pre-processed data resulting from steps 11-13.

15. Perform outlier detection on the first 2 principal components using the FIEmspro function *outl.det*, (see Figure 8.8).

16. Investigate samples identified as potential outliers at a given threshold (*p=0.975* is adequate for most problems) and eventually remove these samples from **X** and **Y** datasets. The outliers will normally lie in the area above the distance cut-off line and outside the confidence ellipse.

17. Repeat the outlier detection procedure (steps **14 – 16**) on each of the different sample sub-groups corresponding to any experimental factors that are commonly found to be major sources of variance.

CAUTION Any outlying samples identified in PCA (or LDA, see step 18) should be investigated for potential mislabelling or typographical errors in for example run sequence lists. Any remaining samples thought to be significant outliers can be removed from the data matrix to be used for classification or feature selection tasks with the caveat that sufficient numbers of class replicates are retained to maintain a data set structure as close as possible to the original experimental design.

Steps 18 to 29: Unsupervised clustering or a supervised classification task with pre-processed data.

Steps 18 to 19: Perform unsupervised multivariate modelling.

18. Load appropriately pre-processed data (i.e. at minimum the outlying samples should be removed and the data log transformed) and perform Principal Components Analysis (PCA) using all selected sample classes using R function *prcomp*. A clustering or classification task can be performed using all the data available or can be carried out using samples corresponding to a predefined-selection of classes to address a specific biological question. This step can apply to multiple and two class problems.

19. Produce a PCA scores plot of the first two principal components (or more depending on the experimental design complexity) to investigate the origin of the main sources of variability (see Figure 8.9 in Anticipated Results).

20. Optional: Examination of PC loading plots can provide potential indications at the signal level if the natural clustering of samples cannot be explained by any of the meta-data available.

Steps 21 to 24: Supervised multivariate discrimination on the overall data.

21. Perform Linear Discriminant Analysis (LDA) using FIEmspro function *nlda*.

22. Check the results of LDA by printing out the confusion matrix and statistics, plotting the LDA scores on the first two or more dimensions and plotting the hierarchical clustering of the group centres in the LDA space if the problem comprises more than two classes (see Figure 8.10 in Anticipated Results).

23. Perform Random Forest (RF) classification using function *randomForest*, implemented in the *random Forest* package.

24. Check the results of RF classification by printing out the confusion matrix and plotting sample margin from the RF model (see Figure 8.11 in Anticipated Results).

Steps 25 to 30: Assessment of multivariate discrimination robustness.

25. Choose one of the re-sampling strategies, including leave-one-out cross-validation, cross-validation, randomised validation (holdout) and bootstrap by using the FIEmspro function *valipars*. Please note that leave-one-out cross-validation should only be used if there are insufficient class replicates to generate suitable independent training and test data sets. Randomised validation and bootstrap have been found to be particularly reliable with FIE-MS data containing a minimum of 10-18 class replicate fingerprints each containing up to 2000 variables. Please note that as data paucity is often a problem it is not always possible to construct independent test and training data sets to carry out optimal model validation.

CRITICAL Steps 25 to 31 must be repeated several times to estimate the classifier variance by setting a sufficient number of iterations while calling *accest* (usually 10-20).

26. Generate the vector (indices) of samples used for model training associated with the chosen resampling method using FIEmspro function *trainind*.

27. Perform RF and LDA analysis with the chosen resampling strategies to assess model performance by using *accest*.

28. Check results such as error rate and the confusion matrix (see Figure 8.12 in anticipated results).

29. Check average margins calculated on the left out samples (RF only) and, if a two class problem, AUC values (for any method). Please note that for two class problems an average RF margin around 0.2 is an indication that the classifier might lie on a threshold for robust interpretation, whilst models with a margin above 0.4 are generally considered good candidates for adequate generalizability and can be used with reasonable confidence for informative feature selection [12, 37]. Models with lower margins are not often robust and are likely to present difficulties for feature ranking and lack generalizability. Similarly, an AUC value > 0.9 is normally an indication of a robust (adequate) classifier [12, 37].

30. Optional: Assessment of discrimination results significance metrics and of experimental factors (see Section 8.3.7) on discrimination abilities of classification model can be performed by following the advice given in Section 8.7.

Steps 31 to 34: Applying feature ranking.

31. Perform feature ranking with all training sample replicates using several feature selection methods implemented in package FIEmspro, such as Random Forest (*fs.rf*), ANOVA (*fs.anova*) and AUC (*fs.auc*). Please note that feature ranking aims to determine which variables are significantly explanatory of biological differences between sample classes. The most generalizable models are produced and the most accurate feature rankings are obtained when the overall biological question is broken down into relevant pair-wise (binary) comparisons of classes.

32. Perform feature ranking based on re-sampling procedure by using FIEmspro function *feat.rank.re* (see Table 8.3 and Table 8.4).

33. Compute rank stability for each feature in each method by using FIEmspro function *fs.mrpval*.

34. Summarize the results by FIEmspro function *fs.summary* (see Table 8.4). Please note that in binary comparisons using Random Forest we have shown that selected features with an Importance Score of greater than 0.0025-0.003 are likely to have significant explanatory power [12, 37]. Similarly, variables with AUC values > 0.9 or FDR corrected p value (from any classical univariate statistical test) lower than 10^{-4} can be considered discriminatory in pair-wise comparisons as a first estimate [12, 37].

8.5 OPTIONS and TROUBLE SHOOTING

The optional steps included in the protocol relate firstly to situations where signal intensity values, particularly in large data sets, are suspected to contain elements of regular variance that do not relate to the biological questions under consideration and can effectively confound data mining. Such factors are discussed in detail in Section 5and can commonly include: signal baseline 'drift' within individual fingerprints (often more of a problem in negative ion data); gradual signal intensity shifts between fingerprints resulting from changes for example in ionization or detector efficiency; occasional large TIC divergence from mean in fingerprints resulting from, for example, changes in extraction efficiency of individual samples. With sufficiently large data sets such regular variance may not be an important issue, but as data paucity is often a problem it is commonplace to perform data 'normalisations' and/or transformations to counteract such features in the data during pre-processing.

A second group of options relate to performing significance tests on the initial classification models generated using default parameterisations. In high dimensional data containing elements of often subtle variance it is important to be aware that powerful multivariate modelling techniques will always strive to drive discrimination between classes and try to achieve high accuracy. Purely mathematical significance or accuracy measurements may be misleading if it cannot be certain that the model is highlighting features that are directly related to the biological questions under consideration. To develop confidence in the data therefore it is important to perform appropriate cross-validation tests to determine whether classification models are robust. There are no 'hard and fast' rules for appropriate significance metrics but, as described previously [12, 37] we suggest that the best policy is to assess model significance in relation to the biological significance of the outcome in any new type of experiment with a novel matrix

8.5.1 Baseline correction

- Select one of a few representative samples.
- Run the baseline correction algorithm *onebc* in package FIEmspro with various different window sizes and lower quantile values to evaluate the most adequate set of parameters (*wsize*=50 and *qtl*=0.1 are adequate starting points).
- Apply the baseline correction procedure on the overall data by using FIEmspro function *multibc* for the selected parameters of *wsize* and *qtl*. (see Figure 8.4 for example results).

8.5.2 Imputation of low/zero values

Perform imputation of low values if data shows high numbers of zero or very low values.

- This can be achieved either by increasing all signals to a pre-determine threshold.

CAUTION If baseline correction has been applied previously then it is possible that some variables will have negative values and so thresholding could cause problems

- Alternatively, an imputation method can be used based on a simple *k*-nearest neighbour (k-NN) algorithm implemented in FIEmspro function *koptimp*. This function assists in the determination of an optimum number of nearest samples to be considered to interpolate missing data by comparing the imputation root mean squared error (RMSE) for different values of *k* . To get a better RMSE estimate, this procedure is repeated several times. RMSE distributions at each *k* is plotted in form of a box plot. The characteristics of the graph 'elbow' (i.e. position where curve starts to level off rapidly; see Table 8.2 and Figure 8.5 in Anticipated Results) is the most convenient way to determine the optimum *k* (i.e. trade-off between under- and over-fitting).

8.5.3 Sample normalisation

- Plot the sample TIC against experimental factors of interest such as information relating to sampling day and replicates by using functions *sum* and *boxplot*.
- Perform sample normalisation corrected with the chosen experimental factor if the TIC shows significant class-dependence by using FIEmspro function *preproc* with method of *TIC* and a factor (*y*) specifying the experimental factor, otherwise perform

TIC normalisation by *preproc* without specifying the experimental factor (see Figure 8.7 for Anticipated Results).

8.5.4 Assessing effect of experimental factors on discrimination abilities of classification model

- Assess LDA and RF performances with one constrained cross validation strategy where the experimental factor under scrutiny is used to form the folds (for e.g. samples from the first analytical batch are in fold one and samples from the second analytical batch in fold 2 etc…)
- Assess LDA and RF performances with cross validation with similar stratification of the classes as above but with random split. Repeat this step 20-50 times with different random partitioning.
- Test the null hypothesis that the accuracies obtained by the constrained CV procedure do come from the same distribution of the accuracies calculated on the random data partitioning.

8.5.5 Trouble shooting in ‘R’ environment

Most classification algorithms implemented in the R environment treat the measured values of data variables and sample class information differently. The data type for the variables must be in a matrix or data frame and the class information represented as a vector or factor. If an error is encountered, the error information is prompted in the R console and it is indicated what kind of error has happened. The most likely errors relate to data format or inconsistent length in the list of variables or class information, or un-identified arguments passed into the functions.

8.6 ANTICIPATED RESULTS WITH MODEL DATA

To provide a context, most of the examples shown as anticipated results relate to the processing and analysis of FIE-MS data representing a time course (5 sampling points plus a healthy control) of *Brachypodium distachyon* (a model grass species) responding to attack by the fungal pathogen *Magnaporthae grisae.* Biological variance is relatively high in this data set and it is expected that the majority of sample classes should display significant differences in metabolome, especially after lesions become visible and fungal biomass increases. Anticipated results illustrating some of the optional steps are demonstrated in the context of other large G02 data sets either developed on a different mass spectrometry instrument where baseline drift was a problem or where class differences in metabolome were likely to be much more subtle. All the data shown was generated using FIEmspro using a sample workflow available at the following URL (http://users.aber.ac.uk/jhd).

8.6.1 Data loading

The data set used for analysis should be provided in two separate ASCII files (spaced .txt or column .csv separated as produced by Excel). The first file (X-data) consists of the measured values of each variable and the second one (Y-data) is the experimental factors or response vectors, including information such as injection order, class label and batch (see Table 8.1). The Y-file also contains information used for converting raw data *.cdf files into the X matrix using function *fiems_ltq_main* in FIEmspro (columns 'pathcdf' and 'filecdf'). The two files loaded into R should be in the format of a matrix or data frame. The columns in the first file should consist of the *m/z* (variables) and their measured intensity values, whilst the rows represent the samples. The columns of the second file contain the experimental factor information labels (or values) while the rows correspond to the same sample indices as in the X-data file. During data analysis each individual Y-data column is extracted and converted into an R factor and then passed to the analytical algorithm along with the X-data matrix (or data frame). For example, to undertake a discrimination process, we need to extract the Y-column information labelled 'class'. After loading the data into R, the consistency of data dimensions in the two data sets should be checked, for example, the number of rows in the two data sets must be the same, and otherwise the data analysis procedure will generate an error message.

8.6.2 Preliminary data assessment

When performing data analyses upon instrument raw data output it is important that the data analyst has a good understanding of the methods used to generate samples and machine runs, as well as having an understanding of key features in the experimental design (for example the number of classes within the dataset, number of replicates and any relationship between runs/samples). This is especially relevant if the data analyst was not the person who performed the experiment, which is often the case. To do this, the data analyst must thoroughly assess the accuracy of the data matrix against the metadata to determine the number of meta-classes present within the matrix, to make sure that all the variable names are correct and avoid confusion when interpreting feature ranking lists. Another source of confusion may come from the fact that the metadata information has been spuriously loaded in a different order to the fingerprint matrix. Finally, the data analyst must be aware of potential interactions between planned or unplanned experimental factors. Typically, initial experimental design and any changes to the original plan occurring during the overall data collection phase must be clearly stated before starting the data analysis process. One must also ensure that each treatment (including sample class) is fairly well randomised across the dataset, regardless of analytical batches or order of injection. In the present example this is achieved by examining the columns '**name.org**' and '**injorder**' in Table 8.1b in which it is obvious that samples from different growth trays and treatments have been injected in a random order.

8.6.3 Data intensity checking using sample Total Ion Count (TIC)

By definition the total ion count of a spectrum is the sum of all the *m/z* signal intensities. As an easy diagnostic measure, the TIC can provide an estimation of factors that may affect the overall intensity of the run such as gradual instrument drift (e.g. resulting from loss of sensitivity of the ion source), or step changes in instrument characteristics after maintenance. Also, an examination of the TIC can reveal suspicious samples where unusually low or especially high signal intensities in some runs may be due to contamination or poorly extracted samples.

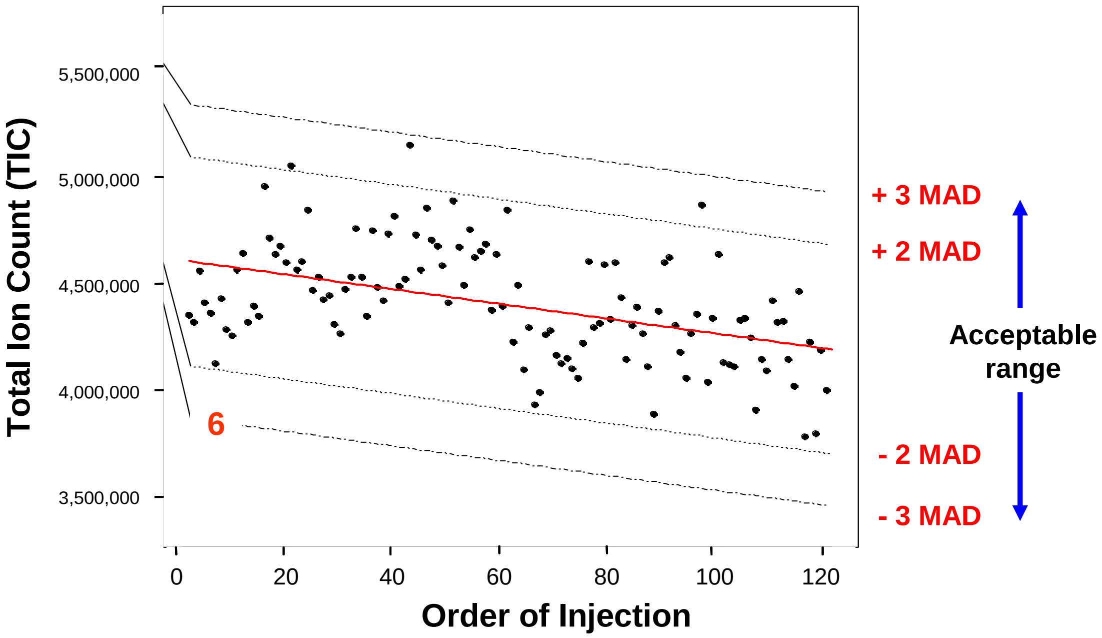


Figure 8.2 TIC checking for potential outlying samples in FIE-MS data using FIEmspro function *ticstats*. Plot of total ion count against order of injection of samples in FIE-MS data representing a time course of *Brachypodium distachyon* infected with the rice blast fungus. The red line indicates the fitted regression model. Potential outlying samples are indicated as red numbers based on a TIC outside of an acceptable range of +/-3 median absolute deviation of the residuals from the fitted model (MAD).

A robust regression can be built to model the effect of the injection order on the TIC of each sample as shown in Figure 8.2 by using FIEmspro function *ticstats*. As a conservative rule, any sample for which the residual deviates more than 2 to 3 times from the median absolute deviation (MAD) of the residuals (*e.g.* sample 6 in Figure 8.2) must be examined manually to identify the origin of the different intensity behaviour and then potentially removed before further statistical analysis if corrective measures do not improve the individual fingerprint. Further assessment of outlying samples is discussed later. In any case where a linear relationship (i.e. gradually changing TIC in sample set) is observed between the injection order and sample TIC, this dependency will be removed by TIC normalisation. If other structure related to the order of injection is noticed, for example an “analytical batch” effect (i.e. a step change in TIC at beginning or in middle of an injection series as shown in Figure 8.3), the user must identify its potential origin (*e.g.* changes in machine calibration,operator or mobile phase) and finally create a new experimental factor (batch) where each step change in level corresponds to the start of a new batch.

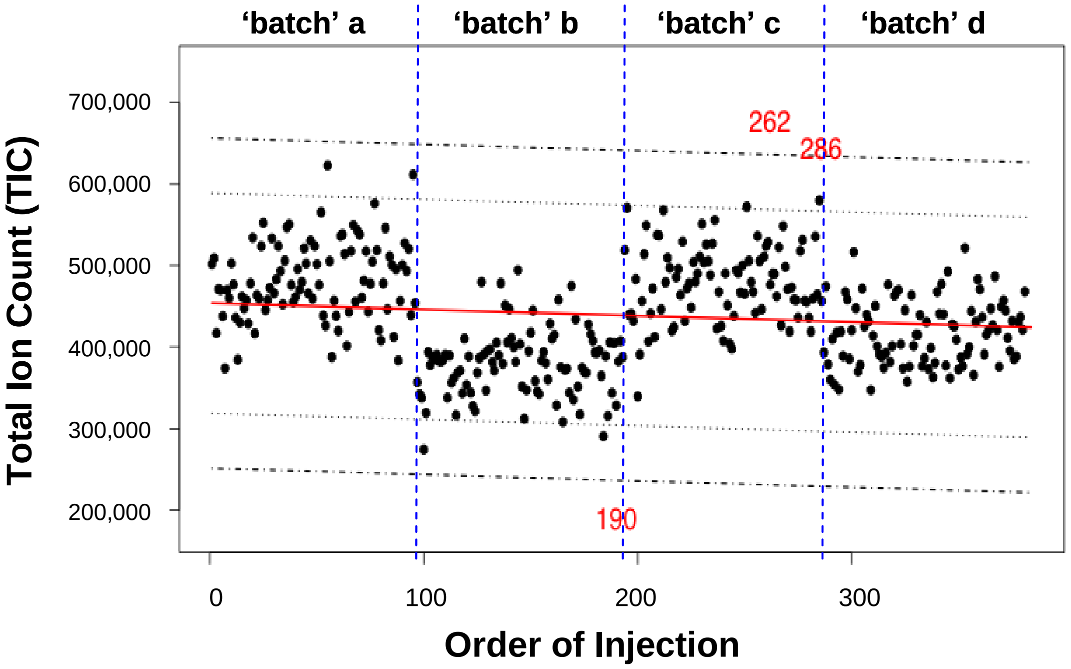


Figure 8.3 TIC checking for potential batch effects in FIE-MS data using FIEmspro function *ticstats*. Plot of total ion count against order of injection of samples in FIE-MS data representing 5 different potato cultivars from the G02 2001 FIE-MS data set analysed in 4 batches over a period of 6 months. The red line indicates the fitted regression model. Blue dotted lines indicate positions in injection sequence where experimental batch variance is evident. Potential outlying samples are indicated as red numbers based on a TIC outside of an acceptable range of +/-3 median absolute deviation of the residuals from the fitted model (MAD).

8.6.4 Baseline correction

Despite preliminary efforts in fingerprint pre-processing (i.e. removing sample background and residual electronic noise), a *post hoc* baseline correction may be necessary to remove systematic and *m/z*-dependent artefacts caused by 'chemical noise'. Chemical noise is matrix dependant and attributable to clusters of ionized matrix molecules hitting the detector during early portions of the FIE-MS infusion and/or detector overload from large molecules. Typical examples displaying baseline problems include samples dominated by large molecules such as soluble long chain carbohydrates or high molecular weight complex lipids. One possible consequence of baseline drift includes the possibility that the baseline may become discriminatory if it appears to be characteristic of specific classes or subgroups of samples from the same class. A second possibility is that important information may be obscured in areas affected by baseline problems. Rapid diagnostics for initial evaluation of baseline bias include the examination of the PCA loading plots and simple non-parametric testing using the raw data to check whether the related statistics are also *m/z* dependent.

As in other contexts (e.g. proteomics), there is not a general and adequate solution accepted to solve a baseline problem and more research is required to understand baseline behaviour. A simple consensual approach consists in fitting a monotone local minimum curve to each fingerprint by using FIEmspro functions *onebc* and *multibc*. Basically, the fingerprint is divided into equally spaced *m/z* intervals and a local minimum intensity value is returned as the baseline estimate for this region. Finally, the whole fingerprint baseline is computed by linear interpolation based on pairs made of the centre of the interval and its corresponding local minima (Figure 8.4). Intervals are in the order of 30-70 amu as a trade off between the removal of relevant chemical (small interval) or under-estimation bias when using larger intervals. Rather than using the minimum value of an interval, it is also judicious to use the value corresponding to a low quantile (typically probability of 0.1) to avoid any spurious estimates due to zeros or abnormally low signals. Baseline corrected fingerprints can be examined visually to determine the effect of this corrective procedure (green plot in Figure 8.4) and the improvement of the data quality assessed by plotting the AUC of the original raw data against the baseline corrected AUC to determine the quality of fit characteristics at higher intensities. Due to the inherent nature of FIE-MS profiles, earlier and later parts of the fingerprint may be excluded to avoid any downward bias of the baseline estimation because of many consecutive zeros.

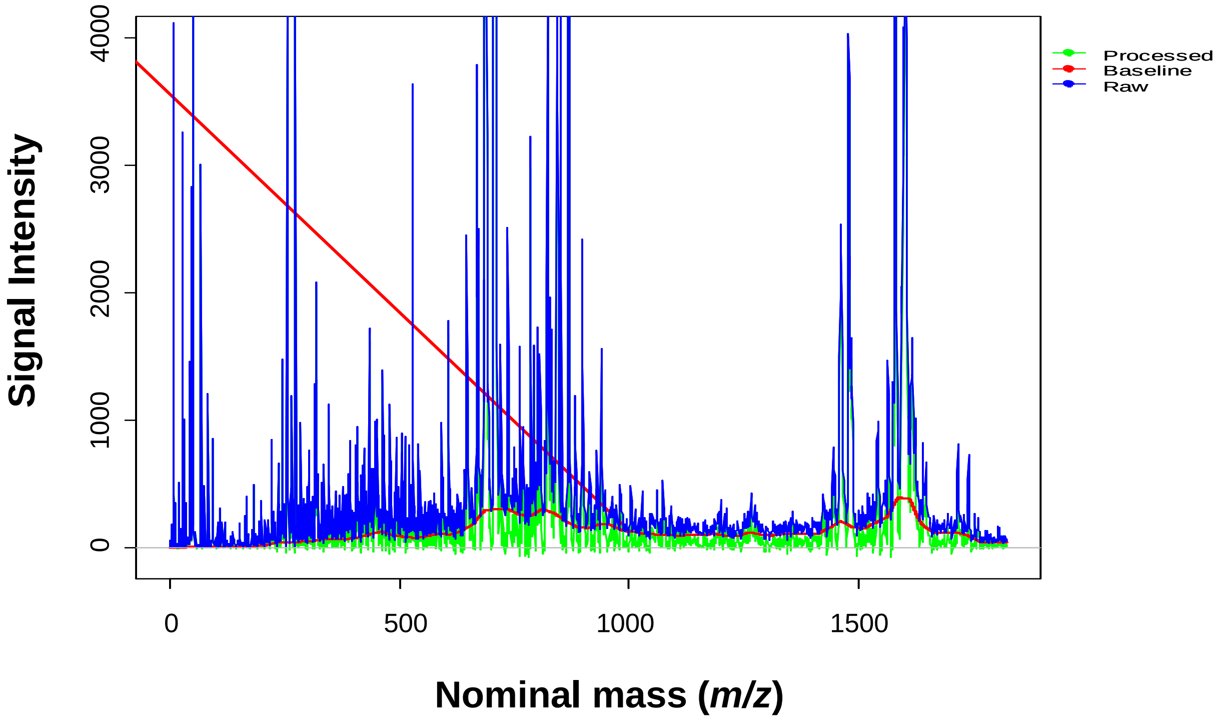
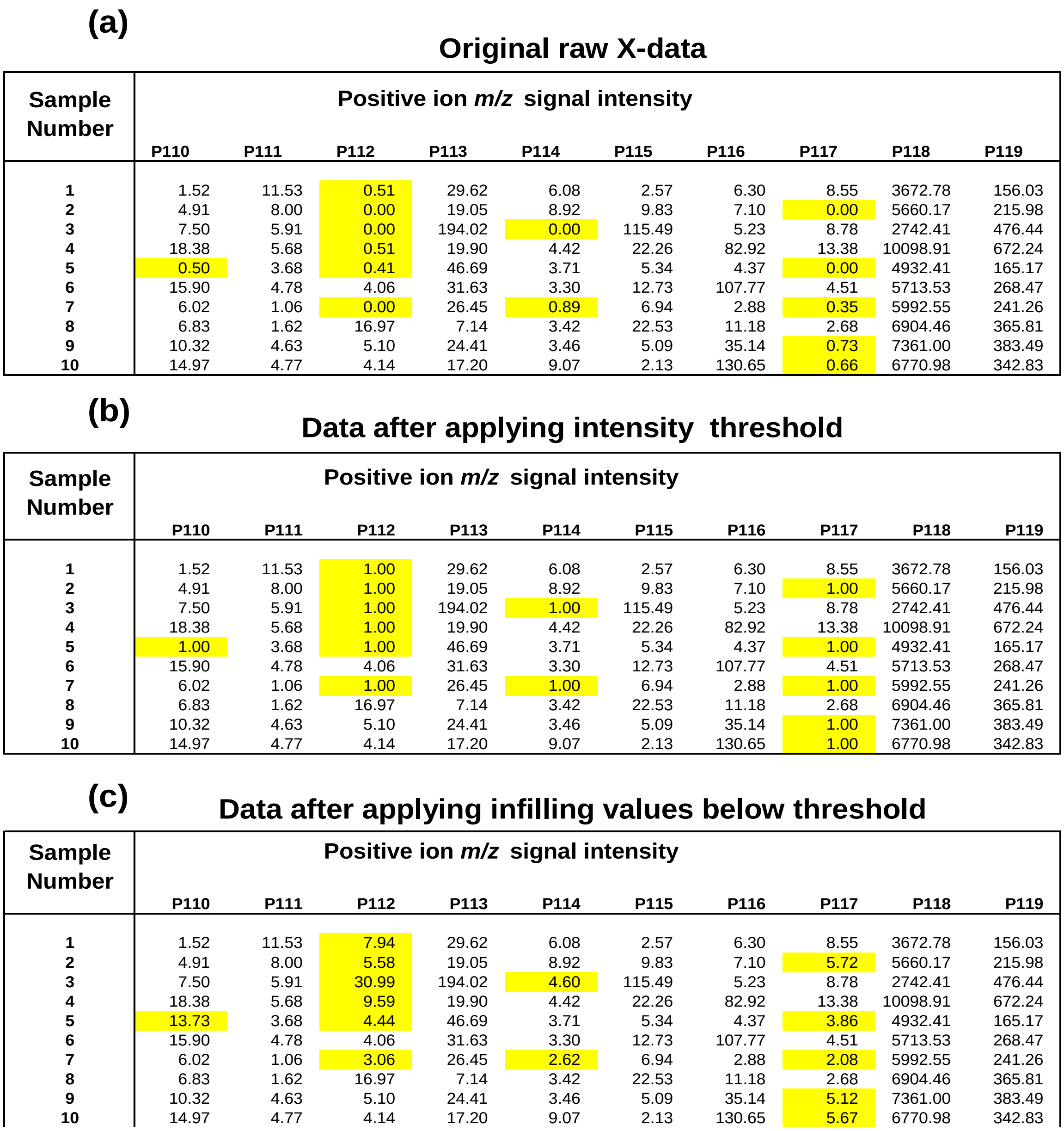


Figure 8.4 Use of baseline correction to improve FIE-MS fingerprint data representing pathogen challenged *Brachypodium distachyon* plants. The different colours indicate signals before (blue) and after (green) baseline correction using the FIEmspro function *multibc*. The original baseline is coloured red.

8.6.5 Imputation of low and negative intensity signals

Some variables may exhibit zero values (or even negative values after baseline correction is applied, see Figure 8.4); this effect is often predominant in the early and later phases of an individual infusion profile (see Figure 5.1 in Section 5). In most instances, variables that exhibit a majority (e.g. 70%) of zero values across the whole data set should be removed before further analysis. For the remaining signals in FIE-MS data, zero values should typically concern no more than a maximum of roughly 3% of the total number of cells in the matrix (Table 8.2a).



(a)

Original raw X-data

Sample Number	Positive ion *m/z* signal intensity									
	P110	P111	P112	P113	P114	P115	P116	P117	P118	P119
1	1.52	11.53	0.51	29.62	6.08	2.57	6.30	8.55	3672.78	156.03
2	4.91	8.00	0.00	19.05	8.92	9.83	7.10	0.00	5660.17	215.98
3	7.50	5.91	0.00	194.02	0.00	115.49	5.23	8.78	2742.41	476.44
4	18.38	5.68	0.51	19.90	4.42	22.26	82.92	13.38	10098.91	672.24
5	0.50	3.68	0.41	46.69	3.71	5.34	4.37	0.00	4932.41	165.17
6	15.90	4.78	4.06	31.63	3.30	12.73	107.77	4.51	5713.53	268.47
7	6.02	1.06	0.00	26.45	0.89	6.94	2.88	0.35	5992.55	241.26
8	6.83	1.62	16.97	7.14	3.42	22.53	11.18	2.68	6904.46	365.81
9	10.32	4.63	5.10	24.41	3.46	5.09	35.14	0.73	7361.00	383.49
10	14.97	4.77	4.14	17.20	9.07	2.13	130.65	0.66	6770.98	342.83

(b)

Data after applying intensity threshold

Sample Number	Positive ion *m/z* signal intensity									
	P110	P111	P112	P113	P114	P115	P116	P117	P118	P119
1	1.52	11.53	1.00	29.62	6.08	2.57	6.30	8.55	3672.78	156.03
2	4.91	8.00	1.00	19.05	8.92	9.83	7.10	1.00	5660.17	215.98
3	7.50	5.91	1.00	194.02	1.00	115.49	5.23	8.78	2742.41	476.44
4	18.38	5.68	1.00	19.90	4.42	22.26	82.92	13.38	10098.91	672.24
5	1.00	3.68	1.00	46.69	3.71	5.34	4.37	1.00	4932.41	165.17
6	15.90	4.78	4.06	31.63	3.30	12.73	107.77	4.51	5713.53	268.47
7	6.02	1.06	1.00	26.45	1.00	6.94	2.88	1.00	5992.55	241.26
8	6.83	1.62	16.97	7.14	3.42	22.53	11.18	2.68	6904.46	365.81
9	10.32	4.63	5.10	24.41	3.46	5.09	35.14	1.00	7361.00	383.49
10	14.97	4.77	4.14	17.20	9.07	2.13	130.65	1.00	6770.98	342.83

(c)

Data after applying infilling values below threshold

Sample Number	Positive ion *m/z* signal intensity									
	P110	P111	P112	P113	P114	P115	P116	P117	P118	P119
1	1.52	11.53	7.94	29.62	6.08	2.57	6.30	8.55	3672.78	156.03
2	4.91	8.00	5.58	19.05	8.92	9.83	7.10	5.72	5660.17	215.98
3	7.50	5.91	30.99	194.02	4.60	115.49	5.23	8.78	2742.41	476.44
4	18.38	5.68	9.59	19.90	4.42	22.26	82.92	13.38	10098.91	672.24
5	13.73	3.68	4.44	46.69	3.71	5.34	4.37	3.86	4932.41	165.17
6	15.90	4.78	4.06	31.63	3.30	12.73	107.77	4.51	5713.53	268.47
7	6.02	1.06	3.06	26.45	2.62	6.94	2.88	2.08	5992.55	241.26
8	6.83	1.62	16.97	7.14	3.42	22.53	11.18	2.68	6904.46	365.81
9	10.32	4.63	5.10	24.41	3.46	5.09	35.14	5.12	7361.00	383.49
10	14.97	4.77	4.14	17.20	9.07	2.13	130.65	5.67	6770.98	342.83

Table 8.2 Dealing with zero values and low intensity signals in FIE-MS data

A conservative strategy to improve data quality consists of calculating an intensity value representing a "limit of detection" signal to ‘in fill’ cells with missing values to avoid the effect of zero values skewing further data analysis (Table 8.2b). This being said, quite often imputation of very small values is more of a computational convenience (very few multivariate techniques can intrinsically accommodate missing data points) than a significant contribution to the improvement of model predictive abilities. Obviously it is important that this treatment of the data should also not remove nor create any new spurious knowledge. A value of 1 could be judicious to use at it is a rough estimate of the lowest "machine error";

however this will also bound further log transformation to 0. Further increasingly elaborate techniques can be applied to treat the missing value problem in a more formal way. It is often the case that the mechanism underlying an

apparently random pattern of missing values cannot be identified clearly and could be caused by a combination of technical reasons (detection threshold or sample preparation) and inherent properties of metabolomic profiling (molecular complexity, dynamism of metabolic networks and other experimental variations). Several techniques have been proposed in the literature to handle this issue which consider the presence of empty cells

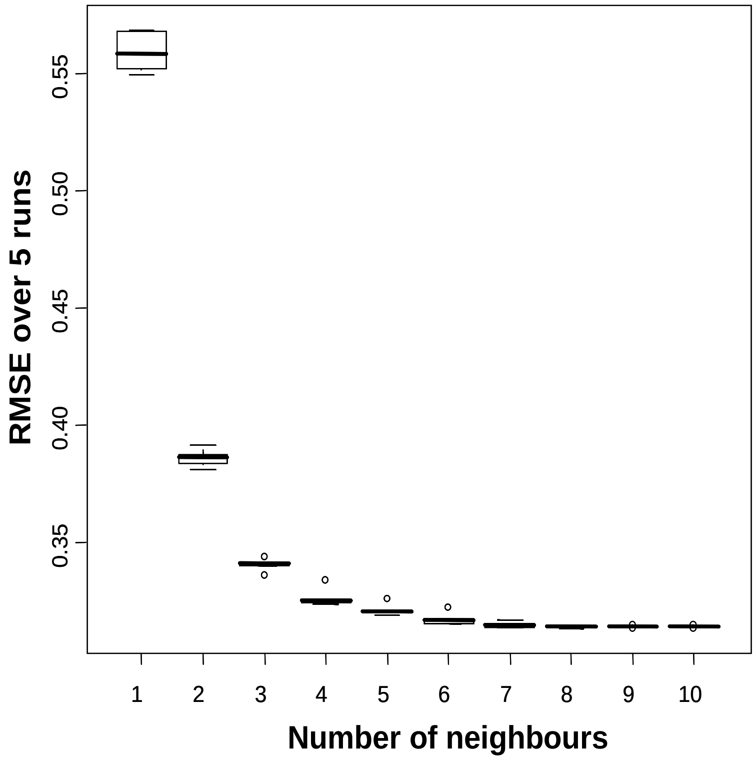


Figure 8.5. Evaluation of an optimum number of k with the FIEmspro function *koptimp*. Box plot depicting the imputation error (RMSE) at various number of (k) neighbors used for the imputation for several random attributions of missing values. The RMSE corresponds to difference between the imputed value and the true intensity of data points selected randomly. Characteristics of the graph 'elbow' is the most convenient way to determine an optimum number of (k) neighbors (i.e. trade-off between under and overfitting). In this case the adequate number of k is 6.

to be caused by data "missing at random" (MAR), which is probably the case in most situations. In the present protocol, we have adopted a non-parametric *k*-nearest neighbours based technique to impute low/missing information. The imputation relies on interpolating a missing data point *xij* (sample *i*, variable *j*) from the the *k* nearest samples to sample *i* that do not contained any missing information in their column *j*. To determine the efficiency of the method to obtain reasonable solutions, one can evaluate the error between imputed values obtained by artificially assigning random missing cells in the matrix and their original true values (Figure 8.5).

The FIEmspro function *koptimp* is a wrapper function to determine this error for several values of *k* and several random assignment. As shown in Figure 8.5, we vary the number of the nearest neighbors from 1 to 10 and each *k* is executed 10 times. From the results of the root mean squared errors (RMSEs) presented we can determine the optimal number of nearest neighbors to be 6 in the present case. This neighbour values is applied to the final imputation of missing or low values in FIE-MS data. Recently, a comprehensive R package *pcaMethods* (http://bioconductor.org/packages/2.0/bioc/html/pcaMethods.html) has been released to cover various PCA-based algorithms for imputing missing data in transcriptomics data and can optionally be applied with similar performance.

8.6.6 Logarithmic transformation of data

Metabolites are present at a huge dynamic range of concentrations in samples and thus FIE-MS data reflects this situation by the fact that signal variance appears apparently much greater for *m/z* typically exhibiting higher intensities when all signals are measured on the same scale in the raw data (Figure 8.6a). For purposes of statistical analysis, the logarithmic transformation of signal intensity can be made in order to alleviate the dependency of the variance with the intensity. This family of transformations seems to be a consensus step in many reported mass spectrometry and metabolomics applications despite a lack of background proof that it is always helpful in every case. Essentially, log transformation converts multiplicative errors into additive errors so that, ideally, variance becomes more constant across the range of signal intensity on the logarithm scale and thus variance is stabilised (Figure 8.6b).

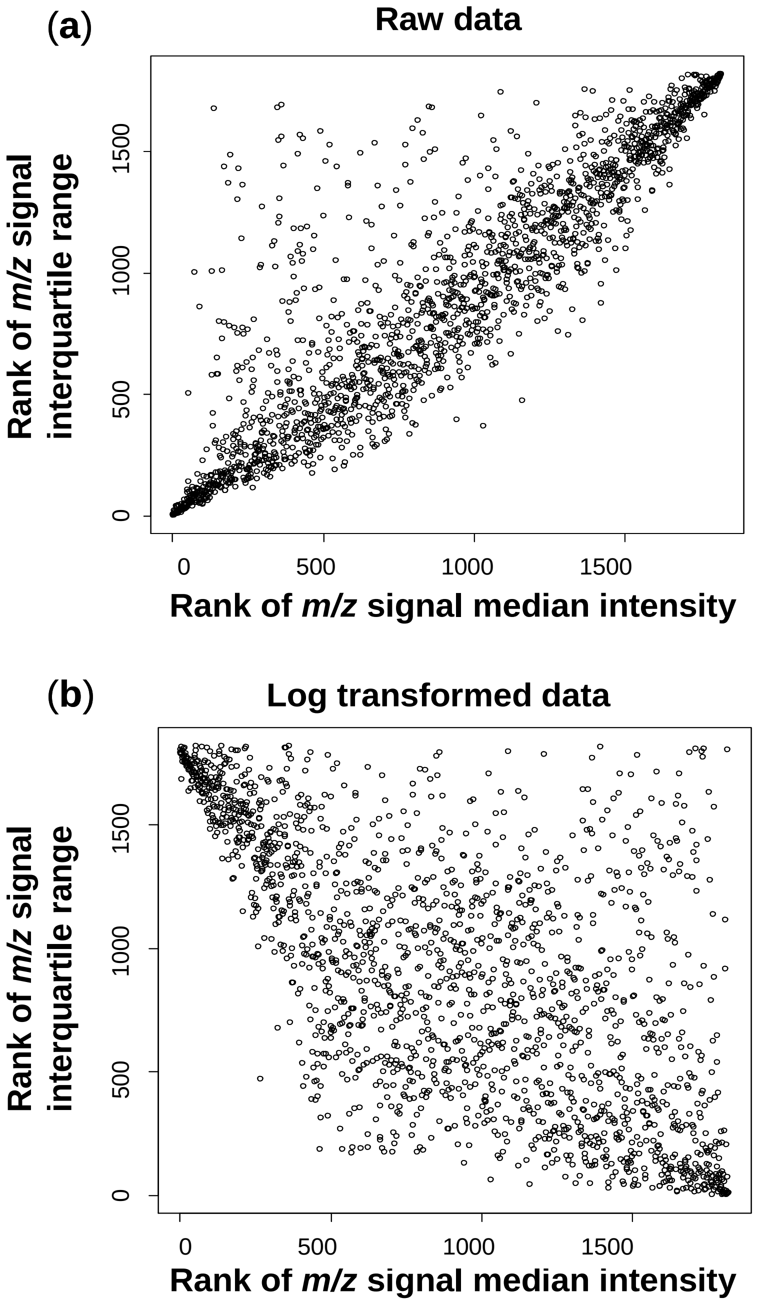


Figure 8.6. Effect of log transformation on the signal variance to signal intensity dependency within FIE fingerprints representing the metabolome of *Brachypodium distachyon* challenged with the rice blast fungus. (a) Plot of rank of the interquartile range versus the rank of the median of each *m/z* signals shows a clear dependency between signal intensity and degree of variance. (b) Prior data log-transformation partially removes this dependency.

Additional variance stabilisation models can be used such as generalised log transformation, however they require parameter optimisation. One drawback of log transformation is that it makes the visualisation difficult and harder to differentiate noise from background. For some statistical techniques such as PCA, there is a basic choice as to whether one uses the data as they come or whether to reduce each signal to unit variance. The former strategy somewhat over-weights the influence of the major signals, while the latter is prone to giving excessive influence to variables that contribute a lot of (or mainly) noise. Log transformation of the variables can therefore be an optimal strategy between the two: our default approach is to first take the logarithm of the signals and not reduce to unit variance further.

8.6.7 Data normalisation between samples

Purpose of normalisation is to remove inter-spectrum sources of variability that come mainly from different sample concentration, loss of sensitivity of the detector over time or degradation of certain samples. One common way to normalise metabolomics profile data is the use of one or more internal standards of "known" concentration, which is not possible in FIE-MS fingerprints and so the most widely used solution consists of a so called global normalisation by rescaling each measurement within a spectrum by a constant factor, such as the sum of all the spectra intensities (Total Ion Count) as shown in Figure 8.7 (a) and (b). This method of TIC normalisation is often systematically used in a blind way and probably adequate to achieve a more or less substantial improvement of model quality in most situations. However, by nature, TIC normalisation also has an impact on non-parametric statistical strategies in contrast to commonly used monotonous transformations such as log transformation or range scaling that will only affect parametric techniques. An assumption behind such normalization approaches is that the overall intensity captures a sort of average of both up and down changes in concentration related to biological treatment and assumes that these potentially informative changes are negligible regarding to the overall experimental variability (so that the TIC is dominated by the "stable" features). In addition to introducing relationships between variables that do not normally exist, the default use of TIC normalisation can lead to the generation of spurious knowledge (false positives) if the overall sample intensity is class/factor dependent; this treatment can therefore sometime consequently bias discrimination abilities in data modelling. In such a situation, compensation for TIC class dependency can be performed by removing the difference between the average TIC of the corresponding class and the average TIC calculated from every samples. Figure 8.7 (c) and (d) illustrate a better concordance of raw data AUC with class corrected normalised data and the effects on TIC range.

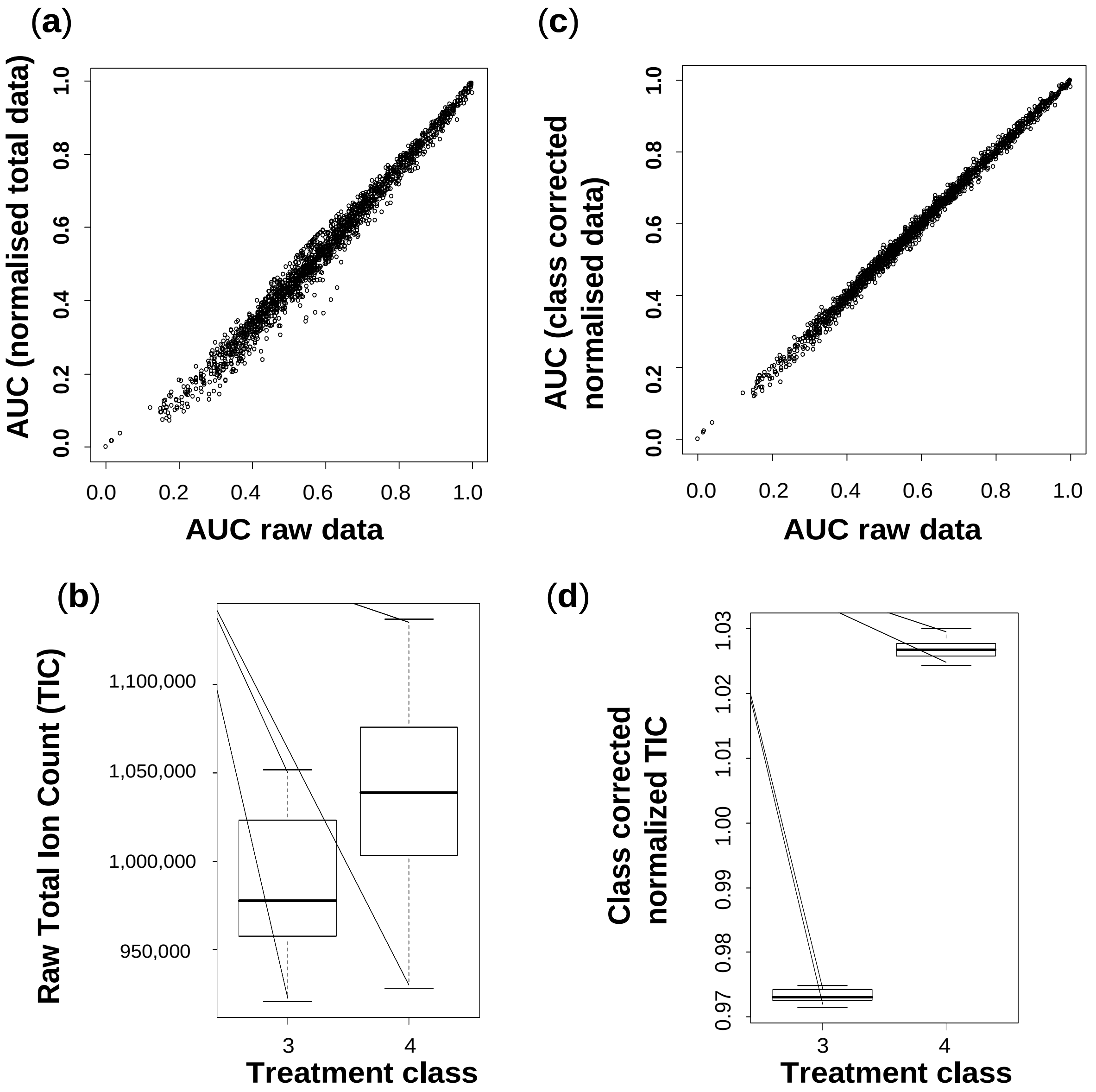


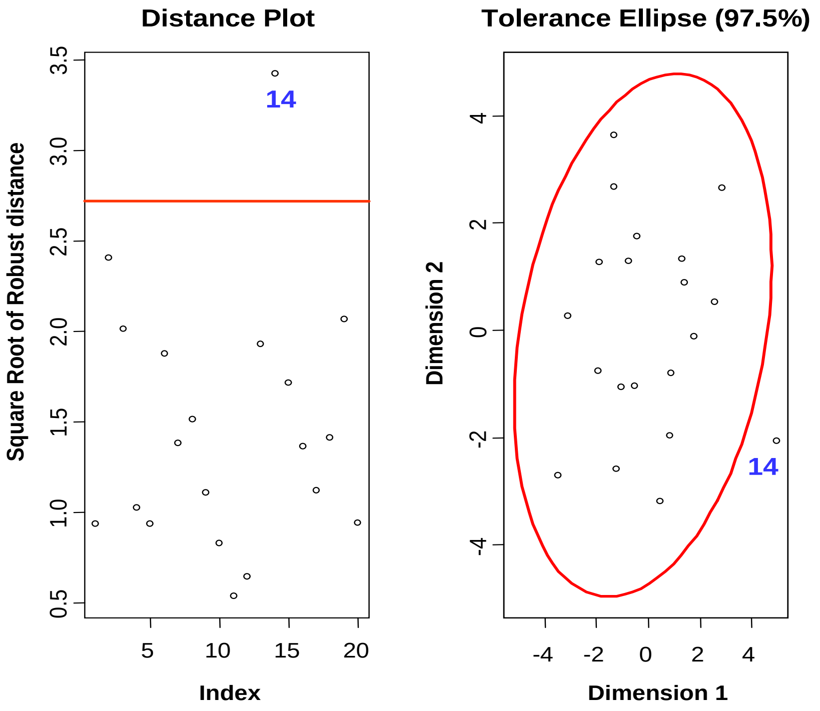
Figure 8.7. Effect of TIC normalisation on fingerprint data representing *Brachypodium distachyon* leaves after either 3 or 4 days infection with rice blast fungus. (a) Plot of *m/z* signal AUC calculated using the raw data and AUC calculated with the TIC normalised data. (b) TIC distributions in both classes. (c) Plot of *m/z* signal AUC calculated from the raw data and AUC calculated with the class corrected TIC normalised data. (d) Distributions of sample sum of the intensities in both classes after class corrected TIC normalisation. (c) and (d) illustrate that the information resulting from 4 days of infection is kept in the overall intensity and that no spurious knowledge is created as opposed to the global TIC normalisation shown in (a).

8.6.8 Sample outlier detection

Despite an initial inspection of distribution of sample TIC, subsequent outliers may not be detected unless they exhibit gross intensity differences from the mean. To our knowledge outlier detection in a metabolomics context is mentioned in various proposed data analysis workflows but no explicit studies have been described in detail because both detection and interpretation of outliers in high dimensional space is a rather difficult and ill-defined problem. Due to inherent constraints associated with the dual problems of small sample size and dimensionality, metabolomics data becomes so sparse so that every sample can be considered as outlying under some circumstances.

Application of conventional univariate techniques to detect outliers is not realistic as we end up removing most of the samples. More importantly, outliers can also encapsulate important information related to instrument artefacts, substructures in the data, important biological behaviour or simply mislabelling. In such circumstances it is thus useful to have access to metadata so that possible reasons for deviating from a "normal" behaviour can be assessed as well as corrected for if possible. Rather than concentrating on individual variables that could potentially lead to the rejection of all the samples, outlier detection can be approached from a different angle that consists of looking at the homogeneity of the samples in the main directions of variance. Standard methods are based on calculating (robust) Mahalanobis distances between each sample and the data centre in a given reduced space such as that derived from Principal Component Analysis (PCA). PCA produce fewer directions of uncorrelated dimensions (called Principal Components) with decreasing variance so that each PC will explain different dimensions in the data that could be related to a different source of variability. Due to the inherent difficulties in defining outliers, inclusion of the first few dimensions only is almost always sufficient to compute Mahalanobis distances. However in more complex designs implicating various factors and/or multiple levels, different contributions to the overall variation modelled by PCA may be confounded in such a reduced space. In such situation, the initial dataset must be decomposed into smaller problems to analyse potential reasons for outlier behaviour. Finally, when an outlier sample is suspected, a rapid examination of the loadings must be performed to help to explain the underlying origin of the deviant behaviour. In our experience in such samples are often highlighted in a series of correlated signals corresponding to sample contamination (e.g. ‘splash over’ from other sample classes during vacuum drying or contamination from laboratory chemicals or plasticware).

In order to avoid removal of important biological information and sample size reduction, a loose significance threshold or confidence level (normally 0.975 or 0.99) is employed to determine a cut-off value of the Mahalanobis distances. Any samples whose Mahalanobis distances are larger than the cut-off are identified as the outliers. As an example shown in Figure 8.8, the fingerprint data of *Brachypodium distachyon* samples in class 3 are analysed by a PCA and the first two principal components are used for outlier detection using FIEmspro *outl.det*. In this instance sample 14 is identified as potential outliers under the confidence level of 0.975. The same sample is also highlighted in the right panel of Figure 8.8, where the outlier is located outside of the 2-dimension tolerance ellipse plot whilst all other samples sit inside the ellipse.



Fi **esenting *Brachypodium distachyon* leaves after 3 days infection with rice blast fungus using Principal Components Analysis (*pccomp*) and FIEmspro function *outl.det*.** (a) Scatter plot of sample Mahalanobis distances and (b) scores plot of the first two Principal Components including tolerance ellipse used to evaluate outlying samples. Outliers, such as sample 14 (in blue) are explicitly highlighted in both graphs. Mahalanobis distances encapsulate both distance to the centre of first two main directions of variability and the dispersion of this reduced space. Assuming that Mahalanobis distances follows a Chi-squared distribution, the ellipse of confidence and distance threshold can be determined (drawn in red).

8.6.9 Assessment of the natural relationships between sample classes in multivariate space

Unsupervised modelling procedures, such as Principal Components Analysis, are often referred to generically as clustering methods and allow a quick inspection of the underlying total data structure in an experiment according to the main directions of variation. As a result of both inherent complexity of FIE-MS fingerprints (50-2000 Da range) and sophisticated and/or multifactorial experimental designs, unsupervised techniques will produce overlapping class 'groups' from which the distinction of explanatory signals responsible for class differences is rather difficult. However, techniques like PCA may outline specific sample similarity relationships in situations where sample class labels do not necessarily relate to metabolomic phenotype and are thus useful for class discovery; any samples grouping together can be objectively considered to be similar. In Figure 8.9, the first two scores plots derived from a Principal Components Analysis illustrates that sample grouping is related to the infection time course of *Brachypodium distachyon* challenged with the rice blast fungus. In general each class grouping is relatively fuzzy, reflecting the large degree of biological variance in the experiment. Samples representing healthy leaves and samples taken 24hours after infection overlap considerably. The PC1 dimension reflects disease progression with samples representing the later time points in the experiment lying increasingly to the left of the plot.

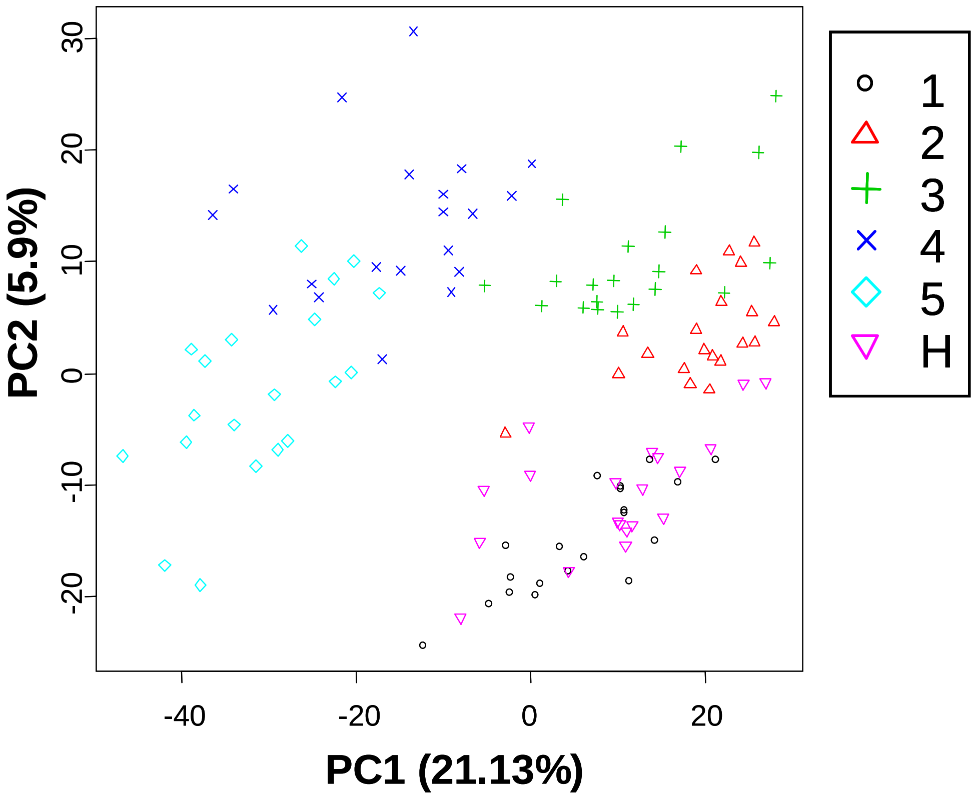
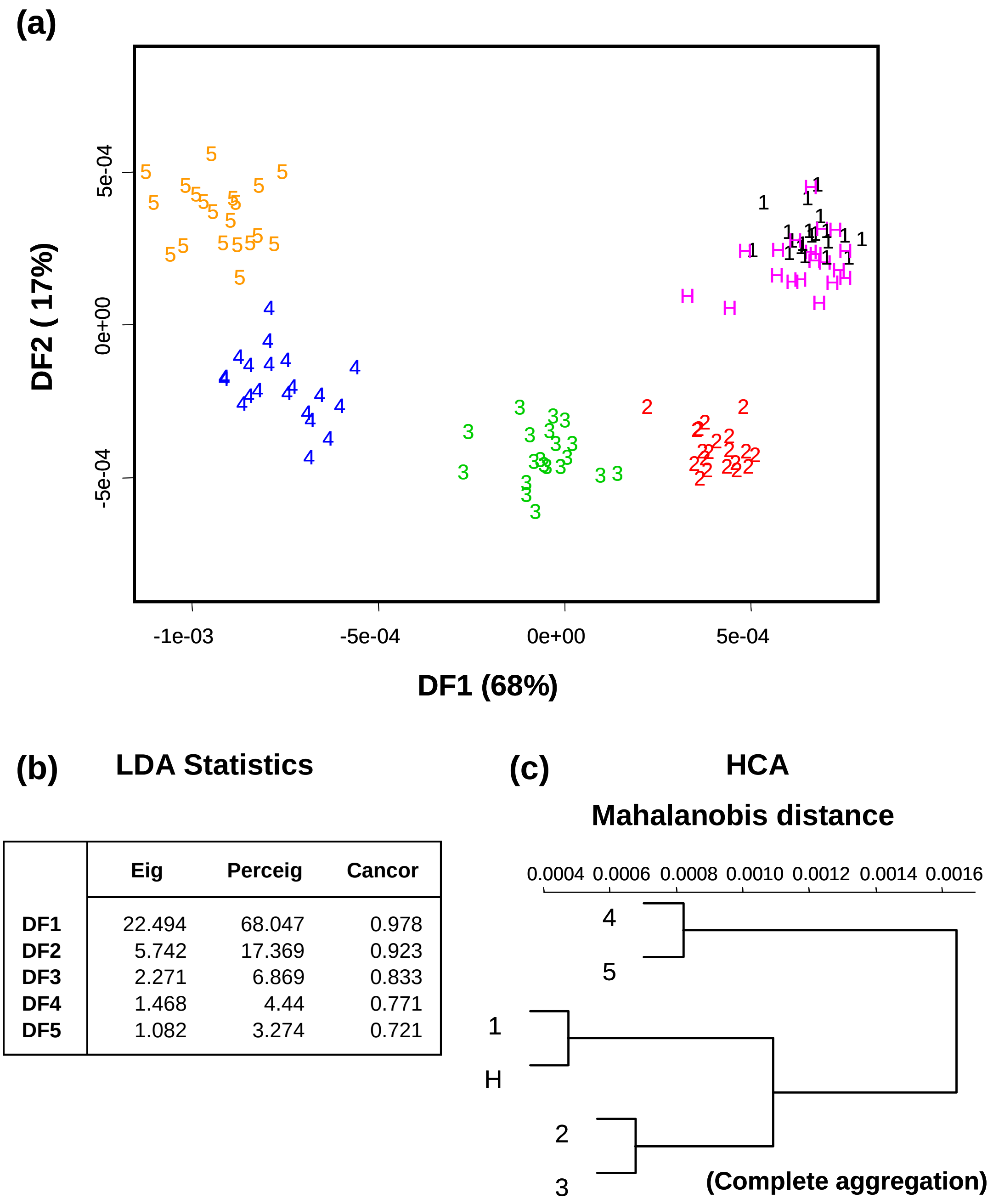


Figure 8.9 Principal Components Analysis (PCA) of FIE-MS fingerprints representing a time course of *Brachypodium distachyon* infected with the rice blast fungus. All the samples are mapped on the first 2 major dimensions of variability (PC1 and PC2). H = healthy leaves and the numbers 1-5 relate to days post-infection. The variability encapsulated in each direction is respectively 21 and 6 %.

8.6.10 Multivariate discrimination of sample classes

In typical metabolomics studies involving designated case and control subjects, a discrimination task is often a binary classification problem, whereas those involving more than two classes, such as an analysis of many plant genotypes or an examination of different phases of disease progression are known as multi-class problems in the machine learning literature. Despite the fact that the machine learning techniques used in this protocol can cope with multi-class problems, decomposition of the general problem into a series of pairwise (binary) classifiers may have advantages in many situations. The construction of multi-class models usually requires optimisation of complex decision boundaries that may become difficult to interpret and necessitate powerful algorithms to avoid inefficiency associated with a large number of output classes. Generally, such classifiers do not exploit natural groupings of classes at the price of classifier robustness and transferability. In addition to generating more understandable conclusions using common two class algorithms, the decision boundaries are usually simpler to interpret in binary problems. As opposed to a global feature selection performed using all groups together, local feature selection discriminating two classes can result in selection of more explanatory variables and a better association between features and class groupings. From a biological point of view, the decomposition of a complex experiment into pairwise problems may be more expressive and sensible than considering all class together (for example one may be interested in comparing a control group against each case subtype, rather than differences between case subtypes). Identification of meaningful grouping from a multi-class problem can simply be approached by looking at the confusion matrix obtained by cross-validation for example or by investigating the way group centres are organised in LDA space. At a certain cost, computation of all possible binary classifiers will highlight pairs of groups that are closely related. For example, combination of binary models margin and non-linear mapping has been successfully applied to summarize genotypic relatedness in a large functional genomics problem [12, 37].

Figure 8.10 illustrates the multivariate discrimination of *Brachypodium distachyon* infection time course data using the FIEmspro function *nlda* for linear discriminant analysis (LDA). Figure 8.10(a) shows a plot of the LDA scores of the all six classes in



(b) LDA Statistics

	Eig	Perceig	Cancor
DF1	22.494	68.047	0.978
DF2	5.742	17.369	0.923
DF3	2.271	6.869	0.833
DF4	1.468	4.44	0.771
DF5	1.082	3.274	0.721

(c) HCA

Mahalanobis distance

(Complete aggregation)

Figure 8.10. Linear Discriminant Analysis (LDA) and Hierarchical Cluster Analysis (HCA) of FIE-MS representing disease progression in *Brachypodium distachyon* plants infected with the rice blast fungus. (a) Mapping of the samples on the two main axis of between group variation (Discriminant Function, DF). (b) LDA model statistics for the 5 DFs (**Eig**, eigenvalue; **Perceig**, percentage of explained between group variance; **Cancor**, canonical correlation. (c) HCA using complete aggregation of the class centers in the LDA space.

the first two linear discriminant functions (DFs). The classes 2, 3, 4, and 5 have a wide discriminant boundary while classes 1 and H are overlapping. Statistical measures derived from the LDA model such as eigenvalues, percentage of variance explained and canonical correlations for each discriminant function are given in Figure 8.10 (b). In such analyses Eigen values greater than 2 (equivalent to canonical correlation greater than 0.8) in any discriminant dimension are generally associated with adequate separability. This illustration also shows that the first two DFs dominate the analysis carried out by the LDA. Figure 8.10(c) illustrates a hierarchical cluster analysis (HCA) based on group means of the projected data in the LDA space. It clearly provides insight into group relatedness, all six classes can be classified as two sub-groups, one of which is divided into further two sub-sub-groups.

Figure 8.11(a) is a confusion matrix giving the results of a classification analysis using the Random Forest (RF) algorithm in which 20 samples of each sample class were provided and analysed using method *randomForest*. The number of correct class predictions (highlighted in yellow) show that classification accuracy is very high with major confusion occurring only between healthy leaves (class H) and infected leaves prior to symptom formation 24 hours (class 1) after challenge with pathogen.

(a) RF Confusion Matrix

		Predicted Class						Classification
		1	2	3	4	5	H	error
True Class	1	18	0	0	0	0	2	0.100
	2	0	19	1	0	0	0	0.050
	3	0	2	18	0	0	0	0.100
	4	0	0	0	20	0	0	0.000
	5	0	0	0	1	19	0	0.050
	H	3	0	0	0	0	17	0.150

(b) RF Margins

Binary comparison	Margin
1_H	0.117
2_5	0.860
3_4	0.631

Figure 8.11. Examples of Random Forest statistics derived from classification of FIE-MS fingerprint data representing *Brachypodium distachyon* plants during a time course of infection with rice blast. (a) Confusion matrix representing the predicted class according to the actual class of a set of samples. Numbers in the matrix diagonal (highlighted in yellow) correspond to the number correct predictions for each class. (b) Typical RF margin results in three pair-wise (binary) comparisons.

Classification accuracy is the most generic way to approach model generalizability but is not always adapted to examine the prospects for efficient mining of explanatory variables. In essence, a high accuracy can be obtained at one extreme with very simple classifiers containing only 4 or 5 significant variables, but it is possible that an equally accurate prediction can be made where classifiers are extremely complex and rather opaque to further interpretation. The average sample margin computed from the votes of the left out samples (later described as margin solely) value is defined as the difference between the score for the true class and the largest score of the rest of the classes (i.e. the most probable misclassification). The margin values are thus a better indication of how robust and potentially generalizable any classifier might be. Figure 8.11(b) illustrates margin values for 3 binary comparisons of *Brachypodium* data and as expected the margins between healthy leaves and the 24h post-infection time point are < 0.2, whereas other comparisons have much higher values. Classifiers with margins above 0.3-0.4 are generally adequate for accurate feature selection in most binary comparisons using a similar experimental design (~ 2000 *m/z,* 10-15 replicates).

Whenever possible it is important to validate the classification results by use of a training and test set where there are sufficient numbers of sample replicates. To obtain an unbiased accuracy estimation, a stratified 5-fold cross-validation can be applied and the process repeat 10 times. Figure 8.12 (a) illustrates an aggregated summary of Random Forest confusion matrices after cross validation with just 10 repetitions (more may be required in other experimental situations). In our experience the statistical output of several classification algorithms (typically RF, LDA and SVM) should be checked for concordance and to reinforce our confidence in concluding about the validity of the experiment and technology to answer the biological question. As a minimum default we would usually inspect accuracies and AUC statistics for at least RF and LDA that generally show good correspondence (Figure 8.12b).

(a) RF Confusion Matrix

		Predicted Class					
		1	2	3	4	5	H
True Class	1	174	0	0	0	0	26
	2	0	190	10	0	0	0
	3	0	18	176	6	0	0
	4	0	0	0	200	0	0
	5	0	0	0	11	189	0
	H	43	0	0	0	0	157

(b) Validated model statistics

Binary comparisons	Random Forest		LDA	
	Acc.	AUC	Acc.	AUC
1_H	0.828	0.899	0.773	0.868
2_5	1.000	1.000	0.988	1.000
3_4	0.990	1.000	0.978	1.000

Figure 8.12. Validation of discrimination models using FIE-MS fingerprint data that describe disease progression in *Brachypodium distachyon* infected with the rice blast fungus. (a) Aggregated summary of the Random Forest (RF) confusion matrix following stratified 5x cross validation, repeated 10 times. This shows that errors are mainly made between classes H and 1 as well as between infected leaves 2 and 3 days post-inoculation. (b) Comparative overview of the average accuracies (Acc.) and AUCs in RF and LDA models when discriminating samples from 3 and 4 days post-inoculation.

8.6.11 Assessing discrimination results

Deciding on the effectiveness of the model to describe the biological problem at hand deserves careful consideration. Firstly, accuracies and related performance measures obtained on various and different classifiers must converge to similar conclusions. Secondly, the observed error rate can be variable in data sets with only a small number of sample replicates and so the extent of the classifier performance variability is also important to assess in order to derive adequate conclusions regarding model generalizability. Commonly, a definition of model significance is formulated by comparing the statistics under consideration to a null hypothesis stating that this measure is not relevant to the problem under investigation. Unless perfect classifiers are sought for predictive purposes, the highest performances must be obtained and discussed in parallel with the biological relevance of the findings. A significance 'threshold' statistic can be evaluated by comparing two sample classes that are known biologically to display no significant difference at the level of the metabolome.[12] This is not a feature of the present data derived from an analysis of disease progression in infected *Brachypodium* leaves where sample classes are generally well discriminated. Thus principles relating to the assessment of significance threshold are examining in Section 8.7 below with different data sets.

Assessing results of experiments involving field-grown crops is expected to be more difficult than assessing well-controlled experiments involving clear metabolic differences. Empirically for a two class problem, accuracies and AUC greater than 0.9, margins above 0.3 and LDA eigenvalues greater than 2 may be sufficient to accept a model. For models that do not fulfil these criteria further computations are necessary to investigate model properties at lower thresholds. It is crucial to assess the potential influence of experimental factors other than the class of interest, on the classifier performance. As an example in a typical case-control experiment, the design may also include additional metadata information such as field plot or experimental conditions such as the lab origin. It is therefore of interest to check whether one of those factors could affect the characteristics of the model describing the initial outcome under study: can I correctly predict genotype class of one subpopulation (e.g. potato tubers grown in one season) from a model that uses the rest of the population (e.g. potato tubers grown in a different growing season)? This can be answered by constraining the cross validation in such a way so that every sample coming from the same sub-population are gathered together in the same fold (leave-one-class-out cross validation [LOCO-CV] as opposed to random leave-several-samples-out cross validation). Inspection of the confusion matrix can reveal sub-groups badly described by the model. Additional random partitioning is performed with identical data stratification to test whether the accuracy calculated with constrained LOCO-CV falls within the limits of a random cross validation.

8.6.12 Feature ranking

The goal of many metabolomics experiments is to identify metabolite signals responsible for differences between experimental treatments or between classes (explanatory variables). This can be a rather difficult task and obviously requires the use of data analysis methods that explicitly display variables in the final model. This can be as contrasted with so called 'black box' techniques that could excel in predictive tasks but fail to let the user know which features are being used for class discrimination [2, 11, 51]. In addition, FIE-MS data tend to display significant variance and if only small replicate numbers are available then it is difficult to be certain that any variables highlighted by a specific feature ranking method have significant explanatory power. A pragmatic solution to this problem is to use several feature ranking methods that use rather different algorithms and then derive a consensus ranking of all variables. Following this the validation of any feature ranking can be undertaken by looking at the stability of individual features across multiple ranked lists by resampling the original training data.

Table 8.3 displays the feature ranking results (using feature *feat.rank.re* in FIEmspro) in comparisons of leaf samples from *Brachypodium distachyon* taken either 3 or 4 days after infection with rice blast. Potentially explanatory variables (*m/z*) are listed in rank order after Random Forest, AUC or ANOVA analyses. Colour coding of individual variables reveals that, although the rank order of potentially explanatory variables is not identical, it is clear that the list of top 20-30 variables selected by each test are the same. The RF Importance Scores in this analysis are plotted against variable rank and compared with those derived from two further binary comparisons in Figure 8.13.

Feature Rank	Random Forest		Area under Curve		Analysis of Varience	
	m/z	RF-Importance Score (mean)	*m/z*	AUC value (mean)	*m/z*	ANOVA F-value (mean)
1	P749	0.00727	P749	1.00000	P366	79.60337
2	P162	0.00569	P541	0.99739	P203	69.28016
3	P541	0.00563	P355	0.99480	P213	66.74535
4	P911	0.00553	P162	0.99421	P749	65.88820
5	P750	0.00489	P750	0.99182	P367	64.85521
6	P395	0.00453	P911	0.99165	P162	59.19472
7	P877	0.00436	P395	0.99137	P365	57.99994
8	P367	0.00411	P403	0.98681	P403	57.87106
9	P366	0.00404	P366	0.98366	P215	53.50559
10	P403	0.00393	P751	0.98202	P750	53.35232
11	P355	0.00366	P203	0.98149	P538	52.87783
12	P213	0.00350	P213	0.98141	P537	52.75824
13	P193	0.00345	P338	0.98060	P541	50.90727
14	P203	0.00337	P877	0.97980	P949	45.93575
15	P716	0.00327	P385	0.97945	P525	44.90099
16	P219	0.00324	P591	0.97825	P338	43.87574
17	P538	0.00322	P538	0.97801	P911	42.35207
18	P591	0.00321	P367	0.97690	P591	41.47723
19	P338	0.00313	P193	0.97670	P354	41.42447
20	P215	0.00307	P537	0.97509	P395	40.85095
21	P933	0.00302	P707	0.97239	P355	40.69740
22	P751	0.00288	P716	0.97200	P219	39.60786
23	P365	0.00285	P354	0.96824	P707	37.81833
24	P1823	0.00272	P1701	0.96693	P699	37.31595
25	P537	0.00261	P1735	0.96623	P385	35.66861
26	P385	0.00260	P219	0.96312	P716	35.63186
27	P1701	0.00235	P949	0.96225	P878	34.93570
28	P588	0.00210	P365	0.96157	P877	33.47345
29	P949	0.00206	P895	0.96103	P356	33.20625
30	P356	0.00196	P215	0.95938	P185	32.88024
31	P699	0.00193	P525	0.95662	P384	32.39959
32	P1688	0.00191	P699	0.95492	P178	32.31093
33	P354	0.00187	P356	0.95434	P1701	31.32736
34	P1686	0.00179	P161	0.95354	P373	31.30342
35	P525	0.00176	P1700	0.95283	P895	31.27754
36	P161	0.00175	P588	0.95160	P540	30.92370
37	P931	0.00171	P946	0.95061	P942	30.84320
38	P384	0.00169	P1702	0.94821	P715	30.07023
39	P909	0.00165	P185	0.94641	P193	29.94777
40	P1702	0.00163	P627	0.94511	P1729	29.85122
	(….)					
270	P381	0.00014	P233	0.84535	P397	12.56789
271	P708	0.00014	P901	0.84527	P923	12.54686
272	P698	0.00014	P1744	0.84519	P423	12.53922
273	P555	0.00014	P1707	0.84448	P1711	12.45882
274	P676	0.00014	P443	0.84422	P1697	12.45676
275	P1925	0.00014	P827	0.84407	P1537	12.43650
	(….)					

Table 8.3 Feature ranking in binary comparisons using RF Importance Scores, AUC and ANOVA. Signals ranks are calculated with Random Forest, AUC and ANOVA strategies from leaf samples taken either 3 or 4 days after infection of *Brachypodium distachyon* with rice blast. Potentially explanatory variables (*m/z*) are shown in rank order together with statistics after analysis with. The top 20 variables in each list are colour coded and cross referenced to other lists to allow visualisation of ranking by different techniques.

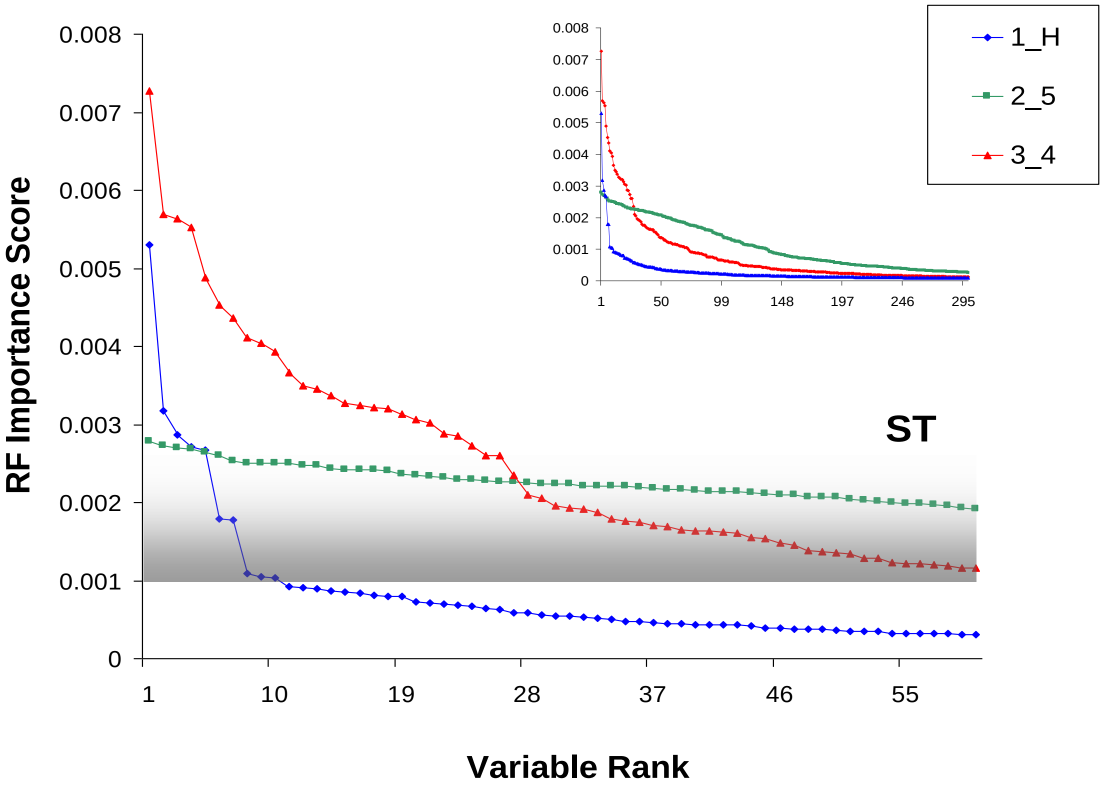


Figure 8.13. Relationship between Random Forest Importance Score and variable ranking for explanatory value in binary comparisons of *Brachypodium distachyon* leaves taken at different times after infection with the rice blast fungus. ST = a significance threshold for explanatory variables. The shaded box below ST indicates a range of variables that still may have potentially significant explanatory value (with decreasing likelihood) and which may also be worth deeper investigation. Inset shows an extended ranking demonstrating that values fall to near zero in all cases.

These data indicate clearly that healthy leaves display few (< 5) detectable differences in metabolome compared to leaves 24h post-inoculation (H_1). Comparing leaves just before and after the appearance of visible symptoms (3_4) reveals a much larger number of potentially explanatory variables (~ 25-40), whilst an extreme comparison of heavily diseased leaves with a pre-symptomatic phase of pathogen establishment (2_5) highlighted a large number (~ 60) of potentially explanatory metabolome signals. In such models comparing samples with extremely different metabolomes it is not unusual for a large number of variables to 'compete' for high ranking, which tends to reduce the maximum mean importance scores. A significance threshold for explanatory variables is difficult to discern without further validation and the shaded box indicates an importance score range of signals that still may have potentially significant explanatory value (with decreasing likelihood as shading become darker) and which may also be worth deeper investigation. One should be aware that lowering such thresholds would not only increase substantially the number of putative compounds but also augment significantly the possibility of false positive signal to enter subsequent interpretation. The inset shows an extended ranking demonstrating that values fall to near zero in all cases.

8.6.13 Validation of feature ranking results

The process of deriving a ranked list of potentially explanatory variables from FIE-MS data must be treated with extra care for several reasons. In the context of FIE-MS fingerprints, appropriate corrections must be implemented to take into account that up to a couple of thousand hypotheses are being tested. Amongst other approaches, the False Discovery Rate (FDR) [68, 69] approach appears to be a suitable practice in our context to correct *p*-values. By definition, the FDR of a set of prognosis represents the expected proportion of false predictions in this set. An updated quantity, q-value, is directly computed from the original *p*-value, obtained by a classic statistical test, to describe the "expected proportion of false positives incurred when calling that feature significant" [68, 69]. For example, if the statistical test returns 20 signals at a q-value of 0.1, 18 *m/z* variables are expected to be correct. Both underlying behaviour of metabolic systems and FIE-MS fingerprints are quite complex, variable and essentially poorly understood and the SFR produces mathematical models with many free parameters so that no statistical method can be effective under all circumstances. It is thus recommended to derive conclusions using several techniques preferably differing by their mechanism and underlying assumptions.

Potentially Explanatory Variables	Statistics from Welch t-test with 100 bootstraps					
	fs.welch	pval	pval.ad	**AvgRk**	SdevRk	**mrpval**
P236	5.691427	0.000002	0.003731	**11.80**	32.1986	**0.000249**
P1812	4.890290	0.000025	0.009203	**13.29**	16.3363	**0.000322**
P141	5.073186	0.000011	0.006429	**18.08**	30.0275	**0.000564**
P274	5.282550	0.000006	0.005019	**21.28**	36.8511	**0.000776**
P257	4.890105	0.000020	0.009056	**25.12**	39.2522	**0.001083**
P371	4.358766	0.000096	0.029163	**30.92**	43.4194	**0.001449**
P232	4.451570	0.000128	0.033324	**33.39**	46.5090	**0.001698**
P615	4.129826	0.000222	0.050568	**49.22**	57.9107	**0.003419**
P305	4.177756	0.000255	0.051641	**57.74**	98.4975	**0.004378**
P210	3.837525	0.000499	0.075716	**71.80**	90.7021	**0.006266**
P187	3.740230	0.000700	0.090997	**74.26**	92.8030	**0.006676**
P1930	3.972184	0.000309	0.054765	**75.73**	83.4347	**0.006896**
P489	3.642667	0.001198	0.121218	**84.10**	84.1333	**0.008455**
P717	3.952605	0.000331	0.054765	**86.95**	109.9155	**0.008843**
P712	3.778591	0.000563	0.078823	**90.46**	111.7649	**0.009627**
P1008	3.646543	0.000806	0.097855	**97.09**	123.5465	**0.011032**
P633	3.653903	0.000996	0.113384	**99.79**	139.5846	**0.011398**
P1642	3.215622	0.002786	0.174955	**128.25**	147.3454	**0.017870**
P1039	3.488691	0.001566	0.142549	**132.02**	204.4971	**0.018785**
P536	3.265035	0.002327	0.165567	**138.13**	158.1812	**0.020227**
P438	3.321025	0.001990	0.156919	**140.84**	179.7139	**0.020783**
P685	3.394884	0.001696	0.147095	**142.90**	181.8547	**0.021274**
P472	3.512672	0.001163	0.121218	**143.99**	204.4315	**0.021501**
P166	3.184337	0.003587	0.187629	**145.80**	211.5667	**0.021999**

Table 8.4 Assessing the significance of feature ranking by re-sampling using a Welch t-test with bootstrap analysis

Table 8.4 gives the results of a feature ranking validation analysis using a Welch t-test including a bootstrap re-sampling method (using features *fs.mrpval*.and *fs.summary* in FIEmspro). In this example the features were ranked for explanatory value for the problem of discriminating healthy leaves from infected leaves 24 hours after challenge of *Brachypodium distachyon* with the rice blast pathogen. The first three columns are the Welch t-values, p-values and adjusted p-values by FDR for each variable using all data sets. The last columns are the resampling-based results of the average feature rank, standard derivation of feature rank and modified r-test p-values (mrpval) [40]. The curve shown in Figure 8.13 (H_1) demonstrates that the RF Importance Scores dropped extremely, leaving rapidly only the top five variables above the previously suggested nominal significance boundary of 0.0025-0.003 [12]. The data in Table 8.4 concur with these observations and suggest that a threshold value of *mrpval* of around 0.001 may provide an adequate guide for variable significance in this instance.

REFERENCES

1. Somorjai, R.L., Dolenko, B. & Baumgartner, R. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics* **19,** 1484-1491 (2003).
2. Berrar, D., Bradbury, I. & Dubitzky, W. Avoiding model selection bias in small-sample genomic datasets. *Bioinformatics* **22,** 1245-50 (2006).
3. Braga-Neto, U.M. & Dougherty, E.R. Is cross-validation valid for small-sample microarray classification? *Bioinformatics* **20,** 374-380 (2004).
4. Lyons-Weiler, J. et al. Assessing the Statistical Significance of the Achieved Classification Error of Classifiers Constructed using Serum Peptide Profiles, and a Prescription for Random Sampling Repeated Studies for Massive High-Throughput Genomic and Proteomic Studies. *Cancer Informatics* **1**, 53-77 (2005).
5. Broadhurst, D.I. & Kell, D.B. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* **2**, 171-196 (2006).
6. Saghatelian, A. & Cravatt, B.F. Global strategies to integrate the proteome and metabolome. *Current Opinion in Chemical Biology* **9**, 62-68 (2005).
7. Ein-Dor L., Zuk O. & Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Nat. Acad Sci USA* **103**, 5923–5928 (2006).
8. Dıaz-Uriarte, R. Supervised methods with genomic data: a review and cautionary view. *Data analysis and visualization in genomics and proteomics.* Wiley, New York**,** pp193-214 (2005).
9. Fawcett, T. *ROC graphs: Notes and practical considerations for data mining researchers*. Technical report HPL-2003-4. HP Laboratories, Palo Alto, CA, USA. Available at http://www.hpl.hp.com/techreports/ 2003/HPL-2003-4.pdf. (2003).
10. Mukherjee, S., Roberts, S.J. & van der Laan, M.J. Data-adaptive test statistics for microarray data *Bioinformatics* **21**, 108-114 (2005).
11. Sima, C. & Dougherty, E. R. What should be expected from feature selection in small-sample settings**.** *Bioinformatics* **22**, 2430-2436 (2006).
12. Enot, D.P., Beckmann, M., Overy, D. & Draper, J. Predicting interpretability of metabolome models based on behavior, putative identity, and biological relevance of explanatory signals. *Proc Natl Acad Sci U S A* **103**, 14865-14870 (2006).
13. Kell, D.B., Darby, R.M. & Draper, J. Genomic computing. Explanatory analysis of plant expression profiling data using machine learning. *Plant Physiology* **126**, 943-951 (2001).
14. Catchpole, G.S. et al. Hierarchical metabolomics demonstrates substantial compositional similarity between genetically modified and conventional potato crops. *Proc Natl Acad Sci U S A* **102**, 14458-14462 (2005).
15. Goodacre, R., Vaidyanathan, S., Dunn, W.B., Harrigan, G.G. & Kell, D.B. Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends in Biotechnology* **22**, 245-252 (2004).
16. Bino, R.J. et al. Potential of metabolomics as a functional genomics tool. *Trends in Plant Science* **9**, 418-425 (2004).
17. Fiehn, O. et al. Metabolite profiling for plant functional genomics. *Nature Biotechnology* **18**, 1157-1161 (2000).
18. Sumner, L.W., Mendes, P. & Dixon, R.A. Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry* **62**, 817-836 (2003).
19. Nicholson, J.K. & Wilson, I.D. Understanding 'global' systems biology: Metabonomics and the continuum of metabolism. *Nature Reviews Drug Discovery* **2**, 668-676 (2003).
20. Roessner, U., Wagner, C., Kopka, J., Trethewey, R.N. & Willmitzer, L. Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant Journal* **23**, 131-142 (2000).
21. Tolstikov, V.V. & Fiehn, O. Analysis of highly polar compounds of plant origin: Combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. *Analytical Biochemistry* **301**, 298-307 (2002).
22. Beckmann, M., Enot, D.P., Overy, D.P. & Draper, J. Representation, comparison and interpretation of metabolome fingerprint data for total composition analysis and quality trait investigation in potato cultivars. *J. Agricultural and Food Chemistry* **55**, 3444-3451 (2007).
23. Dear, G.J., James, A.D. & Sarda, S. Ultra-performance liquid chromatography coupled to linear ion trap mass spectrometry for the identification of drug metabolites in biological samples. *Rapid Communications in Mass Spectrometry* **20**, 1351-1360 (2006).

24. Wagner, C., Sefkow, M. & Kopka, J. Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. *Phytochemistry* **62**, 887-900 (2003).

25. Jonsson, P., Gullberg, J., Nordstrom, A., Kusano, M., Kowalczyk, M., Sjostrom, M. and Moritz, T. A strategy for identifying differences in large series of metabolomic samples analyzed by GC/MS. *Anal Chem* **76**, 1738-45 (2004).

26. Vorst, O., de Vos, C.H.R., Lommen, A., Staps, R.V., Visser, R.V., Bino, R. J., and Hall, R.D. A non-directed approach to the differential analysis of multiple LC–MS-derived metabolic profiles. *Metabolomics* **1,** 169-180 (2005).

27. Ward, J.L., Harris, C., Lewis, J. & Beale, M.H. Assessment of H-1 NMR spectroscopy and multivariate analysis as a technique for metabolite fingerprinting of Arabidopsis thaliana. *Phytochemistry* **62**, 949-957 (2003).

28. Allen, J. et al. High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nature Biotechnology* **21**, 692 - 696 (2003).

29. Scholz, M., Gatzek, S., Sterling, A., Fiehn, O. & Selbig, J. Metabolite fingerprinting: detecting biological features by independent component analysis. *Bioinformatics* **20**, 2447-2454 (2004).

30. Aharoni, A. et al. Nontargeted metabolome analysis by use of Fourier Transform Ion Cyclotron Mass Spectrometry. *Omics* **6**, 217-234 (2002).

31. Smedsgaard, J. & Frisvad, J.C. Using direct electrospray mass spectrometry in taxonomy and secondary metabolite profiling of crude fungal extracts. *Journal of Microbiological Methods* **25**, 5-17 (1996).

32. Dunn, W.B., Bailey, N.J. & Johnson, H.E. Measuring the metabolome: current analytical technologies. *Analyst* **130,** 606-625 (2005).

33. Beckmann, M., Parker, D., Enot, D.P., Duval, E. & Draper, J. High throughput, non-targeted metabolite fingerprinting using nominal mass Flow Injection Electrospray Mass Spectrometry**.** *Nature Protocols* **3**, (2008).

34. Overy, D.P., Enot, D.P., Tailliart, K., Jenkins, H., Parker, D., Beckmann, M. & Draper, J. Explanatory signal interpretation and metabolite identification strategies for nominal mass FIE-MS metabolite fingerprints. *Nature Protocols* **3,** 471-485 (2008).

35. Parker, D., Beckmann, M., Enot, D.P., Overy, D.P., Rios, Z.C., Gilbert, M., Talbot, N. & Draper, J. Rice blast infection of *Brachypodium distachyon* as a model system to study dynamic host/pathogen interactions. *Nature Protocols* 3, 435-445 (2008).

36. Enot, D.P., Beckmann, M. & Draper, J. Detecting a difference - assessing generalisability when modelling metabolome fingerprint data in longer term studies of genetically modified plants. *Metabolomics* **3**, 335-347 (2007).

37. Enot, D.P. & Draper, J. Statistical measures for testing substantial equivalence of GM plant genotypes in a multivariate context. *Metabolomics* **3**, 349-355 (2007).

38. Jain, A.K., Murty, M.N. & Flynn, P.J. Data clustering: a review. *ACM Computing Surveys (CSUR)* **31,** 264-323 (1999).

39. Manly, B.F.J. *Multivariate Statistical Methods: A Primer*. Chapman & Hall/CRC (2004).

40. Zhang, C., Lu, X. & Zhang, X. Significance of Gene Ranking for Classification of Microarray Samples. *EEE/ACM Transactions on Computational Biology and Bioinformatics* **3**, 312-320 (2006).

41. Ransohoff, D.F. Rules of evidence for cancer molecular-marker discovery and validation. *Nature Reviews Cancer* **4**, 309-313 (2004).

42. Davis, C.A., Gerick, F., Hintermair, V., Friedel, C.C., Fundel, K., Küffner, R. & Zimmer, R. Reliable gene signatures for microarray classification: assessment of stability and performance. *Bioinformatics* **22**, 2356-2363 (2006)**.**

43. Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K. & Zhao, H. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics* **19,** 1636-1643 (2003).

44. Cristianini, N. & Shawe-Taylor, J. *'An introduction to support vector machines and other kernel-based learning methods'*, Cambridge University Press (2000).

45. Zhu, C., Kitagawa, H. & Faloutsos, C. Example-based outlier detection for high dimensional datasets. *IPSJ Digital Courie*r **1**, 234-243 (2005)

46. Craig, A., Cloarec, O., Holmes, E., Nicholson, J.K & Lindon, J.C. Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Anal Chem* **78**, 2262-2267 (2006).

47. Hastie, T., Tibshirani, R. & Friedman, J. *The elements of statistical learning: data mining, inference, and prediction*. Springer (2001).

48. Good, P. *Permutation tests: A practical guide to resampling methods for testing hypotheses*. Springer series in statistics (2000).

49. Efron, B. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association* **78,** 316-331 (1983).

50. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **21,** 3940-3941 (2005).

51. Fu, W.J., Carroll, R.J. & Wang, S. Estimating misclassification error with small samples via bootstrap cross-validation. *Bioinformatics* **21,** 1979-1986 (2005).

52. Thomaz, C.E., Boardman, J.P., Hill, D.L.G., Hajnal, J.V., Edwards, D.D., Rutherford, M.A., Gillies, D.F. & Rueckert, D. Using a Maximum Uncertainty LDA-Based Approach to Classify and Analyse MR Brain Images. *Lecture notes in computer science***:** *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2004*, pp291-300, Springer Berlin (2004).

53. Yang, J. & Yang, J. Why can LDA be performed in PCA transformed space? *Pattern Recognition* **36,** 563-566 (2003).

54. Breiman, L. Random Forests. *Machine Learning* **45,** 5-32 (2001).

55. Zar, J.H. Biostatistics. 2nd edn. Englewood Cliffs, New Jersey: Prentice-Hall (1984).

56. Dietterich, T.G. Ensemble methods in machine learning. *Lecture Notes in Computer Science* **1857,** 1-15 (2000).

57. Vaidyanathan, S., Kell, D.B. & Goodacre, R. Flow-injection electrospray ionization mass spectrometry of crude cell extracts for high-throughput bacterial identification. *J Am Soc Mass Spectrom* **13**, 118-128 (2002).

58. Roessner, U., Luedemann, A., Brust, D., Fiehn, O., Linke, T., Willmitzer, L. & Fernie, A. Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* **13,** 11-29 (2001).

59. Mazzella, N. et al. Use of electrospray ionization mass spectrometry for profiling of crude oil effects on the phospholipid molecular species of two marine bacteria. *Rapid Communications in Mass Spectrometry* **19**, 3579-3588 (2005).

60. Favretto, D., Piovan, A., Filippini, R. & Caniato, R. Monitoring the production yields of vincristine and vinblastine in *Catharanthus roseus* from somatic embryogenesis. Semiquantitative determination by flow-injection electrospray ionization mass spectrometry. *Rapid Communications in Mass Spectrometry* **15**, 364-369 (2001).

61. Rashed, M.S., Al-Ahaidib, L.Y., Aboul-Enein, H.Y., Al-Amoudi, M. & Jacob, M. Determination of L-pipecolic acid in plasma using chiral liquid chromatography-electrospray tandem mass spectrometry. *Clinical Chemistry* **47**, 2124-2130 (2001).

62. Overy, S.A. et al. Application of metabolite profiling to the identification of traits in a population of tomato introgression lines. *Journal of Experimental Botany* **56**, 287-296 (2005).

63. Goodacre, R., York, E.V., Heald, J.K. & Scott, I.M. Chemometric discrimination of unfractionated plant extracts analyzed by electrospray mass spectrometry. *Phytochemistry* **62**, 859-863 (2003).

64. Koulman, A. et al. High-throughput direct-infusion ion trap mass spectrometry: a new method for metabolomics. *Rapid Communications in Mass Spectrometry* **21**, 421-428 (2007).

65. Martinez, A.M. & Kak, A.C. PCA versus LDA. *IEEE Transactions on: Pattern Analysis and Machine Intelligence,* **23,** 228-233 (2001).

66. Windeatt, T. Vote counting measures for ensemble classifiers. *Pattern Recognition* **36**, 2743-2756 (2003).

67. R_Development_Core_Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-900007-900050, URL http://www.R-project.org (2006)

68. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **57**, 289–300. (1995).

69. Storey, J. D. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**, 479–498 (2002).

70. Enot, D., Lin, W., Beckmann, M., Parker, D. Overy, D., & Draper, J. Press (2008). Pre-processing, classification modelling and feature selection using Flow Injection Electrospray Mass Spectrometry (FIE-MS) metabolite fingerprint data. *Nature Protocols,* **3,** 446-470.

71. Steinfath, M, Groth, D., Lisec, J. and Selbig, J. (2008) Metabolotite profile analysis: from raw data to regression and classification. *Physiologia Plantarum*, **132,** 150-161.

72 Van den Berg, R.A., Hoefsloot, H. C. J., Westerhuis, J.A., Smilde, A.K. and van der Werf, M. J. (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* **7**, 142.

73. Parsons, H. M., Ludwig, C., Gunther, U.L and Viant, M.R. (2007) Improved classification accuracy in 1- and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation. *BMC Bioinformatics*, **8**, 234.

G03012 Final Report - Section 9

John Draper/David Enot/Wanchang Lin

Understanding the output of standardised multivariate methods to analyse metabolomics data from the G02 programme (part of Task 04/03)

9.0 INTRODUCTION

A major aspect of the G03012 project was to develop standardized metabolomics data mining tools in a single software platform. Applying multivariate data mining algorithms to any data set will obviously produce models, but as with any type of statistical testing there is the problem of deciding whether the modelling conclusions are of significance in the context of the biological problem being addressed. The data in the G02 programme was collected in order to compare GM and non-GM crop genotypes or compare *Arabidopsis* mutants with different biochemical lesions in which the nature of the expected compositional difference was known beforehand (i.e. the transgene or mutation target was usually a defined enzyme). Thus differences between genotypes were expected to be both discrete (possibly quite large) and interpretable (i.e. related to the biochemical alteration). In the context of the G03 project effectively interest was focused on the GO2 data relating to 'control' genotypes consisting of traditionally bred potato cultivars as the objective was to develop a standardised methodology to compare the global chemical composition of food raw materials. In this context there is no previous literature providing guidance on how compositionally 'similar' potato cultivars might be and so a first major issue was to develop methodology to assess both the biological significance and mathematical reliability of the classification results along with the relevance of the ranked list of selected explanatory signals. A decision was made early on in the programme that these problems are best contextualised by designing experiments with the limited G02 data to answer a series of questions that would be of interest to any regulatory authority:

- Test software with dataset from different sources (ESI-MS and GC-MS) *Do we have to parameterize for different data sets?*

- What kind of approach to data pre-processing is appropriate to cope with variability and scale of data representing field grown crops? *How can we demonstrate that data alignment methods are appropriate for the characteristics of a large scale data set?*

- Is the sample replication strategy adequate for the problem at hand? *How can we be certain that we have enough samples replicates?*

- What kind of similarity measures derived from multivariate modelling prove to be the most robust when comparing food raw materials? *Thus, can we define threshold values below which it is likely there are effectively no significant compositional differences between food raw materials?*

- Can we demonstrate that our feature selection tools are generating models with greatest biological as well as statistical significance? *In other words do the compositional differences between different cultivars make biological sense?*

- Can we develop a protocol with potential for the large scale comparison of the chemical composition of related genotypes? *Is there scope for using a metabolomics database to compare food raw material compositions in the future?*

These questions have been addressed in the G03 programme by examining the output of the standardized data mining tools described in Section 8 of the report mainly when used to analyse the G02 data relating to traditional potato cultivars. In some instances the experimental conclusions were validated using further G02 data sets representing a collection of *Arabidopsis* genotypes analyzed in Aberystwyth. The majority of these questions are considered in a series of publications using the data produced by the G02 programme which can be consulted for detail (Enot and Draper, 2007; Enot, Beckmann and Draper, 2007; Beckmann et al., 2007).

9.1 Assessing a data alignment strategy in a large scale, multi-batch analysis of field-grown potato cultivars experiment

Data pre-processing including data normalisation and re-scaling methods have a large influence on metabolomics data analysis. When a simple, high throughput sampling technique is used which does not easily take into account for example sample weight (e.g potato tuber slices were used in many G02 experiments), water content or metabolite extraction efficiency then it is to be expected that some samples will be more dilute than others and so data requires normalisation within a batch. Similarly, in a large scale or long term experiment in which data is generated in different temporal batches it is crucial that any differences in instrument sensitivities experiment are normalised. However the choice of the most appropriate method is not a trivial process. The widely used methods including: range scaling, Pareto scaling, auto scaling and logarithmic transformation [71, 72, 73.] The utmost goal of the data pre-processing is to minimise the ‘technical and environmental’ variation and to maximise meaningful ‘biological’ differences [73]. In the context of the G03 programme it was important to decide how to assess the effectiveness of any approach to data normalisation. Previous researchers have approached this issue by using the fact that the biological variance is responsible for any significant discrimination between different categorical groups of samples (classes) and looking for improvements in classification following different pre-processing methods. A common practise has been to carry out a range of different data-pre-processing methods and then to visualise the data for improvements in either sample clustering or sample discrimination in scores plots derived from multivariate modelling techniques such as PCA or LDA. Different researchers in the area of metabolomics appear to have different preferences for data pre-processing methods according to the context of their own research. Unfortunately, these finding are not always consistent between projects and sometimes appear in conflict. For example, the effects of several data pre-processing methods, such as centering, auto-scaling, log transformation and range scaling for the metabolomics data analysis have been investigated previously [72] and it was concluded that the range scaling and auto scaling performed better than the other methods. Other studies [71] compared the performance of so-called variance stabling techniques, generalised logarithmic transformation (glog), auto scaling and Pareto scaling based on NMR metabolomics data. The results showed that glog method had higher classification accuracies compared to the unscaled, auto-scaled and Pareto scaled data. Yet other researchers [73] preferred methods based on unit variance and unit vector normalisation for metabolomics data pre-processing and based on the visualisation of data by PCA, the latter is better than the former.

In the context of developing a standardised methodology for metabolomics data analysis we consider data normalisation do be an optional part of the process (see Section 8.??) based on an empirical experimental demonstration that the data quality is actually improved. Different data pre-processing approaches have different impacts on the data set and each method has its advantage and disadvantage. Hence the selection of suitable data pre-processing should be tested on the practical data set being analysed and the appropriate one can be determined based on some criteria or performance measurements. The R package FIEmspro developed for the G03 programme implements eleven data pre-processing methods, ie. Centering, auto scaling, range scaling, Pareto scaling, vast scaling, level scaling, log transformation, log 10 transformation, square root transformation (sqrt), Inverse hyperbolic sine transformation (asinh) and TIC normalisation. A strategy to select appropriate approaches based on performance evaluation is described below using three large scale G02 data sets representing the compositional analysis of 6 field-grown conventional potato tubers Agria (Ag), Desiree_1 (De1) , Desiree_2 (De2) , Linda (Li), Granola (Gr), Solara (So). The data sets chosen included:

A) **Aber2001-sub**: a GC-MS analysis of tubers from a harvest in 2001 generated on an Agilent GC-MS quadropole instrument which measured 86? peaks in chromatographic data from data from analysis of 96 replicates of 6 cultivars.

B) **ESI2001_pos1:** positive ion mode nominal mass data (*m/z* 50-1000) of a FIE-MS analysis of tubers from a harvest in 2001 generated on Micromass LCT instrument which measured approximately 950 signals in data from analysis of 96 replicates of 6 cultivars.

C) **ESI2001_neg1:** negative ion mode nominal mass data (*m/z* 50-1000) of a FIE-MS analysis of tubers from a harvest in 2001 generated on Micromass LCT instrument which measured approximately 950 signals in data from analysis of 96 replicates of 6 cultivars.

Data normalisation was carried out using the eleven methods outlined above and model output are shown and compared to the original data (raw data). The evaluation criteria included visualisation of sample natural grouping data by PCA, sample class separation in discrimination tests using PC-LDA and classification accuracies based on data re-sampling procedures. It should be noted that the visual inspection approaches are only ever qualitative and may have value only in situation where obvious clustering is expected.

9.1.1 Effect of normalisation on feature variance

Fig.9.1 shows the data variance structure after different data pre-processing approaches are applied to a single sample of GC-MS data. The sub-figure *raw* indicates that some peaks in the raw data are very dominant in terms of peak area. Scaling by *log, log10* and *asinh* displays similar patterns in which the strong peaks have been scaled down,l while some weak peaks are scaled up. The *center* method shifts the variable by removing the variable mean and has very limit effects of stabilising feature variance. The TICnorm results from the normalisation of each peak in the data against the intensity of the total ion current (TIC) and has no effects on variable variance. As such TIC normalisation can be performed after the transformation by one of variance scaling such as *log*, *log10* and *asinh*.

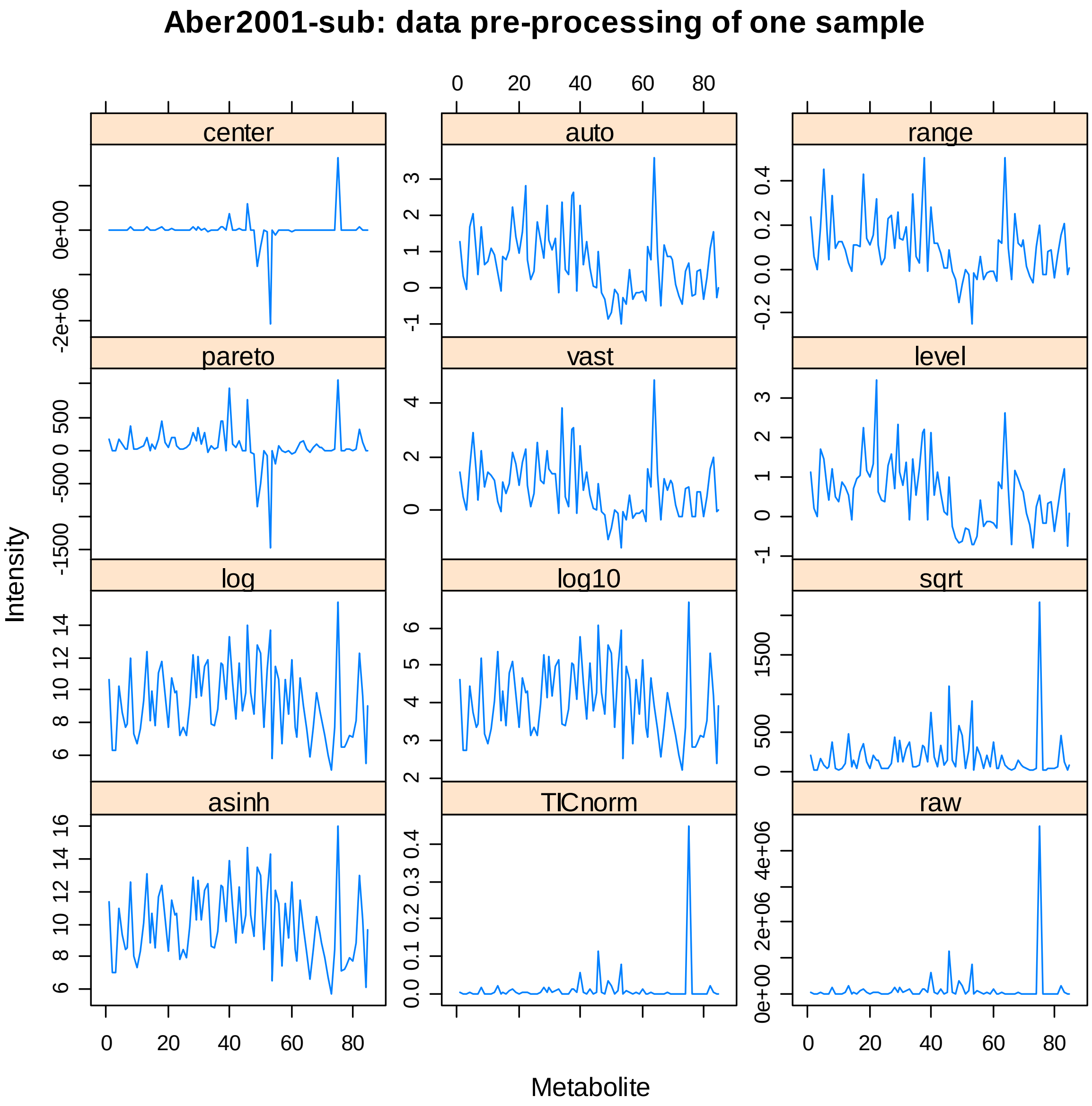


Fig. 9.1 Data structure of one sample after data pre-processing on Aber2001-sub data set

Application of the different normalisation approaches to FIE-MS data (Figure 9.2 and Figure 9.3) shows that scaling by *log*, *log10* and *asinh* all result in similar patterns: the strong peaks have been scaled down while some weak peak scaled up. The behaviour of *sqrt* is also close to them. TICnorm and raw data are very similar to each other.

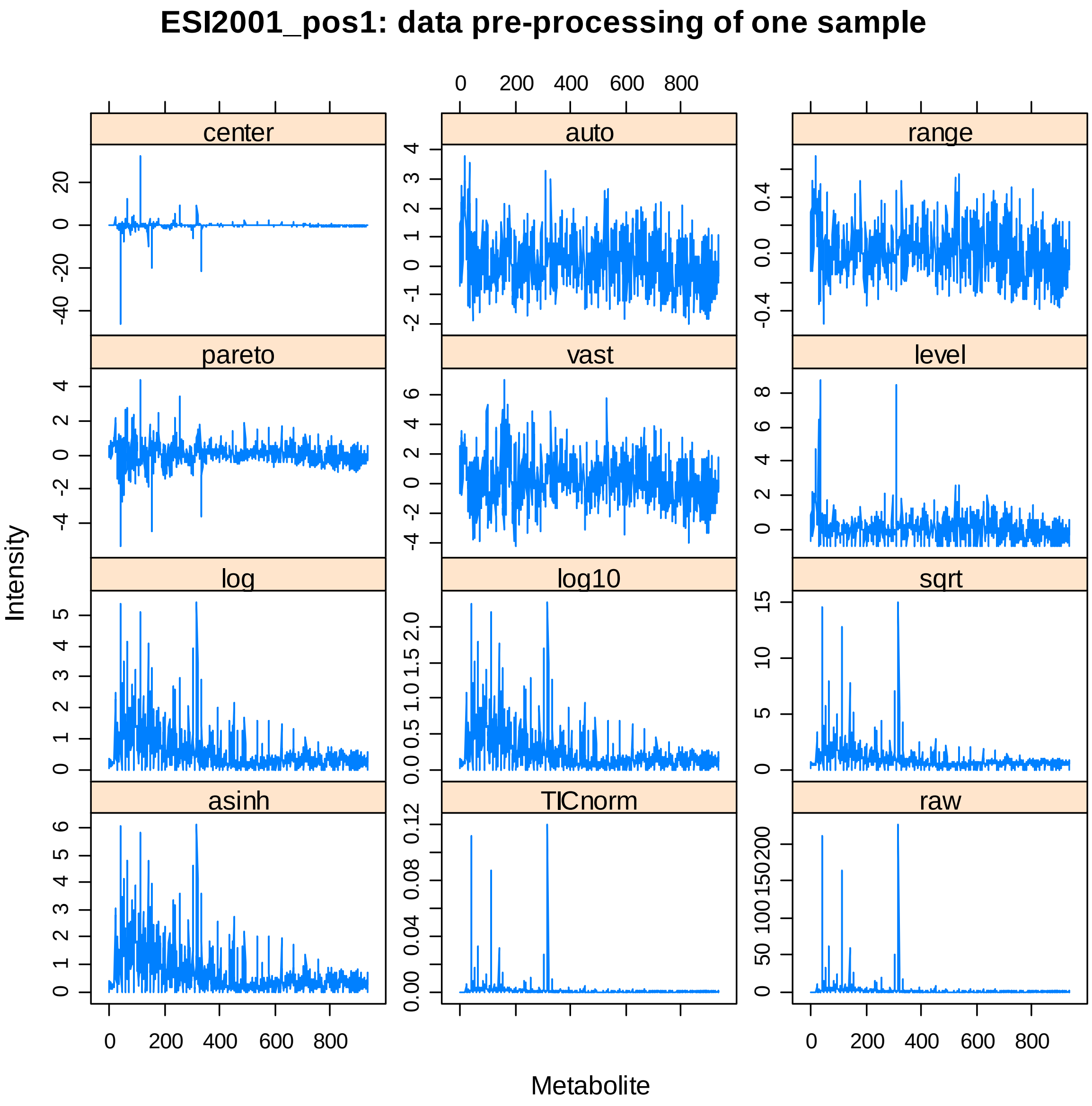


Fig. 9.2: Data structure of one sample after data pre-processing on ESI2001 positive mode

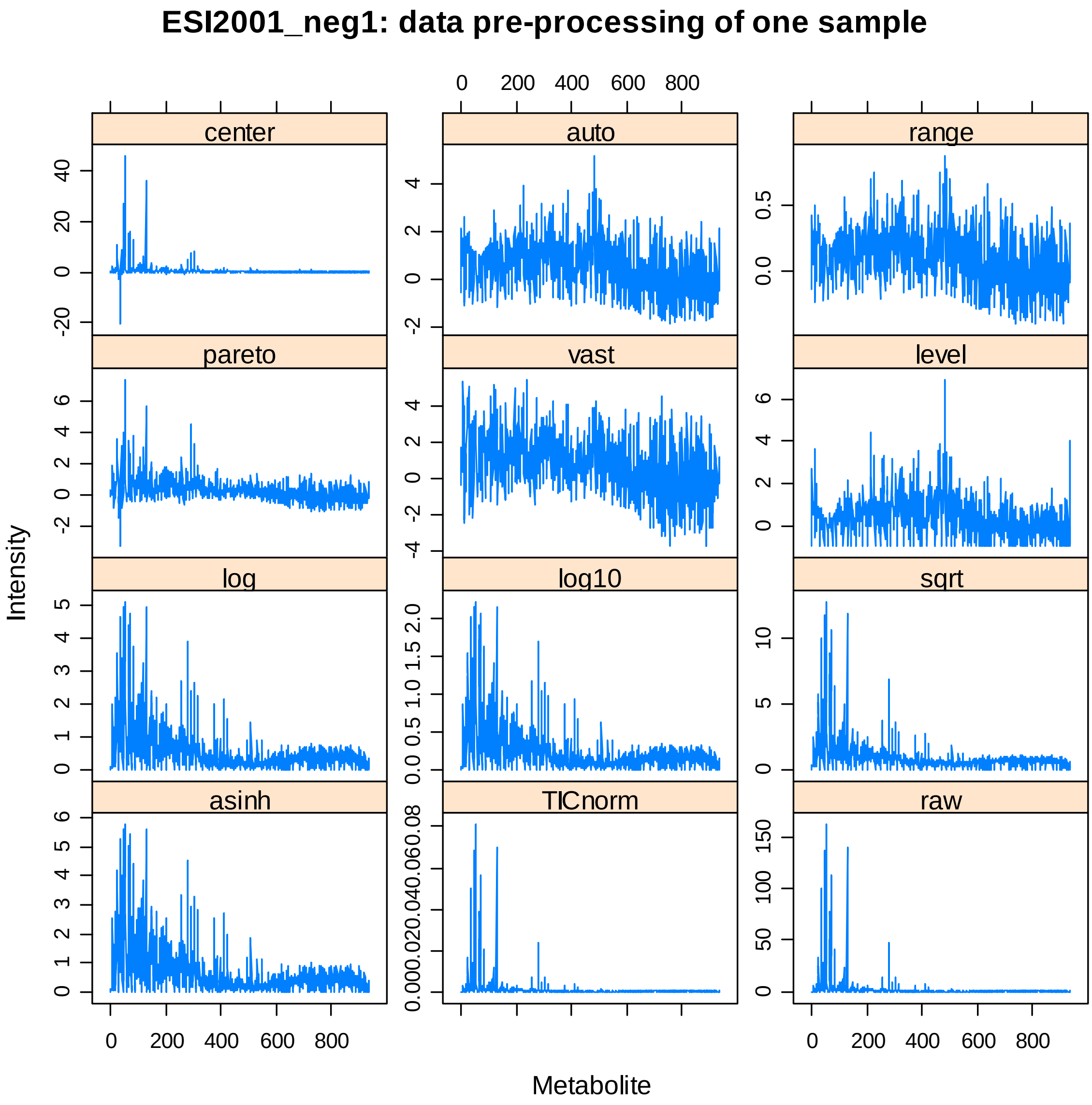


Figure. 9.3: Data structure of one sample after data pre-processing on ESI2001 negative mode

Application of the different normalisation approaches to FIE-MS data (Fig. 9.2 and Figure 9.3) shows that scaling by *log*, *log10* and *asinh* all result in similar patterns: the strong peaks have been scaled down while some weak peaks scaled up. The behaviour of *sqrt* is also close to them. TICnorm and raw data are very similar to each other.

9.1.2 Effect of normalisation process on sample clustering in PC

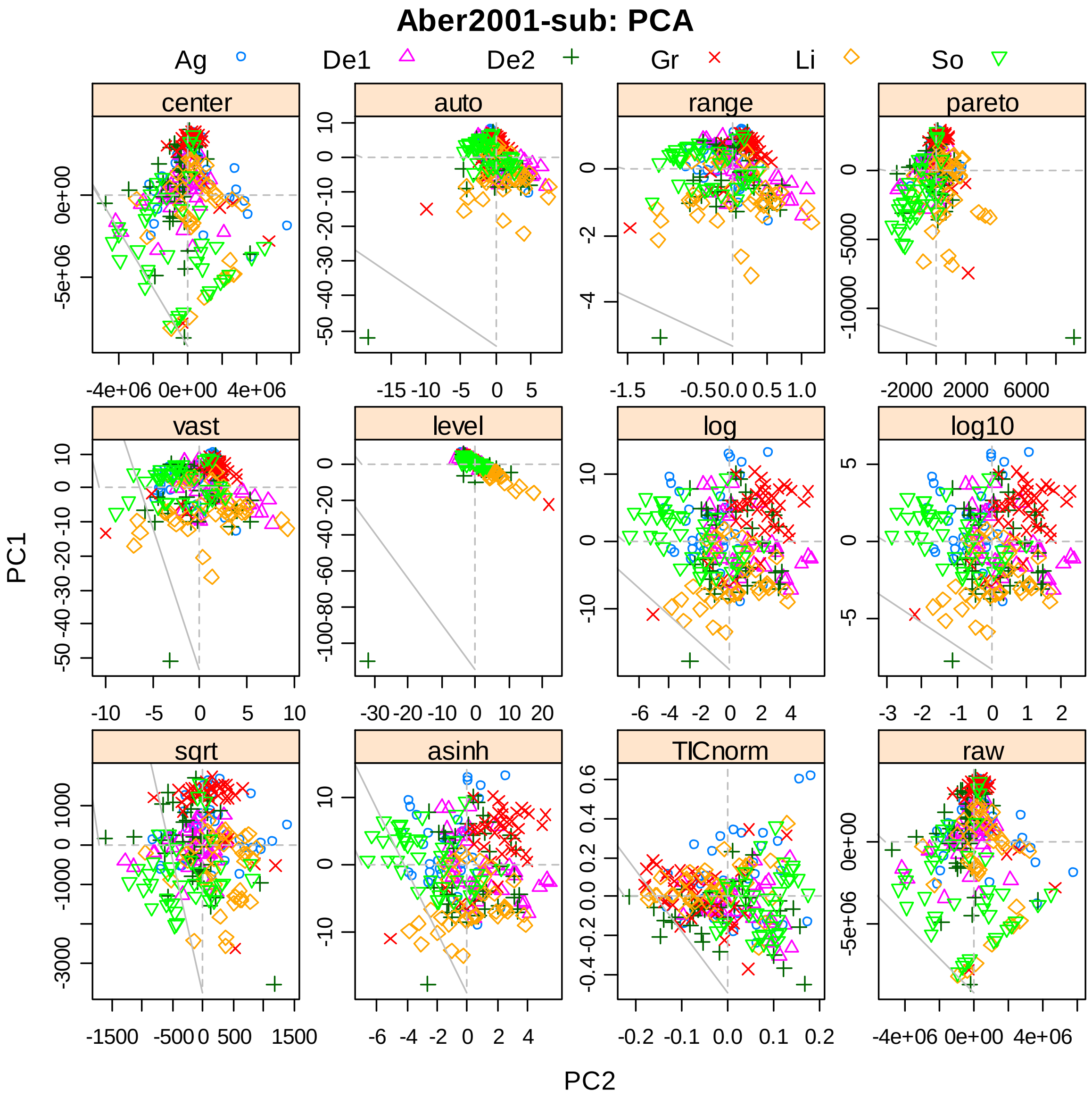


Figure. 9.4: PCA plot (PC1 vs PC2) on Aber2001-sub

Fig.9.4 illustrates the effects of PCA using different data pre-processing methods in GC-MS data. Inspecting the results of normalisation by *log*, *log10* and *asinh*, it can be seen that there is some clustering around cultivars *Gr*, *Li* and *So,* although some overlap still exists among the three cultivars. No obvious class clusters were found in the output of other methods.

Inspection of the PCA scores plots from both ESI-MS positive and negative ion data subjected to different normalisation routines (Fig.9.5 shows positive ion data) indicated that it is difficult to assess the effects of data pre-processing using PCA.

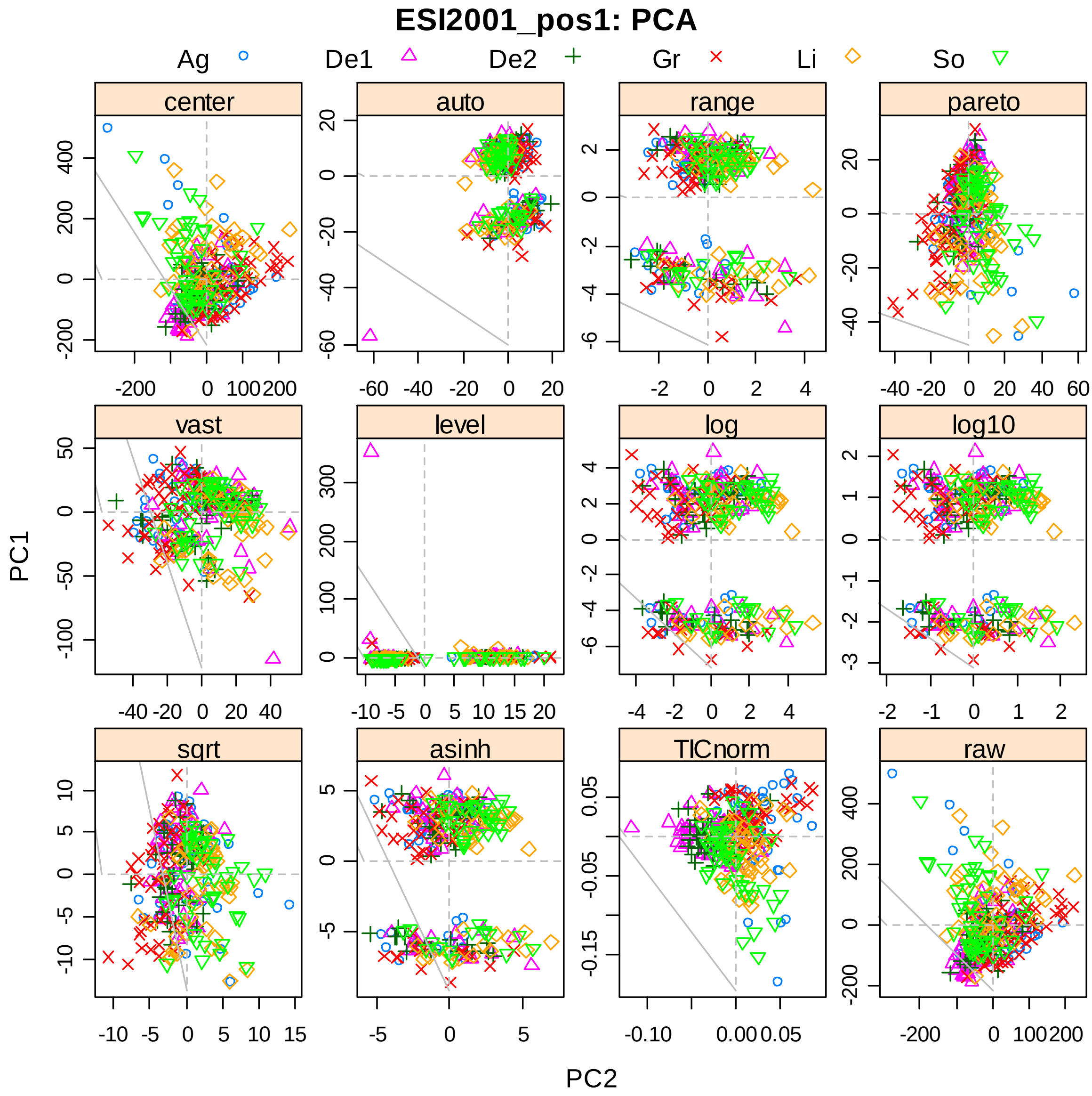


Fig. 9.5: PCA plot (PC1 vs PC2) on ESI2001 positive mode

9.1.3 Effect of normalisation method on sample discrimination in LDA

Fig.9.6 shows the PC-LDA scores (DF1 vs DF2) when GC-MS data is modelled using LDA. The effects of *log*, *log10* and *asinh* outperform other data pre-processing methods by the visualisation in which *Ag*, *Gr* and *So* separate well between them and between any other single cultivar.

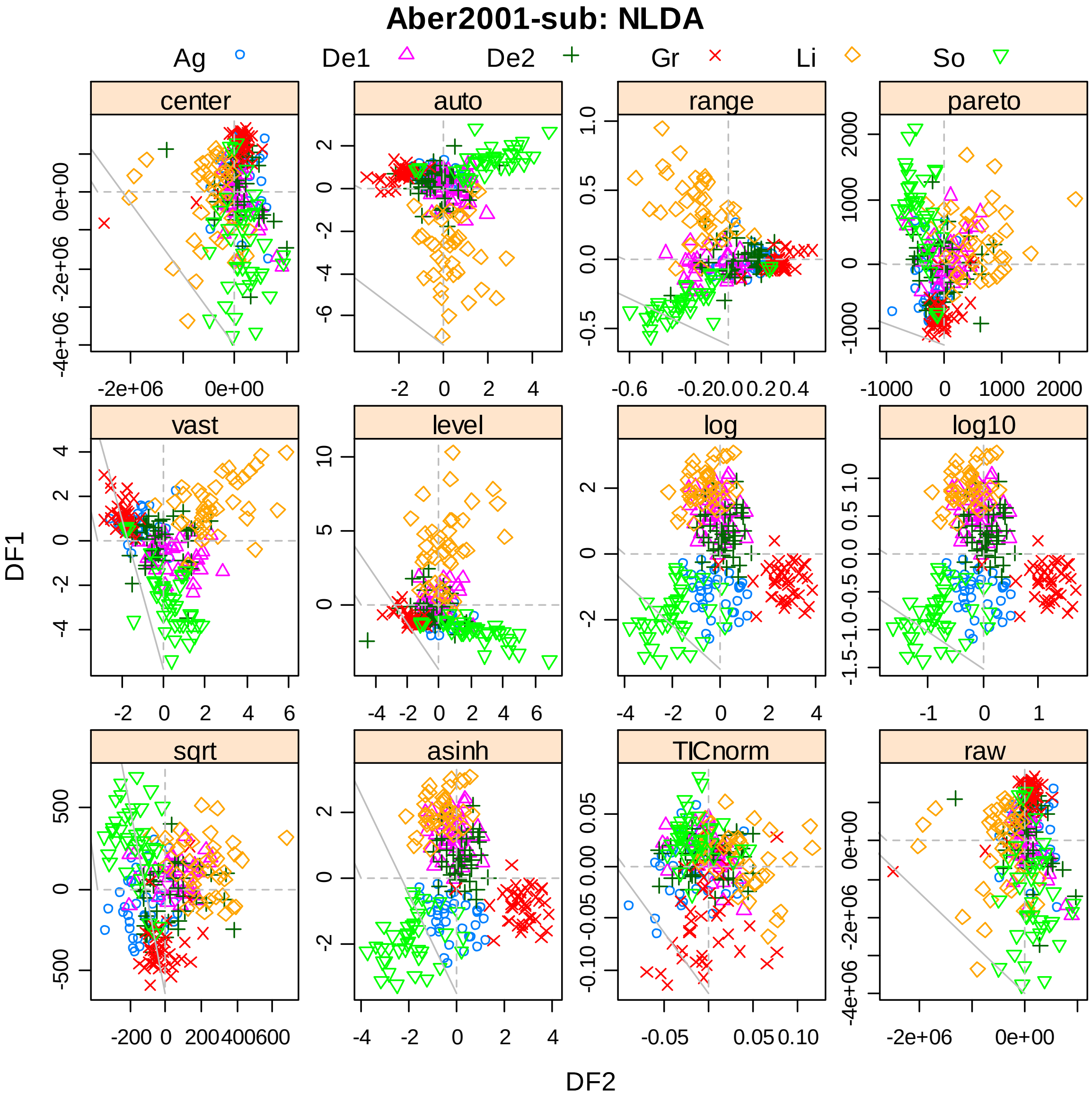


Figure. 9.6: LDA plot (DF1 vs Df2) of Aber2001_sub GC-MS data.

Fig.9.7 shows the PC-LDA scores (DF1 vs DF2) when ESI-MS data has been subjected to different pre-processing regimes. The pattern of *log*, *log10* and *asinh* are broadly the same. The effect of *sqrt* is similar to *log*, *log10* and *asinh*, all of which exhibit good separation between cultivars *Gr*, *Li and So.* The normalisation methods had a similar effect on sample discrimination in negative ion mode data.

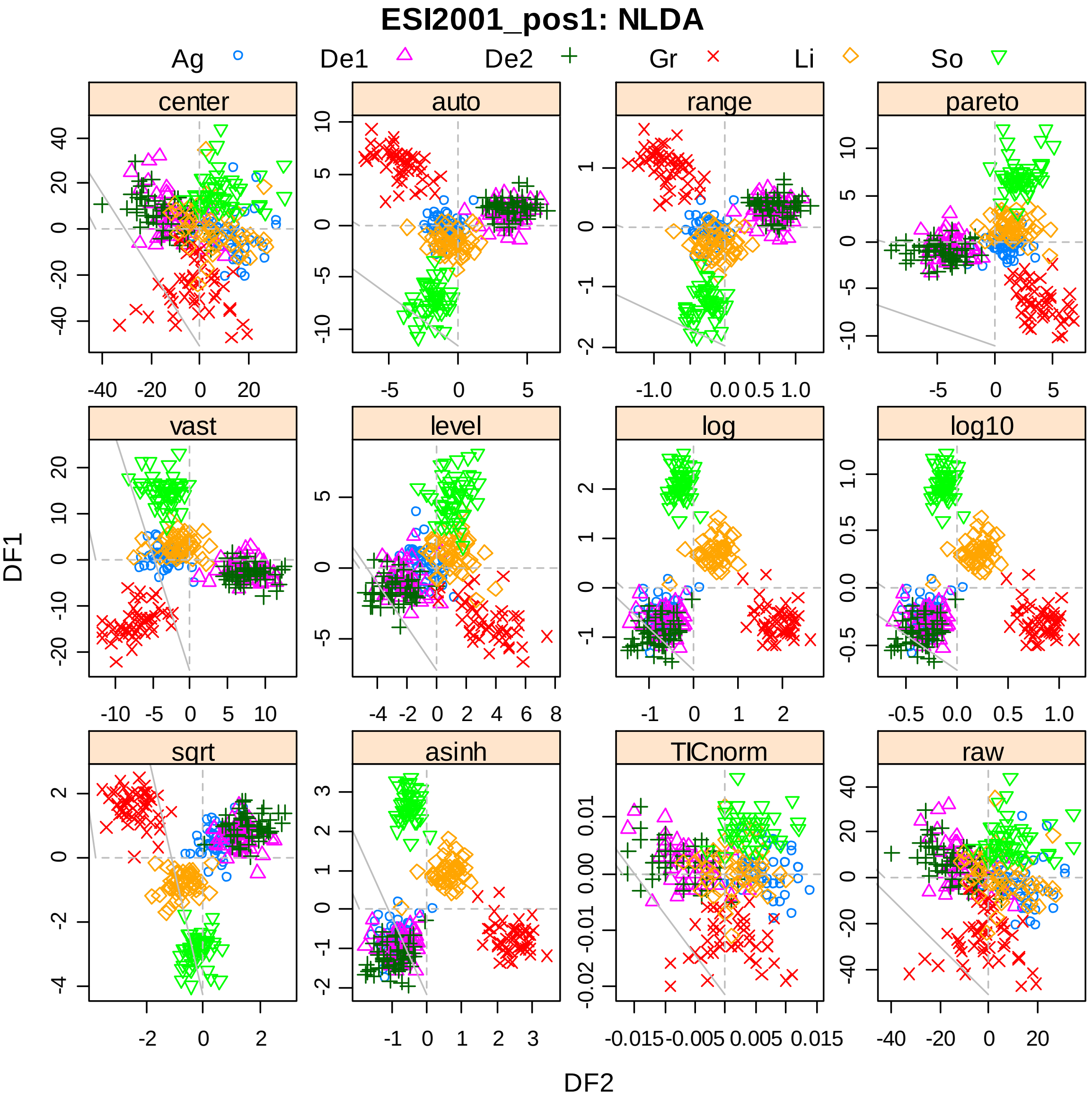


Fig. 9.7: PC-LDA plot (DF1 vs DF2) on ESI2001 positive ion mode data

9.1.4 Assessing normalisation method by performance of data mining algorithms

Sample classification was performed using LDA and Random Forest and the classification accuracy was calculated after cross validation with result presented as the accurate rate (1.0 = 100% correct classification of all samples).

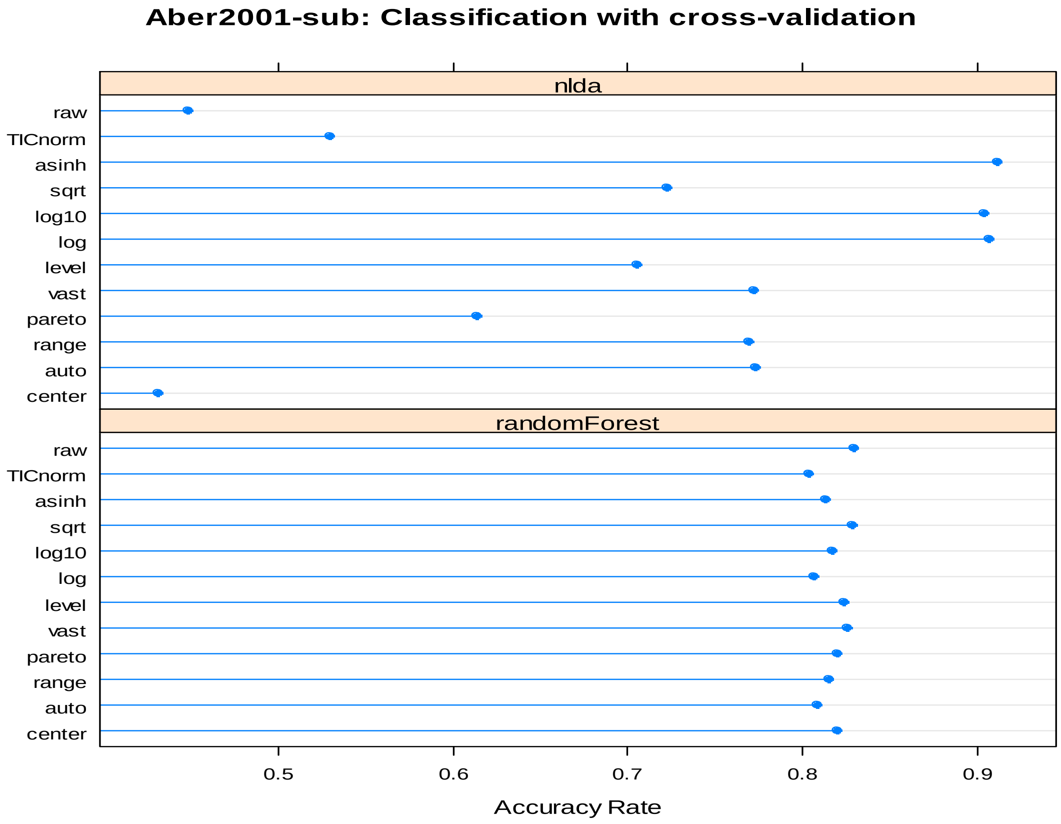


Figure 9.8: Classification on Aber2001-sub GC-MS data

Figs. 9.8 -9.10 illustrate the effects of different data pre-processing methods in this quantitative way. Fig. 9.8 indicates that three methods, *log*, *log10* and *asinh* are competitive when applied to GC-MD data and perform better than other approaches by the classifier *nlda*. Note especially the rather poor performance using just the raw data indicating the importance of scaling pre-processing. These results are consistent with visual inspection of PCA and PC-LDA. It is interesting that *randomForest* is insensitive to different methods, although their classification accuracies are lower than *log*, *log10* and *asinh* by *nlda*, roughly by 10%.

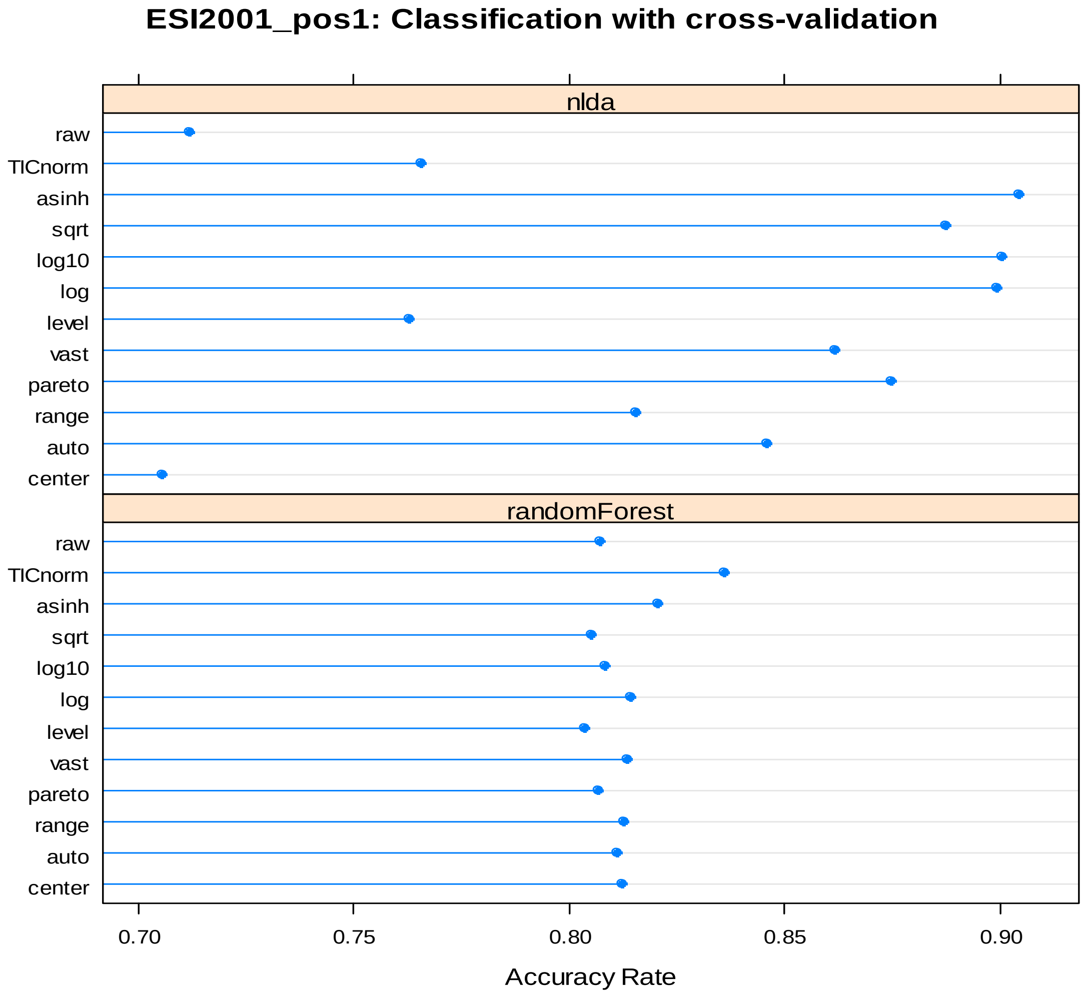


Figure 9.9: Classification on ESI2001 positive mode data

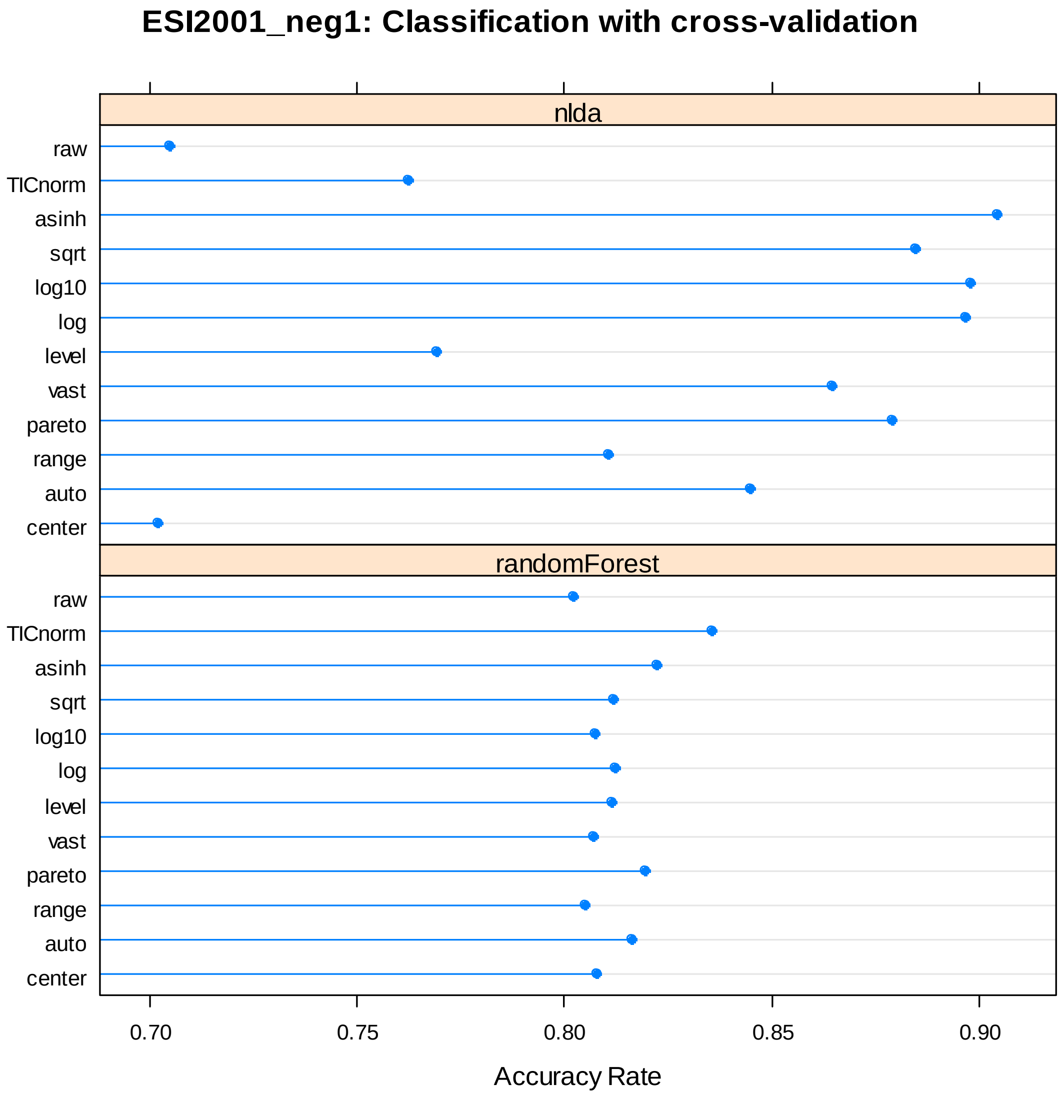


Figure 9.10: Classification on ESI2001 negative mode data

Fig. 9.9 and 9.10 shows that *asinh*, *log10* and *log* outperform others with respect to classification of ESI-MS data by *nlda*. The performance of *sqrt* is also close to *log*, *log10* and *asinh*. Again the results of Random Forest modelling show that it is quite robust to the data pre-processing method although normalisation of *TICnorm* is slightly better than others.

9.1.5 Conclusions relating to assessment of data pre-processing approach

The validation based on modelling of three different large data sets are that in general data scaling by logarithmic transformation (log or log10) and inverse hyperbolic sine transformation (*asinh*) are very competitive. However, since *TICnorm* is a normalisation approach operating on each individual sample, it can be used independently after the transformation by one of *log*, *log10* and *asinh.* Apart from classification accuracies other indicators of algorithm performance can be used to reveal any improvement /detriment of data pre-processing choice to the classification process. Classification 'distance measures' such as **eigenvalues** and modelling 'sensitivity' measures such as **margins** and '**area under curve'** (AUC) have been shown to be particularly valuable and are described in detail in the next section.

One interesting output from this section of the project was a demonstration of the relative insensitivity of Random Forest modelling to the data pre-processing method. This is probably a function of the fact that Random Forest builds thousands of decision trees using random combinations of only a small number of variables each time, and hence any highly intensity or highly variable signals will only effect a very small proportion of the total ensemble of decisions trees used to generate the final model. Associated with this behaviour is also the fact that Random Forest analysis is much less likely to result in 'over-fit' models which is an important reason why classification accuracy and sensitivity measures derived from this technique are particularly valuable (see section 9.2).

9.2 Assessment of discrimination results significance metrics

Testing for similarity between plant genotypes usually refers to the application of statistical routines that aim to show whether there are, somehow, *significant* differences. The conclusions drawn from statistical results (for example deciding an appropriate confidence level) can be somewhat arbitrary and context-dependent as different experts may give different interpretations of the degree of significance. Importantly, there can also be a gap between the mathematical significance and the biological relevance of the effect detected. In the context of food raw material analysis a major conclusion of the G02 project was that only data analysis methods in which individual variables are explicitly highlighted in multivariate models should be used in order to judge whether such 'explanatory' features are associated with changes to predicted areas of biochemistry. Only with such approaches could the 'biological significance' of any differences be tested. In the G02 project we described the use of several data analysis methods (Linear Discriminant Analysis, Support Vector Machine and Random Forest) to classify plant and food raw material samples.

A range of supervised data modelling methods will achieve high classification accuracies. However, he characteristics of the underlying probability distribution of the data and more precisely, an examination of class complexity in the original input space must also be considered in conjunction to the overall predictive power of any model. For most supervised learning techniques applied to high dimensional problems, the decision boundary properties cannot directly be used because it usually involves an optimisation process that tends to 'overfit' the training data (hence it is necessary to use external samples to assess robustness of the model rather than the resubstitution error). When using projection based techniques such as PLS-DA or PC-LDA, additional estimation of the number of components to be used in the modelling process has to be carried out, which might affect discrimination. For these reasons, we propose different approaches to test the similarly of plant raw materials in the original multivariate space (i.e. without prior feature selection), with an optimal use of the available information under realistic data paucity constraints and which can provide general and meaningful metrics for future comparisons. Several approaches are outlined below.

Classifier error estimation

Estimation of classifier accuracy has received considerable interest in the bioinformatics community and particularly in the omics literature. Strategies recommended for transcriptomics experiments can be directly applied in a metabolomics context as both data structures share similar characteristics. In problems related to small sample size, validation approaches are based on repetitive sub-sampling of the original training data to build the model and then use some averaging of the left-out examples to estimate the classifier error. Amongst different resampling techniques, bootstrap based techniques are known to offer an adequate trade off between bias and variance of the error estimation [47, 49] The general idea is that the variance of an estimate based on the left-out samples is a good approximation of the true variance of the original population. Although bootstrap is widely used to assess statistical accuracy in data mining and machine learning experiments it will often display a bias towards upward accuracy [49]. To minimize this problem of overfitting, a '0.632+' estimator has been proposed which is designed to be a less-biased compromise between upward and downward accuracy estimation. B632, utilizes a resubsitution error (i.e fitting the training data) which is added to correct the bias inherent to the error computed by weighted average of the left out samples (bootstrap zero estimator). Despite its computational cost, bootstrap error provides a less variable estimate than an error counting only techniques (e.g. cross validation) where possible misclassification changes are increments corresponding to the inverse of total number of samples.

Receiver operating characteristic

Receiver operating characteristic (ROC) curves can be used as an alternative measure of the predictive abilities of any binary classifier [9, 50, 51]. ROC curves display the relationship between sensitivity (true-positive rate) and specificity (false-positive rate) across all possible threshold values that define the decision boundary. The most common way to summarize the ROC curve is to compute the area under the curve (AUC). As a single value measure, the AUC specifies the probability that the decision boundary assigns a higher value to a positive sample than a negative one, both chosen randomly. AUC takes a value of 0.5 when the samples from both classes are uniformly distributed across the decision boundary and a value of 1 when the decision boundary can incontestably discriminate both groups. One advantage of both accuracy and AUC is that they can be calculated regardless of the algorithm used for data analysis.

Eigenvalues

Linear Discriminant Analysis (LDA; also known as Discriminant Function Analysis in the metabolomics literature) is a supervised method that computes new directions (canonical variates or discriminant functions) in which the groups are best separated [39]. The aim of LDA is to find discriminant functions that maximise between-class separability (S_B) and minimise within-class variability (S_W). The eigenvalue of the eigensystem of ($S^{-1}_W S_B$) can be used to measure similarity of both sample replicates, and, more importantly, different classes: the greater the distance between classes (large S_B) and the more compact the classes are (small S_W), the better the classes are separable. However, due to sample size problems (as are common in metabolomics), S_W is always singular and/or unstable and the inversion of S_W cannot be possible without initial data dimensionality reduction using for example Principal Components Analysis [53] . To avoid bias in the eigenvalue estimation introduced by the selection of the number of components, we chose the two step procedure proposed by [52] .

Margin concept in ensemble methods

As opposed to a single model that aims to find the best hypothesis, ensemble methods are techniques that generate multiple models by running a base algorithm many times [56, 66]. One example is Random Forest (RF) that uses the standard decision tree algorithms as the base learner [54]. Prediction of new samples is done commonly by determining the winner class from the votes on the overall ensemble of models. Therefore, confidence in attributing a sample to a designated class can be deduced from the difference between the score (averaged number of votes) for the true class and the largest score of the rest of the classes. This is defined as the sample margin and measures the extent to which the average number of votes for the right class exceeds the average vote for any other class (i.e. the most probable misclassification). The larger the margin, the higher is the confidence that an example belongs to the actual class. In the G03 project, the margin of a classifier is presented as the mean of all of margins calculated used training data in the ensemble of RF models comparing classes.

9.2.1 Assessment of similarity measures

To illustrate the behaviour of the statistical measures outlined above (i.e. 0.632 bootstrap accuracy, eigenvalue, AUC and margin), an initial validation was conducted on a G02 data set consisting of a heterogeneous set of 27 *Arabidopsis* genotypes providing wide coverage of modelling situations ranging from comparisons with no/little metabolic differences, single gene mutation/insertion effects on isolated biochemical pathways to huge ecotype divergences [12]. To simulate a situation with a realistic sample size, each measure was calculated on a training set comprising 18 plant replicates per genotype whereas 12 different plant replicates were left out of the training data to provide a validation set. Features of different statistical similarity measures are illustrated in Figure 9.11 using all possible 351 pairwise comparisons between the 27 *Arabidopsis* genotypes. Both Random Forest (RF) margins and LDA eigenvalues calculated on the training data provide sufficient information regarding the generalisability and the sensitivity/specificity relationships of the models (Figures 9.11a and 9.11b). In the light of these results, several remarks can be made. Although LDA and RF algorithms differ by the principles by which they operate, voting confidence in RF is higher when the between-class/within-class ratio is maximised (Figure 9.11c). One of the claims regarding the fact that RF does not overfit is apparent in Figure 9.11d where there is a direct correlation between the confidence in the prediction votes calculated on the training and those of the unseen data. It is thus concluded that RF margin calculations and eigenvalues are able to distinguish between models for which both bootstrap accuracy and AUC are reaching their maximal value of 1. This property is of particular interest in studies where meaningful relative distances between genotypes are sought.

9.2.2 Developing a baseline for model significance (biological versus permutation)

Most supervised classification/discrimination methods will achieve high accuracy, but it is not always apparent how robust or generalizable any specific model might be.

Thus, the decision whether to accept a supervised multivariate model for feature selection and further investigation of a biological phenomenon is an important step in a data mining workflow. Essentially, due to the feature dimensionality and sample sparsity issues associated with data containing substantial sources of variance, a

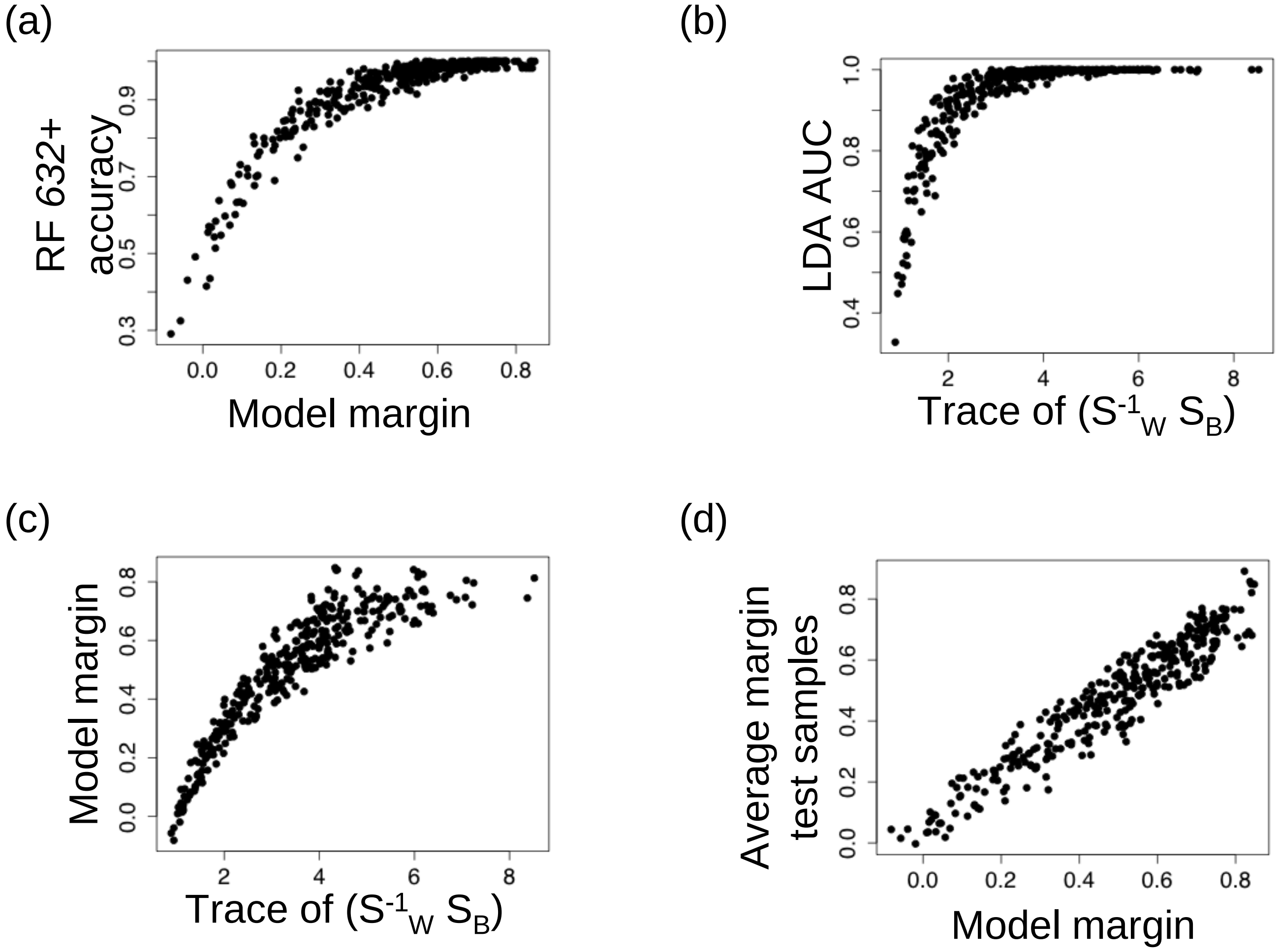


Figure 9.11: Relationships between various statistical measures computed from 351 binary classifiers: a) average margin of the training samples versus the *0.632+* bootstrap accuracy in Random Forest (RF) models; (b) eigenvalue versus area under curve of the LDA model; (c) eigenvalue of the LDA classifier versus average RF margin of the training samples; (d) average RF margin of the training samples versus average RF margin of the unseen samples.

threshold metric for determining model acceptance must be developed alongside the simple presentation of modelling accuracy [1-4, 12, 36, 37, 54]. This is even more important when the statistics of interest result values are somewhere between what is expected of a purely random distribution and a near perfect value. Classical tests for assessing the significance of model properties are rather limited unless the quantity under study is known to satisfy a particular statistical distribution and has also an adequate sample size.

We consider that a significance ’threshold’ statistic can be evaluated by comparing two sample classes that are known to display no biologically significant difference at the level of the metabolome [12, 36]. Thus in the G03 project we have explored the possibility of developing a meaningful **significant difference threshold** for model classification metrics or feature selection by defining distance measures or sensitivity metrics (e.g. 'margins', AUC values, Importance Scores) in comparisons were there is effectively no phenotypic difference between genotypes. These results could then be compared with model characteristics in comparisons between increasingly different genotypes. The G02 data contained two types of sample classes which allowed us to compare either a wild type organism to mutant lines with no metabolic phenotype (*Arabidopsis*), or examining replicate plots of the same plant genotypes grown in the same field (potato Desiree genotypes) which were used to define a meaningful baseline for model significance. One advantage of this approach to develop a 'biologically significant baseline' is that it is data driven, encapsulating sample size effects, poor knowledge about the fingerprint characteristics and experimental variance. As with any other approaches, this cannot avoid confounding effects that may inadvertently highlight features that have no biological significance as a result of unplanned experimental factors (e.g. sample contamination or instrument measurement drift) which contribute artifactual regular variance characteristics in the data. When such data driven null hypothesis information is not available, permutation tests can be used to determine the impact of the chosen statistic by computing all its possible values by permutation of the sample labels (or several 1000s random rearrangements if processing time is prohibitive) [4, 48]. In permutation testing, p-value is defined as the fraction of times one obtains a value as extreme as the original statistical quantity calculated from the un-permuted data.

To illustrate these points comparisons have been made (using Random Forest and SVM modeling) between four *Arabidopsis thaliana* mutants: *pgm1* (deficient in the isozyme phosphoglucomutase required for starch synthesis), *fah1-2* (mutation of ferulate-5-hydroxylase, an enzyme functioning in cell wall metabolism), *vtc1* (mutation of GDP-mannose pyrophosphorylase in the ascorbate synthesis pathway) and *amt14* (ammonium transporter defective mutant) and their progenitor line Columbia (*Col2*) are extracted from the overall data. In a previous study [12] we had already demonstrated that *amt14* did not have any appreciable phenotypic differences when compared to the wild types line, presumably as this particular mutation is compensated for by other functional ammonium transporters. Thus any statistical measures derived from classifiers attempting to discriminate the two lines will represent a threshold for significance. In contrast *pgm-1* had large, pleotrophic phenotypic alterations, whilst *fah1-2* and *vtc-1* displayed distinctive, but more contained, metabolic differences from the progenitor ecotype.

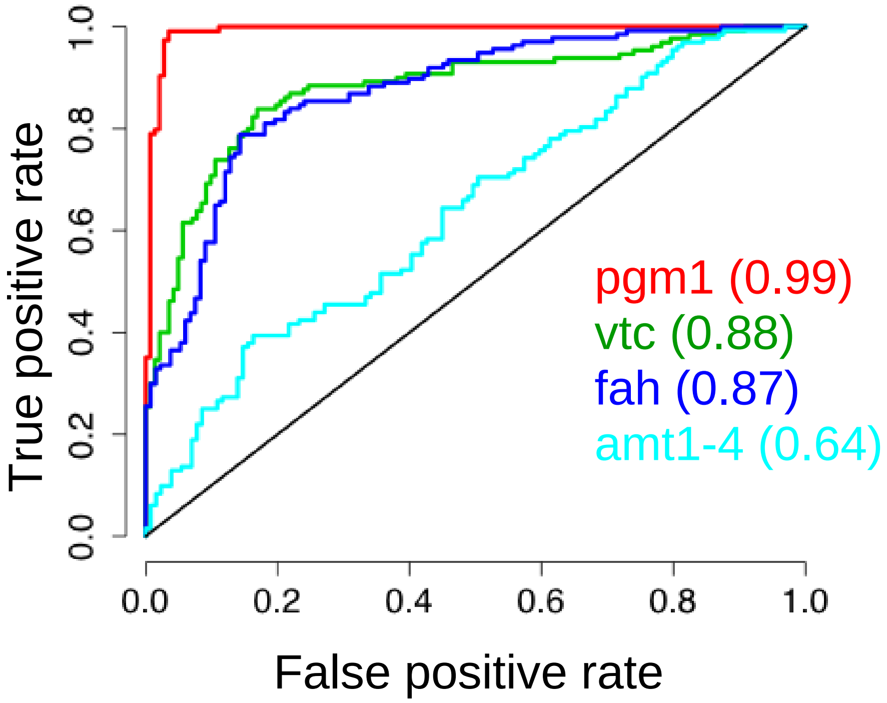


Figure 9.12: Receiver operator characteristics curve (ROC) for a selected number of linear Support Vector Machines (SVM) models. Area under the curve is given alongside the genotype name. The diagonal line is the ROC curve equivalent to random guesses.

ROC curves derived from resampling the SVM models of the four comparisons are gathered in Figure 9.12. Whilst, the RF model margins of each comparison are given alongside the distribution of the null hypothesis obtained by a permutation test in Figure 9.13. It can be seen in both figures that the model statistical measures are ordered according to the perturbation level expected in the biological outcome: thus the *pgm1-Col2* comparison exhibits logically the highest ‘dissimilarity’ statistics as this mutant is effectively starch deficient and transcriptomic studies have highlighted

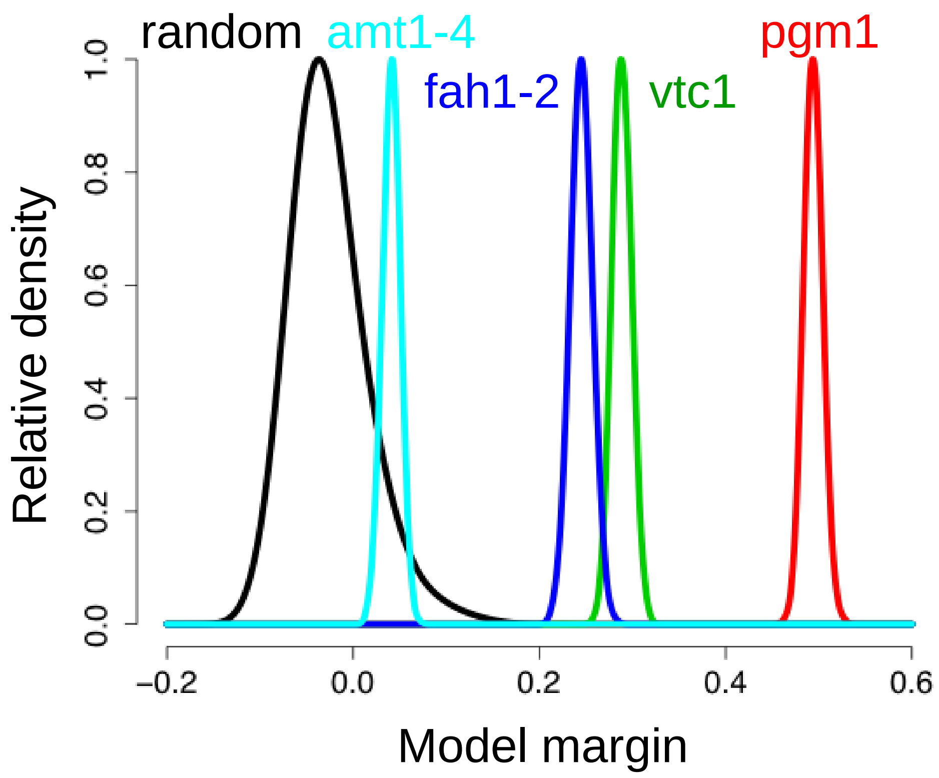


Figure 9.13: Illustration of significance testing of the RF margin by class label permutation.

expression increases in over 4000 genes. *Vtc*1 or *fah1*-2 represent mutations in two discrete secondary metabolic pathways and as such relatively fewer metabolome differences are detectable by metabolome fingerprinting, resulting in slightly weaker, but none the less significant, modeling statistics. In all three of these comparisons, the statistical measures are higher than the *null hypothesis* comparison between *Col-2* and *amt1-4* in which the ammonium salt transporter mutation did not affect the measured metabolome. In the comparison of *Col-2* and *amtl-4* the RF model margins overlap significantly the reference distribution (labeled random in Figure 9.13) obtained by permuting the class labels and the area under the ROC curve (AUC=0.64) is below the accepted limit of 0.8 for defining a good model. In addition, to relate to models in a more meaningful manner, this simple example demonstrates that the statistical “zero difference” (i.e. AUC≈0.5, margin≈0) regardless of sample size can be reformulated within a biological context without major mathematical redefinition of the problem under study .

9.2.3 Conclusions

Although a range of metabolomics level analytical chemistry tools for assessing compositional differences between any food raw materials exist, there have been few attempts to explore what kinds of statistical metrics are suitable to quantify compositional similarity. The effect of data dimensionality, often combined with sample paucity, can generate problems for model validation; classifier accuracy alone can be misleading unless validated by sufficient resampling of the available data and AUC assessments. We do not advocate that “one data mining technique fits all” because of the variety of applications and biological questions addressed by a metabolomics approach. A range of significance metrics are important to evaluate in order to decide whether a model is worth pursuing and in the future there is a need for standardised metrics so that everyone can compare results. Currently we suggest that margin measures and scatter matrices eigenvalues in conjunction with estimates of classification accuracy and model sensitivity provide complimentary and appropriate metrics in any specific compositional comparisons.

9.3 Testing model ‘biological’ significance by interpretation of selected explanatory features.

As indicated in section 9.2 with the analysis of a large number of *Arabidopsis* genotypes it is possible to use specific modelling outputs to evaluate the degree of similarity between any number of different crop varieties in the same species. As mentioned earlier, with previous work on GM and mutant plants there is a strong expectation that these will exhibit diverse and substantial compositional difference. Clearly new crop cultivars/varieties are registered traditionally following the demonstration that they have new, or unique combinations of, botanical characteristics. Many food raw material quality characteristics associated with flavour, aroma and food-processing properties for example are a function of global composition. However, in the context of a crops such as potato tubers there were no previous studies which analysed global quantitative or qualitative compositional differences between individual cultivars. A major goal of the G03 programme was firstly to refine a technology base which would allow the global composition of raw food material to be meaningfully compared. A second important objective was to determine whether such differences had biological significance and if possible, whether they could be linked to quality parameters. In the present section we demonstrate how the modelling methods described above can be use to interpret the compositional differences between 6 traditional potato using the pre-processed (log10 transformed and TIC normalised) FIE-MS and GC-MS data described in Section 9.

9.3.1 Potato Cultivar Classification by FIE-MS Fingerprinting

The G02 programme produced a large data set comprising FIE-MS fingerprints representative of tuber extracts of field-grown the potato cultivars Agria, Désirée, Granola, Linda and Solara. Two closely related populations of Désirée, that were independently propagated (De1 and De2) were included in the sample set; comparison of these two genotypes (which were considered compositionally equivalent) would provide a baseline metrics for both cultivar discrimination and feature selection. With considerable environmental variability (4 different field plots) as well as technical variability (6 analysis batches) the ESI-MS 2001 FIE-MS data already contains significant levels of variance. We have already demonstrated that Random Forest (RF) was insensitive to data pre-processing and overfitting and thus is was chosen as the most robust method for both classification and feature selection. RF margin values were used as an indication of model ;'sensitivity' robustness and the RF Importance Score was used to rank features for significance in the modelling process. An initial pair-wise comparison of the positive ion FIE-MS fingerprints, representing each cultivar, generated models with high classification accuracy and high sensitivity (**Table 9.1**). These data suggested that surprisingly large differences in composition exist between tubers of individual cultivars.

Pair-wise comparison	**Classification accuracy (%)**	**Model margin**
Ag_De1	100	0.62
Ag_Gr	97	0.61
Ag_Li	97	0.44
Ag_So	100	0.64
De1_Gr	100	0.66
De1_Li	100	0.42
De1_So	100	0.68
Gr_Li	100	0.59
Gr_So	100	0.74
Li_So	100	0.58

*The model margin is presented as a measure of Sensitivity

Table 9.1: Model statistics in pair wise comparison of potato cultivars

9.3.2 Identifying important features that discriminate between cultivars

Feature selection was performed using Random Forest modelling and all ESI-MS signals were ranked by Importance Score. Figure 9.14 shows a plot of Importance Score against Feature Rank in a binary comparison of the two Desiree cultivars. From this initial comparison it is concluded that a Random Forest Importance Score value of around 0.004 was likely to be an adequate threshold for feature significance. Random Forest *Importance Scores* in all pair-wise comparisons levelled off before variable 30 in the ranked lists, with anywhere between 8-24 *m/z* signals having a value greater than the significance threshold of around 0.004 (Figure 9.15). Permutation testing revealed that out of 1000 signals, only a small number of variables had significant discriminatory potential (Figure 9.16). Analysis of negative ion data yielded comparable results (data not shown).

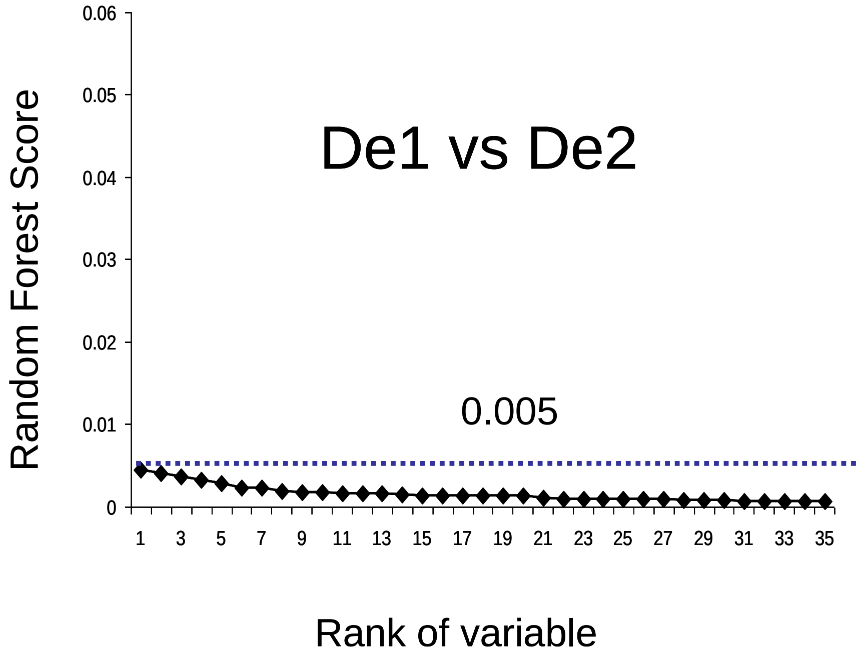


Figure 9.14: Plot of RF Importance Score against Feature Rank in a binary comparison of the two Desiree cultivars.

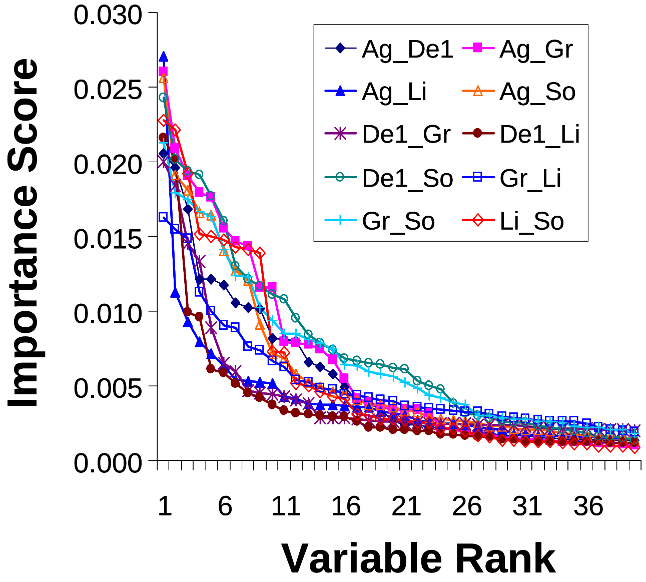


Figure 9.15: Relationship between RF importance sore and feature rank in all pairwise comparisons of potato cultivars

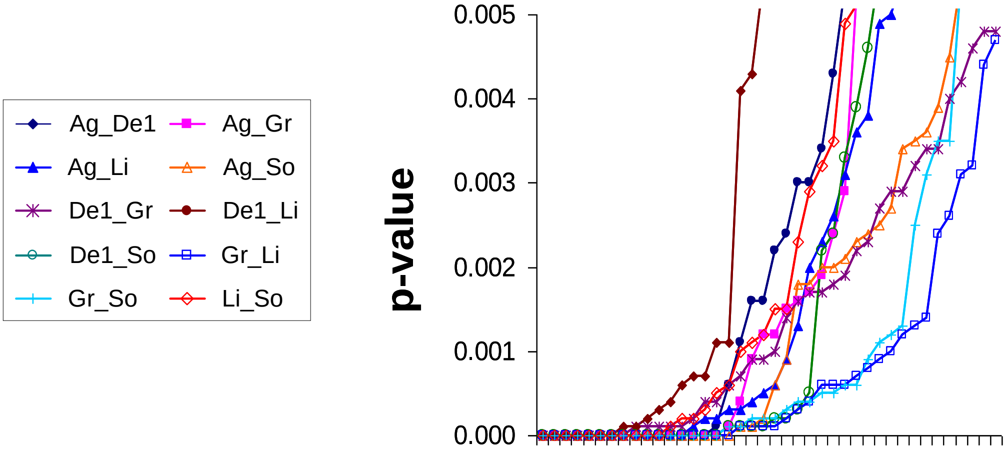
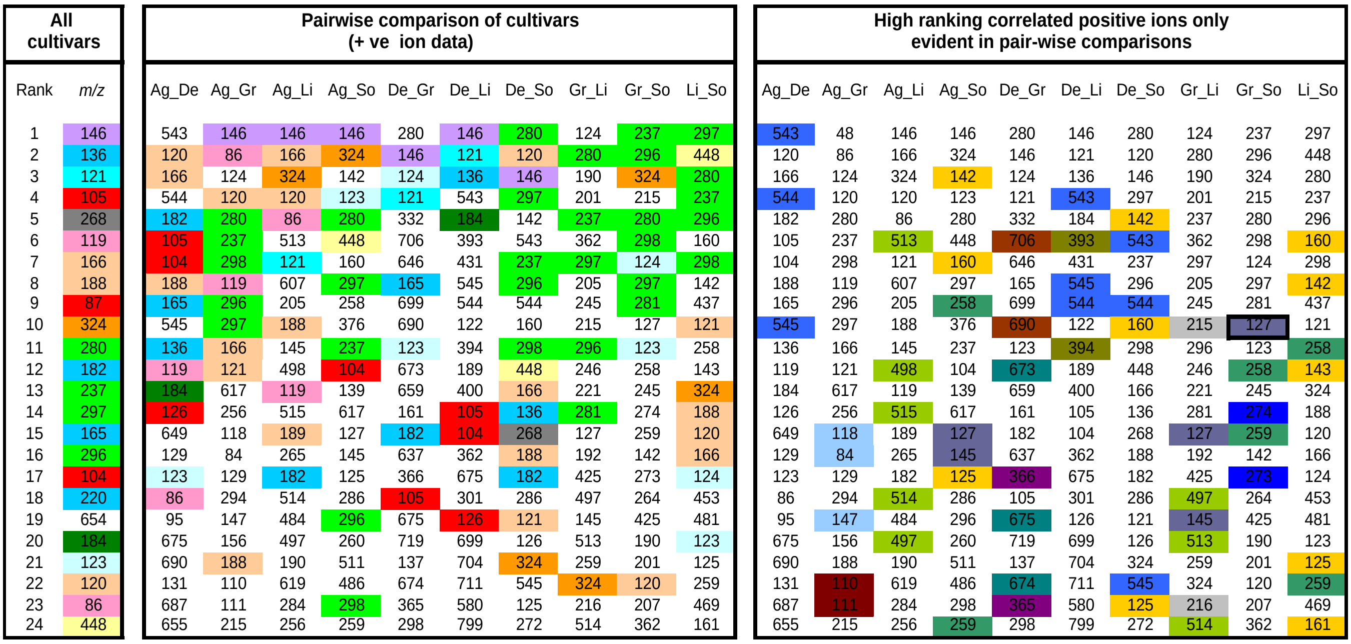


Figure 9.16: Significance of signals highlighted by Random Forest in pair-wise comparison of potato cultivars after data permutation. A *p*-value of around 0.002 is considered to be a reasonable threshold for significance as above this level p-values increased dramatically with RF rank.

9.3.3 Interpretation of metabolome differences detected by FIE-MS (+ ve ion data) in potato cultivars.

For each cultivar combination, the behaviour of the top 24 variables selected by Random Forest was examined in detail. The top 24 variables for discrimination of all cultivars (left panel, Figure 9.17) included many that were potentially related (e.g. 104/105 and 296/297 could represent isotopes). In the pair-wise comparisons these putatively linked (colour coded) signals clustered and were even more highly ranked (centre panel, Figure 9.17). Each pair-wise model contained specific combinations of such groupings and, importantly, further signals potentially linked to the same metabolites were additionally highly ranked, strengthening the possibility that they were derived from the same metabolite. Furthermore, a range of potentially related ions not present in the top 24 signals discriminating all cultivars were evident in specific subsets of the pair-wise comparisons (right panel, Figure 9.17).



All cultivars

Rank	*m/z*
1	146
2	136
3	121
4	105
5	268
6	119
7	166
8	188
9	87
10	324
11	280
12	182
13	237
14	297
15	165
16	296
17	104
18	220
19	654
20	184
21	123
22	120
23	86
24	448

Pairwise comparison of cultivars
(+ ve ion data)

Ag_De	Ag_Gr	Ag_Li	Ag_So	De_Gr	De_Li	De_So	Gr_Li	Gr_So	Li_So
543	146	146	146	280	146	280	124	237	297
120	86	166	324	146	121	120	280	296	448
166	124	324	142	124	136	146	190	324	280
544	120	120	123	121	543	297	201	215	237
182	280	86	280	332	184	142	237	280	296
105	237	513	448	706	393	543	362	298	160
104	298	121	160	646	431	237	297	124	298
188	119	607	297	165	545	296	205	297	142
165	296	205	258	699	544	544	245	281	437
545	297	188	376	690	122	160	215	127	121
136	166	145	237	123	394	298	296	123	258
119	121	498	104	673	189	448	246	258	143
184	617	119	139	659	400	166	221	245	324
126	256	515	617	161	105	136	281	274	188
649	118	189	127	182	104	268	127	259	120
129	84	265	145	637	362	188	192	142	166
123	129	182	125	366	675	182	425	273	124
86	294	514	286	105	301	286	497	264	453
95	147	484	296	675	126	121	145	425	481
675	156	497	260	719	699	126	513	190	123
690	188	190	511	137	704	324	259	201	125
131	110	619	486	674	711	545	324	120	259
687	111	284	298	365	580	125	216	207	469
655	215	256	259	298	799	272	514	362	161

High ranking correlated positive ions only
evident in pair-wise comparisons

Ag_De	Ag_Gr	Ag_Li	Ag_So	De_Gr	De_Li	De_So	Gr_Li	Gr_So	Li_So
543	48	146	146	280	146	280	124	237	297
120	86	166	324	146	121	120	280	296	448
166	124	324	142	124	136	146	190	324	280
544	120	120	123	121	543	297	201	215	237
182	280	86	280	332	184	142	237	280	296
105	237	513	448	706	393	543	362	298	160
104	298	121	160	646	431	237	297	124	298
188	119	607	297	165	545	296	205	297	142
165	296	205	258	699	544	544	245	281	437
545	297	188	376	690	122	160	215	127	121
136	166	145	237	123	394	298	296	123	258
119	121	498	104	673	189	448	246	258	143
184	617	119	139	659	400	166	221	245	324
126	256	515	617	161	105	136	281	274	188
649	118	189	127	182	104	268	127	259	120
129	84	265	145	637	362	188	192	142	166
123	129	182	125	366	675	182	425	273	124
86	294	514	286	105	301	286	497	264	453
95	147	484	296	675	126	121	145	425	481
675	156	497	260	719	699	126	513	190	123
690	188	190	511	137	704	324	259	201	125
131	110	619	486	674	711	545	324	120	259
687	111	284	298	365	580	125	216	207	469
655	215	256	259	298	799	272	514	362	161

Figure 9.17: The top 24 *m/z* ranked by *Importance Score* in different RF models. The left panel is a combined model involving all cultivars and *m/z* considered to be related are color coded. The middle panel shows the *m/z* ranked by Random Forest in pair-wise comparisons of potato cultivars; all signals previously selected in the combined model are similarly color coded. The right panel illustrates new sets of potentially related ions found only in pair-wise comparisons.

A correlation analysis performed using variables with *p*-values < 0.02 similarly identified clusters of signals typical of individual cultivar comparisons (Figure 9.18). Essentially the same behaviour was identified in negative ion data.

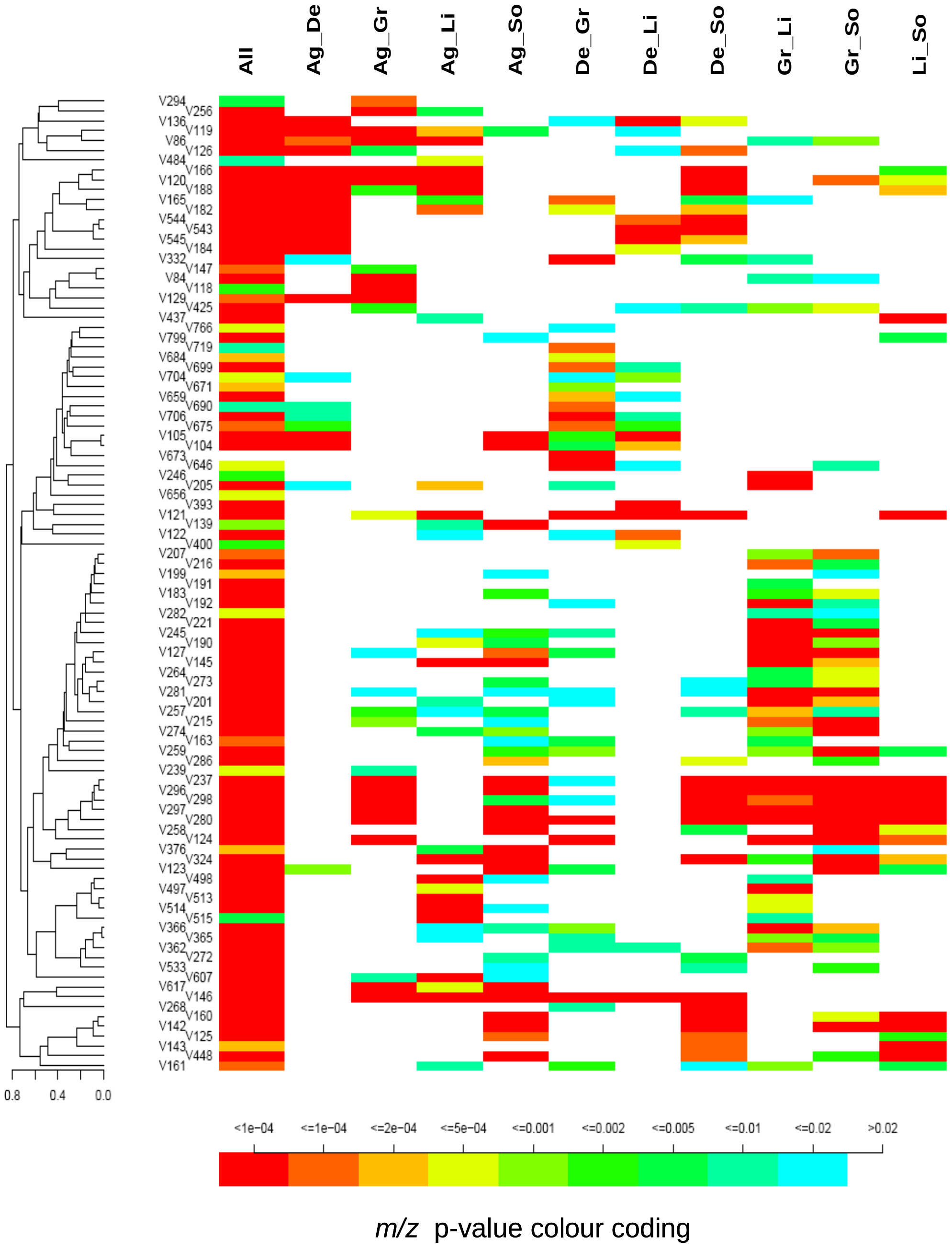


Figure 9.18: Correlation analysis of signals (positive ion data) using a subset of significant (*P* <= 0.01) variables selected by a Random Forest model comparing all of the five cultivars. A 'heatmap' representation is shown where each variable is colour coded according to significance in model.

Interpretation of the correlated FIE-MS with ARMec FIE-MS data annotation tool developed in the G03 project using the relationship between Désirée and other cultivars as an example, it is evident (Figure 9.19) that many of such groups of signals potentially represent isotopes, salt adducts and fragments of known metabolites.

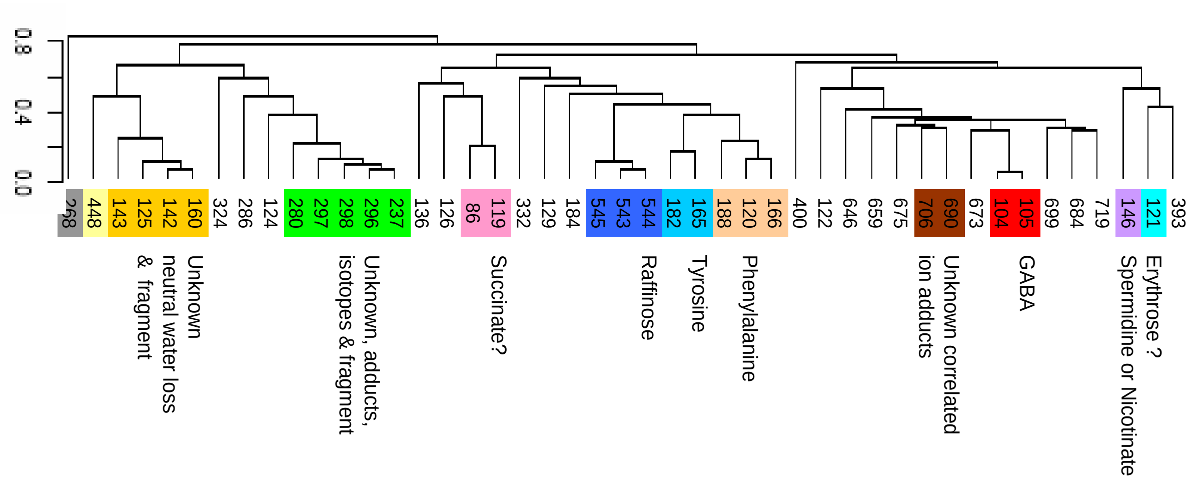


Figure 9.19: Section of a correlation analysis in a comparison of Desiree with all other cultivars. Signals thought to be related to the same metabolite (adducts and isotopes) are colour coded as in Figure 9.18 and potential annotations from the ARMeC database are included.

Some metabolites were highlighted in both ionisation modes (e.g. raffinose and tyrosine), whilst others were found only in negative or positive ion data (e.g. aspartate, gluconate, leucine/isoleucine, GABA, quinate and chlorogenate). In all instances the related *m/z* were found in the top 20-30 variables ranked by RF suggesting that an *Importance Score* threshold of 0.004 was adequate.

The GC-MS metabolite profiling data relating to the same set of potato cultivars was used for Random Forest analysis. Pair-wise comparison of the GC-MS cultivar data generated higher importance scores in comparison to the much more highly dimensional FIE-MS data, with values levelling off between rank 8-15 (Figure 9.20).

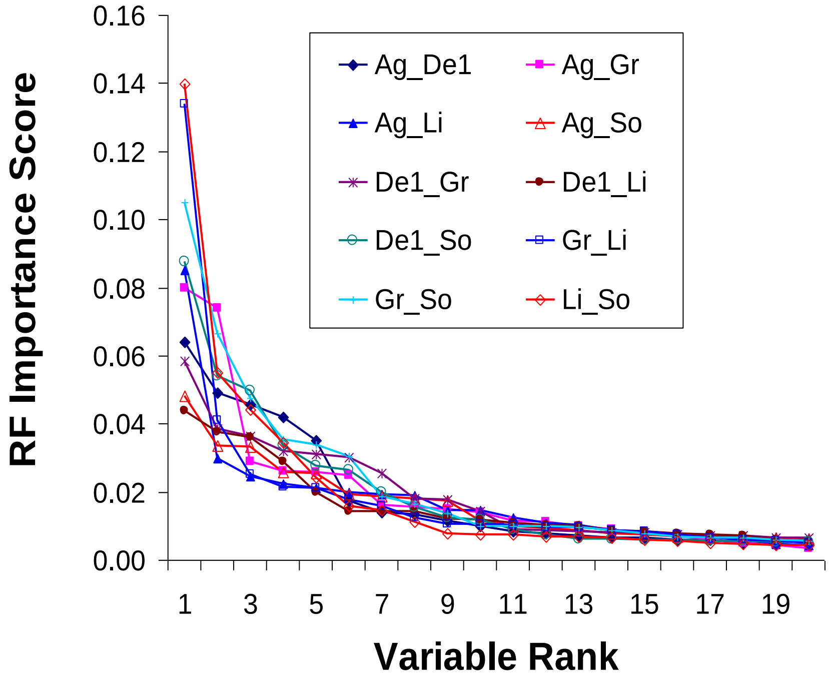


Figure 9.20: Relationship between *Importance Score* and variable rank in Random Forest analysis of GC-MS data representing pair-wise comparisons of 5 potato cultivars

All models representing pair-wise comparisons of GC-MS profiles had a high classification accuracy (>92%) and variables with *p*-values > 0.001 were only evident below rank 17 in all models. Where peak identity was known, the highlighted GC-MS variables were matched against the list of explanatory metabolites predicted by high throughput FIE-MS analysis (Figure 9.21A). Using compositional comparisons with Désirée as an example, it can be seen that explanatory peaks (colour-coded) corresponding to tyrosine, raffinose, phenylalanine, GABA, gluconate, isoleucine, leucine, methionine and aspartate were generally all highly ranked in the same cultivar-specific pattern in the FIE-MS data (Figure 9.21B).

A

Rank	De_Ag	De_Gr	De_Li	De_So
1	Tyrosine-2	U0014	Quinic acid	Tyrosine-2
2	Raffinose	Raffinose	U049	Isoleucine
3	L-Phenylalanine	Quinic acid	Gluconic acid	U-082
4	U-130	U019	U-130	Quinic acid
5	U-088	L-Aspartic-acid	U0014	L-Phenylalanine
6	Glucose-2	U-088	U055	U-094
7	Tyrosine-1	Phosphate	U-129	Lysine
8	GABA	Leucine	Inositol	Leucine
9	Gluconic acid	U-130	U-128	L-Threonine
10	U-124	Gluconic acid	U-073	U-130
11	Lysine	Inositol	U-091	U055
12	Isoleucine	Glucose-2	Tyrosine-2	Gluconic acid
13	Methionine	U-108	Methionine	Raffinose
14	U-094	Tryptophan	U-093	L-Serine-1
15	Inositol	GABA	U016	Tryptophan

B

Rank	De_Ag		De_Gr		De_Li		De_So	
	neg.	pos.	neg.	pos.	neg.	pos.	neg.	pos.
1	180	543	377	280	195	146	130	280
2	360	120	317	146	176	121	180	120
3	164	166	379	124	353	136	181	146
4	195	544	318	121	165	543	163	297
5	503	182	378	332	354	184	164	142
6	223	105	353	706	533	393	360	543
7	163	104	354	646	355	431	353	237
8	404	188	319	165	98	545	118	296
9	148	165	88	699	477	544	354	544
10	567	545	215	690	135	122	477	160
11	147	136	191	123	621	394	539	298
12	159	119	101	673	279	189	243	448
13	354	184	380	659	148	400	176	166
14	353	126	179	161	96	105	503	136
15	129	649	242	182	163	104	443	268
16	609	129	192	637	503	362	212	188
17	539	123	218	366	391	675	417	182
18	377	86	120	105	158	301	147	286
19	378	95	533	675	534	126	470	121
20	636	675	132	719	111	699	116	126
21	635	690	122	137	267	704	448	324
22	522	131	130	674	189	711	405	545
23	116	687	189	365	443	580	203	125
24	130	655	144	298	173	799	218	272

Figure 9.21: Confirmation of FIE-MS fingerprint models by GC-MS. (A) Random forest ranking of GC-MS metabolite peaks used to discriminate between different potato cultivars. (B) Nominal mass *m/z* are colour coded as in (A) revealing ARMec annotations of highly ranked FIE-MS signals in pair-wise comparisons of potato cultivars

9.3.4 Relationship between metabolite content and potato cultivar characteristics.

A good correspondence was evident between high RF ranking and the relative concentrations of the selected metabolites in individual cultivars (**Figure 9.22).** For example, tyrosine was present at a considerably higher concentration in a Désirée background than in either Solara or Agria and was the most highly ranked metabolite signal for discrimination between these cultivars.

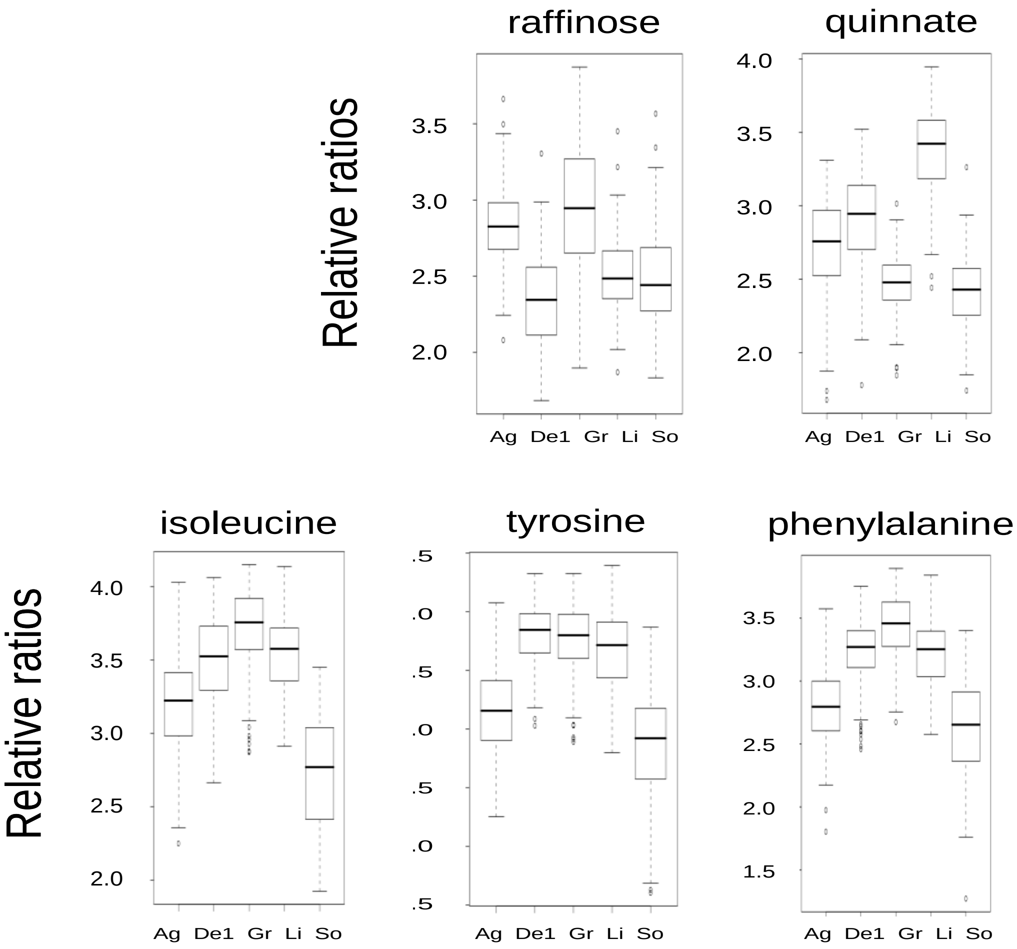


Figure 9.22. Box plots of the concentration (relative ratios) measured by GC-MS of selected explanatory metabolites in 5 potato cultivars.

Many of the identified metabolites that contributed significantly to compositional differences between the potato cultivars are linked closely to quality traits in potato tubers (Figure 9.23A). Isoleucine, leucine, tyrosine and phenylalanine are all known to contribute to flavour and aroma properties of cooked potatoes [74, 75] . For example isoleucine/leucine and tyrosine

A

	Agria	Desiree	Granola	Linda	Solara
After cooking blackening	None-Trace	Trace-Little	Trace-Little	N/D	None-Trace
Taste	Good-Excellent	Moderate-Good	Moderate-Good	Good	Good
French fry suitability	Good-V. Good	Moderate-Good	Poor-Moderate	N/D	N/D
Frying colour	Pale	Medium	N/D	N/D	N/D

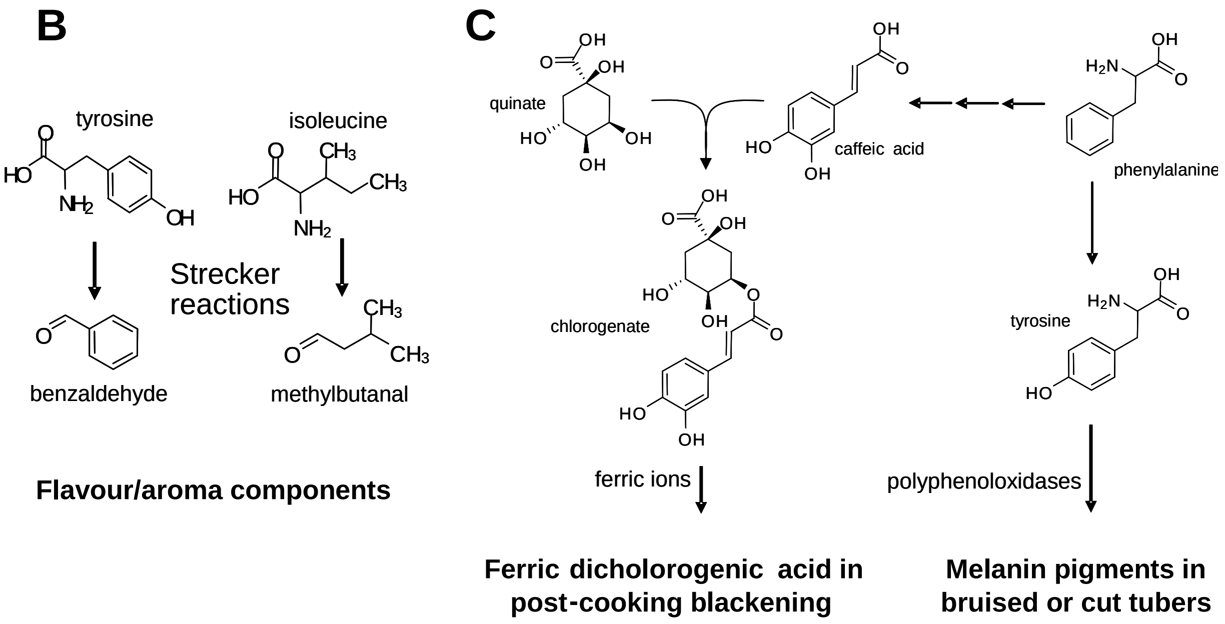


Figure 9.23: Relationship between explanatory metabolites and potato quality traits

major substrates for Strecker reactions (Figure 9.23B) that produce volatile aldehydes (methylbutanals and benzaldehyde respectively) which contribute ‘almond’ and ‘toasted/sweet’ aromas to boiled potatoes. Free tyrosine, is also a major substrate for polyphenoloxidases [76] which are responsible for undesirable melanin biosynthesis in mechanically damaged potatoes (Figure 9.23C). Tyrosine is relatively low in cultivars such as Solara and Agria which are suitable for slicing and frying, and high in Désirée and Granola which are unsuitable (Figure 9.22). Similarly, chlorogenic acid (and by implication its precursors phenylalanine, quinate and caffeate) is linked to non-enzymatic reactions associated with post-cooking blacking[77] in cultivars such as Granola and Désirée (Figure 9.23C).

9.4 Optimising the replication strategy in large scale metabolomics experiments

It is well known that the number of replicates has a major impact on the outcome of any multivariate data mining experiments. The number of replicates required to generate robust classification models essentially is a function of the degree of compositional differences between sample classes and the level of biological and technical variance. Hence the more similar the biological classes are, or the more variance contained in the data, the greater the number of replicates will be needed to develop robust models. Despite this obvious fact it is not at all common to find published data in which a replication strategy was validated prior to sampling; in many published experiments biological variability is represented by only 3 or 4 replicate samples. In order to investigate the behaviour of statistics with a different sample size, the potato cultivar data from the G02 project, ESI2001, was used to generate artificial data. To conduct experiments on the sample size effect our strategy consisted of generating a very large number of pseudo samples to 1) avoid bias introduced by reusing the identical original samples many times and 2) look at model characteristics when a number of replicates larger than the number at hand (in this case 48). The R code used to generate 2000 pseudo samples per class is shown below:

```
>artif3 <- function(mat, cl, nd = 1000, cutoff = 0.99,
+ noise = 1.05) {
+ nlev = nlevels(cl)
+ newmat <- newcl <- NULL
+ for (k in 1:nlev) {
+ omat = mat[cl == levels(cl)[k], ]
+ pc = prcomp(omat)
+ lpc = which(cumsum(pc$sdev^2/sum(pc$sdev^2)) >=
+ cutoff)[1]
+ meg = apply(pc$x[, 1:lpc], 2, mean)
+ seg = apply(pc$x[, 1:lpc], 2, sd)
+ tmp = matrix(rnorm(n = nd * lpc, mean = meg,
+ sd = seg * noise), ncol = lpc, byrow = T)
+ newm = sweep(t(pc$rotation[, 1:lpc] %*%
+ t(tmp)), 2, -pc$center)
+ newmat = rbind(newmat, newm)
+ newcl = c(newcl, rep(levels(cl)[k], nd))
+ }
+ newcl = factor(newcl)
+ return(list(mat = newmat, cl = newcl))
+ }
```

Additionally, the original dataset is used as an independent test set to calculate accuracies and average sample margin in an unbiased way. Note that as the approach of generating artificial data is expected to give slightly optimistic results the variance has been inflated slightly (by 5%).

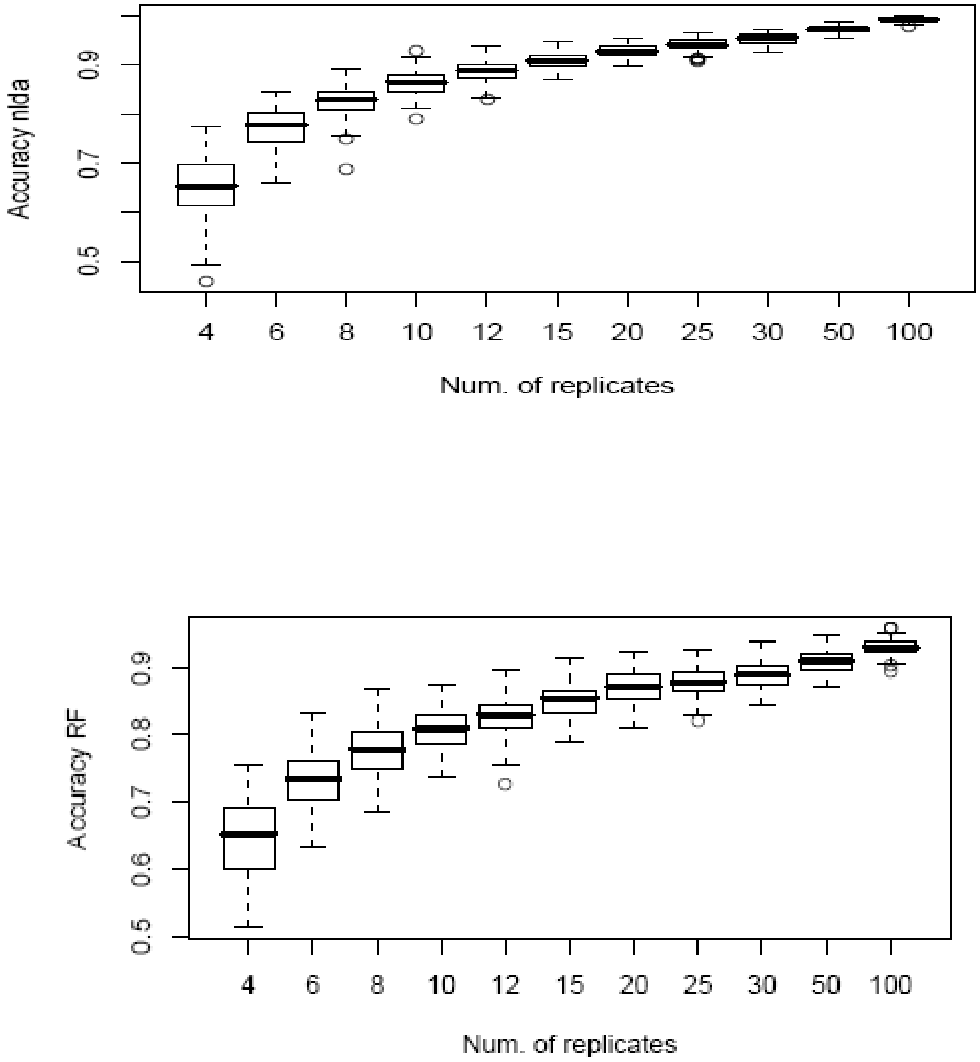
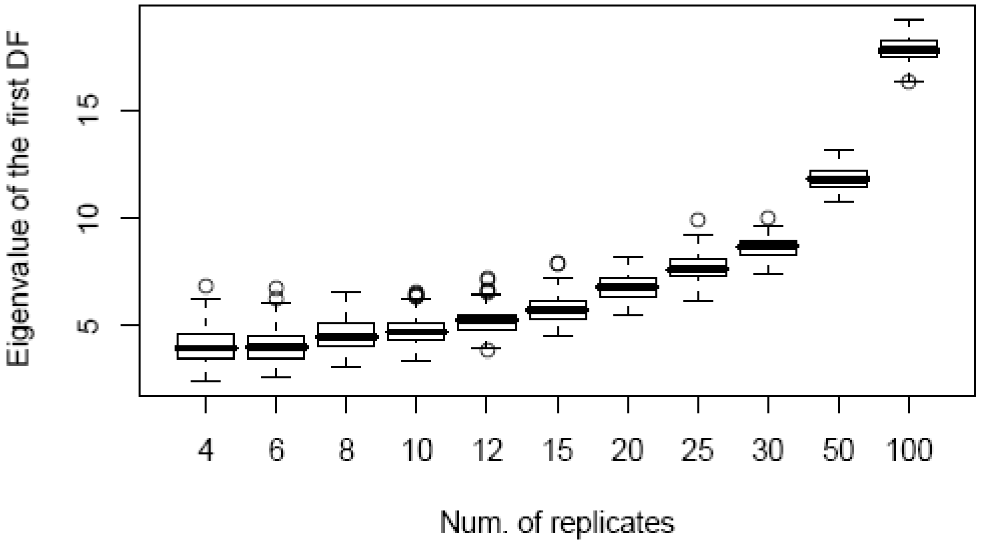


Figure 9. 24: Independent test set accuracies for NLDA and RF models built with different number of replicates in the training data.

9.4.1 Replication strategy in a multiple class context

For different number of replicates (num. Replicates = 4, 6, 8, 10, 12, 15, 20, 25, 30, 50 and 100) per class, 100 models discriminating the 6 potato cultivar classes were generated using Linear Discriminat Analysis (nlda) and Random Forest (RF, ntree =1000) as described in Section 8 of the GO3 report. The classification accuracies on the independent test set (i.e. original ESI dataset), the eigenvalue of the first DF alongside its percentage of variance explained and the average sample margins calculated on the actual training data and the external test are summarised in the Figure 9.24 – 9. 26.



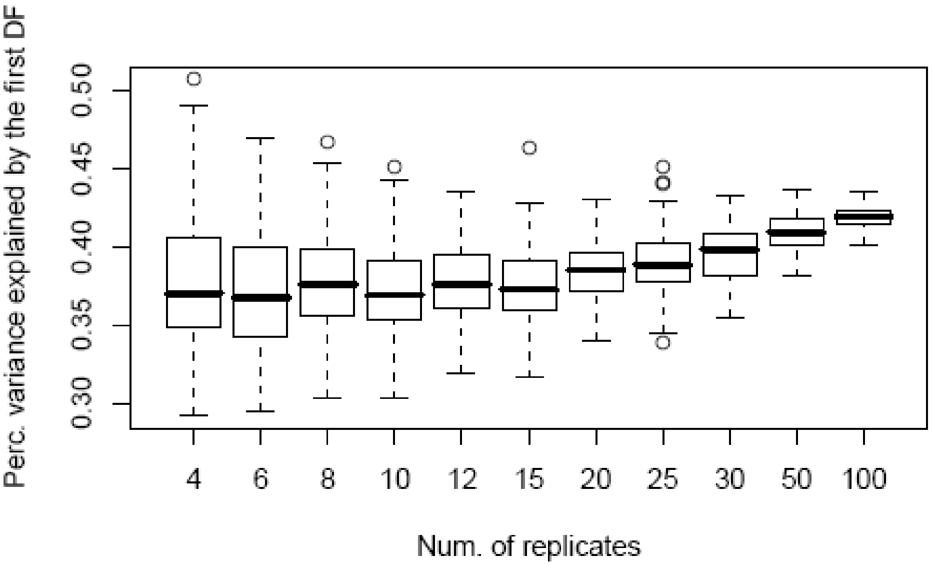
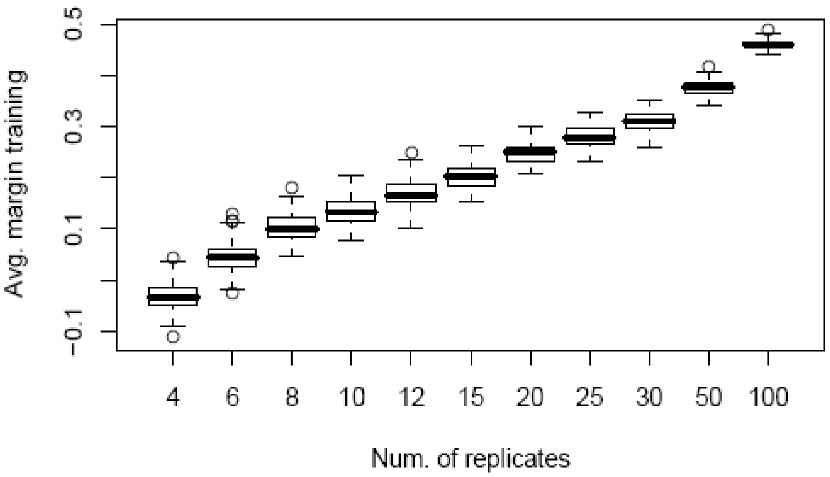


Figure 9.25: Eigenvalue and percentage of variance explained in the first DF when different number of replicates are used in the training data

As expected, classification accuracies on the independent test set are improved when more replicates are used to build the classifiers (Figure 9.24). Eigenvalues on the first DF also increase as the number of replicates increases (Figure 9.25). Interestingly, the percentage of variance explained by the first DF is highly variable when only a few samples are used to train the classifier (Figure 9.25). The error associated with the accuracy estimation is also reduced with larger training sets. This is also translated by greater test sample margins (Figure 9.26). In this particular case, a suitable number of replicates giving robust models is somewhere around 15 to 20 samples. In terms of the model characteristics such as Eigenvalue or average training sample margin, the associated statistical measure are significantly improved with greater sample size. These date illustrate that one can be more confident regarding the predictive abilities of the model based solely on the information available from the training data.



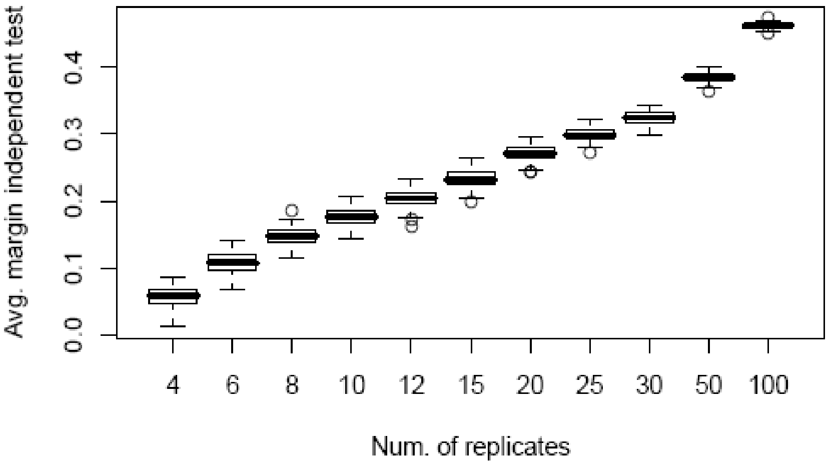


Figure 9.26: Average sample margin in the training set and independent set when different number of replicates are used in the training data

9.4.2 Replication strategy in a two class context

Using the same pseudo data as before, 5 pairwise comparisons involving Desiree1 and the other cultivars (Agria, Linda, Granola, Solara and Desiree2) were examined to evaluate the impact of the number of replicates on the model characteristics in a two class context. Note that the comparison between De1 and De2 is not ”biologically meaningful” as differences observed are encapsulating experimental variability rather than true genotypical biological information (see Section 9.3).

As the number of replicates is increased in a two class problem, the evolution of the independent test set accuracies (RF and NLDA), eigenvalue on the DF and average sample margin is given in the plots below (Figure 9.27-9.29):

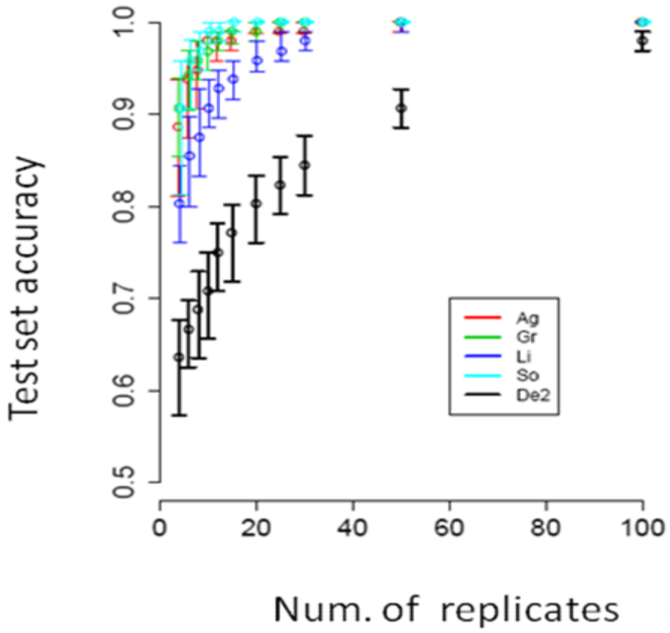


Figure 9.27: Independent test set accuracies for NLDA pairwise models built with different number of replicates in the training data

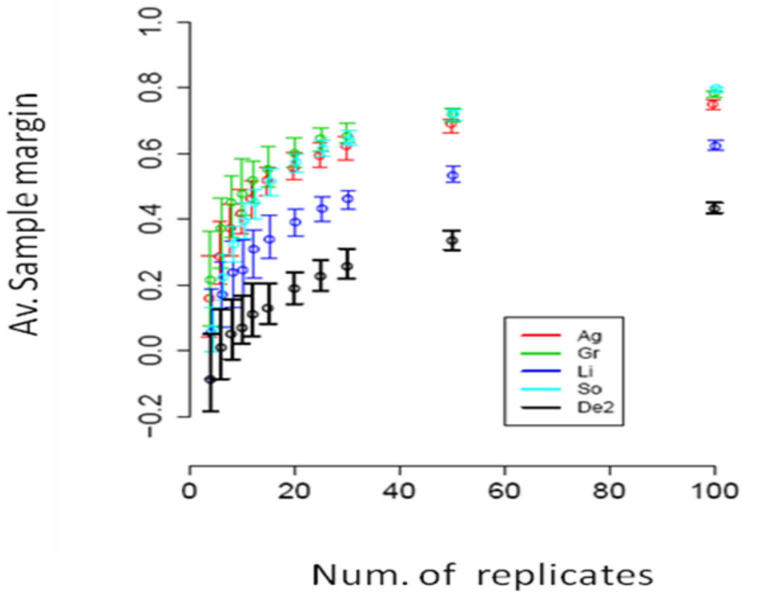


Figure 9.28 Average sample margin in the training set for various pair-wise comparisons when different number of replicates are used in the training data

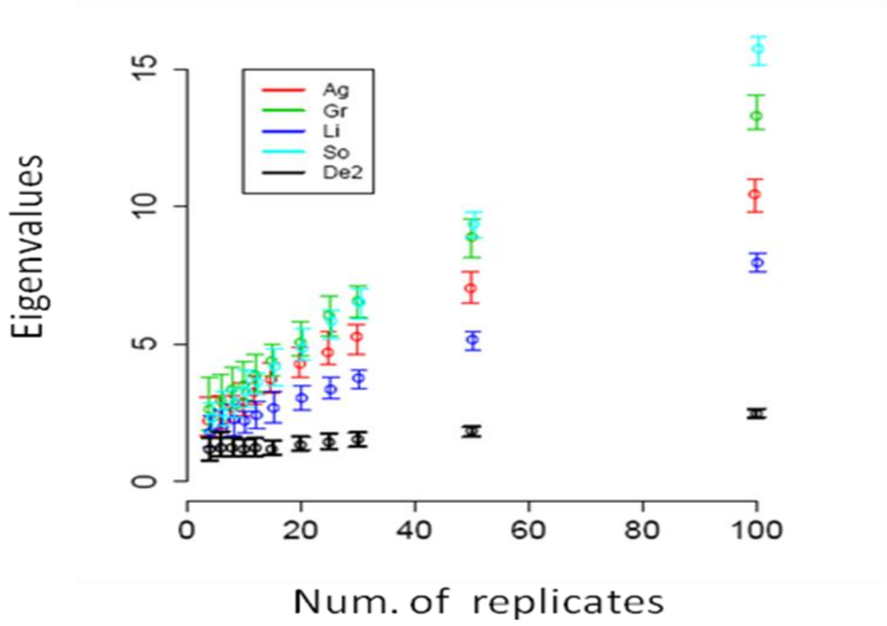
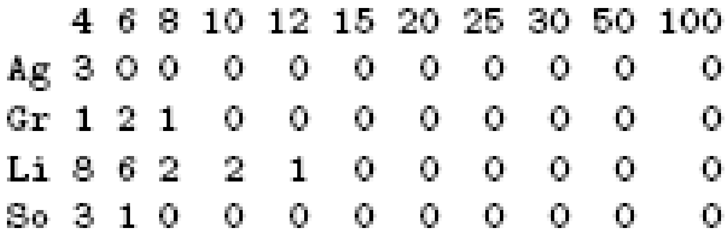


Figure 9.29 Eigenvalue for various pair-wise comparisons when different number of replicates are used in the training data.

The number of times a given statistics (respectively LDA eigenvalue and average sample margin) is greater or equal to the one observed for the De1-De2 comparison for different number of training samples is given in Table 9.2.

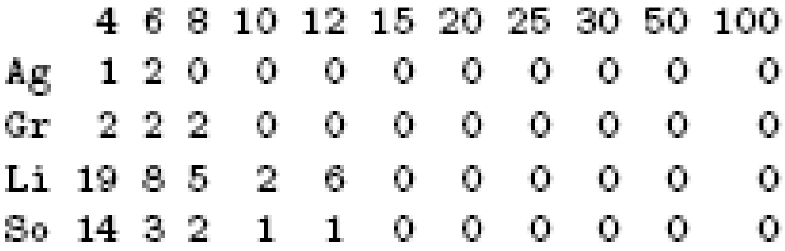
Table 9.2 Effect of replication on detecting biologically significant differences

LDA Eigenvalues



	4	6	8	10	12	15	20	25	30	50	100
Ag	3	0	0	0	0	0	0	0	0	0	0
Gr	1	2	1	0	0	0	0	0	0	0	0
Li	8	6	2	2	1	0	0	0	0	0	0
So	3	1	0	0	0	0	0	0	0	0	0

RF margins



	4	6	8	10	12	15	20	25	30	50	100
Ag	1	2	0	0	0	0	0	0	0	0	0
Gr	2	2	2	0	0	0	0	0	0	0	0
Li	19	8	5	2	6	0	0	0	0	0	0
So	14	3	2	1	1	0	0	0	0	0	0

Without a great deal of significance testing, Table 9.2 illustrates that the risk of rejecting a valid comparison (such as De1-Li) using one ”null hypothesis” when the number of training sample is small. This effect is even more pronounced when the average training sample margin is used as a comparator. This naive approach can be used to determine the adequate number of replicates.

9.4.3 Conclusions regarding optimization of sample size

Using a restricted number of samples the results from supervised multivariate analysis must be treated with extra care unless biological classes with expected massive differences are studied. In most situations, tools and associated statistics in use routinely (margin and eigenvalue) are not suitable to give a robust answer for a scenario when only 4 replicates/class have been analysed as the risk of accepting/rejecting a model under those conditions is quite high considering the actual value of the statistics and the error associated with its estimation. In the context of the G03 project this precludes any multivariate modelling of the LC-MS, GC-MS and ESI-MS data from the 24 potato cultivars from the NOVRISK project which contain only 4 sample replicates (2 from each of 2 harvest years).

From the experiments described above, a suitable number of replicates must exceed 15-20 representatives. This study shows that model significance metrics are not only sample size (in the sense of num. of features versus num. of sample) dependent but also algorithm dependent illustrating that alternative approaches may tackle more efficiently this problem. Although the use of margin/eigenvalue could be seen as rather exotic approaches to estimate class separability, estimates of classification accuracy absolutely do not give reliable answers in a small sample size setting.

Analysis of the learning curves (i.e. evolution of the accuracy with regard to training data sample size) reveal some interesting aspects. The shape of the curve can help to determine a reasonable number of replicates to obtain satisfactory significance thresholds. It also illustrates that there is no need to collect more samples once optimal learning has been achieved, thus preventing larger amounts of variability associated with technical issues entering the analysis. This is particularly evident with the large GC-MS G02 data sets generated in Aberystwyth and Golm in which batch problems effect modelling efficiency. However, a major drawback with this approach to determine a correct sample size is that the dataset at hand must be large enough in the first place to allow a realistic modelling of the relation between N and the statistics under study so that generated data are as close as possible to what we might expect from a true experimental sampling. Thus with a new sample matrix we would advocate performing a pilot run with a few genotypes/classes with large replication in order to perform this type of analysis before designing larger experiments with a reduced number of replicates.

9.5 A proposed database strategy for the comparison of cultivar global composition based on a ranked list of explanatory variables

To provide a more useful framework for compositional comparisons between larger numbers of plant genotypes and to provide possibilities for data integration between laboratories a single metabolome representation will be required as a comparator. Clearly if only one genotype is used as a comparator then any genotypes that are compositionally similar will generate very poor models in pair-wise comparisons and *vice versa* any very different cultivars will form very strong models. We considered that an artificial population of samples that captured the compositional variability representative of all 5 cultivars (ie. ‘mastermix’) would provide a common comparator for all genotypes and produce models of similar robustness.

9.5.1 Validating the method used to generate a Mastermix sample population

In pilot experiments an electronic ‘mastermix’ population of FIE-MS fingerprints was generated as outlined below. The applied term "electronic mastermix" essentially represents a new class of potato within the cultivar matrix. This technique is known in the microarray community as "pooling", where individual samples are combined either "*in vitro*" or "*in silico*" to reduce inherent subject-to-subject variability. In this manner, changes observed in the pooled samples are more representative of population changes. In microarray experiments, samples with the same outcome of interest (i.e. same class) are pooled to explain differences between treatments rather than delineating conclusions from the global population itself. In our case, the objective was to form a virtual new potato class; wherein, the new comparator encapsulated the overall metabolic diversity of the potato cultivar population. Our strategy centred on randomly pooling a single FIE-MS fingerprint from each class where the median intensity, at all *m/z* of the selected fingerprints, was used to create a virtual comparator sample. In order to restrict the inevitable reduction in variance, a sample was used only once when generating the electronic mastermix.

To assess the validity of the approach, we conducted experiments in which the mastermix was formed on training data generated by: (1) a single random selection of 32 samples from each class, or, (2) by bootstrapping of the original data so that classes may have been unbalanced. Test data were used to assess the predictive power and both simulations were repeated 100 times. Out-of-bag samples accuracies and classifier margins were used to compute statistics relative to the training set. Samples not used to form the mastermix were then used to assess the predictive power of each pair-wise classifier. A modified version of the Kendall concordance test was applied to compare the lists of the top k ranked features. In this study, we used k=10 and k=20 as they corresponded roughly to the "expected" number of significant features. Statistics for each pair-wise (i.e. mastermix vs. cultivar X comparison) Random Forest model are summarized in Table 9.3 (1) . As expected, the averaged statistics were slightly lower when more perturbation was induced into the training data (Table 9.3 (2). However, accuracies on unseen data were more satisfactory in all cases, illustrating that the combination of multiple samples originating from diverse cultivars did not degrade the predictive power of the RF models. Confidence in attributing the test sample was fairly high, corresponding to that expected from the training set with a good degree of concordance between the top 10 and 20 ranked features.

Table 9.2 Performance of Random Forest models using an electronic "mastermix" in binary comparisons with potato cultivars

(1) Training data generated by a single random selection of 32 samples from each class.

		Accuracy		Margin		Concordance		
		Mean	SD	Mean	SD	k	Mean	SD
Agr	Train	95.67	1.77	0.493	0.019	10	0.797	0.043
	Test	100.00	0.00	0.582	0.021	20	0.724	0.034
De1	Train	92.56	1.90	0.448	0.023	10	0.705	0.051
	Test	99.69	1.37	0.524	0.028	20	0.672	0.037
De2	Train	95.14	2.22	0.431	0.022	10	0.745	0.053
	Test	100.00	0.00	0.428	0.025	20	0.709	0.036
Gra	Train	95.39	1.18	0.637	0.015	10	0.753	0.057
	Test	100.00	0.00	0.584	0.015	20	0.738	0.035
Lin	Train	91.06	2.36	0.485	0.018	10	0.701	0.062
	Test	87.50	1.26	0.410	0.019	20	0.714	0.042
Sol	Train	99.83	0.49	0.663	0.014	10	0.818	0.061
	Test	100.00	0.00	0.639	0.014	20	0.821	0.030

(2) Training data generated by bootstrapping of the original data so that classes may have been unbalanced

		Accuracy		Margin		Concordance		
		Mean	SD	Mean	SD	k	Mean	SD
Agr	Train	95.90	1.84	0.515	0.031	10	0.757	0.050
	Test	95.49	4.77	0.508	0.068	20	0.686	0.033
De1	Train	90.26	3.18	0.437	0.032	10	0.639	0.046
	Test	90.89	9.68	0.416	0.079	20	0.602	0.034
De2	Train	94.14	3.23	0.396	0.033	10	0.699	0.051
	Test	98.13	3.82	0.413	0.063	20	0.656	0.038
Gra	Train	97.33	1.40	0.627	0.027	10	0.690	0.070
	Test	96.36	4.70	0.637	0.071	20	0.708	0.045
Lin	Train	90.70	3.26	0.447	0.032	10	0.614	0.062
	Test	89.61	8.33	0.428	0.094	20	0.605	0.045
Sol	Train	99.89	0.43	0.640	0.024	10	0.737	0.076
	Test	100.00	0.00	0.639	0.053	20	0.782	0.030

9.5.2 Random forest pair-wise comparisons of cultivar FIE-MS data with an electronic mastermix population of FIE-MS fingerprints.

Importance score ranking by RF revealed that 10-20 *m/z* signals were highly significant to discriminate each cultivar from the mastermix population (**Figure 9.30A**). To confirm the robustness of this approach 100 mastermix populations were generated by random pooling of fingerprints and the pairwise comparison of each cultivar performed using RF. In all instances the explanatory signals identified and their rank order showed high correspondence in relation to importance score (**Figure 9.30B**).

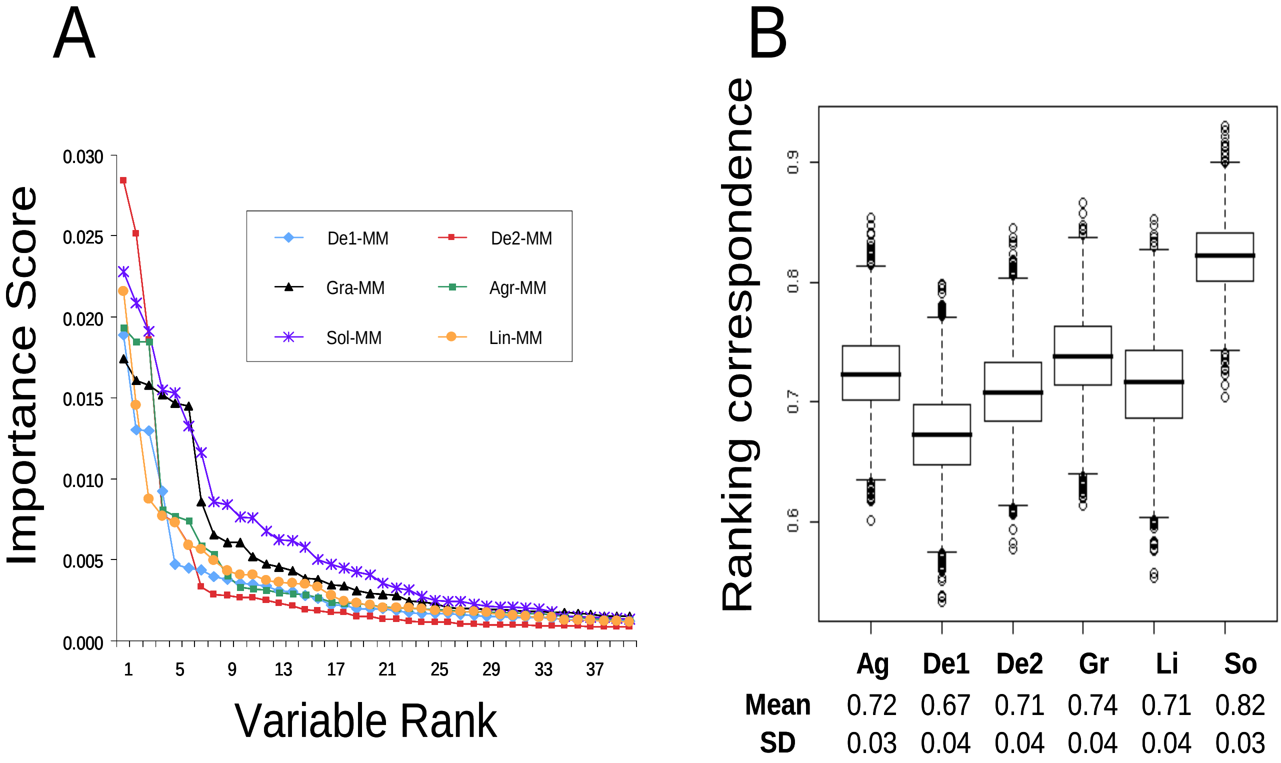


Figure 9.30: Use of a 'mastermix' reference population to generate a unique cultivar composition representation based on explanatory variables ranked by Random Forest. (**A**) Relationship between *Importance Score* and variable rank in a Random Forest analysis of FIE-MS data involving the pairwise comparison of each cultivar with a mastermix fingerprint population. (**B**) Signal ranking correspondence in Random Forest pairwise comparisons of cultivar FIE-MS fingerprints with 100 different randomly-generated 'mastermix' populations.

Figure 9.31A shows in detail the rank order of explanatory signals selected in a RF analysis model comparing all potato cultivars against the Mastermix polulation.. Representative (colour coded) combinations of these highlighted signals populate the top 10 rank positions in pairwise comparisons of individual cultivars with the mastermix population (Figure 9.31.B). Interestingly, samples representing two independently propagated clones of the cultivar Désirée (De1 and De2) exhibited high correspondence (Figure 9.31C).

A

RF Rank	*m/z*
1	146
2	136
3	121
4	105
5	268
6	119
7	166
8	188
9	87
10	324
11	280
12	182
13	237
14	297
15	165
16	296
17	104
18	220
19	654
20	184
21	123
22	120
23	86
24	448
25	221
26	141
27	696
28	95
29	712
30	257

All cultivars

B

	Agria	Granola	Linda	Solara
RF Rank	*m/z*	*m/z*	*m/z*	*m/z*
1	120	146	121	280
2	146	124	146	146
3	166	280	263	142
4	184	296	362	237
5	188	121	247	160
6	182	237	513	448
7	306	298	514	296
8	562	294	189	324
9	118	297	261	297
10	256	216	161	124

C

	Desiree 1	Desiree 2
RF Rank	*m/z*	*m/z*
1	121	268
2	543	146
3	146	136
4	268	182
5	544	527
6	136	121
7	182	184
8	125	673
9	722	528
10	786	352

Individual cultivars compared to Mastermix population

Figure 9.31: Signal ranking correspondence in Random Forest pairwise comparisons of cultivar FIE-MS fingerprints with 100 different randomly-generated 'mastermix' populations. (**A**) Top 30 ranked explanatory variables identified in a Random Forest model comparing FIE-MS fingerprints of all five potato cultivars. (**B & C**) Top ten ranked explanatory *m/z* identified in the Random Forest pairwise comparison of cultivar FIE-MS fingerprints with a 'mastermix' fingerprint population.

9.5.3 Conclusions in relation to a future strategy for the compositional comparison of raw food materials

Both representation and subsequent comparison of global chemical composition in plant breeding germplasm and food raw materials are difficult tasks. Additionally, important quality assessments by sensory panels rely heavily on subjective data rather than quantitative measures. Against this background there is an urgent need to develop a data structure and a data analysis strategy that will allow robust, high-throughput comparative assessment of metabolite content. In pair-wise comparisons between cultivars, we show that highly interpretable models generated by the Random Forest analysis of FIE-MS fingerprints can both accurately classify potato tubers and also explicitly identify significant compositional differences. From a utility perspective, the metabolites we predicted to be responsible for compositional differences between potato cultivars by both FIE-MS and GC-MS were indeed associated with important quality traits. Interestingly, several unknown signals were also highly significant, providing scope for the discovery of novel chemical attributes potentially relating to quality characteristics of individual cultivars. In the future we expect that FIE-MS fingerprinting can be used effectively to generate a meaningful and comprehensive representation of compositional differences within/between complex biological samples. The high-throughput nature and the requirement of only small sample volumes, easily allow for the generation of sufficient replicate measurements to ensure for a rigorous generalisability assessment. Unlike many other powerful data mining approaches which strive to produce classifiers comprising largely non-redundant features, the approach we describe using Random Forest allows all variables an equal chance of being both explicitly identified and highly ranked. By comparing FIE-MS data representing individual genotypes to a common ‘mastermix’ fingerprint population, chemical ‘bar codes’ based on *Importance Score* ranking of explanatory variables (e.g. top 25-30 *m/z* in adequate models) can be generated and might form part of a future database strategy regarding the chemical composition of foodstuffs. For example RF ranking could be used to assess compositional similarity between batches of food raw materials, or form part of a phenotyping strategy to evaluate large genotype populations encountered in either plant breeding or functional genomics experiments. From the perspective of directed plant breeding, a similar approach could be envisaged to link metabolome fingerprints to complex quality traits and to identify genetic sources of significant compositional novelty.

REFERENCES

1. Somorjai, R.L., Dolenko, B. & Baumgartner, R. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics* **19,** 1484-1491 (2003).
2. Berrar, D., Bradbury, I. & Dubitzky, W. Avoiding model selection bias in small-sample genomic datasets. *Bioinformatics* **22,** 1245-50 (2006).
3. Braga-Neto, U.M. & Dougherty, E.R. Is cross-validation valid for small-sample microarray classification? *Bioinformatics* **20,** 374-380 (2004).
4. Lyons-Weiler, J. et al. Assessing the Statistical Significance of the Achieved Classification Error of Classifiers Constructed using Serum Peptide Profiles, and a Prescription for Random Sampling Repeated Studies for Massive High-Throughput Genomic and Proteomic Studies. *Cancer Informatics* **1**, 53-77 (2005).
5. Broadhurst, D.I. & Kell, D.B. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* **2**, 171-196 (2006).
6. Saghatelian, A. & Cravatt, B.F. Global strategies to integrate the proteome and metabolome. *Current Opinion in Chemical Biology* **9**, 62-68 (2005).
7. Ein-Dor L., Zuk O. & Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Nat. Acad Sci USA* **103**, 5923–5928 (2006).
8. Dıaz-Uriarte, R. Supervised methods with genomic data: a review and cautionary view. *Data analysis and visualization in genomics and proteomics.* Wiley, New York**,** pp193-214 (2005).
9. Fawcett, T. *ROC graphs: Notes and practical considerations for data mining researchers*. Technical report HPL-2003-4. HP Laboratories, Palo Alto, CA, USA. Available at http://www.hpl.hp.com/techreports/ 2003/HPL-2003-4.pdf. (2003).
10. Mukherjee, S., Roberts, S.J. & van der Laan, M.J. Data-adaptive test statistics for microarray data *Bioinformatics* **21**, 108-114 (2005).
11. Sima, C. & Dougherty, E. R. What should be expected from feature selection in small-sample settings**.** *Bioinformatics* **22**, 2430-2436 (2006).
12. Enot, D.P., Beckmann, M., Overy, D. & Draper, J. Predicting interpretability of metabolome models based on behavior, putative identity, and biological relevance of explanatory signals. *Proc Natl Acad Sci U S A* **103**, 14865-14870 (2006).
13. Kell, D.B., Darby, R.M. & Draper, J. Genomic computing. Explanatory analysis of plant expression profiling data using machine learning. *Plant Physiology* **126**, 943-951 (2001).
14. Catchpole, G.S. et al. Hierarchical metabolomics demonstrates substantial compositional similarity between genetically modified and conventional potato crops. *Proc Natl Acad Sci U S A* **102**, 14458-14462 (2005).
15. Goodacre, R., Vaidyanathan, S., Dunn, W.B., Harrigan, G.G. & Kell, D.B. Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends in Biotechnology* **22**, 245-252 (2004).
16. Bino, R.J. et al. Potential of metabolomics as a functional genomics tool. *Trends in Plant Science* **9**, 418-425 (2004).
17. Fiehn, O. et al. Metabolite profiling for plant functional genomics. *Nature Biotechnology* **18**, 1157-1161 (2000).
18. Sumner, L.W., Mendes, P. & Dixon, R.A. Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry* **62**, 817-836 (2003).
19. Nicholson, J.K. & Wilson, I.D. Understanding 'global' systems biology: Metabonomics and the continuum of metabolism. *Nature Reviews Drug Discovery* **2**, 668-676 (2003).
20. Roessner, U., Wagner, C., Kopka, J., Trethewey, R.N. & Willmitzer, L. Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant Journal* **23**, 131-142 (2000).
21. Tolstikov, V.V. & Fiehn, O. Analysis of highly polar compounds of plant origin: Combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. *Analytical Biochemistry* **301**, 298-307 (2002).
22. Beckmann, M., Enot, D.P., Overy, D.P. & Draper, J. Representation, comparison and interpretation of metabolome fingerprint data for total composition analysis and quality trait investigation in potato cultivars. *J. Agricultural and Food Chemistry* **55**, 3444-3451 (2007).
23. Dear, G.J., James, A.D. & Sarda, S. Ultra-performance liquid chromatography coupled to linear ion trap mass spectrometry for the identification of drug metabolites in biological samples. *Rapid Communications in Mass Spectrometry* **20**, 1351-1360 (2006).

24. Wagner, C., Sefkow, M. & Kopka, J. Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. *Phytochemistry* **62**, 887-900 (2003).

25. Jonsson, P., Gullberg, J., Nordstrom, A., Kusano, M., Kowalczyk, M., Sjostrom, M. and Moritz, T. A strategy for identifying differences in large series of metabolomic samples analyzed by GC/MS. *Anal Chem* **76**, 1738-45 (2004).

26. Vorst, O., de Vos, C.H.R., Lommen, A., Staps, R.V., Visser, R.V., Bino, R. J., and Hall, R.D. A non-directed approach to the differential analysis of multiple LC–MS-derived metabolic profiles. *Metabolomics* **1,** 169-180 (2005).

27. Ward, J.L., Harris, C., Lewis, J. & Beale, M.H. Assessment of H-1 NMR spectroscopy and multivariate analysis as a technique for metabolite fingerprinting of Arabidopsis thaliana. *Phytochemistry* **62**, 949-957 (2003).

28. Allen, J. et al. High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nature Biotechnology* **21**, 692 - 696 (2003).

29. Scholz, M., Gatzek, S., Sterling, A., Fiehn, O. & Selbig, J. Metabolite fingerprinting: detecting biological features by independent component analysis. *Bioinformatics* **20**, 2447-2454 (2004).

30. Aharoni, A. et al. Nontargeted metabolome analysis by use of Fourier Transform Ion Cyclotron Mass Spectrometry. *Omics* **6**, 217-234 (2002).

31. Smedsgaard, J. & Frisvad, J.C. Using direct electrospray mass spectrometry in taxonomy and secondary metabolite profiling of crude fungal extracts. *Journal of Microbiological Methods* **25**, 5-17 (1996).

32. Dunn, W.B., Bailey, N.J. & Johnson, H.E. Measuring the metabolome: current analytical technologies. *Analyst* **130,** 606-625 (2005).

33. Beckmann, M., Parker, D., Enot, D.P., Duval, E. & Draper, J. (2008) High throughput, non-targeted metabolite fingerprinting using nominal mass Flow Injection Electrospray Mass Spectrometry**.** *Nature Protocols,* **3**, 486-504.

34. Overy, D.P., Enot, D.P., Tailliart, K., Jenkins, H., Parker, D., Beckmann, M. & Draper, J. (2008) Explanatory signal interpretation and metabolite identification strategies for nominal mass FIE-MS metabolite fingerprints. *Nature Protocols*, **3**, 471-485.

35. Parker, D., Beckmann, M., Enot, D.P., Overy, D.P., Rios, Z.C., Gilbert, M., Talbot, N. & Draper, J.(2008) Rice blast infection of *Brachypodium distachyon* as a model system to study dynamic host/pathogen interactions. *Nature Protocols*, **3**, 435-445.

36. Enot, D.P., Beckmann, M. & Draper, J. Detecting a difference - assessing generalisability when modelling metabolome fingerprint data in longer term studies of genetically modified plants. *Metabolomics* **3**, 335-347 (2007).

37. Enot, D.P. & Draper, J. Statistical measures for testing substantial equivalence of GM plant genotypes in a multivariate context. *Metabolomics* **3**, 349-355 (2007).

38. Jain, A.K., Murty, M.N. & Flynn, P.J. Data clustering: a review. *ACM Computing Surveys (CSUR)* **31,** 264-323 (1999).

39. Manly, B.F.J. *Multivariate Statistical Methods: A Primer*. Chapman & Hall/CRC (2004).

40. Zhang, C., Lu, X. & Zhang, X. Significance of Gene Ranking for Classification of Microarray Samples. *EEE/ACM Transactions on Computational Biology and Bioinformatics* **3**, 312-320 (2006).

41. Ransohoff, D.F. Rules of evidence for cancer molecular-marker discovery and validation. *Nature Reviews Cancer* **4**, 309-313 (2004).

42. Davis, C.A., Gerick, F., Hintermair, V., Friedel, C.C., Fundel, K., Küffner, R. & Zimmer, R. Reliable gene signatures for microarray classification: assessment of stability and performance. *Bioinformatics* **22**, 2356-2363 (2006)**.**

43. Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K. & Zhao, H. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics* **19,** 1636-1643 (2003).

44. Cristianini, N. & Shawe-Taylor, J. *'An introduction to support vector machines and other kernel-based learning methods'*, Cambridge University Press (2000).

45. Zhu, C., Kitagawa, H. & Faloutsos, C. Example-based outlier detection for high dimensional datasets. *IPSJ Digital Courie*r **1**, 234-243 (2005)

46. Craig, A., Cloarec, O., Holmes, E., Nicholson, J.K & Lindon, J.C. Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Anal Chem* **78**, 2262-2267 (2006).

47. Hastie, T., Tibshirani, R. & Friedman, J. *The elements of statistical learning: data mining, inference, and prediction*. Springer (2001).

48. Good, P. *Permutation tests: A practical guide to resampling methods for testing hypotheses*. Springer series in statistics (2000).

49. Efron, B. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association* **78,** 316-331 (1983).

50. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **21,** 3940-3941 (2005).

51. Fu, W.J., Carroll, R.J. & Wang, S. Estimating misclassification error with small samples via bootstrap cross-validation. *Bioinformatics* **21,** 1979-1986 (2005).

52. Thomaz, C.E., Boardman, J.P., Hill, D.L.G., Hajnal, J.V., Edwards, D.D., Rutherford, M.A., Gillies, D.F. & Rueckert, D. Using a Maximum Uncertainty LDA-Based Approach to Classify and Analyse MR Brain Images. *Lecture notes in computer science***:** *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2004*, pp291-300, Springer Berlin (2004).

53. Yang, J. & Yang, J. Why can LDA be performed in PCA transformed space? *Pattern Recognition* **36,** 563-566 (2003).

54. Breiman, L. Random Forests. *Machine Learning* **45,** 5-32 (2001).

55. Zar, J.H. Biostatistics. 2nd edn. Englewood Cliffs, New Jersey: Prentice-Hall (1984).

56. Dietterich, T.G. Ensemble methods in machine learning. *Lecture Notes in Computer Science* **1857,** 1-15 (2000).

57. Vaidyanathan, S., Kell, D.B. & Goodacre, R. Flow-injection electrospray ionization mass spectrometry of crude cell extracts for high-throughput bacterial identification. *J Am Soc Mass Spectrom* **13**, 118-128 (2002).

58. Roessner, U., Luedemann, A., Brust, D., Fiehn, O., Linke, T., Willmitzer, L. & Fernie, A. Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* **13,** 11-29 (2001).

59. Mazzella, N. et al. Use of electrospray ionization mass spectrometry for profiling of crude oil effects on the phospholipid molecular species of two marine bacteria. *Rapid Communications in Mass Spectrometry* **19**, 3579-3588 (2005).

60. Favretto, D., Piovan, A., Filippini, R. & Caniato, R. Monitoring the production yields of vincristine and vinblastine in *Catharanthus roseus* from somatic embryogenesis. Semiquantitative determination by flow-injection electrospray ionization mass spectrometry. *Rapid Communications in Mass Spectrometry* **15**, 364-369 (2001).

61. Rashed, M.S., Al-Ahaidib, L.Y., Aboul-Enein, H.Y., Al-Amoudi, M. & Jacob, M. Determination of L-pipecolic acid in plasma using chiral liquid chromatography-electrospray tandem mass spectrometry. *Clinical Chemistry* **47**, 2124-2130 (2001).

62. Overy, S.A. et al. Application of metabolite profiling to the identification of traits in a population of tomato introgression lines. *Journal of Experimental Botany* **56**, 287-296 (2005).

63. Goodacre, R., York, E.V., Heald, J.K. & Scott, I.M. Chemometric discrimination of unfractionated plant extracts analyzed by electrospray mass spectrometry. *Phytochemistry* **62**, 859-863 (2003).

64. Koulman, A. et al. High-throughput direct-infusion ion trap mass spectrometry: a new method for metabolomics. *Rapid Communications in Mass Spectrometry* **21**, 421-428 (2007).

65. Martinez, A.M. & Kak, A.C. PCA versus LDA. *IEEE Transactions on: Pattern Analysis and Machine Intelligence,* **23,** 228-233 (2001).

66. Windeatt, T. Vote counting measures for ensemble classifiers. *Pattern Recognition* **36**, 2743-2756 (2003).

67. R_Development_Core_Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-900007-900050, URL http://www.R-project.org (2006)

68. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **57**, 289–300. (1995).

69. Storey, J. D. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**, 479–498 (2002).

70. Enot, D.P., Lin, W., Beckmann, M., Parker, D., Overy, D.P. and Draper, J. (2008) Pre-processing, classification modelling and feature selection using Flow Injection Electrospray Mass Spectrometry (FIE-MS) metabolome fingerprint data. *Nature Protocols*, **3**, 446-470.

71. Steinfath, M, Groth, D., Lisec, J. and Selbig, J. (2008) Metabolotite profile analysis: from raw data to regression and classification. *Physiologia Plantarum*, **132,** 150-161.

72. Van den Berg, R.A., Hoefsloot, H. C. J., Westerhuis, J.A., Smilde, A.K. and van der Werf, M. J. (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* **7**, 142.

73. Parsons, H. M., Ludwig, C., Gunther, U.L and Viant, M.R. (2007) Improved classification accuracy in 1- and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation. *BMC Bioinformatics*, **8**, 234.

74. Martin, F.; Ames, J. M. (2001) Formation of Strecker aldehydes and pyrazines in a fried potato model system. *J. Agric. Food Chem*., **49**, 3885-3892.

75. Duckham, S. C.; Dodson, A. T.; Bakker, J.; Ames, J. M. (2001) Volatile flavour components of baked potato flesh. A comparison of eleven potato cultivars. *Nahrung*,, **45,** 317-323.

76. Corsini D. L.; Pavek J. J.; Dean, B. (1992) Differences in free and protein-bound tyrosine among potato genotypes and the relationship to internal blackspot resistance. *Am. Potato J*., **69,** 423-435.

77. Dao L.; Friedman, M. (1992) Chlorogenic acid content of fresh and processed potatoes determined by ultraviolet spectrophotometry. *J. Agric. Food Chem*. , **40,** 2152-2156.

ASSESSMENT REPORT

Name & Address of Research Contractor	University of Wales (Aberystwyth) Institute of Biological Sciences Edward Llwyd Building Aberystwyth SY23 3DA	**Date of Assessment**	12 February 2008
		Date of Assessment Plan	3 February 2008
		Funding Body	FSA
Assessment Location	As above	**Contact**	Prof John Draper
Assessment Criteria	Joint Code of Practise for Research	**UKAS Assessor**	Rachel Oakley
Research Contractor Representatives	**Project Representatives** Prof John Draper Dr Nigel Hardy Manfred Beckman David Enot	**Project Number**	G03012
		Project Reference	Development of unified data models and data pre-processing strategies and the generation of meaningful, standardised statistical analyses of metabolome variability in crop plants
Report Issued by	Rachel Oakley	**Recommendations**	3
Report Issued Date	22/2/08	**Report Acknowledged Method**	Email

1. Executive Summary

The project has been well planned and has been effectively managed. The project team was cooperative and open to ideas. From the discussions during the audit they demonstrated a good awareness and understanding of the requirements of the Joint Code of Practise and systems implemented were compliant with those requirements in a majority of areas. Only a small number of recommendations were made.

Overall this project has been conducted to a high standard.

2. Scope

The purpose of this visit was to carry out an audit of the FSA funded research project - *Development of unified data models and data pre-processing strategies and the generation of meaningful, standardised statistical analyses of metabolome variability in crop plants* which is being undertaken by the University of Aberystwyth in order to assess their compliance with the requirements of Joint Code of Practise for Research.

The assessment involved discussions with key members of the project team at the University of Aberystwyth. Much of the work has involved data collection and analysis although some practical work has also been conducted at the University. The audit included discussion about the status and progress of the project, project personnel, records of on-going reviews, staff recruitment, training and on-going development as well an inspection of the facilities and equipment used to conduct the practical work. Discussions were held with key members of the project team in order to understand their roles and activities on this project.

3. Overview of Research Contractor

The project is being carried out by the Institute of Biological Sciences in conjunction with the Computer Science department. Both departments operate within the University of Aberystwyth.

Most of the work on this project has been done by the University, although there has been one collaborator Scottish Crop Research Institute (SCRI), who has been involved to some extent in the collection, pre-processing and statistical analysis of data. The collaborator has been made aware of the requirements of Code of Practise for Research. The University has had close contact with SCRI throughout the project through regular review meetings.

University staff employed on this project have been apointedd on the basis of their previous experience from working on similar projects at the University. Other new staff have been employed specifically to work on this project and have appropriate experience.

The project has largely progressed according to the identified timescales. Some changes have been made and some objectives abandoned following review of information obtained, but these have been agreed with the FSA and are appropriately documented.

4. Observations & Assessment Findings

4.1 Responsibilities

4.1.1 Project Personnel

The project is being jointly conducted by the Institute of Biological Sciences and Computer Sciences departments at the University of Aberystwyth. The Project Coordinator is Prof John Draper who has been managing the project and he is ably supported by other project team members.
The project team has a high level of experience. All were aware of their responsibilities in relation to their role within the project and the requirements of the JCoPR.

4.1.2 Subcontractors

SCRI has been collaborating in some aspects of this project and have been closely involved in the on-going review meetings.

No other subcontractors have been used, except for those organisations who were requested to supply information to assist with the literature search. Their input has been limited to the supply of information.

4.2 Competence

4.2.1 CVs of Project Personnel

Project personnel clearly have a high level of experience. Experience and qualifications are documented. Many of the personnel working on the project were already working at the University and had worked on similar projects in the past and therefore had appropriate experience. The University has a personnel policy, which ensures that any new staff employed are subjected to appropriate assessment and the laboratory was able to provide evidence to show this policy had been followed during the recruitment of staff to work on the project. The recruitment system enables fair and thorough assessment to ensure the appropriate candidate is selected.

4.2.2 Training Records

The University has specific policies in relation to training of staff, which the laboratory was able to demonstrate were being followed.

Induction training is conducted and this ensures that Health and Safety and other administrative issues are covered.

Training of staff is planned and reviewed on an on-going basis to ensure that skills relevant to the area of work are obtained and competency maintained. There is a good system for staff review and continual development.

Training records are generally well-documented and where it is appropriate competency is assessed through use of e.g. QC data/materials, which enables effective demonstration of competence.

4.3 Project Planning

4.3.1 Risk Assessment

Potential risks had been taken into account at the project planning stage. This identified a potential risk with obtaining information from the external organisations, which the laboratory was reliant on to progress some of the key objectives of the project. Adequate provision was made to minimise this risk and in the event this did not materialise and the information was provided by organisations when requested, enabling the project to progress as planned.

4.3.2 Project Plan

Project planning has been done effectively and the project plan is clearly detailed showing the objectives, responsibilities and anticipated timescales. On-going planning and review is well-organised and the internal reviews enable progress to be closely monitored.

Two major external reviews have been conducted according to plan involving the FSA at 6 and 18 months and this has enabled further on-going review. Following these reviews some aspects of the project were reviewed and a decision made to abandon them. This was agreed with the FSA at the time and records are held.

4.3.3 Approved Procedures for Sampling Materials

No specific requirements for this project. Potato material supplied by SCRI has been tested and is currently being stored frozen, although if further samples were required fresh material would need to be obtained.

4.3.4 Ethical Approval

None required for this project.

4.4 Quality Control

4.4.1 Auditing & Assessment Procedures

There is no formal auditing system as such, except those relating to Health and Safety activities, which are audited by the University. There are good systems in place for on-going review of progress against project objectives, which allows for any issues to be identified and appropriate corrective action taken and this is considered sufficient for this project.

4.4.2 Internal Project Reviews

Progress on the project has been monitored through use of internal review meetings, which has involved the project team and representatives from the SCRI. These meetings are recorded and show an in-depth review which has enabled issues to be identified and appropriately followed-up.

4.4.3 Publication Policy & Authorisation Procedures

The laboratory has published a number of standardised protocols in relevant journals. These publications have been agreed with the FSA via the review meetings. The laboratory is aware of the need to agree any future publications with the FSA beforehand.

4.5 Health & Safety

The laboratory has to follow the University wide policy for Health and Safety and there are clear guidelines regarding these requirements. It is apparent that safety assessments are done, however some gaps were noted in the records held by the laboratory, although these will have been held by the Departmental Safety Officer. It is still advisable for the laboratory to ensure copies of risk assessments relevant to their areas of work are located in the laboratory to enable easy access by all relevant staff. Evidence of staff training in safety requirements was not documented, although lab confirmed that in practise staff would follow correct safety procedures, however the records are needed to support this.

4.6 Handling of Samples & Materials

The only samples handled by the laboratory are those received from the SCRI. Records for those samples tested by the laboratory are satisfactory.

4.7 Facilities & Equipment

The equipment used by the laboratory to analyse the potato samples suppled by SCRI appeared to be well-maintained. It was a little unclear as to whether the requisite calibration checks were always being done on the FIEMS, but for the level of accuracy required the checks are in place are probably adequate. It would be useful to ensure that there is a clear record to show when the necessary checks are done to provide evidence that the equipment was operating satisfactorily when in use. Other more sensitive equipment available was subjected to more regular calibrations and records of these were held.

Other equipment used for testing appears fit for purpose. Where checks are done these should be recorded where possible to support the effective performance of the equipment, e.g. pipettes. Lab should consider whether increased checking of balances (e.g. with check weights) is necessary particularly where they are being used to weigh out critical measures.

4.8 Documentation of Procedures & Methods

4.8.1 Validated Standard Operating Procedures

Practical work done by the University is subject to QC checking where appropriate and this supports the methods in use.

Where separate operating procedures are used in the laboratory, these could be better controlled. Laboratory records could be improved to provide evidence that procedures have been followed in practise, e.g. particularly in relation to equipment checks, where requirements are documented in the protocols, but there are no records generated by the lab to support this.

4.8.2 Document Control Procedures

Document control practises would benefit from review particularly where standard operating procedures are held, e.g. which cover use of equipment. If documents are formally issued then this helps when further changes are needed and the document needs to be re-issued.

4.9 Research/Work Records

4.9.1 Experimental Records, Sampling records, Project related records

Appropriate records are being maintained in a suitable format. Most records are stored electronically and there are mechanisms in place to protect this data. The systems in place ensure good traceability.

4.9.2 Data Management & Archiving Procedures

This appears to be under effective control. Most data is stored electronically and protection of this information is controlled to ensure appropriate security and back-up.

The laboratory has an excellent web based system where key documents relevant to the project such as review meeting minutes, project plan are being stored. This is accessible by all members of the project team and provides a very effective mechanism for communicating information about the project across the team.

Retention times of documents has been agreed with the FSA. Records need to be stored where they can be readily accessed following the completion of the project.

5. References

Joint Code of Practice for Research

6. Appendices

Improvement Action Report & Recommendations prepared by Rachel Oakley

G02 Data Selection

Sponsor: G03R0002 project

Reference: $Source: /
dcs/fsaweb/G03CVS/g03r0002rep/management/planning/g02dataselection/doc.xml,v $

Version: $Revision: 1.2 $

Date: $Date: 2005/11/28 16:43:47 $

Author: Nigel Hardy

Table of Contents

1. Document objectives

Deliverable 01.01 will be:

> *... a catalogue of the labs, experiments and technologies in use in selected G02 projects, focusing on potato and tomato ...*

This document seeks to structure the assembly of that catalogue through a provisional checklist of issues to be considered.

This document should be considered a draft and comments and additions are welcomed. It should not restrict collection of information and comments about potentially useful G02 data, though in view of the timescales, it might help in prioritising tasks.

The result of this cataloguing exercise will be the deliverable document [1].

2. Information about a data set

For each data set under consideration or selected the following classes of information will help the process. A form is available (Appendix 1) to assist with collecting this information.

Project. A name for the project under which data were collected. "G02006" would be an example.

Experiment. A name or reference for the experiment so that (along with the Project) it can be identified. Experiment should be broadly interpreted to include "trial", "observations" and other procedures.

Laboratory. The institution, site and/or laboratory responsible for the experiment.

Contact. The person best able to discuss data availability, structure and meaning.

Description. A few sentences to explain the structure and intention of the experiment.

Rationale. The rationale for (perhaps) including this experiment in the G03 project.

Species. The species or organism concerned. Cultivars, strains, ecotypes, transgenics etc. should be mentioned.

Growth. Indication of growth conditions (field, greenhouse, growth room etc.) including sequences (e.g. germination in one environment followed by transplantation etc.).

Harvest. Brief description of how samples were obtained.

Handling. Brief description of whether and how samples were transported and stored.

Preparation. Brief indication of the sample preparation methods used (typically in the laboratory).

Analysis machine. The machine used for the metabolome estimation. Indicate the technology (e.g. GC-MS), the maker and model. Any special techniques or procedures should be noted.

Type of result. The nature of the available results (e.g. "peak list", "spectrum", absolute or relative measurements).

Format of results. The format(s) in which the results are available (e.g. "Excel spreadsheets", CSV, "GenStat").

Reference. Available papers, reports etc. where the experiment is described or reported in more detail. G02 final reports will be very suitable, but please include page or section numbers as appropriate.

Respondent. The person completing this form.

1. G02 data cataloguing

This proforma table may assist in collecting information about selected G02 data sets. Explanation of categories is give above (Section 2).

It consists of two sections. If an experiment involves, for example, more than one metabolome estimate method (e.g. GC-MS and FT-IR) complete Part A in full and a copy of Part B for each alternative, referring back to the complete Part A Project and Experiment to avoid repetition.

1. Part A

Item	**Details**
Project	
Experiment	
Laboratory	
Contact	
Description	
Rationale	
Species	
Growth	
Harvest	
Handling	

Item	Details

2. Part B

<table>
<tr><td>Project (From Part A)</td><td colspan="2">Experiment (From Part A)</td></tr>
<tr><td><b>Item</b></td><td colspan="2"><b>Details</b></td></tr>
<tr><td>Preparation</td><td colspan="2"></td></tr>
<tr><td>Analysis machine</td><td colspan="2"></td></tr>
<tr><td>Type of result</td><td colspan="2"></td></tr>
<tr><td>Format of results</td><td colspan="2"></td></tr>
<tr><td>Other comments</td><td colspan="2"></td></tr>
<tr><td>Reference</td><td colspan="2"></td></tr>
<tr><td>Respondent</td><td colspan="2"></td></tr>
<tr><td>Date</td><td colspan="2"></td></tr>
</table>

Bibliography

[1] *G02 data catalogue*. Nigel Hardy. http://sirius.dcs.aber.ac.uk:9095/g03r0002/docs/requirements/g02data/.

G02 data catalogue

Sponsor: G03R0002 project

Reference: $Source: /dcs/fsaweb/G03CVS/g03r0002rep/requirements/g02data/doc.xml,v $

Version: $Revision: 1.7 $

Date: $Date: 2006/07/13 11:35:21 $

Authors: Nigel Hardy
Helen Jenkins

Table of Contents

1. Document objectives

This document will be deliverable 01.01.

2. The process of assembly

A document was produced ([1] Version 1.2, 2005/11/28 16:43:47) defining a limited set of details required about each experiment for potential inclusion in the study. This was circulated to three parties: i) Institute of Biological Sciences, UWA (IBS, UWA); ii) Scottish Crop Research Institute (SCRI); iii) (RHUL)

Note

Need details of RHUL

.

Help was offered in completing the form included in that document, via e-mail or 'phone conferences.

The appendices to this document are a collection of those responses.

3. Summary of Responses

3.1. Sources

Data have been received from the following:

- **G02001 Project.** Through SCRI
  - E-mails from Derek Stewart suggested a listof 7 experiments with 6 data sets each.
  - The 'phone conference of 2006/02/16 (http://sirius.dcs.aber.ac.uk:9095/g03r0002/docs/management/minutes/2006-02-16_phone.html) identified some lines from one experiment with five data sets.
  - The document e-mailed by Susan Verrall 2006-13-03 confirmed the lines from the one experiment and increased the data sets to six.
  - At the meeting on 2006-04-20, it was agreed that the FT-IR data set would be dropped (http:// sirius.dcs.aber.ac.uk:9095/ g03r0002/ docs/ management/ minutes/ 2006-04-20.html. This is reflected here.
  - A spreadsheet of data fields sent by Susan Verrall 2006-05-12 (http:// sirius.dcs.aber.ac.uk:9095/g03r0002/data/SCRI/) contained information about 2 SCRI LC-MS experiments. This is reflected here.
- **G02006 Project.** Through UWA, IBS
  - The meeting of 2005/11/28 (http://sirius.dcs.aber.ac.uk:9095/g03r0002/docs/management/minutes/2005-11-28.html) identified 5 data sets. This is reflected here, as two experiments (2001 and 2003) with 3 and 2 data sets respectively. .
  -

    Note

    The GM lines will presumably need to be removed from the G02006 data sets.

Note

Data are expected from RHUL.

3.2. Data Sets

This table sumarises the "Part B" responses. It shows the experiments with their associated data sets.

Project	Experiment	Data Set
G02001	FSA 2002 (Appendix 2)	GC-MS Aqueous extraction and derivitisation (Appendix 2, Section 2)
		GC-MS Non-polar extraction and derivitisation (Appendix 2, Section 3)
		LC-MS extraction (Appendix 2, Section 4)
		LC-MS extraction (set 2) (Appendix 2, Section 5
		LC-MS extraction - Ian Calhoon via Derek (Appendix 2, Section 6)
		NMR extraction - Ian Calhoon -via Derek (Appendix 2, Section 7)
G02006	2001 Field Trial (Appendix 3)	GC-TOF (Appendix 3, Section 2)
		GC Quadrupole (Appendix 3, Section 3)
		DIMS (Appendix 3, Section 4)
	2003 Field Trial (Appendix 4)	GC-TOF (Appendix 4, Section 2)
		DIMS (Appendix 4, Section 3)

1. The questions

The following is a list of the topics specified in [1]

1. Information about a data set

Project.

Experiment.

Laboratory.

Contact.

Description.

Rationale.

Species.

Growth.

Harvest.

Handling.

Preparation.

Analysis machine.

Type of result.

Format of results.

Reference.

Respondent.

2. G02001

1. Part A

Item	Details
Project	G02001
Experiment	FSA 2002: Transcriptome, proteome and metabolome analysis to detect unintended effects in genetically modified potato.
Laboratory	Scottish Crop Research Institute
Contact	Dr Derek Stewart mailto:Derek.Stewart@scri.ac.uk
Description	A fully replicated block experiment to assess if metabolomics can identify changes in GM potatoes. A range of GM lines with the relevant controls were grown in the field along with commercially grown cultivars and landrace species.
Rationale	Rationale for metabolomics only: To ensure the technologies were robust enough to measure several hundreds of small molecules quickly (a technology called metabolomics) rather than targeted analysis of only some key components. Also to test their effectiveness in detecting unintended effects using a range of GM crop materials. The sub-set cultivars will be used in this instance.
Species	Lines 110 to 124 are modern cultivars of potato with known introgression for disease resistance from a variety of wild species. Tetraploids (4 sets of chromosomes). • *S. tuberosum* • line 110, DESIREE • line 111, RECORD • line 112, P.DELL • line 113, SHELAGH • line 114, STIRLING • line 115, TORRIDON • line 116, GLENNA • line 117, MORAG • line 118, EDEN • line 119, M.PIPER • line 120, P.JAVELIN • line 121, CARA • line 122, P.CROWN • line 123, BRODICK • line 124, BARBARA • line 149, 91.MT.46 E 15, Multi trait high dry matter potato cultivar (in National List Trials). • line 150, PINK FIR APPLE, Salad potato modern cultivar. • line 151, G.WONDER, High dry matter, low yielding • line 154, LUMPERS, Old cultivar. • line 155, FORTYFOLD, Old cultivar. • line 156, ANYA, Salad potato modern cultivar. • *S. phureja* • line 152, INCA SUN, Diploid (in National List Trials). Adapted to long days. • line 153, MAYAN GOLD, Diploid (in National List Trials). Adapted to long days.
Growth	5 tubers of each independent GM and control cultivar were planted in rows in a total of four randomised plots within the field. (20 plants per line in total). Plants were grown with standard potato management practices. (Taken from the G02001 final report p.7)
Harvest	Burn down date - first haulm half dose sulphuric acid 28-08-2002 - second haulm, half dose sulphuric acid 03-09-2002. Fungicide/aphicide 04-09-2002.

Item	Details
	Haulm pulverised 11-09-2002. Harvest date 19-09-2002. Tubers stored at 10°C for a minimum of 2 weeks in the dark to harden off. Sampling performed following *Standard Operating Procedure (SOP) For the Sub-Sampling Of Tubers from Bags/Sacks* - see G02001 final report p.39.
Handling	See sample preparation protocols for LC-MS analysis and GC-MS analysis in the G02001 final report pages 41 and 45.

2. Part B - GC-MS Aqueous extraction and derivitisation

Project: G02001	Experiment: FSA 2002
Item	**Details**
Preparation	See *Extraction of Freeze-dried Potato for GC-MS Analysis* and *Polar Fraction Derivatisation* in the GC-MS Extraction and GC-MS Derivatisation AQ worksheets of the spreadsheet at http:// sirius.dcs.aber.ac.uk:9095/ g03r0002/ data/ SCRI/ also documented as *Extraction of freeze-dried potato* and *Derivatisation of polar fraction* on pages 45 and 46 of the G02001 final report.
Analysis machine	ThermoElectron Tempus GC-TOF-MS See *Analysis of polar and non-polar samples by GC-MS* on pages 47-49 of the G02001 final report also partially documented as *GC-MS machine method* in the GC-MS Machine Method worksheet of the spreadsheet at http:// sirius.dcs.aber.ac.uk:9095/g03r0002/data/SCRI/
Type of result	• absolute area • absolute peak height • relative response to internal standard
Format of results	• Excel spreadsheet containing a sheet per peak • `.raw` Excalibur output • Genstat format sheet
Other comments	
Reference	
Respondent	• Susan Verrall mailto:Susan.Verrall@scri.ac.uk • Tom Shepherd
Date	14 March 2006

3. Part B - GC-MS Non-polar extraction and derivitisation

Project: G02001	Experiment: FSA 2002
Item	**Details**
Preparation	See *Extraction of Freeze-dried Potato for GC-MS Analysis* and *Non-Polar Fraction Derivatisation* in the GC-MS Extraction and GC-MS Derivatisation NP worksheets of the spreadsheet at http:// sirius.dcs.aber.ac.uk:9095/ g03r0002/data/SCRI/ also documented as *Extraction of freeze-dried potato* and *Derivatisation of non-polar fraction* on pages 45 and 46 of the G02001 final report.
Analysis machine	ThermoElectron Tempus GC-TOF-MS See *Analysis of polar and non-polar samples by GC-MS* on pages 47-49 of the G02001 final report also partially documented as *GC-MS machine method* in the GC-MS Machine Method worksheet of the spreadsheet at http:// sirius.dcs.aber.ac.uk:9095/g03r0002/data/SCRI/
Type of result	• absolute area

Item	**Details**
	• absolute peak height • relative response to internal standard
Format of results	• Excel spreadsheet containing a sheet per peak • `.raw` Excalibur output • Genstat format sheet
Other comments	
Reference	
Respondent	• Susan Verrall mailto:Susan.Verrall@scri.ac.uk • Tom Shepherd
Date	14 March 2006

4. Part B - LC-MS extraction

Project: G02001	Experiment: FSA 2002
Item	**Details**
Preparation	See *Extraction of potato samples for LCMS analysis* in the Potato LC-MS 4ml Method worksheet of the spreadsheet at http:// sirius.dcs.aber.ac.uk:9095/ g03r0002/ data/ SCRI/ also documented as the *Sample extraction* protocol on p.41 of the G02001 final report.
Analysis machine	Finnigan LCQ DECA 3D ion trap API source See *LC conditions*/*Tune method*/*MS Method* on pages 41-43 of the G02001 final report also partially documented as *Potato Instrument Method* in the Nigel_polar_MET_MSMS worksheet of the spreadsheet at http:// sirius.dcs.aber.ac.uk:9095/g03r0002/data/SCRI/.
Type of result	• absolute area • absolute peak height • relative response to internal standard
Format of results	• Excel spreadsheet containing a sheet per peak • `.raw` Excalibur output • Genstat format sheet
Other comments	Technician - Michael Anderson
Reference	
Respondent	• Susan Verrall mailto:Susan.Verrall@scri.ac.uk • Julie Sungurtas
Date	9 March 2006

5. Part B - LC-MS extraction (set 2)

Project: G02001	Experiment: FSA 2002
Item	**Details**
Preparation	See *LC-MS potato extraction original method* in the LC-MS Potato Extraction Orig worksheet of the spreadsheet at http:// sirius.dcs.aber.ac.uk:9095/ g03r0002/data/SCRI/
Analysis machine	Finnigan LCQ DECA 3D ion trap API source See *LC conditions*/*Tune Method*/*MS Method* on pages 41-43 of the G02001 final report also partially documented as *Potato Instrument Method* in the Nigel_polar_MET_MSMS worksheet of the spreadsheet at http:// sirius.dcs.aber.ac.uk:9095/g03r0002/data/SCRI/.

Item	**Details**
Type of result	• absolute area • absolute peak height • relative response to internal standard
Format of results	• Excel spreadsheet containing a sheet per peak • `.raw` Excalibur output • Genstat format sheet
Other comments	Technician - Julie Sungurtas
Reference	
Respondent	• Susan Verrall mailto:Susan.Verrall@scri.ac.uk • Julie Sungurtas
Date	9 March 2006

6. Part B - LC-MS extraction - Ian Calhoon via Derek

Project: G02001	Experiment: FSA 2002
Item	**Details**
Preparation	
Analysis machine	
Type of result	
Format of results	
Other comments	
Reference	
Respondent	
Date	

7. Part B - NMR extraction - Ian Calhoon -via Derek

Project: G02001	Experiment: FSA 2002
Item	**Details**
Preparation	
Analysis machine	
Type of result	
Format of results	
Other comments	
Reference	
Respondent	
Date	

3. G02006, 2001

1. Part A

Item	Details
Project	G02006
Experiment	2001
Laboratory	MPIMPP/IBS, UWA
Contact	Manfred Beckman mailto:meb@aber.ac.uk
Description	"a context for determining whether transgenic potatoes displayed alterations in metabolite composition outside the range exhibited normally by conventional cultivars." [4].
Rationale	
Species	*Solanum tuberosum*. Cultivars: Désiré, Agria, Linda, Granola, Solara. GM construtcts based on Désiré.
Growth	Field conditions in a block design.
Harvest	Approximately 48 tubers were selected at random from each of four randomly arranged field blocks
Handling	Stored at 4°C for 4 weeks before sample preparation.

2. Part B - GC TOF-MS

ProjectG02006	Experiment2001
Item	**Details**
Preparation	Potato tuber disks (fresh weight, 200mg each) were excised from 3mm below the tuber peel, perpendicular to the main tuber axis. Immediately after cutting, disks were frozen in liquid nitrogen and kept frozen at -80°C before extraction. Tuber slice homogenization and extraction in 1ml of prechilled water/methanol/chloroform (2:5:2, vol/vol/vol).
Analysis machine	GC TOF-MS. [5].
Type of result	
Format of results	
Other comments	
Reference	[4]
Respondent	NwH [mailto:nwh@aber.ac.uk]
Date	2005/12/15 and subsequently updated until ...

3. Part B - GC Qudrupole-MS

ProjectG02006	Experiment2001
Item	**Details**
Preparation	
Analysis machine	Quadrupole
Type of result	
Format of results	
Other comments	

Item	Details
Reference	
Respondent	
Date	

4. Part B - FIE-MS

ProjectG02006	Experiment2001
Item	**Details**
Preparation	Randomized extracts were diluted 1:50 in water/methanol (60:40, vol/vol), and aliquots of 40µl were injected into a flow of 100µlmin1 water/methanol (60:40, vol/vol) with a Waters Alliance 2690 liquid chromatography (LC) system.
Analysis machine	FIE-MS was performed by using an LCT mass spectrometer (Micromass, Manchester, U.K.). Randomized extracts were diluted 1:50 in water/methanol (60:40 vol/vol), and aliquots of 40µl were injected into a flow of 100µl min1 water/methanol (60:40 vol/vol) using a Waters Alliance 2690 liquid chromatography (LC) system. The flow was split before the ion source to maintain a flow of 50µl min1. Data were collected in positive mode and negative mode every second (0.9-s scan time and 0.1-s interscan delay) for 2 min per sample from m/z 65-1,000. Ionization conditions were set to 3,000-V capillary voltage, 80°C source temperature, 120°C desolvation temperature, 100-V RF lens, 30-V sample cone voltage, and 10-V extraction cone voltage. The raw ion intensity data were binned to nominal mass. (taken from http://www.pnas.org/cgi/content/full/0503955102/DC1#ST)
Type of result	
Format of results	
Other comments	
Reference	
Respondent	NwH [mailto:nwh@aber.ac.uk]
Date	2005-12-15 and subsequently updated until ...

4. G02006, 2003

1. Part A

Item	Details
Project	G02006
Experiment	2003
Laboratory	MPIMPP/IBS, UWA
Contact	Manfred Beckman mailto:meb@aber.ac.uk
Description	
Rationale	
Species	*Solanum tuberosum*. See *G02006, 2001*, Part A
Growth	See *G02006, 2001*, Part A
Harvest	See *G02006, 2001*, Part A
Handling	See *G02006, 2001*, Part A

2. Part B- GC TOF-MS

ProjectG02006	Experiment2003
Item	**Details**
Preparation	See *G02006, 2001*, Part B - GC TOF-MS
Analysis machine	TOF See *G02006, 2001*, Part B - GC TOF-MS
Type of result	
Format of results	
Other comments	
Reference	See *G02006, 2001*, Part B - GC TOF-MS
Respondent	NwH [mailto:nwh@aber.ac.uk]
Date	2005/12/15 and subsequently updated until ...

3. Part B - FIE-MS

ProjectG02006	Experiment2003
Item	**Details**
Preparation	See *G02006, 2001*, Part B - GC TOF-MS
Analysis machine	FIE-MS. See *G02006, 2001*, Part B - FIE-MS
Type of result	
Format of results	
Other comments	
Reference	See *G02006, 2001*, Part B - GC TOF-MS
Respondent	NwH [mailto:nwh@aber.ac.uk]
Date	2005/12/15 and subsequently updated until ...

5. SCRI

To be completed

6. RHUL

To be completed

Bibliography

[1] *G02 Data Selection*. Nigel Hardy. http://sirius.dcs.aber.ac.uk:9095/g03r0002/docs/management/planning/g02dataselection/.

[2] *Final Report: Transcriptome, proteome and metabolome analysis to detect unintended effects in genetically modified potatoes*. Scottish Crops Research Institute. G02001. Food Standards Agency. 135pp.

[3] *G02006 Final report*. NEED DETAILS.

[4] Catchpole, Gareth S. and Beckmann, Manfred and Enot, David P. and Mondhe, Madhav and Zywicki, Britta and Taylor, Janet and Hardy, Nigel and Smith, Aileen and King, Ross D. and Kell, Douglas B. and Fiehn, Oliver and Draper, John. *Hierarchical metabolomics demonstrates substantial compositional similarity between genetically modified and conventional potato crops*. *PNAS*. 0503955102. 2005.

[5] Weckwerth, Wolfram and Loureiro, Marcelo Ehlers and Wenzel, Kathrin and Fiehn, Oliver. *Differential metabolic networks unravel the effects of silent plant phenotypes*. *PNAS %R 10.1073/pnas.0303415101*. 101. 20. 7809-7814. 2004.

G03 Database Design

Reference: $Source: /dcs/fsaweb/G03CVS/g03r0002rep/design/Summer06/doc.xml,v $

Version: Draft

Edition: $Revision: 1.4 $ Edition

Date: $Date: 2006/07/26 14:14:06 $

Author: Dr Helen Jenkins

Table of Contents

1. Introduction

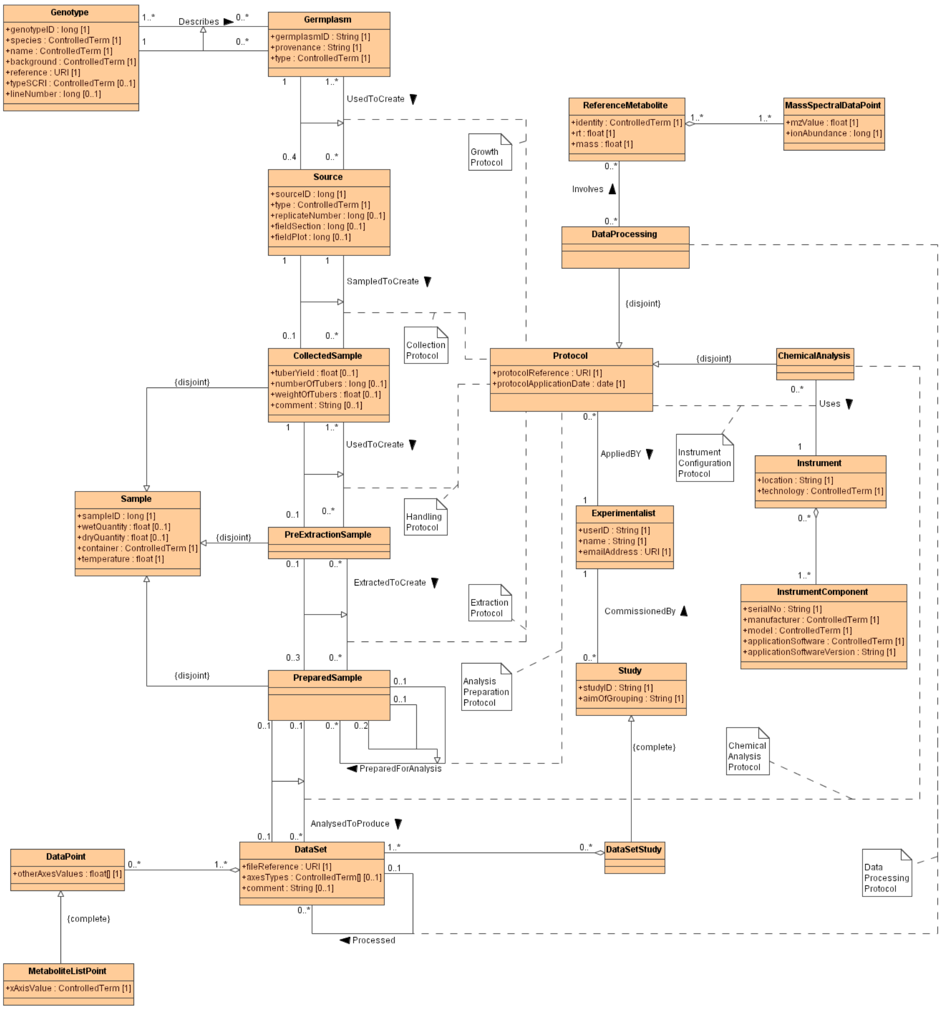
This document contains the design for the ArMet-compliant database for the FSA G03R0002 project. The design contained herein is based on description of the GC-MS and LC-MS experiments carried out under FSA project G02001 by the SCRI.

2. Database design

The design for the database is based on the updated design for ArMet [1]. This design will not be repeated here; it is assumed that the reader is familiar with the contents of the ArMet re-design document.

The G03R0002 specific extensions and restrictions to the core ArMet design are depicted in the diagram below and described in the following sub-sections.

Figure 1. G03 extensions and restrictions to ArMet



2.1. Describing experiments

The *Study* and *DataSetStudy* classes

The FSA project G02001 involved a number of experiments on potato led by SCRI. Four of these experiments have been used to generate this design:

- GC-MS measuring non-polar metabolites.
- GC-MS measuring polar metabolites.
- LC-MS.
- LC-MS (set 2).

These four experiments will each be represented by an ArMet *DataSetStudy*. Each *DataSetStudy* will group the data sets that were produced by the experiment it represents. Each experiment description includes contact information for the SCRI project leader. To support this an extension to the ArMet design is required in the form of an association between *Study* and an *Experimentalist*. This part of the diagram is described below:

Classes. The *Study* and *DataSetStudy* classes in the diagram are equivalent to the classes of the same names in the ArMet design.

Attributes. The attributes of the *Study* and *DataSetStudy* classes in the diagram are equivalent to those of the same name and class in the ArMet design.

Associations. The DataSetStudy::DataSet and Study::DataSetStudy associations in the diagram are equivalent to the same associations in the ArMet design. The Study::Experimentalist association extends the ArMet design and is described below:

Table 1. Additional Associations

Association	Description
Study::Experimentalist 0..*::1..1	A Study must be associated with an Experimentalist to provide the contact details of the person responsible for the experiment. An Experimentalist may be registered without association with a particular Study, but may be associated with multiple Studies for which he is responsible.

2.2. Describing protocols

The *Protocol* and *Experimentalist* classes

The G03R0002 database will implement neither the protocol repository nor the attribute repository standard extensions to the design for ArMet. Documents describing the protocols followed during the experiments included in the database will be maintained on the project website and referenced from the database. Investigation into providing these protocol documents in a searchable format will be carried out if time permits. This part of the diagram is described below:

Classes. The *Protocol* and *Experimentalist* classes in the diagram are equivalent to the classes of the same names in the ArMet design.

Attributes. The attributes of the *Protocol* and *Experimentalist* classes in the diagram are equivalent to those of the same name and class in the ArMet design.

Associations. The Protocol::Experimentalist association in the diagram is equivalent to the same association in the ArMet design.

2.3. Biological source material

The *Genotype*, *Germplasm* and *Source* classes

During FSA G02001 potates were grown from seed/seed tuber stocks (*Germplasm*), each of a single *Genotype*. Each different *Germplasm* stock was used to plant out four separate field plots (*Source*).

Extensions and restrictions to the ArMet design to support the G03R0002 biological source material descriptions are described below:

Classes. The *Genotype*, *Germplasm* and *Source* classes in the diagram are equivalent to the classes of the same name in the ArMet design.

Additional Attributes. The attributes of the *Genotype*, *Germplasm* and *Source* classes that extend those in the ArMet design are described below:

Table 2. Additional Attributes

Attribute	Description
Genotype::typeSCRI	The type of the genotype. (Optional, qualifier = "genotype type")
Genotype::lineNumberSCRI	A unique line number for the genotype assigned by SCRI. (Optional)
Source::replicateNumber	An SCRI replicate number assigned to the item of source material. (Optional)
Source::fieldSection	The SCRI field section in which the item of source material was cultivated. (Optional)
Source::fieldPlot	The SCRI field plot in which the item of source material was cultivated. (Optional)

Use of Pre-existing Attributes. Specific G03 usage of the core ArMet attributes of the

Genotype, *Germplasm* and *Source* classes is described below:

Table 3. Pre-existing Attributes

Attribute	Description
Germplasm::germplasmID	An auto-generated unique identifier for the germplasm.

Additional Associations. The additional Genotype::Germplasm and Germplasm::Source associations shown in the diagram are described below:

Table 4. Additional Associations

Association	Description
Genotype::Germplasm 1..1::0..*	A specialisation of the core Genotype::Germplasm association to restrict Germplasm to a single Genotype.
Germplasm::Source 1..*:0..4	A specialisation of the core Germplasm::Source association to restrict the number of items of Source material created from an item of Germplasm to a maximum of 4.

Use of Pre-existing Associations. Specific G03 usage of the core ArMet Genotype::Germplasm and Germplasm::Source associations is described below:

Table 5. Pre-existing Associations

Attribute	Description
Genotype::Germplasm 1..*::0..*	The ArMet core Genotype::Germplasm association should be retained (by way of a view) to enable ArMet core querying over the G03 database.
Germplasm::Source 1..*:0..*	The ArMet core Germplasm::Source association should be retained (by way of a view) to enable ArMet core querying over the G03 database.

Association Classes. The Germplasm::Source *Protocol* association class is equivalent to the same association class in the ArMet design.

2.4. Samples

The *Sample*, *CollectedSample*, *PreExtractionSample* and *PreparedSample* classes.

From each field plot one biological replicate (*CollectedSample*) was taken. Each *CollectedSample* was freeze-dried and milled to produce one *PreExtractionSample* from which material was taken for all analyses.

Three aliquots were taken from each *PreExtractionSample*. One aliquot was extracted for GC-MS analysis whilst the remaining two aliquots were extracted for LC-MS analysis. Each aliquot was extracted to produce a single *PreparedSample*.

Each *PreparedSample* for GC-MS analysis underwent fractionation to produce two further *PreparedSamples*, one containing non-polar metabolites and one containing polar metabolites. Each fraction then underwent derivatisation to produce a further *PreparedSample* ready for analysis.

Extensions and restrictions to the ArMet design to support the G03R0002 sample descriptions are described below:

Classes. The *Sample*, *CollectedSample*, *PreExtractionSample* and *PreparedSample* classes in the diagram are equivalent to the classes of the same name in the ArMet design.

Additional Attributes. The attributes of the *Sample*, *CollectedSample*, *PreExtractionSample* and *PreparedSample* classes that extend those in the ArMet design are described below:

Table 6. Additional Attributes

Attribute	Description
CollectedSample::tuberYield	The conbined weight of all of the tubers harvested from the SCRI field plot from which the CollectedSample was taken. (Optional)
CollectedSample::numberOfTubers	The number of tubers sampled to create the CollectedSample. (Optional)
CollectedSample::weightOfTubers	The combined weight of the tubers sampled to create the CollectedSample. (Optional)
CollectedSample::comment	A user defined comment on the creation of the CollectedSample. (Optional)

Use of Pre-existing Attributes. Specific G03 usage of the core ArMet attributes of the *Sample*, *CollectedSample*, *PreExtractionSample* and *PreparedSample* classes is described below:

Table 7. Pre-existing Attributes

Attribute	Description
CollectedSample::wetQuantity, CollectedSample::dryQuantity	Each G03 CollectedSample should record a value for dryQuantity only.
PreExtractionSample::wetQuantity, PreExtractionSample::dryQuantity	Each G03 PreExtractionSample should record a value for dryQuantity only.

Additional Associations. The additional Source::CollectedSample, CollectedSample::PreExtractionSample, PreExtractionSample::PreparedSample and PreparedSample::PreparedSample associations shown in the diagram are described below:

Table 8. Additional Associations

Association	Description
Source::CollectedSample 1..1:0..1	A specialisation of the core Source::CollectedSample association to restrict each CollectedSample to a single item of Source material.
CollectedSample::PreExtractionSample 1..1:0..1	A specialisation of the core CollectedSample::PreExtractionSample association to restrict each PreExtractionSample to a single CollectedSample.
PreExtractionSample::PreparedSample 0..1:0..3	A specialisation of the core PreExtractionSample::PreparedSample association to restrict the number of PreparedSamples that may be produced by extraction from a single PreExtractionSample to a maximum of 3.
PreparedSample::PreparedSample 0..1:0..2	A specialisation of the core PreparedSample::PreparedSample association to restrict the number of PreparedSamples that may be produced from a single parent PreparedSample to a maximum of 2 (note that this association is used to describe fractionation and derivatisation of

Association	Description
	GC-MS extracts).

Use of Pre-existing Associations. Specific G03 usage of the core ArMet Source::CollectedSample, CollectedSample::PreExtractionSample, PreExtractionSample::PreparedSample and PreparedSample::PreparedSample associations is described below:

Table 9. Pre-existing Associations

Association	Description
Source::CollectedSample 1..1:0..*	The ArMet core Source::CollectedSample association should be retained (by way of a view) to enable ArMet core querying over the G03 database.
CollectedSample::PreExtractionSample 1..*:0..*	The ArMet core CollectedSample::PreExtractionSample association should be retained (by way of a view) to enable ArMet core querying over the G03 database.
PreExtractionSample::PreparedSample 0..*:0..*	The ArMet core PreExtractionSample::PreparedSample association should be retained (by way of a view) to enable ArMet core querying over the G03 database.
PreparedSample::PreparedSample 0..1:0..*	The ArMet core PreparedSample::PreparedSample association should be retained (by way of a view) to enable ArMet core querying over the G03 database.

Association Classes. The *Protocol* association class on the Source::CollectedSample, CollectedSample::PreExtractionSample, PreExtractionSample::PreparedSample and PreparedSample::PreparedSample associations is equivalent to the association class of the same name on the same associations in the ArMet design.

2.5. ChemicalAnalysis

The *ChemicalAnalysis*, *Instrument* and *InstrumentComponent* classes

Each final *PreparedSample* is subjected to one instance of chemical analysis.

Extensions and restrictions to the ArMet design to support the G03R0002 chemical analysis description are described below:

Classes. The *ChemicalAnalysis*, *Instrument* and *InstrumentComponent* classes in the diagram are equivalent to the classes of the same name in the ArMet design.

Attributes. The attributes of the *ChemicalAnalysis*, *Instrument* and *InstrumentComponent* classes in the diagram are equivalent to those of the same name and class in the ArMet design.

Additional Associations. The additional PreparedSample::DataSet association shown in the diagram is described below:

Table 10. Additional Associations

Association	Description
PreparedSample::DataSet 0..1::0..1	A specialisation of the core PreparedSample::DataSet association to restrict each PreparedSample to a single DataSet.

Use of Pre-existing Associations. Specific G03 usage of the core ArMet PreparedSample::DataSet association is described below:

Table 11. Pre-existing Associations

Attribute	Description
PreparedSample::DataSet 0..1::0..*	The ArMet core PreparedSample::DataSet association should be retained (by way of a view) to enable ArMet core querying over the G03 database.

Association Classes. The ChemicalAnalysis::Instrument *Protocol* association class and the PreparedSample::DataSet *ChemicalAnalysis* association class are equivalent to the association classes of the same name on the same associations in the ArMet design.

2.6. Data sets

The *DataSet*, *DataPoint*, *MetaboliteListPoint*, *DataProcessing*, *ReferenceMetabolite* and *MassSpectralDataPoint* classes

Chemical analysis produces one raw data file (Excalibur .RAW) file per sample. This contains the TIC and spectra for the sample. Data processing is performed on this raw file to produce a metabolite peak list for the sample. Data processing is mostly performed under automation in the Excalibur software, but some manual checking of the peaks identified by Excalibur is performed on the output.

Extensions and restrictions to the ArMet design to support the G03R0002 data sets are described below:

Additional Classes. The classes that extend those in the ArMet design are described below:

Table 12. Additional Classes

Class	Description
DataProcessing	The protocol followed to create a metabolite peak list from the raw data set for a sample. A specialisation of Protocol.
ReferenceMetabolite	A reference description for a metabolite used when manually checking the results of automatic peak identification during data processing.
MassSpectralDataPoint	A data point within a mass spectrum.

Additional Attributes. The attributes of the *DataSet*, *DataPoint*, *MetaboliteListPoint*, *DataProcessing*, *ReferenceMetabolite* and *MassSpectralDataPoint*classes that extend those in the ArMet design are described below:

Table 13. Additional Attributes

Attribute	Description
DataSet::comment	A user-defined comment on the data set. (Optional)
ReferenceMetabolite::identity	The chemical identity for the reference metabolite. (Required, Primary Key, qualifier = "metabolite identifier")
ReferenceMetabolite::rt	The expected retention time for the reference metabolite. (Required)
ReferenceMetabolite::mass	The mass of the characteristic ion for the reference metabolite. (Required)
MassSpectralDataPoint::mzValue	The mass-to-charge ratio for the data point. (Required, Partial Primary Key)

Attribute	Description
MassSpectralDataPoint::ionAbundance	The absolute abundance of ions measured for the mass-to-charge ratio. (Required, Partial Primary Key)

Additional Associations. The additional DataProcessing::ReferenceMetabolite and ReferenceMetabolite::MassSpectralDataPoint associations shown in the diagram are described below:

Table 14. Additional Associations

Association	Description
DataProcessing::ReferenceMetabolite 0..*:0..*	A ReferenceMetabolite may be defined without association with a particular DataProcessing protocol, but may be associated with many DataProcessing protocols in which it is used. A DataProcessing protocol may be defined without association with a ReferenceMetabolite, but may be associated with multiple ReferenceMetabolites that it employs.
ReferenceMetabolite::MassSpectralDataPoint 1..*:1..*	A MassSpectralDataPoint must be associated with at least one ReferenceMetabolite whose spectrum it appears in, but may be associated with multiple ReferenceMetabolites whose spectra it appears in. A ReferenceMetabolite must be associated with at least one MassSpectralDataPoint to describe its mass spectrum, but may be associated with multiple MassSpectralDataPoints that comprise its mass spectrum.

Association Classes. The DataSet::DataSet *DataProcessing* association class is equivalent to the DataSet::DataSet *Protocol* class in the ArMet design.

2.7. A note on data types and controlled vocabularies

The following definitions should be followed for the non-basic data types used within the design:

Table 15. Non-basic data types

Data type	Definition
ControlledTerm	This type is equivalent to the *ControlledTerm* class described in the ArMet design.
URI	An ASCII value that conforms to the Network Working Group RFC 2396.
Date	A date value that conforms to ISO 8601.

The following table documents additional controlled vocabularies that should be provided in implementations for G03.

Table 16. Additional G03 controlled vocabularies

Qualifier	Values
genotype type	cultivar

Bibliography

[1] Helen Jenkins. *ArMet Re-design*. University of Wales, Aberystwyth Technical Report. 2006.

G03 Metabolite Library Logical Design

Reference: $Source: /dcs/fsaweb/G03CVS/g03r0002rep/design/MetaboliteLibrary/logicalDesign.xml,v $
Version: Draft
Edition: $Revision: 1.8 $ Edition
Date: $Date: 2007/01/15 13:36:47 $

Author: Dr Helen Jenkins

Table of Contents

1. Introduction

This document contains the logical design for the metabolite library for the FSA G03R0002 project. The design contained herein is based on data examples provided by and discussions held with David Overy.

2. Metabolite library description

The following list provides a description of the data that will be held in the library:

- Each metabolite in the library will be associated with:
  1. A primary chemical identity.
  2. A chemical formula.
  3. Its molecular weight.
  4. Zero or one chemical structure diagram.
  5. For each class (plants, mammals, fungi, bacteria) for which the metabolite is designated a primary metabolite, one or more references to supporting information.
  6. A flag to indicate whether or not the metabolite has been measured from a standard.
  7. Zero or more chemical synonyms.
  8. Zero or more chemical pathways.
  9. Zero or more references to further information.
  10. Zero or more descriptions of biological source in which the metabolite may be found. For each biological source description that is associated with a metabolite there will be one or more references to evidence for the association.
- On submission to the database each metabolite will be assigned a set of predicted adducts. These adducts will be identified by a formula and a mass. Each prediction will be annotated with information on the ESI ion mode in which it is predicted that the adduct may be measured and the accuracy of the mass value attributed to the adduct.
- Each metabolite may also be associated with a set of measured adducts. These adducts will also be identified by a formula and parent ion mass. Each measurement will be annotated with information on the ESI ion mode and the accuracy of the mass measurement.
- Each measured adduct may optionally be associated with an ion tree providing the full results of measurement by ESI-MS$_n$.

Additional constraints on the metabolite library:

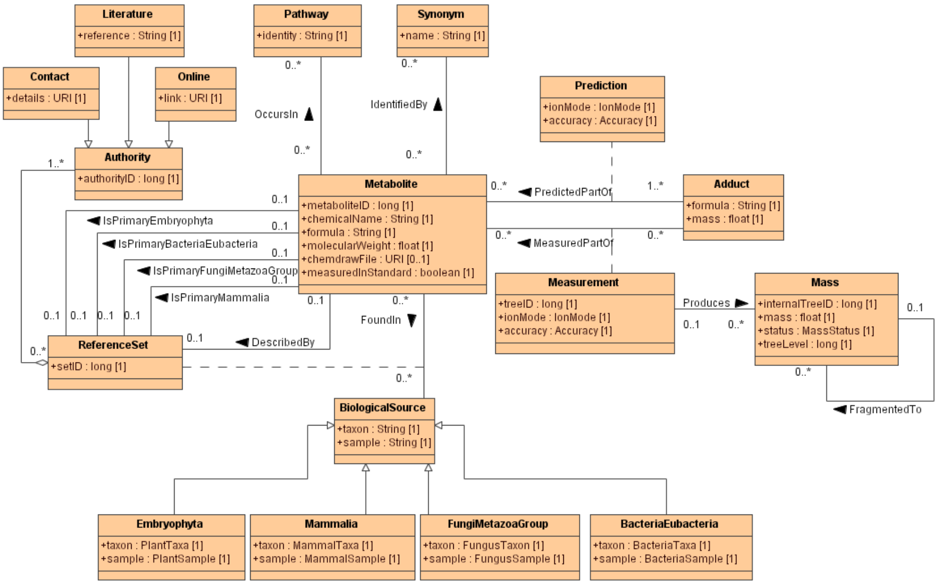
- References to additional information/supporting evidence may be literature refer-

ences, online references or contact details for experimentalists.
- Biological source are described by a taxon designation and a sample description. They are broadly classified into plants (embryophyta), mammals (mammalia), fungi (fungiMetazoaGroup) and bacteria (bacteriaEubacteria).
- The same metabolite description including its set of predicted adducts and (optionally) measured adducts will be associated with all items of biological source in which the metabolite is expected to occur. To confirm: Once an adduct has been measured for a metabolite the resulting ion tree will be associated with that metabolite and, thereby, with all of the biological source with which that metabolite is associated, regardless of the provenance of the sample from which it has been measured.
- As an adduct when produced from different metabolites may fragment in different ways, ion trees should be associated with metabolite/adduct relationships and not simply with adducts.
- A particular ion may appear multiple times within an ion tree. In addition, a particular ion may appear multiple times within a generation of a tree, i.e. as a fragment of different ions in the previous generation.

3. Database design

The design for the metabolite library is depicted in the diagram below and described in the following sub-sections.

Figure 1. Metabolite Library Design



Classes. The classes in the diagram are defined below:

Table 1. Classes

Class	Description
Metabolite	A metabolite that is held in the library.
Synonym	A metabolite synonym.
Pathway	A metabolite pathway.
Authority	A reference to further information on a metabolite.
Online	A specialisation of Authority to represent online references.
Literature	A specialisation of Authority to represent literature references.

Class	Description
Contact	A specialisation of Authority to represent experimentalist contact details.
ReferenceSet	A set of authorities.
BiologicalSource	Some biological source material.
Embryophyta	A specialisation of BiologicalSource to represent plants.
Mammalia	A specialisation of BiologicalSource to represent mammals.
FungiMetazoaGroup	A specialisation of BiologicalSource to represent fungi.
BacteriaEubacteria	A specialisation of BiologicalSource to represent bacteria.
Adduct	An adduct of a metabolite.
Prediction	The conditions under which it is predicted that a particular Adduct (formula,mass) may be produced from a particular Metabolite.
Measurement	The conditions under which a particular Adduct (formula,mass) was measured from a particular Metabolite.
Mass	A daughter ion produced from a particular Adduct of a particular Metabolite using ESI-MS$_n$ under particular Measurement conditions.

Attributes. The attributes of these classes are described below:

Table 2. Attributes

Attribute	Description
Metabolite::metaboliteID	A unique auto-generated identifier for the metabolite. (Required, Primary Key)
Metabolite::chemicalName	The primary chemical identity for the metabolite. (Required).
Metabolite::formula	The chemical formula for the metabolite. (Required)
Metabolite::molecularWeight	The molecular weight of the metabolite. (Required, Derived)
Metabolite::chemdrawFile	A reference to a Chemdraw file containing a depiction of the chemical structure of the metabolite. (Optional)
Metabolite::measuredInStandard	A flag to indicate that the metabolite has been measured from a standard. (Required, Default = *false*)
Synonym::name	A metabolite identity or synonym. (Required, Primary Key)
Pathway::identity	An identifier for a metabolite pathway. (Required, Primary Key)
Authority::authorityID	A unique auto-generated identifier for a reference. (Required, Primary Key)
Online::link	An online reference. (Required)
Literature::reference	A literature reference. (Required)
Contact::details	The email address for an experimentalist. (Required)

Attribute	Description
ReferenceSet::setID	A unique auto-generated identifier for the reference set. (Required, Primary Key)
BiologicalSource::taxon	The taxonomic designation for a biological source. (Required, Partial Primary Key)
BiologicalSource::sample	A type of sample for the biological source. (Required, Partial Primary Key)
Embryophyta::taxon (overrides BiologicalSource::taxon)	The taxonomic designation for the plant.
Embryophyta::sample (overrides BiologicalSource::sample)	A type of plant sample.
Mammalia::taxon (overrides BiologicalSource::taxon)	The taxonomic designation for the mammal.
Mammalia::sample (overrides BiologicalSource::sample)	A type of mammal sample.
FungiMetazoaGroup::taxon (overrides BiologicalSource::taxon)	The taxonomic designation for the fungus.
FungiMetazoaGroup::sample (overrides BiologicalSource::sample)	A type of fungus sample.
BacteriaEubacteria::taxon (overrides BiolgicalSource::taxon)	The taxonomic designation for the bacteria.
BacteriaEubacteria::sample (overrides BiologicalSource::sample)	A type of bacteria sample.
Adduct::formula	The chemical formula for the adduct. (Required, Partial Primary Key)
Adduct::mass	The mass of the adduct. (Required, Partial Primary Key)
Prediction::ionMode	The ion mode in which it is predicted that the Adduct may be produced from the Metabolite. (Required)
Prediction::accuracy	The accuracy of the mass value for this prediction. (Required)
Measurement::treeID	A unique auto-generated identifier for the measurement. (Required, Primary Key)
Measurement::ionMode	The ion mode under which the Adduct was measured from the Metabolite. (Required)
Measurement::accuracy	The accuracy of the mass value as measured. (Required)
Mass::internalTreeID	An auto-generated unique identifier for the ion within the ion tree in which it appears. (Required, Partial Primary Key)
Mass::mass	The mass value of the ion. (Required)
Mass::status	The status of the mass within the ion tree. (Required)
Mass::treeLevel	The level of the ion tree at which this mass occurs. (Required)

Note

- Metabolite::molecularWeight is derived from Metabolite::formula and a chart of the weights of individual atoms available from David Overy.
- The primary keys for the Online, Literature and Contact tables will be the parent Authority primary key.

- The primary key for the Prediction table will comprise the foreign keys from Adduct and Metabolite.
- The primary key for the Mass table will comprise Mass::internalTreeID and the foreign key from Measurement.
- Values stored for Mass::treeLevel should be > 1.

Associations. The associations in the diagram are described below:

Table 3. Associations

Association	Description
Metabolite::Synonym 0..*::0..*	A metabolite Synonym may be defined without association with a Metabolite or may be associated with multiple Metabolites. A Metabolite may be defined without associated Synonyms, but may be associated with multiple Synonyms.
Metabolite::Pathway 0..*::0..*	A Pathway may be defined without association with a Metabolite, but may be associated with multiple Metabolites that it contains. A Metabolite may be defined without an associated Pathway, but may be associated with multiple Pathways in which it occurs.
Authority::Online Authority::Literature Authority::Contact	Each Authority must be an Online reference, a Literature reference or a Contact. Each Authority may only be associated with one Online reference, Literature reference or Contact.
ReferenceSet::Authority 0..*::1..*	An Authority may be defined without being part of a ReferenceSet, but may belong to multiple ReferenceSets. A ReferenceSet must contain at least one Authority and may contain multiple Authorities.
Metabolite::ReferenceSet 0..1::0..1 (IsPrimaryEmbryophyta, IsPrimaryMammalia, IsPrimaryFungiMetazoaGroup, IsPrimaryBacteriaEubacteria)	A ReferenceSet may be defined for purposes other than primary metabolite designation, but may be used to provide evidence for a single primary metabolite designation. A Metabolite may be defined without designation as a primary metabolite of any class, but may be designated a primary metabolite in any or all of the four classes and be associated with a ReferenceSet to provide evidence for each.
Metabolite::ReferenceSet 0..1::0..1 (DescribedBy)	A ReferenceSet may be defined for purposes other than Metabolite description, but may be used to describe a Metabolite. A Metabolite may be defined without association with a ReferenceSet, but may be associated with one ReferenceSet that describes it.
Metabolite::BiologicalSource 0..*::0..*	A BiologicalSource may be defined without association with a Metabolite, but may be associated with multiple Metabolites that it contains. A Metabolite may be defined without association with a BiologicalSource (where the Metabolite has been measured from a standard and no further information is available), but may be associated with multiple BiologicalSource in which it occurs.
BiologicalSource::Embryophyta	Each BiologicalSource must be associated with an Embryophyta, a Mammalia, a FungiMetazo-

Association	Description
BiologicalSource::Mammalia BiologicalSource::FungiMetazoaGroup BiologicalSource::BacteriaEubacteria	aGroup or a BacteriaEubacteria. Each BiologicalSource may only be associated with one Embryophyta, Mammalia, FungiMetazoaGroup or BacteriaEubacteria.
Metabolite::Adduct 0..*::1..* (predictedPartOf)	An Adduct may be defined without being a predicted part of a Metabolite but may be a predicted part of multiple Metabolites from which it can be produced. A Metabolite must be associated with at least one predicted Adduct and may be associated with multiple predicted Adducts that can be produced from it.
Metabolite::Adduct 0..*::0..* (measuredPartOf)	An Adduct may be defined without being a measured part of a Metabolite but may be a measured part of multiple Metabolites from which it can be produced. A Metabolite may not be associated with any measured Adducts, but may be associated with multiple Adducts that have been measured from it.
Measurement::Mass 0..1::0..* Mass::Mass 0..1::0..*	A Mass must be associated with exactly one of a parent Mass or a Measurement description. Masses that are direct fragments of the parent ion of an Adduct must be associated with the Measurement description for that Adduct. Other Masses within an ion tree must be associated with parent Masses. An Measurement description need not be associated with any Masses (i.e. if the parent ion has been measured, but not fragmented), but may be associated with multiple Masses that represent direct fragments of the parent ion measured. A Mass need not be associated with a Measurement description, but may be associated with exactly one for which it provides additional measurement information. A child Mass need not be associated with a parent mass (i.e. if it is a direct fragment of the parent ion), but may be associated with exactly one parent Mass. A parent Mass need not have children, but may have multiple children.

Note

- A Metabolite that is not associated with any BiologicalSource and is not designated as a primary metabolite for any class must contain the value *true* for its measuredInStandard attribute. Conversely, a Metabolite that contains the value *false* for its measuredInStandard attribute must be associated with at least one BiologicalSource and/or be designated a primary metabolite for at least one class. (A Metabolite may contain the value *true* for its measuredInStandard attribute and at the same time be associated with one or more BiologicalSource and/or be designated a primary metabolite for at least one class.)
- A Metabolite that contains the value *true* for its measuredInStandard attribute must be associated with Adducts that are MeasuredPartsOf it.

3.1. A note on data types

The following definitions should be followed for the non-basic data types used within the design:

Table 4. Non-basic data types

Data type	Definition
URI	An ASCII value that conforms to the Network Working Group RFC 2396.
PlantTaxa	An extendable (following curation) enumeration of the following values: "Arabidopsis thaliana", "Brachypodium distachyon", "Oryzae sativa", "Hordeum vulgare", "Solanum tuberosum"
PlantSample	An extendable (following curation) enumeration of the following values: "leaf", "stem", "root", "shoot", "bark", "flower", "seed", "fruit", "tuber", "rhizome", "bulb", "corm"
MammalTaxa	An extendable (following curation) enumeration of the following values: "Homo sapien", "Canis familiaris", "Rattus"
MammalSample	An extendable (following curation) enumeration of the following values: "blood", "urine", "faeces", "bile", "CSF"
FungusTaxa	An extendable (following curation) enumeration of the following values: "Magnaporthe grisea"
FungusSample	An extendable (following curation) enumeration of the following values: "in vitro", "in situ"
BacteriaTaxa	An extendable (following curation) enumeration. No initial values available.
BacteriaSample	An extendable (following curation) enumeration. No initial values available.
IonMode	An enumeration of the following values: "negative", "positive".
Accuracy	An enumeration of the following values: "accurate", "nominal".
MassStatus	An enumeration of the following values: "major", "minor", "possible".

G03 Metabolite Library Logical Design

Reference: $Source: /dcs/fsaweb/G03CVS/g03r0002rep/design/MetaboliteLibrary/logicalDesign.xml,v $
Version: Draft
Edition: $Revision: 1.8 $ Edition
Date: $Date: 2007/01/15 13:36:47 $
Author: Dr Helen Jenkins

Table of Contents

1. Introduction

This document contains the logical design for the metabolite library for the FSA G03R0002 project. The design contained herein is based on data examples provided by and discussions held with David Overy.

2. Metabolite library description

The following list provides a description of the data that will be held in the library:

- Each metabolite in the library will be associated with:
  1. A primary chemical identity.
  2. A chemical formula.
  3. Its molecular weight.
  4. Zero or one chemical structure diagram.
  5. For each class (plants, mammals, fungi, bacteria) for which the metabolite is designated a primary metabolite, one or more references to supporting information.
  6. A flag to indicate whether or not the metabolite has been measured from a standard.
  7. Zero or more chemical synonyms.
  8. Zero or more chemical pathways.
  9. Zero or more references to further information.
  10. Zero or more descriptions of biological source in which the metabolite may be found. For each biological source description that is associated with a metabolite there will be one or more references to evidence for the association.
- On submission to the database each metabolite will be assigned a set of predicted adducts. These adducts will be identified by a formula and a mass. Each prediction will be annotated with information on the ESI ion mode in which it is predicted that the adduct may be measured and the accuracy of the mass value attributed to the adduct.
- Each metabolite may also be associated with a set of measured adducts. These adducts will also be identified by a formula and parent ion mass. Each measurement will be annotated with information on the ESI ion mode and the accuracy of the mass measurement.
- Each measured adduct may optionally be associated with an ion tree providing the full results of measurement by ESI-MS$_n$.

Additional constraints on the metabolite library:

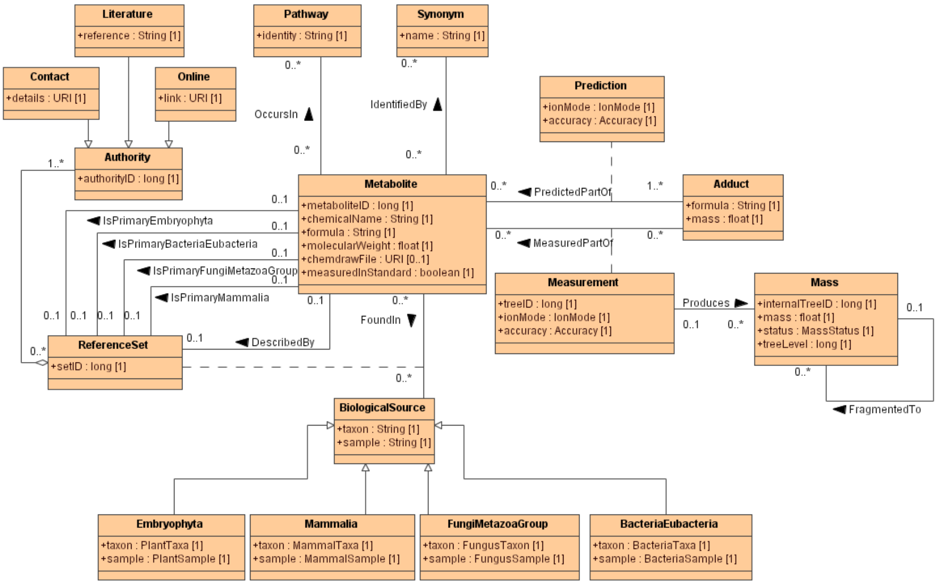
- References to additional information/supporting evidence may be literature references, online references or contact details for experimentalists.

- Biological source are described by a taxon designation and a sample description. They are broadly classified into plants (embryophyta), mammals (mammalia), fungi (fungiMetazoaGroup) and bacteria (bacteriaEubacteria).
- The same metabolite description including its set of predicted adducts and (optionally) measured adducts will be associated with all items of biological source in which the metabolite is expected to occur. To confirm: Once an adduct has been measured for a metabolite the resulting ion tree will be associated with that metabolite and, thereby, with all of the biological source with which that metabolite is associated, regardless of the provenance of the sample from which it has been measured.
- As an adduct when produced from different metabolites may fragment in different ways, ion trees should be associated with metabolite/adduct relationships and not simply with adducts.
- A particular ion may appear multiple times within an ion tree. In addition, a particular ion may appear multiple times within a generation of a tree, i.e. as a fragment of different ions in the previous generation.

3. Database design

The design for the metabolite library is depicted in the diagram below and described in the following sub-sections.

Figure 1. Metabolite Library Design



Classes. The classes in the diagram are defined below:

Table 1. Classes

Class	Description
Metabolite	A metabolite that is held in the library.
Synonym	A metabolite synonym.
Pathway	A metabolite pathway.
Authority	A reference to further information on a metabolite.
Online	A specialisation of Authority to represent online references.
Literature	A specialisation of Authority to represent literature references.

Class	Description
Contact	A specialisation of Authority to represent experimentalist contact details.
ReferenceSet	A set of authorities.
BiologicalSource	Some biological source material.
Embryophyta	A specialisation of BiologicalSource to represent plants.
Mammalia	A specialisation of BiologicalSource to represent mammals.
FungiMetazoaGroup	A specialisation of BiologicalSource to represent fungi.
BacteriaEubacteria	A specialisation of BiologicalSource to represent bacteria.
Adduct	An adduct of a metabolite.
Prediction	The conditions under which it is predicted that a particular Adduct (formula,mass) may be produced from a particular Metabolite.
Measurement	The conditions under which a particular Adduct (formula,mass) was measured from a particular Metabolite.
Mass	A daughter ion produced from a particular Adduct of a particular Metabolite using ESI-MS$_n$ under particular Measurement conditions.

Attributes. The attributes of these classes are described below:

Table 2. Attributes

Attribute	Description
Metabolite::metaboliteID	A unique auto-generated identifier for the metabolite. (Required, Primary Key)
Metabolite::chemicalName	The primary chemical identity for the metabolite. (Required).
Metabolite::formula	The chemical formula for the metabolite. (Required)
Metabolite::molecularWeight	The molecular weight of the metabolite. (Required, Derived)
Metabolite::chemdrawFile	A reference to a Chemdraw file containing a depiction of the chemical structure of the metabolite. (Optional)
Metabolite::measuredInStandard	A flag to indicate that the metabolite has been measured from a standard. (Required, Default = *false*)
Synonym::name	A metabolite identity or synonym. (Required, Primary Key)
Pathway::identity	An identifier for a metabolite pathway. (Required, Primary Key)
Authority::authorityID	A unique auto-generated identifier for a reference. (Required, Primary Key)
Online::link	An online reference. (Required)
Literature::reference	A literature reference. (Required)
Contact::details	The email address for an experimentalist. (Required)

Attribute	Description
ReferenceSet::setID	A unique auto-generated identifier for the reference set. (Required, Primary Key)
BiologicalSource::taxon	The taxonomic designation for a biological source. (Required, Partial Primary Key)
BiologicalSource::sample	A type of sample for the biological source. (Required, Partial Primary Key)
Embryophyta::taxon (overrides BiologicalSource::taxon)	The taxonomic designation for the plant.
Embryophyta::sample (overrides BiologicalSource::sample)	A type of plant sample.
Mammalia::taxon (overrides BiologicalSource::taxon)	The taxonomic designation for the mammal.
Mammalia::sample (overrides BiologicalSource::sample)	A type of mammal sample.
FungiMetazoaGroup::taxon (overrides BiologicalSource::taxon)	The taxonomic designation for the fungus.
FungiMetazoaGroup::sample (overrides BiologicalSource::sample)	A type of fungus sample.
BacteriaEubacteria::taxon (overrides BiolgicalSource::taxon)	The taxonomic designation for the bacteria.
BacteriaEubacteria::sample (overrides BiologicalSource::sample)	A type of bacteria sample.
Adduct::formula	The chemical formula for the adduct. (Required, Partial Primary Key)
Adduct::mass	The mass of the adduct. (Required, Partial Primary Key)
Prediction::ionMode	The ion mode in which it is predicted that the Adduct may be produced from the Metabolite. (Required)
Prediction::accuracy	The accuracy of the mass value for this prediction. (Required)
Measurement::treeID	A unique auto-generated identifier for the measurement. (Required, Primary Key)
Measurement::ionMode	The ion mode under which the Adduct was measured from the Metabolite. (Required)
Measurement::accuracy	The accuracy of the mass value as measured. (Required)
Mass::internalTreeID	An auto-generated unique identifier for the ion within the ion tree in which it appears. (Required, Partial Primary Key)
Mass::mass	The mass value of the ion. (Required)
Mass::status	The status of the mass within the ion tree. (Required)
Mass::treeLevel	The level of the ion tree at which this mass occurs. (Required)

Note

- Metabolite::molecularWeight is derived from Metabolite::formula and a chart of the weights of individual atoms available from David Overy.
- The primary keys for the Online, Literature and Contact tables will be the parent Authority primary key.

- The primary key for the Prediction table will comprise the foreign keys from Adduct and Metabolite.
- The primary key for the Mass table will comprise Mass::internalTreeID and the foreign key from Measurement.
- Values stored for Mass::treeLevel should be > 1.

Associations. The associations in the diagram are described below:

Table 3. Associations

Association	Description
Metabolite::Synonym 0..*::0..*	A metabolite Synonym may be defined without association with a Metabolite or may be associated with multiple Metabolites. A Metabolite may be defined without associated Synonyms, but may be associated with multiple Synonyms.
Metabolite::Pathway 0..*::0..*	A Pathway may be defined without association with a Metabolite, but may be associated with multiple Metabolites that it contains. A Metabolite may be defined without an associated Pathway, but may be associated with multiple Pathways in which it occurs.
Authority::Online Authority::Literature Authority::Contact	Each Authority must be an Online reference, a Literature reference or a Contact. Each Authority may only be associated with one Online reference, Literature reference or Contact.
ReferenceSet::Authority 0..*::1..*	An Authority may be defined without being part of a ReferenceSet, but may belong to multiple ReferenceSets. A ReferenceSet must contain at least one Authority and may contain multiple Authorities.
Metabolite::ReferenceSet 0..1::0..1 (IsPrimaryEmbryophyta, IsPrimaryMammalia, IsPrimaryFungiMetazoaGroup, IsPrimaryBacteriaEubacteria)	A ReferenceSet may be defined for purposes other than primary metabolite designation, but may be used to provide evidence for a single primary metabolite designation. A Metabolite may be defined without designation as a primary metabolite of any class, but may be designated a primary metabolite in any or all of the four classes and be associated with a ReferenceSet to provide evidence for each.
Metabolite::ReferenceSet 0..1::0..1 (DescribedBy)	A ReferenceSet may be defined for purposes other than Metabolite description, but may be used to describe a Metabolite. A Metabolite may be defined without association with a ReferenceSet, but may be associated with one ReferenceSet that describes it.
Metabolite::BiologicalSource 0..*::0..*	A BiologicalSource may be defined without association with a Metabolite, but may be associated with multiple Metabolites that it contains. A Metabolite may be defined without association with a BiologicalSource (where the Metabolite has been measured from a standard and no further information is available), but may be associated with multiple BiologicalSource in which it occurs.
BiologicalSource::Embryophyta	Each BiologicalSource must be associated with an Embryophyta, a Mammalia, a FungiMetazo-

Association	Description
BiologicalSource::Mammalia BiologicalSource::FungiMetazoaGroup BiologicalSource::BacteriaEubacteria	aGroup or a BacteriaEubacteria. Each BiologicalSource may only be associated with one Embryophyta, Mammalia, FungiMetazoaGroup or BacteriaEubacteria.
Metabolite::Adduct 0..*::1..* (predictedPartOf)	An Adduct may be defined without being a predicted part of a Metabolite but may be a predicted part of multiple Metabolites from which it can be produced. A Metabolite must be associated with at least one predicted Adduct and may be associated with multiple predicted Adducts that can be produced from it.
Metabolite::Adduct 0..*::0..* (measuredPartOf)	An Adduct may be defined without being a measured part of a Metabolite but may be a measured part of multiple Metabolites from which it can be produced. A Metabolite may not be associated with any measured Adducts, but may be associated with multiple Adducts that have been measured from it.
Measurement::Mass 0..1::0..* Mass::Mass 0..1::0..*	A Mass must be associated with exactly one of a parent Mass or a Measurement description. Masses that are direct fragments of the parent ion of an Adduct must be associated with the Measurement description for that Adduct. Other Masses within an ion tree must be associated with parent Masses. An Measurement description need not be associated with any Masses (i.e. if the parent ion has been measured, but not fragmented), but may be associated with multiple Masses that represent direct fragments of the parent ion measured. A Mass need not be associated with a Measurement description, but may be associated with exactly one for which it provides additional measurement information. A child Mass need not be associated with a parent mass (i.e. if it is a direct fragment of the parent ion), but may be associated with exactly one parent Mass. A parent Mass need not have children, but may have multiple children.

Note

- A Metabolite that is not associated with any BiologicalSource and is not designated as a primary metabolite for any class must contain the value *true* for its measuredInStandard attribute. Conversely, a Metabolite that contains the value *false* for its measuredInStandard attribute must be associated with at least one BiologicalSource and/or be designated a primary metabolite for at least one class. (A Metabolite may contain the value *true* for its measuredInStandard attribute and at the same time be associated with one or more BiologicalSource and/or be designated a primary metabolite for at least one class.)
- A Metabolite that contains the value *true* for its measuredInStandard attribute must be associated with Adducts that are MeasuredPartsOf it.

3.1. A note on data types

The following definitions should be followed for the non-basic data types used within the design:

Table 4. Non-basic data types

Data type	Definition
URI	An ASCII value that conforms to the Network Working Group RFC 2396.
PlantTaxa	An extendable (following curation) enumeration of the following values: "Arabidopsis thaliana", "Brachypodium distachyon", "Oryzae sativa", "Hordeum vulgare", "Solanum tuberosum"
PlantSample	An extendable (following curation) enumeration of the following values: "leaf", "stem", "root", "shoot", "bark", "flower", "seed", "fruit", "tuber", "rhizome", "bulb", "corm"
MammalTaxa	An extendable (following curation) enumeration of the following values: "Homo sapien", "Canis familiaris", "Rattus"
MammalSample	An extendable (following curation) enumeration of the following values: "blood", "urine", "faeces", "bile", "CSF"
FungusTaxa	An extendable (following curation) enumeration of the following values: "Magnaporthe grisea"
FungusSample	An extendable (following curation) enumeration of the following values: "in vitro", "in situ"
BacteriaTaxa	An extendable (following curation) enumeration. No initial values available.
BacteriaSample	An extendable (following curation) enumeration. No initial values available.
IonMode	An enumeration of the following values: "negative", "positive".
Accuracy	An enumeration of the following values: "accurate", "nominal".
MassStatus	An enumeration of the following values: "major", "minor", "possible".

G03 Metabolite Library Physical Design

Reference: $Source: / dcs/fsaweb/G03CVS/g03r0002rep/design/MetaboliteLibrary/physicalDesign.xml,v $

Version: Draft

Edition: $Revision: 1.13 $ Edition

Date: $Date: 2007/01/15 13:36:47 $

Author: Dr Helen Jenkins

Table of Contents

1. Introduction

This document contains the physical design for the metabolite library for the FSA G03R0002 project. The design contained herein is based on data examples provided by and discussions held with David Overy and implements the metabolite library logical design.

2. Designing base tables

The base tables for the library are written in Oracle SQL DDL. All tables have been normalised and checked for entity and referential integrity. The following notes describe the decisions made during the design of these tables.

Naming conventions. The following naming conventions have been followed in the production of the base tables:

Table 1. Database naming conventions

Element	Naming convention
Tables	The name from the logical design appended with "_t".
Intersection tables	The names (or abbreviations of the names) of the two tables concatenated and appended with "_i". (Intersection table names are documented below.)
Constraints	The table name and attribute or table (where applicable) to which the constraint applies separated by "_" and appended with a description of the constraint. Descriptions are as follows: • Not null constraints: "_notnull" • Check constraints: "_check" • Primary key constraints: "_pk" • Foreign key constraints: "_fk"
Sequences	The name of the table and attribute that the sequence will be used to populate separated by "_" and appended with "_seq".

Element	Naming convention
Triggers	The name of the table on which the trigger will operate appended with a description of the trigger and then appended with "_tr". Common descriptions are as follows, others are documented below: • Automatic generation of primary keys: "_autoPK" • Checks on attribute values prior to population: "_populate" • Implementation of ON UPDATE CASCADE: "_updatecascade"
Stored functions and procedures	Names that describe their function - documented below.
Packages	Names that describe their function - documented below.
Indexes	The table and attribute names to which the index applies separated by "_" and appended with "_ind".

Tables. The following notes describe the implementation of the base tables:

Table 2. Notes on base table design

Table	Notes
Authority_t	This table implements the Authority, Contact, Literature and Online tables from the logical design. Each row in the table contains the Authority::authorityID, Contact::details, Literature::reference and Online::link attributes. The authorityID attribute is mandatory whilst the contactDetails, LiteratureReference and onlineLink attributes are optional. Whilst this table will be sparse, as each row should contain only one type of reference, the number of references to be stored as a whole makes this is an acceptable trade-off for efficiency in data retrieval. An additional mandatory display attribute has been added to the table. This attribute is designed to hold values that should be displayed in the interface as hotlinks to the references. For contacts this will be an identifier for the contact that when selected will open an email application. For online references this will be the name of the resource that is referenced by the URL. For literature references the value of the attribute defaults to "Literature references". It is anticipated that this will link to a separate page populated with the details of the reference(s). The literatureReference attribute has a maximum length of 200 and the display attribute has a maximum length of 50. A sequence and trigger have been implemented to populate the authorityID attribute. A trigger has been implemented to ensure that only one of contactDetails, literatureReference and onlineLink are populated in any one row of the table. This trigger also supplies the default

Table	Notes
	value for the display attribute when a literature reference is provided and checks for the presence of a value for the display attribute when a contact or online reference is provided.
ReferenceSet_t	This table implements the ReferenceSet table from the logical design. A sequence and trigger have been implemented to populate the setID attribute.
RefSetAuthority_i	This table implements the aggregation association between ReferenceSet and Authority from the logical design. ON DELETE CASCADE has been implemented on the foreign key from ReferenceSet.
Metabolite_t	This table implements the Metabolite table from the logical design and the associations between the Metabolite table and the ReferenceSet table. There are five associations between the Metabolite table and the ReferenceSet table which represent the Metabolite's designation as a primary metabolite in one of the four classes and an authoratative description of the Metabolite. These associations are all specified as 0..1::0..1 in the logical design. All of these associations have been implemented as foreign keys to the ReferenceSet table in the Metabolite table so that the meaning of each association with respect to each metabolite may be maintained within the Metabolite table and, thereby, facilitate more efficient searching by Metabolite (which will be more common than by ReferenceSet). Two triggers (Metabolite_setIDupdate_tr and Metabolite_setIDdelete_tr) have been written to delete referenced sets from the ReferenceSet table on update of a setID within the Metabolite table or on deletion of a metabolite from the Metabolite table. The chemicalName attribute will be required to store greek letters. As reading UTF-16 character codes from Oracle database fields into Java is not straightforward (See notes in g03r0002rep/code/Metabolite/Prototyping/CharacterEncoding) it has been decided that the chemicalName field will be implemented as a VARCHAR2 field and the xhtml formatted character codes for the greek letters will be coded into the chemical names prior to storage in this field. The molecularWeight attribute is a derived attribute whose value can be calculated from the value from the formula attribute and the atomic weights available from the periodic table of the elements. As the molecular weight is a very stable attribute of a metabolite is has been decided that it would be most efficient for this derived value to be calculated and stored on insert or update of a metabolite to the database rather than each time that it is retrieved. To implement this an additional table called WeightLookup_t

Table	Notes
	has been implemented that contains two fields: atomicCode (implemented as VARCHAR2(5)) and weight (implemented as NUMBER). Both attributes are required and atomicCode is the primary key. This table will be populated with the atom codes and atomic weights from the periodic table of the elements. To perform the derivation of the molecular weight for a metabolite a Java stored procedure has been written and loaded into the database. The Java class is called MolecularWeight and the method is called calculateWeight. This procedure takes a chemical formula, parses it into atom codes and their quantities, looks up the atom weights in the WeightLookup_t table and calculates the total molecular weight. A wrapper function called derive_weight has been implemented to provide a call interface to this function and a trigger called Metabolite_deriveWeight_tr has been implemented to call the function on insert or update of the table. Note that this implementation relies on the existence of properly formatted atomic codes in the WeightLookup_t table (i.e. an initial uppercase character optionally followed by a single lowercase character) and the provision of chemical formulae that contain the same. A sequence and trigger have been implemented to populate the metaboliteID attribute. The chemicalName attribute has a maximum length of 150 and the formula attribute has a maximum length of 50. An additional optional String attribute called "MID" has been implemented to store the user-defined metabolite identifiers included in the data provided for upload. This attribute has a maximum length of 20. It is included solely to ease data upload.
Synonym_t	This table implements the Synonym table from the logical design. As with the chemicalName attribute of Metabolite_t (described above) the name attribute will be required to store greek letters. As reading UTF-16 character codes from Oracle database fields into Java is not straightforward (See notes in g03r0002rep/code/Metabolite/Prototyping/CharacterEncoding) it has been decided that the name field will be implemented as a VARCHAR2 field and the xhtml formatted character codes for the greek letters will be coded into the chemical names prior to storage in this field. The name attribute has a maximum length of 200.
MetSyn_i	This table implements the Metabolite::Synonym association from the logical design. ON DELETE CASCADE has been implemented on the foreign key from Metabolite_t.
Pathway_t	This table implements the Pathway table from

Table	Notes
	the logical design. As with the chemicalName attribute of Metabolite_t (described above) the identity attribute will be required to store greek letters. As reading UTF-16 character codes from Oracle database fields into Java is not straightforward (See notes in g03r0002rep/code/Metabolite/Prototyping/CharacterEncoding) it has been decided that the identity field will be implemented as a VARCHAR2 field and the xhtml formatted character codes for the greek letters will be coded into the chemical names prior to storage in this field. The identity attribute has a maximum length of 150.
MetPath_i	This table implements the Metabolite::Pathway association from the logical design. ON DELETE CASCADE has been implemented on the foreign key from Metabolite_t.
Class_t	This table forms part of the implementation of the BiologicalSource table from the logical design. It is designed as a domain table for the class attributes of the Taxon_t, Sample_t and BiologicalSource_t tables (described below). Information that is required about a biological source class is an external reference to an online taxonomy entry that describes the class and a value that should be used to represent the class in the library interface. When classes are displayed in the interface the display value should be presented as a hotlink to the external reference. The table contains two attributes: externalReference of type VARCHAR2 which should contain the URI to an online taxonomy entry and display also of type VARCHAR2 which should contain the value used to represent the class in the interface. Both attributes are required and externalReference is the primary key. A trigger has been written that implements ON UPDATE CASCADE on foreign keys to this table in the Taxon_t, Sample_t and BiologicalSource_t tables described below. The externalReference attribute has a maximum length of 200 and the display attribute has a maximum length of 20.
Taxon_t	This table forms part of the implementation of the BiologicalSource table from the logical design. It is designed as a domain table for the taxon attribute of the BiologicalSource_t table (described below). Information that is required about a taxon is an external reference to an online taxonomy entry that provides further information about the taxon, the genus, species and class of the taxon (used for searching - note that a taxon may represent a subspecies, line etc. of the species used in its description) and a value that should be used to represent the taxon in the library interface. When a taxon is displayed in

Table	**Notes**
	the interface the display value should be presented as a hotlink to the external reference. The table contains five attributes: externalReference of type VARCHAR2 which should contain the URI to an online taxonomy entry, genus of type VARCHAR2 which should contain the genus name, species of type VARCHAR2 which should contain the species name, class which is a foreign key to the Class_t table (described above) and display of type VARCHAR2 which should contain the value used to represent the taxon in the interface. All attributes are required and externalReference is the primary key. A trigger has been written that implements ON UPDATE CASCADE on the foreign key to this table in the BiologicalSource_t table described below. The externalReference attribute has a maximum length of 200, the genus attribute has a maximum length of 100, the species attribute has a maximum length of 100 and the display attribute has a maximum length of 200.
Sample_t	This table forms part of the implementation of the BiologicalSource table from the logical design. It is designed as a domain table for the sample attribute of the BiologicalSource_t table (described below). Information that is required about a sample is an external reference to an online ontology entry that provides further information about the sample, the class of the sample (used for searching) and a value that should be used to represent the sample in the library interface. When a sample is displayed in the interface the display value should be presented as a hotlink to the external reference. The table contains three attributes: externalReference of type VARCHAR2 which should contain the URI to an online ontoloty entry, class which is a foreign key to the Class_t table (described above) and display of type VARCHAR2 which should contain the value used to represent the sample in the interface. All attributes are required and externalReference is the primary key. A trigger has been written that implements ON UPDATE CASCADE on the foreign key to this table from the BiologicalSource_t table described below. The externalReference attribute has a maximum length of 200 and the display attribute has a maximum length of 100.
BiologicalSource_t	This table implements the BiologicalSource table from the logical design. The taxon and sample attributes from the logical design have been implemented as foreign keys to the Taxon_t and Sample_t tables (described above) respectively. In addition a foreign key to the Class_t table (described above) has been in-

Table	**Notes**
	cluded as part of the primary key to implement the subclasses of BiologicalSource in the logical design. A trigger has been written that implements ON UPDATE CASCADE on the foreign key to this table from the MetBioSource_i table described below.
MetBioSource_i	This table implements the Metabolite::BiologicalSource association and its association class from the logical design. Two triggers have been written (MetBioSource_setIDupdate_tr and MetBioSource_setIDdelete_tr) to delete referenced sets from the ReferenceSet table on update of refSetID in this table or on deletion of a record from this table. ON DELETE CASCADE has been implemented on the foreign key from Metabolite_t.
Adduct_t	This table implements the Adduct table from the logical design. The formula attribute has a maximum length of 50.
Prediction_i	This table implements the Prediction association table on the Metabolite::Adduct PredictedPartOf association from the logical design. ON DELETE CASCADE has been implemented on the foreign key from Metabolite_t.
Measurement_i	This table implements the Measurement association table on the Metabolite::Adduct MeasuredPartOf association from the logical design. A sequence and trigger have been implemented to populate the treeID attribute. ON DELETE CASCADE has been implemented on the foreign key from Metabolite_t. Note that this table is not in 3NF. The ionMode and accuracy attributes can both be worked out from the metaboliteID, formula and mass attributes. The treeID attribute has been introduced as an auto-generated primary key for this table to enable more efficient searching over the Mass_t table (described below) which uses its foreign key to the Measurement_i table as part of its primary key.
Mass_t	This table implements the Mass table from the logical design. The treeID foreign key to Measurement_i has been implemented as part of the primary key and is therefore not optional as documented in the logical design. It is important that the treeID forms part of the primary key for the Mass_t table for efficiency in retrieval of complete ion trees and for efficiency in retrieval of metabolite information from daughter ion masses. A sequence and trigger have been implemented to populate the internalTreeID attribute. The

Table	Notes
	constraint on the contents of treeLevel has been implemented as a check constraint. ON DELETE CASCADE has been implemented on the foreign key from Measurement_i and on the foreign key to a parent Mass_t. Note that this table is not in 2NF. The mass, status and treeLevel attributes can be worked out from the internalTreeID part of the primary key alone. The treeID foreign key to Measurement_i has been introduced as part of the primary key for efficiency in data retrieval as described above.

Datatypes. The following notes describe how the datatypes in the logical design have been implemented in the base tables:

Table 3. Datatypes

Datatype	Oracle type	Notes
long	NUMBER(38)	In Oracle NUMBER(38) is the longest integer possible.
URI	VARCHAR2(200)	There is no basic URI type in Oracle. Given that the library will have a web-based interface implementation of the URI nature of this data type is deferred to the interface.
String	VARCHAR2	The length of each specific VARCHAR2 attribute has been documented within the table notes above.
float	NUMBER	NUMBER is the floating point data type in Oracle.
boolean	VARCHAR2(1)	There is no basic boolean type in Oracle. These attributes should be implemented with a check constraint to ensure that the populating values are either "N", "n", "Y" or "y".
IonMode	VARCHAR2(8)	These attributes should be implemented with a check constraint to ensure that the populating values are either "positive" or "negative".
Accuracy	VARCHAR2(8)	These attributes should be implemented with a check constraint to ensure that the populating values are either "accurate" or "nominal".
MassStatus	VARCHAR2(8)	These attributes should be implemented with a check constraint to ensure that the populating values are either "major", "minor" or "possible".

3. Business rules

The following table describes the constraints to the metabolite library that have not been implemented as part of the base tables:

Table 4. Additional constraints on the metabolite library

Constraint description	Implementation notes
Mandatory participation of ReferenceSet in the Authority::ReferenceSet association.	This check should be carried out by the data submission/update application.
Mandatory participation of Metabolite in the Adduct::Metabolite association.	This check should be carried out by the data submission/update application.
Each ReferenceSet should be referred to once only as either an authority for a primary metabolite designation, an authority for a metabolite, or an authority for a Metabolite::BiologicalSource association.	Implemented by way of a package called ReferenceSets and four triggers called Metabolite_setcheck_tr, Metabolite_setuse_tr, MetBioSource_setcheck_tr and MetBioSource_setuse_tr.
Each Mass should be either a direct fragment of the parent ion of an Adduct or should contain a value for the parent Mass foreign key. Direct fragments of parent ions should not contain a value for the parent Mass foreign key.	This check should be carried out by the data submission/update application.
A Metabolite that is not associated with any BiologicalSource and is not designated as a primary metabolite for any class must contain the value *true* for its measuredInStandard attribute. Conversely, a Metabolite that contains the value *false* for its measuredInStandard attribute must be associated with at least one BiologicalSource and/or be designated a primary metabolite for at least one class. (A Metabolite may contain the value *true* for its measuredInStandard attribute and at the same time be associated with one or more BiologicalSource and/or be designated a primary metabolite for at least one class.)	This check should be carried out by the data submission/update application.
A Metabolite that contains the value *true* for its measuredInStandard attribute must be associated with Adducts that are MeasuredPartsOf it.	This check should be carried out by the data submission/update application.
The class attribute of a BiologicalSource record must be the same as the class attributes of the Taxon and Sample records referenced by the BiologicalSource record.	This check should be carried out by the data submission/update application.

4. Database optimisation

A number of measures have been taken to optimise searches made over the library database. These are described below.

Handling of derived data. The only derived data item is the Metabolite::molecularWeight attribute which is calculated from the Metabolite::formula attribute and a lookup table of atomic codes and weights. As described above, the formula attribute is a very stable attribute of a metabolite description and so, for efficiency in data retrieval, the value for molecularWeight is calculated and stored in the database on insert or update of Metabolite::formula.

Relaxation of normalisation. As described above the Measurement_i and Mass_t tables are not fully normalised. The Measurement_i table is not in 3NF and the Mass_t table is not in 2NF. This relaxation of normalisation allows the primary key on the Mass_t table to include a simplified foreign key from the Measurement_i table which in turn optimises the retrieval of complete ion trees which is a major transaction requirement.

Indexes. When searching over metabolites the most common searches will be by chemical names, synonyms and formulae. In addition the transaction requirements specify the need to search by pathway and there will be a large number of pathways stored in the database. Therefore, indexes have been implemented on the Metabolite::chemicalName, Metabolite::formula and MetPath_i::identity attributes (the Synonym::name attribute is a primary key and is, therefore, indexed already). As there will be limited numbers of biological source in the database it was not felt necessary to implement any indexes over the tables that describe it.

When searching over ESI-MS$_n$ data the most common searches will be for parent ions and major daughter ions. Therefore, indexes have been implemented on the Prediction::mass, Measurement::mass, Adduct::mass and Mass::mass/Mass::status attributes.

G02R012 Bibliography of Food Metabolite Composition Sources

Matthieu Vignes

References

[1] Ora Smith et al. *Potatoes: Production, Storing, Processing.* AVI, Westport, Connecticut, 2nd edition edition, 1979.

[2] Philip A. Brindle, Paul J. Kuhna, and David R. Threlfall. Accumulation of phytoalexins in potato-cell suspension cultures. *Phytochemistry*, 22(12):2719–2721, December 1983.

[3] Grazyna Lisinska and Waclaw Leszczynski. *Potato Science and Technology.* Kluwer Academic Publishers Group, New-York, July 1989.

[4] Paul M. Harris, editor. *The Potato Crop: The Scientific Basis for Improvement.* Chapman & Hall, 2nd edition edition, 1992.

[5] Alain d'Harlingue, Ali M. Mamdouh, Pierrette Malfatti, Marci-Christine Soulie, and Gilbert Bompeix. Evidence for rihitin biosynthesis in tomato cultures. *Phytochemistry*, 39(1):69–70, May 1995.

[6] Graciela M. Bianchini, Bruce A. Stremer, and Nancy L. Paiva. Induction of early mevalonate pathway enzymes and biosynthesis of end products in potato (*solunum tuberosum*) tubers by wounding and elicitation. *Phytochemistry*, 42(6):1563–1571, August 1996.

[7] Ewen R. Brierley, Philip L.R. Bonner, and Andrew H. Cobb. Aspects of amino acid metabolism in stored potato tubers (cv. pentland dell). *Plant Science*, 127(1):17–24, August 1997.

[8] Karin Engström. Sesquiterpenoid spiro compounds from potato tubers infected with *phoma foveata* and *fusarium* spp. *Phytochemistry*, 47(6):985–990, March 1998.

[9] Karin Engström. Post-infectional metabolites from infected potato tubers. *Phytochemistry*, 49(6):1585–1587, November 1998.

[10] Joseph L. Schafer and Maren K. Olsen. Multiple imputation for multivariate missing-data problems: a data analyst's perspective. *Multivariate Behavioral Research*, 33(4):545–571, November 1998.

[11] Joseph L. Schafer. Multiple imputation: a primer. *Statistical methods in Medical Research*, 8:3–15, February 1999.

[12] Harry A. Kuiper, Hub P.J.M. Noteborn, and Ad A.C.M. Peijnenburg. Adequacy of methods for testing the safety of genetically modified foods. *The Lancet*, 354(9187):1315–1316, October 1999.

[13] P. Srinivas Reddy, T. Charles Kumar, M. Narsa Reddy, C. Sarada, and Pallu Reddanna. Differentail formation of octodecadienoic acid and octadecatrienoic acid products in control and injured/infected potato tubers. *Biochimica et Biophysica Acta*, 1483(2):294–300, January 2000.

[14] Paul D. Allison. Multiple imputation for missing data: a cautionary tale. *Sociological Methods and Research*, 9:301–309, February 2000.

[15] Yan-Hwa Chu, Chao-Lin Chang, and Hsia-Fen Hsu. Flavonoid content of several vegetables and their antioxidant activity. *Journal of the Science of Food and Agriculture*, 50(5):561–566, April 2000.

[16] Ute Roessner, Cornelia Wagner, Joachim Kopka, Richard N. Trethewey, and Lothar Willmitzer. Simultaneous analysis of metabolites in potato tuber by gas chromatographymass spectrometry. *The Plant Journal*, 23(1):131–142, July 2000.

[17] Oliver Fiehn, Joachim Kopka, Peter Dörmann, Thomas Altmann, Richard N. Trethewey, and Lothar Willmitzer. Metabolite profiling for plant functional genomics. *Nature Biotechnology*, 18(11):1157–1161, November 2000.

[18] Glennon J. Rogan, Jeff T. Bookout, David R. Duncan, Roy L. Fluchs, Paul B. Lavrik, Stephen L. Love, Mike Mueth, Tammy Olson, Elizabeth D. Owens, Peter J. Raymond, and James Zalewski. Compositional analysis of tubers from insect and virus resistant potato plants. *Journal of Agricultural and Food Chemistry*, 48(12):5936–5945, December 2000.

[19] Léonie M. Raamsdonk, Bas Teusink, David Broadhurst, Nianshu Zhang, Andrew Hayes, Michael C. Walsh, Jan A. Berden, Kevin M. Brindle, Douglas B. Kell, Jem J. Rowland, Hans V. Westerhoff, Karel van Dam, and Stephen Oliver. A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature Biotechnology*, 19(1):45–50, January 2001.

[20] Olga Troyanskaya, Michael Cantor, GavinSherlock, Pat Brown, Trevor Hastie, Robert tibshirani, David Botstein, and Russ B. Altman. Missing values estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, June 2001.

[21] Jan Szopa, Magdalena Wróbel, Iwona Matysiak-Kata, and Anna Świedrych. The metabolic profile of the 14-3-3 repressed transgenic potato tubers. *Plant Science*, 161(6):1075–1082, November 2001.

[22] OECD. *Consensus document on compositional considerations for new varieties of potatoes: key food and feed nutrients, anti-nutrients and toxicants*, series on safety of novel foods and feeds, no. 4 edition, 2002.

[23] Oliver Fiehn. Metabolomics the link between genotypes and phenotypes. *Plant Molecular Biology*, 48(1–2):155–171, January 2002.

[24] Pedro Mendes. Emerging bioinformatics for the metabolome. *Briefings in bioinformatics*, 3(2):134–145, June 2002.

[25] Joseph L. Schafer and John W. Graham. Missing data: our view of the state of the art. *Psychological Methods*, 7(2):147–177, June 2002.

[26] Robert Hall, Mike Beale, Oliver Fiehn, Nigel Hardy, Lloyd Summer, and Raoul Bino. Plant metabolomics: the missing link in functional genomics strategies. *Plant Cell*, 2002.

[27] Jean A.T. Pennington. Food composition for bioactive food components. *Journal of Food Composition and Analysis*, 15(4):419–434, August 2002.

[28] Janet Taylor, Ross D. King, Thomas Altmann, and Oliver Fiehn. Application of metabolomics to plant genotype discrimination using statistics and machine learning. *Bioinformatics*, 18(Suppl.2):S241–S248, October 2002.

[29] Elaine Holmes and Henrik Antti. Chemometric contributions to the evolution of metabolomics: mathematical solutions to characterising and interpreting complex biological nmr spectra. *The Analyst*, 127(12):1549–1557, December 2002.

[30] Heather Greenfield and David A.T. Southgate. *Food composition data: production, management and use.* FAO, Rome, 2nd edition edition, 2003.

[31] Joseph L. Schafer. Multiple imputation in multivariate problems when imputation and analysis models differ. *Statistica neerlandica*, 57(1):19–35, February 2003.

[32] Maciej Stobiecki, Iwona Matysiak-Kata, RafałFrański, Jacek Skala, and Jan Szopa. Monitoring changes in anthocyanin and steroid alkaloid glycoside content in lines of transgenic potato plants using liquid chromatography/mass spectrometry. *Phytochemistry*, 62(6):959–969, March 2002.

[33] Cornelia Wagner, Michael Sefkow, and Joachim Kopka. Construction and application of a mass spectral and retention time index database generated from plant gc/ei-tof-ms metabolite profiles. *Phytochemistry*, 62(6):887–900, March 2003.

[34] Marianne Defernez and Ian J. Colquhoun. Factors affecting the robustness of metabolite fingerprinting using ^1h nmr spectra. *Phytochemistry*, 62(6):1009–1017, March 2003.

[35] Riitta Puupponen-Pimiä, Suivi T. Häkkinen, Marjukka Aarni, Tapani Suortti, Anna-Maija Lampi, Merja Eurola, Vieno Piironen, Anna Maria Nuutila, and Kirsi-Marja Oksman-Cadentey. Blanching and long-term freezing affect various bioactive compounds of vegetables in different ways. *Journal of the Science of Food and Agriculture*, 83(14):1389–1402, November 2003.

[36] Trond Hellem Bø, Bjarte Dysvik, and Inge Jonassen. Lsimpute: accurate estimation of missing values in microarray data with least square methods. *Nucleic Acids Research*, 32(3):e34, February 2004.

[37] Pär Jonsson, Jonas Gullberg, Anders Nordström, Miyako Kusano, Marius Kowalczyk, Michael Sjöström, and Thomas Moritz. A strategy for identifying differences in large series of metabolomic samples analyzed by gc/ms. *Analytical Chemistry*, 76(6):1738–1745, March 2004.

[38] Ming Ouyang, William J. Welsh, and Panos Georgopoulos. Gaussian mixture clustering and imputation of microarray data. *Bioinformatics*, 20(6):917–923, April 2004.

[39] Merijn Kant, Kai Ament, Maurice W. Sabelis, Michel A. Haring, and Robert C. Schuurink. Differential timing of spider mite-induced direct and indirect defenses in tomato plants. *Plant Physiology*, 135(1):483–495, May 2004.

[40] Barbara Burlingame. Fostering quality data in food composition databases: visions for the future. *Journal of Food Composition and Analysis*, 17(3):251–258, June 2004.

[41] Ilan Levin, Avraham Lalazar, Moshe Bar, and Arthur A. Schaffer. Non gmo fruit factories: Strategies for modulating metabolic pathways in the tomato fruit. *Industrial Crops and Products*, 20(1):29–36, July 2004.

[42] Masami Yokota Hirai, Mitsuru Yano, Dayan B. Goodenowe, Shigehiko Kanaya, Tomoko Kimura, Motoko Awazuhara, Masanori Arita, Toru Fujiwara, and Kazuki Saito. Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in arabidopsis thaliana. *Proceedings of the National Academy of Science*, 101(27):10205–10210, July 2004.

[43] Jeroen Jansen, Huub C.J. Hoefsloot, Hans F.M. Boelens, Jan van der Greef, and Age K. Smilde. Analysis of longitudinal metabolomics data. *Bioinformatics*, 20(15):2438–2446, October 2004.

[44] Danh V. Nguyen, Naisyin Wang, and Raymond J. Carroll. Evaluation of missing value estimation for microarray data. *Journal of Data Science*, 2(4):347–370, October 2004.

[45] Shan han Cheng, Zhen hong Su, Cong hua Xie, and Jun Liu. Effects of variation in activities of starch-sugar metabolic enzymes on reducing sugar accumulation and processing quality of potato tubers. *Scientia Agricultura Sinica*, 37(12):1904–1910, 2006.

[46] Kent F. McCue, Louise V.T. Shepherd, Paul V. Allen, M. Malendia Maccree, David R. Rockhold, Dennis L. Corsini, Howard V. Davies, and William R. Belknap. Metabolic compensation of steroidal glycoalkaloid biosynthesis in transgenic potato tubers: using reverse genetics to confirm the in vivo enzyme function of steroidal alkaloid galactosyltransferase. *Plant Science*, 168(1):267–273, January 2005.

[47] Diogo Camocho, Alberto de la Fuente, and Pedro Mendes. The origin of correlations in metabolomics data. *Metabolomics*, 1(1):53–63, January.

[48] Britta Zywicki, Gareth Catchpole, John Draper, and Oliver Fiehn. Comparison of rapid liquid chromatography-electrospray ionization-tandem mass spectrometry methods for determination of glycoalkaloids in transgenic field-grown potatoes. *Analytical Biochemistry*, 336(2):178–186, January 2005.

[49] Kare L. Nielsen, Karen Grøonkjær, Karen G. Welinder, and Jeppe Emmersen. Global transcript profiling of potato tuber using longsage. *Plant Biotechnology Journal*, 3(2):175–185, March 2005.

[50] Anna L. Lytovchenko, Nicolas Schauer, Lothar Willmitzer, and Alisdair R. Fernie. Tuber-specific cytosolic expression of a bacterial phosphoglucomutase in potato (*solanum tuberosum* l.) dramatically alters carbon partitioning. *Plant and Cell Physiology*, 46(4):588–599, April 2005.

[51] Masami Yokota Hirai, Marion Klein, Yuuta Fujikawa, Mitsuru Yano, Dayan B. Goodenowe, Yasuyo Yamazaki, Shigehiko Kanaya, Yukiko Nakamura Masahiko Kitayama, Hideyuki Suzuki Nozomu Sakurai, Daisuke Shibata, Jim Tokuhisa, Michael Reichelt, Jonathan Gershenzon, Jutta Papenbrock, and Kazuki Saito. Elucidation of gene-to-gene and metabolite-to-gene networks in arabidopsis by intefgration of metabolomics and transcriptomics. *The Journal of Biological Chemistry*, 27(8):25590–25595, July 2005.

[52] Carol L. Bender. The post-genomic era: new approaches for studying bacterial diseases in plants. *Australasian Plant Pathology*, 34(4):471–474, September 2005.

[53] Gareth S. Catchpole, Manfred Beckmann, David P. Enot, Madhav Mondhe, Britta Zywicki, Janet Taylor, Nigel Hardy, Aileen Smith, Ross D. King, Douglas B. Kell, Oliver Fiehn, and John Draper. Hierarchical metabolomics demonstrates substantial compositional similarity between genetically modified and conventional potato crops. *Proceedings of the National Academy of Science*, 102(40):14459–14462, October 2005.

[54] Yury Tikunov, Arjen Lommen, C.H. Ric de Vos, Harrie A. Verhoeven, Raoul J. Bino, Robert D. Hall, and Arnaud Bovy. A novel approach for nontargeted data analysis for metabolomics. large-scale profiling of tomato fruit volatiles. *Plant Physiology*, 139(3):1125–1137, November 2005.

[55] Wolfram Weckwerth and Katja Morgenthal. Metabolomics: from pattern recognition to biological interpretation. *Drug Discovery Today*, 10(22):1551–1558, November 2005.

[56] Bjoern H. Junker, Rene Wuttke, Adriano Nunes-Nesi, Dirk Steinhauser, Nicolas Schauer, Dirk Büssis, Lothat Willmitzer, and Alisdair R. Fernie. Enhancing vacuolar sucrose cleavage within the developing potato tuber has only minor effects on metabolism. *Plant Cell Physiology*, 47(2):277–289, February 2006.

[57] R. George Ratcliffe and Yair Shachar-Hill. Measuring multiple fluxes through plant metabolic networks. *The Plant Journal*, 45(4):490–511, February 2006.

[58] Katia Morgenthal, Wolfram Weckwerth, and Ralf Steuer. Metabolomic networks in plants: transitions from pattern recognition to biological interpretation. *BioSystems*, 83(2-3):108–117, February–March 2006.

[59] Heiko Rischer and Kirsi-Marja Oksman-Caldentley. Unintended effects in genetically modified crops: revealed by metabolomics? *Trends in Biotechnology*, 24(3):102–104, March 2006.

[60] Gianfranco Diretto, Raffaela Tavazza, Ralf Welsch, Daniele Pizzichini, Fabienne Mourgues, Velia Papacchioli, Peter Beyer, and Giovanni Guiliano. Metabolic engineering of potato tuber carotenoids through tuber-specific silencing of lycopene epsilon cyclase. *BMC Plant Biology*, 6(13), June 2006.

[61] Roshani Shakya and Duroy A. Navarre. Rapid screening of ascorbic acid, glycoalkaloids and phenolics in potato using high-performance liquid chromatography. *Journal of Agricultural and Food Chemistry*, 54(15):5253–5360, July 2006.

[62] Marianne Long, David J. Millar, Yukiko Kimura, Georgina Donovan, Jon Rees, Paul D. Fraser, Peter M. Bramley, and G. Paul Bolwell. Metabolite profiling of carotenoid and phenolic pathways in mutant and transgenic lines of tomato: identification of a high antioxidant fruit line. *Phytochemistry*, 67(16):1750–1757, August 2006.

[63] Sofia Moco, Raoul J. Bino, Oscar Vorst, Harrie A. Verhoeven, Joost de Groot, Terris A. van Beek, Jacques Vervoort, and C.H. Ric de Vos. A liquid chromatography-mass spectometry-based metabolome database for tomato. *Plant Physiology*, 414(4):1205–1218, August 2006.

[64] Michael Stumpe, Cornelia Göbel, Kirill Deremchenko, Manuella Hoffmann, Ralf B. Klösgen, Katharina Pawlowski, and Ivo Feussner. Identification of an allene oxide synthase (cyp74c) that leads to formation of α-ketols from 9-hydroperoxydes of linoleic and linolenic acid in below-ground organs of potato. *The Plant Journal*, 47(6):883–896, September 2006.

[65] Ralf Steuer, Katja Morgentahl, Wolfram Weckwerth, and Joachim Selbig. *Metabolomics: methods and protocols*, volume 358 of *Methods in Molecular Biology*, chapter 7 - A gentle guide to the analysis of metabolomic data, pages 105–126. Humana Press, Totowa, September 2006.

[66] Ian J. Colquhoun, Gwénaëlle Le Gall, Katherine A. Elliott, Fred A. Mellon, and Anthony J. Michael. Shall i compare thee to a gm potato. *Trends in Genetics*, 22(10):525–528, October 2006.

[67] Wayne L. Morris, Laurence J.M. Ducreux, Peter Hedden, Steve Millan, and Mark A. Taylor. Overexpression of a bacterial 1-deoxy-d-xylose 5-phosphate synthase gene in potato tubers perturbs the isoprenoid metabolic network: implications for the control of the tuber life cycle. *Journal of Experimental Botany*, 57(12):3007–3018, September 2006.

[68] Mendel Friedman. Potato glycoalkaloids and metabolites: roles in the plant and in the diet. *Journal of Agricultural and Food Chemistry*, 54(23):8655–8681, November 2006.

[69] Satu J. Lehesranta, Howard V. Davies, Louise V.T. Shepherd, Kaisa M. Koistinen, Nathalie Massat, Naoise Nunan, James W. McNicol, and Sirpa O. Kärenlampi. Proteomic analysis of the potato tuber life cycle. *Proteomics*, 22(6):6042–6052, November 2006.

[70] Dae-Won Kim, Ki-Young Lee, Kwang H. Lee, and Doheon Lee. Towards clustering of incomplete microarray data without the use of imputation. *Bioinformatics*, 23(1):107–113, January 2007.

[71] Barbara J. Daniels-Lake, Robert K. Prange, Sonia O. Gaul, Kenneth B. McRae, and Roberto de Antueno. A musty "off" flavor in nova scottia potatoes is associated with 2,4,6-trichloroanisole released from pesticide-treated soils and high soil temperature. *Journal of the American Society for Horticultural Science*, 132(1):112–119, January 2007.

[72] Peter J. Lea, Ladaslav Sodek, Martin A.J. Parry, Peter R. Shewry, and Nigel G. Halford. Asparagine in plants. *Annals of Applied Biology*, 150(1):1–26, February 2007.

[73] Ibolya Stiller, Gábor Dancs, Holger Hesse, Rainer Hoefgen, and Zsófia Bánfalvi. Improving the nutritive value of tubers: elevation of cysteine and gluthathionine contents in the potato clutivar white lady by marker-free transformation. *Journal of Biotechnology*, 128(2):335–343, February 2007.

[74] Satu J. Lehesranta, Kaisa M. Koistinen, Nathalie Massat, Howard V. Davies, Louise V.T. Shepherd, James W. McNicol, Ismail Cakmak, Julia Cooper, Lorna Lück, Sirpa O. Kärenlampi, and Carlo Leifert. Effects of agricultural production systems and their components on protein profils of potato tubers. *Proteomics*, 7(4):597–604, February 2007.

[75] Kent F. Mccue, Paul V. Allen, Louise V.T. Shepherd, Alison Blake, M. Malendia Maccree, David R. Rockhold, Richard G. Novy, Derek Stewart, Howard V. Davies, and William R. Belknap. Potato glycosterol rhamnosyltransferase, the terminal step in triose side-chain biosynthesis. *Phytochemistry*, 68(3):327–334, March 2007.

[76] Gianluca Paventi, Roberto Pizzuto, Gabriella Chieppa, and Salvatore Passarella. L-lactate metabolism in potato tuber mitochondria. *Federation of European Biochemical Societies Journal*, 274(6):1459–1469, March 2007.