2011

Bioinformatics methods for metabolomics based biomarker detection in functional genomics studies

Preeti Bais
Iowa State University

Bioinformatics methods for metabolomics based biomarker detection in functional genomics studies

By
Preeti Bais

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Julie A. Dickerson, Co-major Professor
Basil Nikolau, Co-major Professor
Dianne Cook
Shashi Gadia
Hui-Hsien Chou

Iowa State University
Ames, Iowa
2011

DEDICATION

Dedicated to my daughter Priyanka and my husband Atul

TABLE OF CONTENTS

LIST OF TABLES ix

LIST OF FIGURES x

ACKNOWLEDGEMENTS xi

ABSTRACT xii

CHAPTER 1. INTRODUCTION 1

Thesis Organization 1

CHAPTER 2. METABOLOMICS: GENERAL BACKGROUND 4

What is Metabolomics? 4

Some Examples of Usage of Metabolomics 4

Metabolomics as a Functional Genomics Tool 5

Types of Metabolic Analyses 5

The Metabolomics Workflow 6

Experiment Design 7

Plant Cultivation 8

Extraction 8

Separation and Detection 8

Data Storage 12

Data analysis and Biological Interpretation 13

References 22

CHAPTER 3. PLANTMETABOLOMICS.ORG: A WEB PORTAL FOR PLANT METABOLOMICS EXPERIMENTS 30

Abstract: 30

Introduction 31

Development of Plantmetabolomics.org: Rationale....32

Design Requirements and Functionality 33

PlantMetabolomics Content....35

Experiment Annotation....35

Design of experiments: 36

Experimental Data 38

Tutorials....39

Data Analysis and Visualization of Experimental Data 41

Data Quality Checks: 44

Query Capabilities....45

Conclusions....46

Materials and Methods 47

Normalization and data processing 47

Missing Values:....47

Ratio Plot 47

Error Plots....48

Compound Curation in AraCyc....48

Data Curation....49

Supplemental Data 49

Acknowledgments 50

Supplementary Document S1: Data Base Schema 51

Supplementary Document S2: PM Website Map.....52

Supplementary Document S3: Applications of PM and Availability of Metabolomics Data: Case Study 53

CHAPTER 4. PLANTMETABOLOMICS.ORG:MASS SPECTROMETRY BASED ARABIDOPSIS METABOLOMICS-DATABASE AND TOOLS UPDATE.....56

Abstract 56

Introduction 56

Database Contents.....57

Analysis Tools for Metabolomics 59

Conclusions and Future Developments 64

Availability 64

Funding 65

Acknowledgements 65

References.....66

List of Figures.....67

CHAPTER 5. DATA ANALYSIS PIPELINE IN FUNCTIONAL GENOMICS USING METABOLOMICS AND MACHINE LEARNING 69

Abstract: 69

Introduction 71

Materials and methods 72

Plant Materials 72

Metabolite Detection platforms 72

Metabolomics Analysis Pipeline 74

Data Preprocessing 74

Exploratory Data Analysis 75

Data Preprocessing /Normalization 76

Data Integration across multiple platforms 76

Random Forest (RF) Analysis 76

Incorporating the structurally unknown compounds 77

Results and Discussion 78

Metabolomics platform integration 81

Incorporating the Unknown Metabolites 84

Biological Confirmations and Discoveries 84

Conclusions 87

List of Tables 88

List of Figures 89

Supplementary Documents 89

Acknowledgments 89

References 91

CHAPTER 6. PARTIAL CORRELATION NETWORKS TO PUTATIVELY INDENTIFY UNKNOWN METABOLITES IN NON-TARGETED METABOLOMICS 95

Abstract: 95

Introduction 96

Materials and methods 98

Plant Materials 98

Non targeted GC-TOF analysis 98

Data Analysis 98

Data Normalization 98

Random Forest Classification 98

Correlation Network Analysis 99

Results and Discussion 100

RF classification and biomarker selection 100

GGM Networks 102

Case Study and examples 105

Conclusions 106

Availability of database 106

List of Tables 107

List of Figures 107

Supplementary Documents 107

Acknowledgments 107

References 108

CHAPTER 7. CONCLUSIONS 110

CHAPTER 8. APPENDIX – SUPPLEMENTARY DOCUMENTS 112

Supplementary Document 8.1: Data Quality analysis 112

Supplementary Document 8.2: List Of Genotypes used for Partial Correlation Analysis 116

Supplementary Document 8.3: Applications of biomarker database and putative identification of unknown metabolites: Case Study.....118

LIST OF TABLES

Table 2-1 Example of Metabolomics Data 14

Table 3-1 Experimental set-ups used to generate metabolomics data contained in PM 39

Table 5-1 Platform Summary 73

Table 5-2 Comparison of Platform integration methods – *oxp1* vs. wild type.....82

Table 5-3 Potential Biomarkers for *oxp1* mutation.....83

Table 5-4 RF classification results for *ggt1* mutant using two platform integration methods .85

Table 5-5 Potential Biomarkers for *ggt1* mutation 86

Table 5-6 RF classification results for *ggt2* mutant.....87

Table 6-1 Top 20 highly correlated compounds from GGM networks.....103

Table 6-2 RF classification results of SALK_070569 vs. wild type from 25 Bootstrap runs.104

Table 8-1 Summary of Replicate quality analysis of platforms 114

Table 8-2 List of Genotypes used for Partial correlation Analysis.....116

LIST OF FIGURES

Figure 2-1 Block Diagram of main steps of a typical plant metabolomics workflow..... 7

Figure 2-2 Hierarchical Cluster Analysis (HCA) example 17

Figure 2-3 PCA Example 19

Figure 2-4 Scree Plot Example 20

Figure 3-1 Diagram of the main components of the plantmetabolomics.org portal 34

Figure 3-2 Schematic representation of the process used in generation of metabolite data ..37

Figure 3-3 Ratio Plot 42

Figure 3-4 Metabolite details of Methionine.....43

Figure 3-5 Replicate quality check 45

Figure 3-6 Database Schema 51

Figure 3-7 Website Map.....52

Figure 4-1 Visualization tools 59

Figure 4-2 Data analysis tools.....63

Figure 5-1 Data Analysis Pipeline 74

Figure 5-2 Log2 ratio plot between *oxp1* mutant and wild type samples.....75

Figure 5-3 GSH degradation Pathway.....80

Figure 6-1 RF classification of 70 mutant lines of Arabidopsis with wild type samples 101

Figure 8-1 Replicate quality analysis.....113

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research. I would like to thank Dr.Julie A. Dickerson and Dr. Basil Nikolau for giving me the opportunity to work in this exciting field and for their guidance throughout my research. I would also like to thank members of my committee Dr. Dianne Cook, Dr. Hui-Hsien Chou and Dr. Shashi Gadia for their guidance.

I would like to thank my colleagues at Arabidopsis Metabolomics Consortium and Iowa State University. I would like to give special thanks to Ms. Trish Stauble for making our interdepartmental program feel like home.

Finally, I would like to thank my family including my sisters, my aunt and my parents for their constant encouragement and motivation.

This work was funded by NSF grant #08200823 and MGET (Multidisciplinary Graduate Education Training) grants.

ABSTRACT

The biochemical and physiological functions of a large proportion of the approximately 27,000 protein-encoding genes in the Arabidopsis genome is experimentally undetermined using sequence homology techniques alone. This thesis presents a set of bioinformatics resources including a software platform for data visualization and data analysis that address the key issues in incorporating the metabolomics data for functional genomics studies.

Multiple mass spectrometry based metabolomics platforms are combined together to get better coverage of the metabolome. Different strategies for integrating the metabolomics abundance data from multiple platforms are compared to find the ideal method for biomarker discovery. A new method of putatively identifying unknown metabolites by first order partial correlation networks is proposed that uses the existing data to incorporate structurally unknown metabolites. A comprehensive study of 70 single gene knock mutants vs. wild type samples is performed using Random Forest machine learning algorithm and a biomarker database for each of the 70 mutations is built with the key metabolites including the putative identifications of unknown metabolites.

A proof-of-concept analysis on the oxoprolinase (*oxp1*) and gamma-glutamyl transpeptidase (*ggt1* and *ggt2)* single gene knock-out mutants in the glutathione degradation (GSH) pathway of the *Arabidopsis* confirms the known biology that OXP1 is responsible for conversion of 5-oxoproline (5-OP) to glutamic acid. In addition, *ggt1/ggt2* analysis supports the hypothesis that the GGT genes may not be major contributors for the 5-OP production. Also, the lack of biochemical changes in *ggt2* mutation supports the previous studies of its low level expression in leaf tissues.

The metabolomics database, the biomarker database and the data mining tools are implemented in a web based software suite at www.plantmetabolomics.org.

CHAPTER 1. INTRODUCTION

Metabolomics is an important functional genomics tool and can be used in finding the functions of genes where sequence genomics techniques alone are not adequate. However, the bioinformatics resources and methods for the metabolomics are still very new and under developed. One of the biggest challenges in metabolomics is to integrate multiple platforms as no single analytical technique or platform can cover the whole metabolome of an organism and to understand the role of structurally unknown metabolites which constitute a major part of the detected metabolites in any large scale metabolomics study.

This thesis communicates my contribution to the field of bioinformatics in metabolomics area. This includes database and tool development to integrate multiple metabolomics platforms and development of new techniques for the analysis of the metabolomics data. The main chapters of this thesis are published, or to be submitted manuscripts in peer reviewed journals in plant science and bioinformatics area and discuss my solutions for the above stated problems in metabolomics. The thesis is organized as follows –

Thesis Organization

- **Chapter 1: Introduction** and the contribution of thesis in Metabolomics area.
- **Chapter 2: General Introduction:** Discusses the general background of the available metabolomics technologies and bioinformatics resources.
- **Chapter 3: Plantmetabolomics.org A database for plant metabolomics experiments:** Discusses a web based database, PM (www.plantmetabolomics.org) that combines 11 different analytical platforms to detect ~1400 metabolites in 140 single gene knock out mutants of Arabidopsis. A researcher can use this database and visualization tools to compare biochemical changes due to a mutation to form hypothesis

about the function of a gene of interest. This chapter was published in Plant Physiology in 2010.

- **Chapter 4: Plantmetabolomics.org: Mass Spectrometry based Arabidopsis metabolomics database and tools- Update:** Discusses new multivariate and machine learning data analysis and data visualization tools that were incorporated in PM database in 2011 along with new morphological data. This database along with the new data mining and visualization tools provides a hypothesis building platform for the researchers that are interested in functions of any of the genes contained in the database. This work has been published in Nucleic Acid Research 2012 database issue.
- **Chapter 5: Data analysis pipeline in functional genomics using metabolomics and machine learning:** Discusses the methods and data analysis pipeline to analyze and integrate metabolomics data from multiple platforms for biomarker discovery. These methods are tested on 3 genes of an Arabidopsis pathway to confirm the known biology and provide new knowledge. This chapter is based on a manuscript that is to be submitted to Plant Physiology in 2011.
- **Chapter 6: Partial correlation networks to putatively identify unknown metabolites in non-targeted metabolomics:** Discusses a comprehensive study on 70 mutation lines of Arabidopsis using machine learning and gas chromatography mass spectrometry based metabolomics. A biomarker database is created for the key metabolites for the classification of mutant vs. wild type samples for all of these mutation lines. First order partial correlation networks built across the mutation are used in putatively identifying the potential biomarkers of unknown structures from a machine learning method. This provides a cost effective way of incorporating the unknown metabolites to gain biological insight. This chapter is based on a manuscript that is to be submitted to Plant Methods in 2011.

- **Chapter 7: Conclusions:** Conclusions and significance of this work.

CHAPTER 2. METABOLOMICS: GENERAL BACKGROUND

What is Metabolomics?

Metabolic analysis is the study of small molecules (molecular weight <1,000 Da) in a biological system (Fiehn O. et al. 2000; Hall R. et al. 2002). The biochemical state of an organism is the result of interaction between its genotype (G), its environmental (E), and its homeostasis mechanisms. Living cells respond to environment or genetic perturbation and this response can be measured by quantifying the change in concentration of metabolites. Metabolic information reflects the response of a plant cell to its environment or genetic perturbation more accurately than the sequence or the gene expression analysis as it is the end product of a gene's expression .Metabolomics can be a useful tool in assessing the plant's physiology, growth characteristics, and stress response (Sumner L. et al. 2003).

Some Examples of Usage of Metabolomics

Metabolomics is increasingly used in determining and improving quality traits such as the color, taste and flavor of the plants because these traits are related to metabolic composition. For example, metabolomics analysis was done in tomato using domesticated and undomesticated species to discover the primary and secondary metabolites that contribute to flavor and color (Schauer N. et al. 2005). A combination of genomic analysis with metabolomic profiling identified novel genes involved in the fragrance production in rose petals (Guterman I.et al. 2002). Metabolic composition analysis was performed on the samples from genetically modified and conventional potato tubers to find out if GM potatoes had any undesired or potentially harmful metabolites apart from the targeted changes (Catchpole G. et al. 2005). Metabolic profiling has also been used to explore the degradation process of linoleic acid in stored apples (Beuerle T. et al. 1999).

Metabolomics as a Functional Genomics Tool

There are about 27,000 protein-encoding genes in the Arabidopsis genome whose functions are experimentally undetermined. The functions of these genes are either completely undermined using sequence homology or can only be classified in the broad functional classes. The first category of unknown genes consists of 9000 genes that share no sequence homology of any genes in the sequence database or share sequence homology to genes of unknown functions. The second category consists of approximately 15000 genes. This thesis is part of a multi-disciplinary experimental system that has been developed to generate and evaluate metabolomics data as a tool for deciphering gene function in Arabidopsis by knocking down a single gene and comparing its metabolomics concentration data with the wild type samples while keeping the environment stable. Since the metabolome reflects the final outcome of a genes expression at the molecular level, the comparison of single gene knock outs with the wild type samples may give clues to the functions of a gene. This thesis focuses on the bioinformatics solutions and computational infrastructure development for the same. The following paragraphs introduce the existing technologies and practices along with the challenges in the metabolomics area. The later chapters discuss our proposed solutions to some of these challenges.

Types of Metabolic Analyses

Metabolomic analyses can be divided into targeted and non-targeted analyses. Targeted analysis aims at a selected group of metabolites or pathways and provides precise quantification of those metabolites in a sample. This method requires that the structure of the targeted metabolites is known and that the metabolites are available in purified form before the analysis. Targeted methods cannot detect any novel metabolites in a sample (Aharoni A. et al. 2002).

Non-targeted analyses can be further divided into finger printing and metabolic profiling. Fingerprinting is a high throughput, global analysis of samples which provides a global snapshot without precisely quantifying or identifying all the metabolites in the sample. It is mainly used to discriminate between two samples under different biological conditions. Metabolic profiling analysis is an unbiased comprehensive analysis of all the metabolites of a biological system. The biological system is perturbed and abundances of all metabolites are compared between two types of samples to determine the effect of perturbation. This is more difficult than targeted analysis as the number and classes of the metabolites affected by the perturbation is usually not known and the results are sensitive to bias and selective reporting (Fiehn O. 2002). A two tiered approach is also employed in some recent studies where an initial assessment is first performed with fingerprinting and then a more targeted approach is applied with higher resolution methods. For example, the two tiered approach was applied in comparison study of two closely related potato crops (Catchpole G.et al. 2005).

The Metabolomics Workflow

A typical plant metabolomics experiment consists of the following sequential steps: experimental design, plant cultivation, extraction, separation and detection, and finally data analysis. Figure 2-1 illustrates this flow and shows how data must be collected in a searchable database for ongoing analysis and annotation.

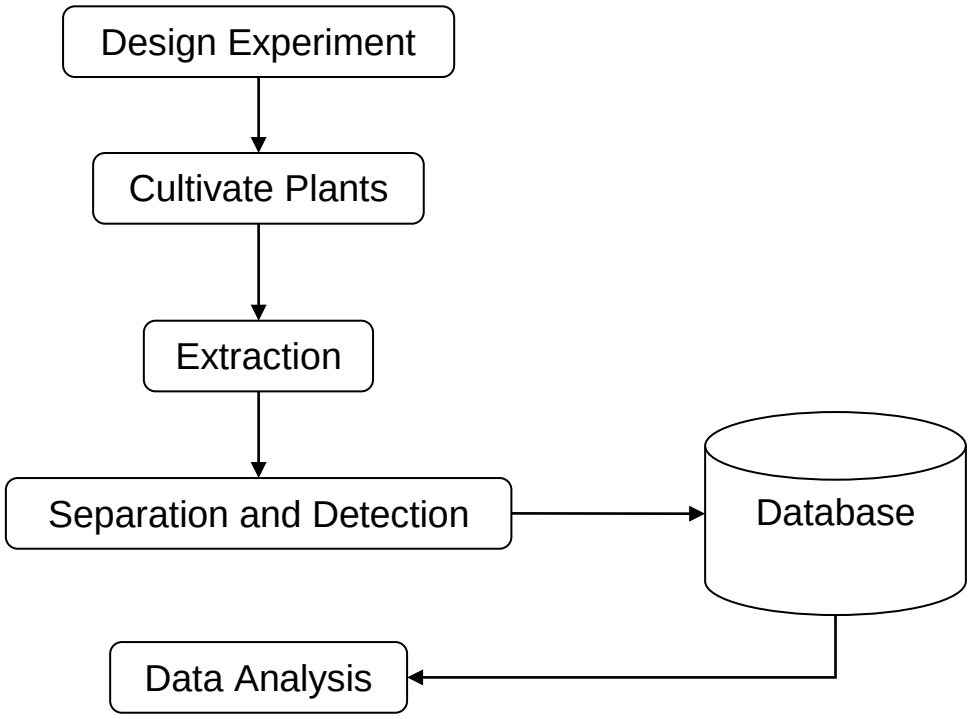

Figure 2-1 **Block Diagram of main steps of a typical plant metabolomics workflow**

Experiment Design

The primary aim of most metabolomics study is to find the difference between the samples that are subjected to genetic or environmental stimuli. Metabolomic experiments tend to be noisy and interpretation of experimental results can be faulty due to bias, inadequate sample size, over fitting and excessive false discovery rate due to multiple hypotheses testing. Powerful multivariate analysis techniques that are commonly used in the high dimensional metabolomics data require adequate number of sample replicates. It is also very important to store the metadata of the experiment design in a structured setup so the experiment results can be tracked and verified. For mass spectrometry based experiments, SetupX and Binbase provide a framework that combines data and biological metadata for steering laboratory work flows and employs automated metabolite annotation (Scholz M. et al. 2007).

Plant Cultivation

To avoid variations among samples in a plant metabolomics experiment, large volume growth chambers are normally suggested to minimize the variance due to soil, temperature and humidity. In case of small volume growth chambers rotation of the pots is suggested and in some recent studies soil less ceramic culture system is also used to maintain the exact supply of plant nutrition and water (Fukusaki E. et al. 2003).

Extraction

First, quenching is performed to freeze the status of the metabolome at a given time and then metabolites are extracted from a biological sample. Different extraction procedures are applied to different classes of metabolites depending on their solubility, stability, interferences in the extraction solvents. A good extraction procedure stabilizes a large number of metabolites without any degradation or modification of the targeted metabolites. There is always a tradeoff between comprehensiveness and metabolite stability because extraction conditions that are ideal for one class of metabolites may degrade other classes of metabolites (Maharjan R. et al. 2003).

Separation and Detection

In a typical metabolomic profiling experiment, multiple biological samples from different stimulus conditions, various time points, and/or genetically distinct cultivars are analyzed to discover biomarkers and their associated biochemical pathways. There are several analytical technologies that are used for separation and detection of individual metabolites from a complex mixture. Each technology has its own advantages and disadvantages therefore a combination of these technologies is commonly applied. Some widely-used analytical methods are described below:

Nuclear Magnetic Resonance (NMR)

NMR is a non-destructive analytical method and can determine the molecular structure along with the quantity of metabolites. NMR has several advantages in comparison with other

analytical technologies for high-throughput metabolite analyses. As a non-destructive method, NMR does not require metabolite derivitization and ionization (Hagel J. et al. 2008). The non-destructive method can be highly automated to achieve high sample throughput. NMR spectra can be obtained in vivo from cultured cells and tissues (Ratcliffe R. et al. 1994; Ratcliffe R. et al. 2001). However, NMR suffers from relatively low sensitivity (Katja D. 2007; Pan Z. et al. 2007), than chromatography-coupled mass spectrometry (Sumner L. et al. 2003).

Mass Spectrometry (MS)

MS is often used as a hyphenated technique where the metabolite mixture is first separated using gas chromatography (GC), liquid chromatography (LC), or capillary electrophoresis (CE) and then analyzed by MS, which produces mass spectrum which is an intensity vs. *m/z* (mass-to-charge ratio) plot representing a chemical analysis. The unknown compounds are identified by comparing its experimental mass spectrum against a library of mass spectra. Manual interpretation or software assisted interpretation of mass spectra are performed if the experimental mass spectrum does not match any spectrum in the database. Recent developments in the instruments have greatly increased the number of metabolites that can be accurately identified and quantified by chromatography-coupled MS. More detailed discussion of the MS technology is provided in the later part of this section.

Separation in MS

- **Gas chromatography – mass spectrometry (GC-MS):** GC-MS involves the separation of volatile, thermally stable analytes by GC and subsequent detection by electron ionization (EI) MS (Hagel J. et al. 2008). It is most suitable for analyzing amino acids, sugars, sugar alcohols, aromatic amines and fatty acids (Roessner U. et al. 2000). GC-MS offers high chromatographic reproducibility and resolution and is lower in cost than the LC-MS or CE-MS that are described below (Kopka J. 2006; Katja D. 2007) , but some large and polar metabolites cannot be analyzed by GC. There are currently many

commercial and public mass spectral reference libraries for GC–MS including the NIST database (http://www.nist.gov/srd/nist1.htm). GC-MS technology requires that samples be either volatile or chemically derivatized (Sumner L.et al. 2003). Recent applications of GC-MS include mutant classification (Messerli G. et al. 2007), functional genomics (Schauer N. et al. 2006), and the integration of metabolite and transcript datasets (Carrari F. et al. 2006; Baxter C. et al. 2007; Fatma K. et al. 2007).

- **Liquid chromatography – mass spectrometry (LC-MS):** LC provides covers a wider range of metabolites than GC. Non-volatile metabolites can also be analyzed because it does not require derivitization. However, it is difficult to compare LC-MS chromatograms between different laboratories because of the variety of LC-MS instrumentations (Moco S. et al. 2006). Mass spectral reference libraries for LC-MS (e.g. METLIN database (Smith C. et al. 2005) are also much fewer than the GC-MS libraries. Some examples of LC-MS include identifying secondary metabolites in roots and leaves of Arabidopsis (von Roepenack-Lahaye E. et al. 2004), and to compare tubers of potato of different genetic origin and developmental stages (Vorst O.et al. 2005).
- **Capillary electrophoresis – mass spectrometry (CE-MS):** CE-MS is useful in detecting charged metabolites because it separates compounds according to their mass-to-charge ratios (Hagel J. et al. 2008). CE-MS has been used to study amino acids, carbohydrates, vitamins, organic acids and inorganic ions (Soga T. et al. 2001). Soga T. et al. (2003) used CE-MS to analyze anionic metabolites, cationic metabolites, and nucleotides and Coenzyme A compounds, achieving a comprehensive coverage of the metabolome (Soga T. et al. 2003). CE-MS is also combined with CE diode array detection to simultaneously determine the main metabolites in rice leaves (Shigeru S. 2004).

Detection

The metabolite detection process involves the following steps.

- **Noise reduction:** Baseline correction and noise reduction are performed on the raw data as a first step.
- **Peak identification and quantification**: After the initial noise reduction, peak detection and deconvolution of overlapped peaks are performed. National Institute of Standards and Technology (NIST) provides the free software called AMDIS (Automatic Mass Spectral Deconvolution and identification System) which extracts individual component spectra from GC/MS data and compares all separated components against a library of compounds (Stein S.1999). WILEY (Palisade Cooperation, Newfield, NY) also provides mass spectrum library of compounds for identifications. Sample preparation and detector sensitivity cause the signal intensity to change over time, therefore internal standards are added to the sample. The variation in intensity of these internal standards is used to normalize between samples (Jonsson P. et al. 2005).
- **Structure elucidation:** Some of the well-known small molecule structure databases which contain the physical-chemical properties of standard compounds are LIGAND database (Goto S.et al. 2002) and the NCI database (Korcok M. 1985). The Golm Metabolome Database contains GC libraries for plant metabolites (Kopka J. et al. 2005).
- **Naming conventions for unknown compounds:** Unbiased analysis of the metabolome discovers many new metabolites and it is very important to name these new metabolites in a consistent way so the results from various experiments can be compared against each other. The novel and unknown compounds from a metabolomics experiment that do not match any known compounds from the above described libraries are given a unique name which combines the information about retention index, polarity among other important features of the mass spectra for that compound (Bino R. et al.

2004). For example BJN-GCMS-CW-20.201 is the unique identifier for an unknown metabolite from Dr. Basil J. Nikolau's (BJN) lab using Gas Chromatography Mass Spectrometry (GCMS) for Cuticle Wax (CW) extraction and with the retention index of 20.201.

Data Storage

The outcome of a metabolomics experiment is a matrix that contains the metabolite abundance data along with the annotations for the metadata of the experiment. It is very important to capture the experimental conditions along with the experiment design information to be able to generate reproducible results. This can be enabled by a good lab information management system (LIMS) for data capture and submission. For examples : SetupX and Binbase provide a framework that combines mass spectrometry data and biological metadata for steering laboratory work flows and employs automated metabolite annotation (Scholz M. et al.2007). Standards for the annotation of metabolomics experiments are still under active development and are based on the recommendations of the Metabolomics Standards Initiative (MSI)(Sansone S. et al. 2007). MIAMET (Minimum Information About a METabolomics experiment) defines necessary information that should accompany the experimental data to make it useful and understandable (Bino R. et al. 2004). MIAMET suggests that each metabolomics experiment should contain information about its design, samples, sample preparation, metabolite extraction and derivation, metabolic profiling design, metabolite measurement and specifications. ArMet is a framework and database model for the description of plant metabolomics experiments. It captures the entire timeline of a plant metabolomics experiments (Jenkins H. et al. 2004).

Examples of plant metabolomics databases:

- The Golm metabolome database (GMD) provides metabolite GC-MS libraries and one set of metabolite profiling experiments for plants (Kopka J. et al. 2005).

- Metabolome Tomato Database (MoTo DB) is an LC-MS based metabolomics of tomato fruit (*Solanum lycopersicum*) (Moco S.et al. 2006).
- Madison Metabolomics Consortium Data base (MMCD) - NMR data base (Cui Q. et al. 2008).
- Platform for Riken Metabolomic (PRIME) is database of multi-dimensional NMR spectroscopy, GC/MS, LC/MS, and CE/MS based metabolomics and provides tools for integration with other omics data (Akiyama K. et al. 2008).
- Human metabolome database (HMDB) provides the most comprehensive database of human metabolites (Wishart D. et al. 2007). METLIN compiles an extensive list of known metabolites and provides their MS/MS spectra with links to KEGG database (Smith C. et al. 2005).

Data analysis and Biological Interpretation

The outcome of a metabolomics experiment is a data matrix that contains the metabolite abundance information for all the metabolites that were detected for a sample under a specific condition. The number of variables (metabolites) is usually much larger than the number of samples as shown in Table 2-1

Table 2-1 Example of Metabolomics Data

	Methionine	Serine	Glu	Metabolite 10	Metabolite 11	Metabolite 12	Metabolite 13
MT1	1.0	23.3	2000				..
MT2	1.5	23	N				..
MT3	1.3	22	2300				..
WT1	100.2	22	2500				..
WT2	111.1	22	2500				..
WT3	100	22.1	2200				..

Table 2-1 Example of Metabolomics Data A sample data matrix in a metabolomics experiment. There are 3 replicates of the mutant samples which are compared against the 3 replicates of wild type samples to assess the effect of mutation. The abundance levels of many metabolites are tested. Sometimes the abundance is below the detection limit of measuring instrument which is shown as an “N” in the above example. A typical metabolomics data contains many features (metabolites) and very few samples. Some of the metabolites have known structures but many have unknown structures and are given a unique identifier.

A typical data analysis process is described below:

Data Preprocessing:

Data preprocessing is performed to improve the overall data quality and prepares the data for statistical analysis with improved accuracy. Some key preprocessing steps are described below

- **Outlier detection:** The outliers are data points which differ from the most of the other data points. This difference can be due to the biological reasons which can lead to discovery of novel pathways (Tjalsma H. 2007). The outliers which may be due to non-biological reasons are then removed before any further statistical analysis. In general, the outliers are the values which are more than 3 standard deviations away from the sample mean. Statistical methods such as principle components analysis (PCA) and independent component analysis (ICA) which are discussed later can be used in outlier detection. No computational method can determine if the outlier is due to biological or non-biological reasons, therefore expert knowledge of the biology is required to decide if the outliers should be kept or not.
- **Missing value estimation:** Missing values can be caused by signals that are below the detection limit of the instrument (True Missing) or because they were not collected. Bioconductor package, *pcamethods*, provides several algorithms for missing value imputation.
- **Data transformation:** The normalization and transformations are performed to remove the non-biological variations and then make the data normally distributed. Metabolomics data can have various sources of uninduced variations like difference in abundance; metabolites in the central metabolism are more stable than the secondary metabolism, and large fluctuations under identical biological conditions, technical variations and heteroscedasticity**.** Range scaling is one of the most commonly used methods to give

equal importance to all the metabolites. Log transformation is also commonly used because it not only provides some scaling but also makes the multiplicative models additive and thus useful in removing heteroscedasticity (van den Berg R. et al. 2006)

Pattern recognition and data mining:

Unsupervised methods do not require any previous knowledge about the groups or classes that the data belongs to and can be used to summarize and find key features in the data and class discovery. Some of the popular unsupervised methods are described below:

Clustering:

Clustering is used to group samples with similar metabolite profiles together. Clustering organizes the data into groups such that objects that are in a cluster are more closely related to each other than with the objects from the other clusters. To measure proximity many distance functions, e.g. Euclidian distance, Mahalanobis distance etc. are used. A good clustering results in the compact clusters that are also distant from other clusters. Some of the traditional clustering algorithms include hierarchical (HCA), k-means and self-organizing maps (SOM). HCA can be performed as an agglomerative methods or a divisive method. The agglomerative methods begin with each observation being considered as separate clusters and then proceeds to combine them until all observations belong to one cluster. Agglomerative methods are more commonly used in metabolomics studies. The divisive methods start with all of the observations in one cluster and then proceeds to partition them into smaller clusters. Some of the algorithms for HCA are average linkage, complete linkage, single linkage and Ward's linkage. Average linkage clustering uses the average similarity of observations between two groups as the distance measure between the two groups. Complete linkage clustering uses the farthest pair of observations between two groups to determine the similarity of the two groups. Single linkage clustering computes the similarity between two groups as the similarity of the closest pair of observations between the two groups. Ward's linkage uses an analysis of variance approach to

evaluate the distances between clusters. For example hierarchical clustering (HCA) was used to study volatile metabolites of 94 tomato genotypes which were obtained with Solid phase micro extraction (SPME-GC-MS). The first stage of clustering was helpful in removing the non-plant contaminant metabolites which were caused by the SPME fiber material. The HCA analysis resulted in two distinct clusters for the plant and non-plant metabolites. After removing the non-plant metabolites from the data, hierarchical clustering was used again to cluster the genotypes (Tikunov Y. et al. 2005). Figure 2-2 shows an example of cluster analysis in of the data from 2010 Arabidopsis Project (www.plantmetabolomics.org) , where wild type samples under 7 different conditions are compared against SALK_021108 mutant under the same 7 conditions. Two different linkage methods are employed in this analysis. We see that both the methods are able to put the most of the first 7 samples (Wild Type) and most of the next 7 samples (mutant) in different clusters. Low light mutant is clustering with low light wild type sample, which should be investigated further.

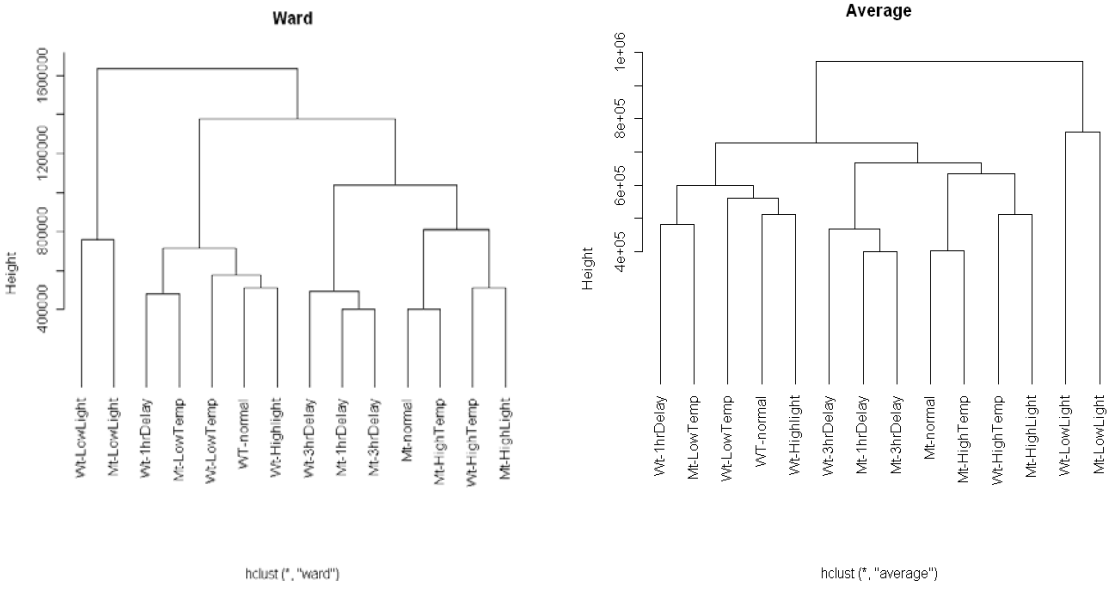

Figure 2-2 Hierarchical Cluster Analysis (HCA) example

Principle Component Analysis (PCA)

PCA transforms the high dimensional data into lower dimensions by finding the linear combinations of the original variables that maximize the variance within the data. The PCs are orthogonal and are ordered according to the variance explained. Therefore, the first PC explains the maximum variance. If the variance in the data reflects the true biological difference then plotting first PC against the second can be used to visualize the separation in the different classes. The original variables that contribute the most to the first few PCs are considered to be the most important. For example, PCA was used to analyze and normalize NMR data , which was followed by supervised discriminant function analysis using a priori information based on spectral replicates (Raamsdonk L. et al. 2001). Figure 2-3 shows the first two principal components of the data from 2010 Arabidopsis Project (www.plantmetabolomics.org) courtesy Dr. Oliver Fiehn's lab at UC Davis. Wild type samples under 7 different environmental conditions are compared against SALK_021108 mutant under the same 7 conditions. Six replicates in each condition were used. The wild type genotype is depicted by letter W and the mutant is depicted by letter M in the figure. We see that the first two PCs were able to separate the wild type samples and the mutant samples pretty well, but there are still some data points that are not separable which may be due to the nonlinear nature of the biological data. A scree plot shown in the figure 6 is generally used to determine the number of PCs. Figure 2-4 shows Scree plot of the PCA analysis from Figure 2-3. This plot shows that the first PC explains more than 30% variance in the data, second PC explains about 29% variance in the data and a combination of the first 5 PCs explain more than 90% of the variance in the data.

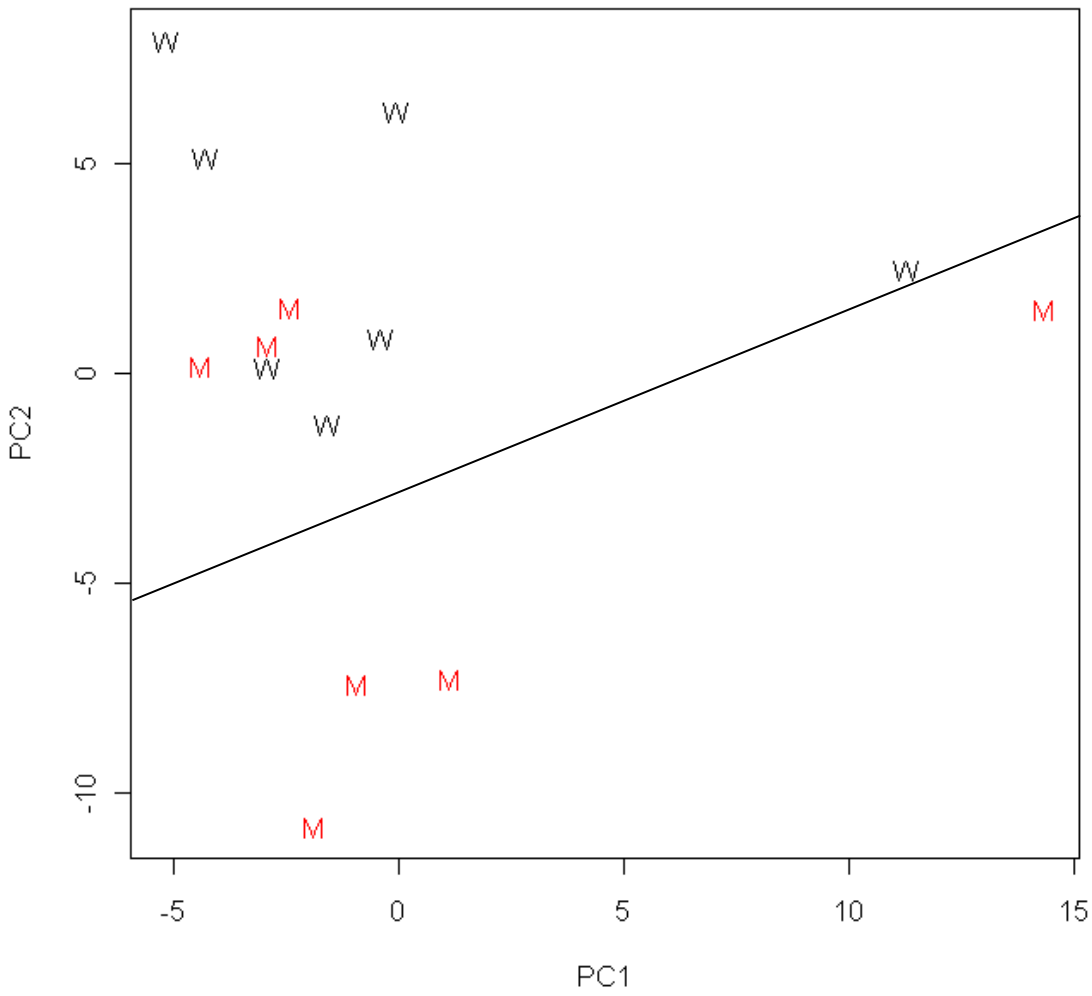

Figure 2-3 **PCA Example**

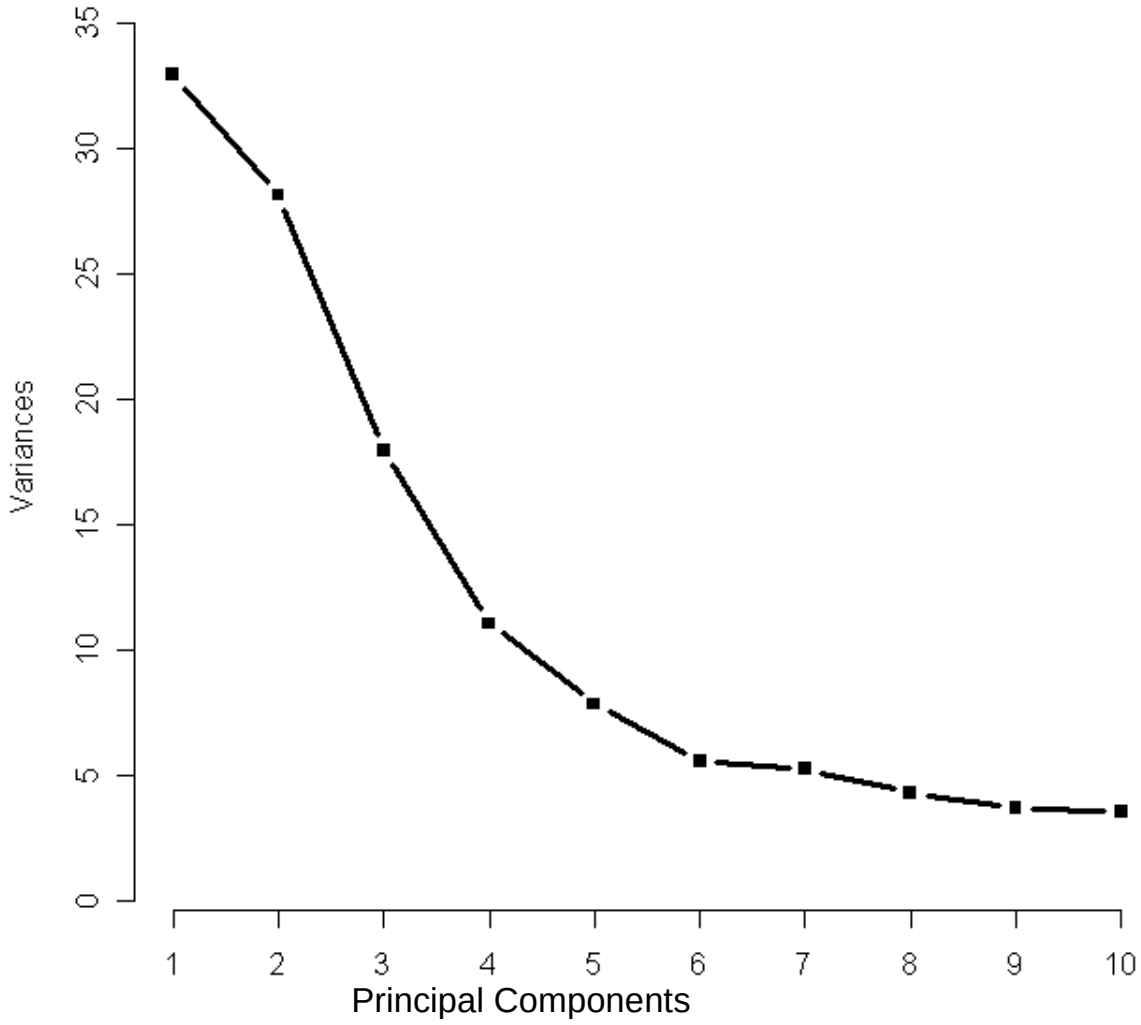

Figure 2-4 **Scree Plot Example**

Classification

PCA and clustering can visualize the separation of the samples according to the treatment factors (classes) but supervised methods are more powerful and can validate the separation numerically. The data is divided into 3 sets which are called training, test and validation sets. First the algorithm is trained using the training data and then class labels are predicted for the test data set. Some of the supervised learning methods are as following: *Support Vector Machine (SVM)* which tries to find the best hyper plane that maximizes the margin of separation between two classes. Decision tree (DT) algorithm branches the data and produces decision boundaries allowing the discovery of which metabolites are important. CART (classification and

regression trees) and C4.5/C5 are the most popular decision tree algorithms. Other popular classification algorithms are *artificial neural networks (ANN)* and Random Forests (RF). ANN performs a good classification but does not explain the model very well, Tree based algorithms provide the rules but do not perform as well.

Pathway Analysis

- ***Pathway databases:*** The first step in pathway reconstruction is comparison with reference pathways like Kyoto Encyclopedia of Genes and Genomes (KEGG)(Kanehisa M.et al. 2004), the ERGO system (Overbeek R. et al. 2003), BioCarta (BioCarta 2009), PathDB and the Roche Applied Science Biochemical Pathways chart which are available in digital form at ExPASy biochemical pathways page (Gasteiger E. et al. 2003). AraCyc (Zhang P. et al. 2005) and MetNetDB (Wurtele E. et al. 2007) are the most comprehensive databases to visualize biochemical pathways for Arabidopsis plants and has recently been expanded to include versions for all plant species (PlantCyc) which have sequenced genomes at the Plant Metabolic Network website (PMN_Team 2009). The software allows querying and the graphical representation of biochemical pathways and expression data MetaCrop (Grafahrend-Belau E. et al. 2008) is a hand curated database that summarizes diverse information about metabolic pathways in crop plants and allows automatic export of information for the creation of detailed metabolic models.
- ***Metabolic Pathway modeling:*** The BRENDA database provides enzyme kinetics and substrate specificity database(Schomburg I. et al. 2004).

References

Aharoni, A., C. H. Ric de Vos, et al. (2002). "Nontargeted metabolome analysis by use of Fourier Transform Ion Cyclotron Mass Spectrometry." Omics **6**(3): 217-34.

Akiyama, K., E. Chikayama, et al. (2008). "PRIMe: a Web site that assembles tools for metabolomics and transcriptomics." In Silico Biol **8**(3-4): 339-45.

Baxter, C. J., H. Redestig, et al. (2007). "The Metabolic Response of Heterotrophic Arabidopsis Cells to Oxidative Stress." Plant Physiol. **143**(1): 312-325.

Beuerle, T. and W. Schwab (1999). "Metabolic profile of linoleic acid in stored apples: formation of 13\(R)-hydroxy-9(Z),11(E)-octadecadienoic acid." Lipids **34**(4): 375-80.

Bino, R. J., C. H. R. de Vos, et al. (2005). "The light-hyperresponsive high pigment-2(dg) mutation of tomato: alterations in the fruit metabolome." New Phytologist **166**(2): 427-438.

Bino, R. J., R. D. Hall, et al. (2004). "Potential of metabolomics as a functional genomics tool." Trends Plant Sci **9**(9): 418-25.

BioCarta. (2009, March 3, 2009). "BioCarta Pathways: interactive graphic models of molecular and cellular pathways." Retrieved March 3, 2009, 2009, from http://www.biocarta.com/genes/index.asp.

Carrari, F., C. Baxter, et al. (2006). "Integrated Analysis of Metabolite and Transcript Levels Reveals the Metabolic Shifts That Underlie Tomato Fruit Development and Highlight Regulatory Aspects of Metabolic Network Behavior." Plant Physiol. **142**(4): 1380-1396.

Catchpole, G. S., M. Beckmann, et al. (2005). "Hierarchical metabolomics demonstrates substantial compositional similarity between genetically modified and conventional potato crops." Proc Natl Acad Sci U S A **102**(40): 14458-62.

Choi, H.-K., Y. H. Choi, et al. (2004). "Metabolic fingerprinting of wild type and transgenic tobacco plants by 1H NMR and multivariate analysis technique." Phytochemistry **65**(7): 857-864.

Choi, Y. H., H. K. Kim, et al. (2004). "Metabolomic Differentiation of Cannabis sativa Cultivars Using 1H NMR Spectroscopy and Principal Component Analysis." Journal of Natural Products **67**(6): 953-957.

Cui, Q., I. A. Lewis, et al. (2008). "Metabolite identification via the Madison Metabolomics Consortium Database." Nat Biotechnol **26**(2): 162-4.

Duran, A. L., J. Yang, et al. (2003). "Metabolomics spectral formatting, alignment and conversion tools (MSFACTs)." Bioinformatics **19**(17): 2283-2293.

Fatma Kaplan, et al. (2007). "Transcript and metabolite profiling during cold acclimation of Arabidopsis reveals an intricate relationship of cold-regulated gene expression with modifications in metabolite content." The Plant Journal **50**(6): 967-981.

Fiehn, O. (2002). "Metabolomics--the link between genotypes and phenotypes." Plant Mol Biol **48**(1-2): 155-71.

Fiehn, O., J. Kopka, et al. (2000). "Metabolite profiling for plant functional genomics." Nat Biotechnol **18**(11): 1157-61.

Fukusaki, E., T. Ikeda, et al. (2003). "A facile transformation of Arabidopsis thaliana using ceramic supported propagation system." J Biosci Bioeng **96**(5): 503-5.

Gasteiger, E., A. Gattiker, et al. (2003). "ExPASy: the proteomics server for in-depth protein knowledge and analysis." Nucleic Acids Research **31**: 3784-3788.

Goto, S., Y. Okuno, et al. (2002). "LIGAND: database of chemical compounds and reactions in biological pathways." Nucleic Acids Res **30**(1): 402-4.

Grafahrend-Belau, E., S. Weise, et al. (2008). "MetaCrop: a detailed database of crop plant metabolism." Nucleic Acids Res **36**(Database issue): D954-8.

Guterman, I., M. Shalit, et al. (2002). "Rose scent: genomics approach to discovering novel floral fragrance-related genes." Plant Cell **14**(10): 2325-38.

Hagel, J. and P. Facchini (2008). "Plant metabolomics: analytical platforms and integration with functional genomics." Phytochemistry Reviews **7**(3): 479-497.

Hall, R., M. Beale, et al. (2002). "Plant metabolomics: the missing link in functional genomics strategies." Plant Cell **14**(7): 1437-40.

Hochberg, Y. and Y. Benjamini (1990). "More powerful procedures for multiple significance testing." Stat Med **9**(7): 811-8.

Jenkins, H., N. Hardy, et al. (2004). "A proposed framework for the description of plant metabolomics experiments and their results." Nat Biotechnol **22**(12): 1601-6.

Jonsson, P., A. I. Johansson, et al. (2005). "High-throughput data analysis for detecting and identifying differences between samples in GC/MS-based metabolomic analyses." Anal Chem **77**(17): 5635-42.

Kanehisa, M., S. Goto, et al. (2004). "The KEGG resource for deciphering the genome." Nucleic Acids Res **32**(Database issue): D277-80.

Katja Dettmer, P. A. A. B. D. H. (2007). "Mass spectrometry-based metabolomics." Mass Spectrometry Reviews **26**(1): 51-78.

Kopka, J. (2006). "Current challenges and developments in GC-MS based metabolite profiling technology." Journal of Biotechnology **124**(1): 312-322.

Kopka, J., N. Schauer, et al. (2005). "GMD@CSB.DB: the Golm Metabolome Database." Bioinformatics **21**(8): 1635-1638.

Korcok, M. (1985). "NCI offering computer database on cancer research." Can Med Assoc J **133**(3): 225-7.

Krishnan, P., N. J. Kruger, et al. (2005). "Metabolite fingerprinting and profiling in plants using NMR." J. Exp. Bot. **56**(410): 255-265.

Luedemann, A., K. Strassburg, et al. (2008). "TagFinder for the quantitative analysis of gas chromatography--mass spectrometry (GC-MS)-based metabolite profiling experiments." Bioinformatics **24**(5): 732-737.

Maharjan, R. P. and T. Ferenci (2003). "Global metabolite analysis: the influence of extraction methodology on metabolome profiles of Escherichia coli." Anal Biochem **313**(1): 145-54.

Manetti, C., C. Bianchetti, et al. (2004). "NMR-based metabonomic study of transgenic maize." Phytochemistry **65**(24): 3187-3198.

Mattoo, A. K., A. P. Sobolev, et al. (2006). "Nuclear Magnetic Resonance Spectroscopy-Based Metabolite Profiling of Transgenic Tomato Fruit Engineered to Accumulate Spermidine and Spermine Reveals Enhanced Anabolic and Nitrogen-Carbon Interactions." Plant Physiol. **142**(4): 1759-1770.

Messerli, G., V. Partovi Nia, et al. (2007). "Rapid Classification of Phenotypic Mutants of Arabidopsis via Metabolite Fingerprinting." Plant Physiol. **143**(4): 1484-1492.

Moco, S., R. J. Bino, et al. (2006). "A Liquid Chromatography-Mass Spectrometry-Based Metabolome Database for Tomato." Plant Physiol **141**(4): 1205-1218.

Moing, A., M. Maucourt, et al. (2004). "Quantitative metabolic profiling by 1-dimensional 1H-NMR analyses: application to plant genetics and functional genomics." Functional Plant Biology **31**(9): 889-902.

Mueller, L. A., P. Zhang, et al. (2003). "AraCyc: A Biochemical Pathway Database for Arabidopsis." Plant Physiol. **132**(2): 453–460.

Oksman-Caldentey, K. M. and D. Inze (2004). "Plant cell factories in the post-genomic era: new ways to produce designer secondary metabolites." Trends Plant Sci **9**(9): 433-40.

Overbeek, R., N. Larsen, et al. (2003). "The ERGO genome analysis and discovery system." Nucleic Acids Res **31**(1): 164-71.

Pan, Z. and D. Raftery (2007). "Comparing and combining NMR spectroscopy and mass spectrometry in metabolomics." Analytical and Bioanalytical Chemistry **387**(2): 525-527.

PMN_Team. (2009). "Plant Metabolic Network (PMN)." Retrieved May 28, 2009, from http://www.plantcyc.org.

Raamsdonk, L. M., B. Teusink, et al. (2001). "A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations." Nat Biotechnol **19**(1): 45-50.

Ratcliffe, R. G. and J. A. Callow (1994). In vivo NMR Studies of Higher Plants and Algae. Advances in Botanical Research, Academic Press. **Volume 20:** 43-123.

Ratcliffe, R. G. and Y. Shachar-Hill (2001). "PROBING PLANT METABOLISM WITH NMR." Annual Review of Plant Physiology and Plant Molecular Biology **52**(1): 499-526.

Robinson, M., D. De Souza, et al. (2007). "A dynamic programming approach for the alignment of signal peaks in multiple gas chromatography-mass spectrometry experiments." BMC Bioinformatics **8**(1): 419.

Roessner, U., J. H. Patterson, et al. (2006). "An investigation of boron toxicity in barley using metabolomics." Plant Physiol **142**(3): 1087-101.

Roessner, U., C. Wagner, et al. (2000). "Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry." Plant J **23**(1): 131-42.

Sansone, S. A., T. Fan, et al. (2007). "The metabolomics standards initiative." Nat Biotechnol **25**(8): 846-8.

Schauer, N. and A. R. Fernie (2006). "Plant metabolomics: towards biological function and mechanism." Trends in Plant Science **11**(10): 508-516.

Schauer, N., D. Zamir, et al. (2005). "Metabolic profiling of leaves and fruit of wild species tomato: a survey of the Solanum lycopersicum complex." J Exp Bot **56**(410): 297-307.

Scholz, M. and O. Fiehn (2007). "SetupX--a public study design database for metabolomic projects." Pac Symp Biocomput: 169-80.

Schomburg, I., A. Chang, et al. (2004). "BRENDA, the enzyme database: updates and major new developments." Nucleic Acids Res **32**(Database issue): D431-3.

Shigeru Sato, T. S. T. N. M. T. (2004). "Simultaneous determination of the main metabolites in rice leaves using capillary electrophoresis mass spectrometry and capillary electrophoresis diode array detection." The Plant Journal **40**(1): 151-163.

Smith, C. A., G. O'Maille, et al. (2005). "METLIN: a metabolite mass spectral database." Ther Drug Monit **27**(6): 747-51.

Smith, C. A., E. J. Want, et al. (2006). "XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification." Anal. Chem. **78**(3): 779-787.

Soga, T. and M. Imaizumi (2001). "Capillary electrophoresis method for the analysis of inorganic anions, organic acids, amino acids, nucleotides, carbohydrates and other anionic compounds." Electrophoresis **22**(16): 3418-25.

Soga, T., Y. Ohashi, et al. (2003). "Quantitative Metabolome Analysis Using Capillary Electrophoresis Mass Spectrometry." Journal of Proteome Research **2**(5): 488-494.

Stein, S. E. (1999). "An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data." Journal of the American Society for Mass Spectrometry **10**(8): 770-781.

Stephen C. Davis, A. A. M. J. D. H. (1999). "Ultrafast gas chromatography using time-of-flight mass spectrometry." Rapid Communications in Mass Spectrometry **13**(4): 237-241.

Styczynski, M. P., J. F. Moxley, et al. (2007). "Systematic Identification of Conserved Metabolites in GC/MS Data for Metabolomics and Biomarker Discovery." Analytical Chemistry **79**(3): 966-973.

Sumner, L. W., P. Mendes, et al. (2003). "Plant metabolomics: large-scale phytochemistry in the functional genomics era." Phytochemistry **62**(6): 817-836.

Tikunov, Y., A. Lommen, et al. (2005). "A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles." Plant Physiol **139**(3): 1125-37.

Tjalsma, H. (2007). "Feature-based reappraisal of the Bacillus subtilis exoproteome." Proteomics **7**(1): 73-81.

van den Berg, R. A., H. C. Hoefsloot, et al. (2006). "Centering, scaling, and transformations: improving the biological information content of metabolomics data." BMC Genomics 7: 142.

Verpoorte, R., Y. Choi, et al. (2007). "NMR-based metabolomics at work in phytochemistry." Phytochemistry Reviews **6**(1): 3-14.

von Roepenack-Lahaye, E., T. Degenkolb, et al. (2004). "Profiling of Arabidopsis Secondary Metabolites by Capillary Liquid Chromatography Coupled to Electrospray Ionization Quadrupole Time-of-Flight Mass Spectrometry." Plant Physiol. **134**(2): 548-559.

Vorst, O., C. H. R. d. Vos, et al. (2005). "A non-directed approach to the differential analysis of multiple LC–MS-derived metabolic profiles." Metabolomics **1**(2): 169-180.

Wishart, D. S., D. Tzur, et al. (2007). "HMDB: the Human Metabolome Database." Nucl. Acids Res. **35**(suppl_1): D521-526.

Wurtele, E., L. Li, et al. (2007). MetNet: Systems Biology Software for Arabidopsis. Concepts in Plant Metabolomics. W. E. Nikolau BJ, Springer**:** 145-158.

Zhang, P., H. Foerster, et al. (2005). "MetaCyc and AraCyc. Metabolic Pathway Databases for Plant Research." Plant Physiology **138**: 27-37.

CHAPTER 3. PLANTMETABOLOMICS.ORG: A WEB PORTAL FOR PLANT METABOLOMICS EXPERIMENTS

Published in Plant Physiology February 2010

Preeti Bais, Stephanie M. Moon, Kun He, Ricardo Leitao, Kate Dreher, Tom Walk, Yves Sucaet, Lenore Barkan, Gert Wohlgemuth, Mary R. Roth, Eve Syrkin Wurtele, Philip Dixon, Oliver Fiehn, B. Markus Lange, Vladimir Shulaev, Lloyd W. Sumner, Ruth Welti, Basil J. Nikolau, Seung Y. Rhee, and Julie A. Dickerson

Abstract:

PlantMetabolomics.org (PM) is a web portal and database for exploring, visualizing and downloading plant metabolomics data. Widespread public access to well-annotated metabolomics datasets is essential for establishing metabolomics as a functional genomics tool. PM integrates metabolomics data generated from different analytical platforms from multiple laboratories along with the key visualization tools such as ratio and error plots. Visualization tools can quickly show how one condition compares to another and which analytical platforms show the largest changes. The database tries to capture a complete annotation of the experiment metadata along with the metabolite abundance data based on the evolving Metabolomics Standards Initiative (MSI). PM can be used as a platform for deriving hypotheses by enabling metabolomic comparisons between genetically unique *Arabidopsis thaliana* populations subjected to different environmental conditions. Each metabolite is linked to relevant experimental data and information from various annotation databases. The portal also provides detailed protocols and tutorials on conducting plant metabolomics experiments to promote metabolomics in the community. PM currently houses *Arabidopsis* metabolomics data generated by a consortium of laboratories utilizing metabolomics to help elucidate the functions of uncharacterized genes. PM is publicly available at http://www.plantmetabolomics.org.

Introduction

In the post genomics era, metabolomics is fast emerging as a vital source of information to aid in solving systems biology puzzles with an emphasis on metabolic solutions. Metabolomics is the science of measuring the pool sizes of metabolites (small molecules of molecular weight ≤1,000 Da), which collectively define the metabolome of a biological sample (Fiehn et al., 2000; Hall et al., 2002). Coverage of the entire plant metabolome is a daunting task as it is estimated that there are over 200,000 different metabolites within the plant kingdom (Goodacre et al., 2004). Although technology is rapidly advancing, there are still large gaps in our knowledge of the plant metabolome.

Despite this lack of complete knowledge and the immense metabolic diversity among plants, metabolomics has become a key analytical tool in the plant community (Hall et al., 2002). This has led to the emergence of multiple experimental and analytical platforms that collectively generate millions of metabolite data points. Because of this vast amount of data, the development of public databases to capture information from metabolomics experiments is vital to provide the scientific community with comprehensive knowledge about metabolite data generation, annotation, and integration with metabolic pathway data. Some examples of these public databases are given below. The Human Metabolome Project contains comprehensive data for more than 2000 metabolites found within the human body (Wishart et al., 2007). The Golm Database is a repository that provides access to mass-spectrometry (MS) libraries, metabolite profiling experiments, and related information from GC-MS (gas chromatography-mass spectrometry) experimental platforms, along with tools to integrate this information with other systems biology knowledge (Kopka et al., 2005). The Madison Metabolomics Consortium Database contains primarily NMR spectra for *Arabidopsis* and features thorough NMR search tools (Cui et al., 2008). SetupX and Binbase provide a framework that combines MS data and

biological metadata for steering laboratory work flows and employs automated metabolite annotation (Scholz and Fiehn, 2007).

A single analytical technique cannot identify and quantify all the metabolites found in plants. Thus, PlantMetabolomics.org (PM) was developed to provide a portal for accessing publicly-available MS-based plant metabolomics experimental results from multiple analytical and separation techniques. PM also follows the emerging metabolomics standards for experiment annotation. PM has extensive annotation links between the identified metabolites and metabolic pathways in AraCyc (Mueller et al., 2003) at The Arabidopsis Information Resource (TAIR); (Rhee et al., 2003) and the Plant Metabolic Network (PMN, www.plantcyc.org), the Kyoto Encyclopedia of Genes and Genomes (KEGG), (Kanehisa et al., 2004), and MetNetDB (Wurtele et al., 2007).

Standards for the annotation of metabolomics experiments are still under active development and the metadata types collected in PlantMetabolomics.org (PM) are based on the recommendations of the Metabolomics Standards Initiative (MSI) (Fiehn et al. 2007) and the Minimal Information for a Metabolomic Experiment, MIAMet (Bino et al., 2004) standards. MSI attempts to capture the complete annotation of metabolomics experiments and includes metadata of the experiments along with the metabolite abundance data. The initial database schema design was guided by the schema proposed in the Architecture for Metabolomics (ArMet) project (Jenkins et al., 2004).

Development of Plantmetabolomics.org: Rationale

The rationale for the development of PM as an information portal is to provide free public access to experimental data along with cross-references to related genetic, chemical and pathway information. The portal also serves as an information resource for the field of metabolomics by

providing tutorials on how to conduct metabolomics experiments. It describes minimum reporting standards (Fiehn, 2007, 2007; Sumner, 2007) for plant metabolomics experiments based on the recommendations of the Metabolomics Standards Initiative (MSI). In addition, PM contains background information about the experimental design and tools that can be used to analyze the collected data (Helsel, 2005). To our knowledge, PM is the only plant metabolomics database that contains data from *Arabidopsis* metabolomics experiments that utilize multiple analytical detectors combined with different separation technologies. These include gas chromatography - mass spectrometry (GC-MS), gas chromatography-time-of-flight mass spectrometry (GC-TOF-MS), capillary electrophoresis-mass spectrometry (CE-MS), ultra high pressure liquid chromatography coupled to a hybrid quadrupole time-of-flight mass spectrometer (UPLC-Q-TOF-MS) and liquid chromatography-mass spectrometry (LC-MS) (Dunn and Ellis, 2005). The statistical analysis and visualization tools are easy to use and aid non-statisticians in the analysis of the effects of different environmental conditions, genetic perturbations and other experimental factors. The information collected within PM can be used to form hypotheses about the roles of genes of unknown function in *Arabidopsis* by comparing the metabolome of a wild-type sample to that of a sample altered by a mutation at a target gene which can provide clues as to the function of that gene. The data (both biological and metabolic) and tools contained within PM, all available to the scientific community, are detailed in this paper.

Design Requirements and Functionality

PM allows users to explore and interpret data sets and put them in a biological context. This requires the integration of relative metabolite abundance along with the metadata of the experimental conditions including growth, harvest and storage conditions of sample tissue, sample extraction, and instrument parameters. We also place an emphasis on ensuring ease of use and providing additional information about each identified metabolite by linking to other data

sources such as AraCyc, KEGG, MetNetDB and PubChem (Figure 3-1). The Metabolomics Standards Initiative (MSI) specifies the minimum amount of metadata from the metabolomics experiments that must be reported so that experiments can be replicated and results can be verified. These minimum data include descriptions of biological study design, sample preparation, data acquisition, data processing, and data analysis procedures. One goal of PM is to fulfill the outlined recommendations by the MSI. Data contributors are required to use the standard data submission spreadsheet templates (available through the portal) to submit metabolomics data. These sheets follow ArMet and MIAMet specifications to capture the metadata of an experiment. PM also includes educational video tutorials to aid metabolomics researchers in quality control.

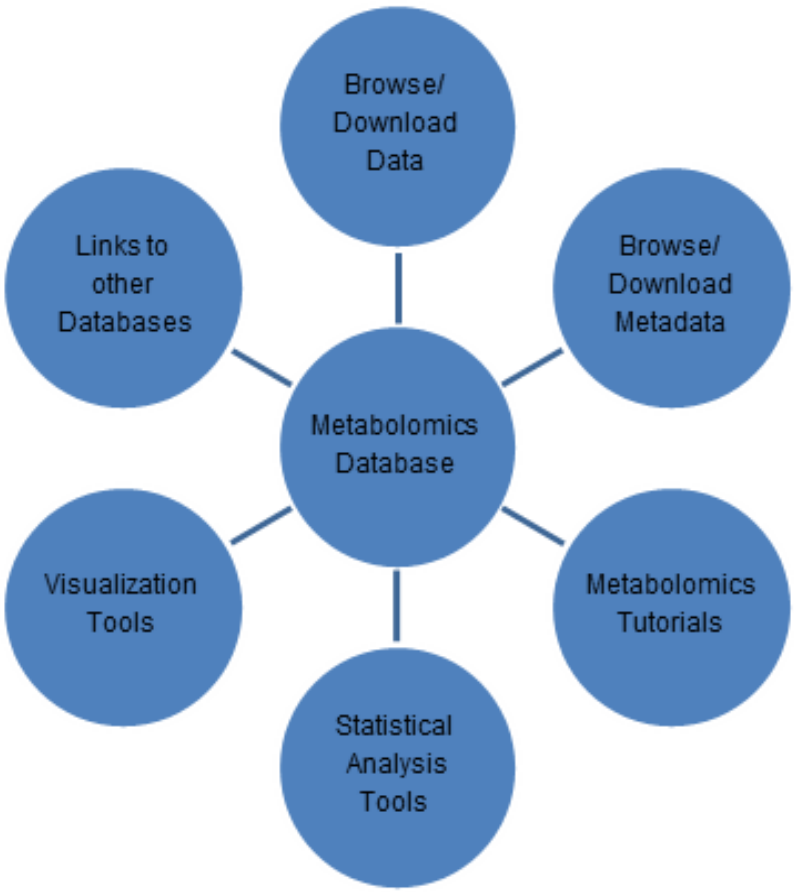

Figure 3-1 **Diagram of the main components of the plantmetabolomics.org portal**

PlantMetabolomics Content

PlantMetabolomics in its current state houses the metabolite data generated from plant metabolomics experiments performed under the *Arabidopsis* 2010 program funded by the NSF. A consortium of metabolomics and metabolite profiling laboratories, in partnership with biochemists, biostatisticians and bioinformaticists, generated the data to formulate hypotheses about *Arabidopsis* genes of unknown function. The consortium employed a strategy of generating *Arabidopsis* material at a single location followed by distribution to all analytical laboratories. Different extraction processes and analytical techniques were used among the laboratories; ultimately providing analysis of roughly 1800 metabolites in each of the experiments conducted, around 400 of which are chemically defined. In total, among all experiments stored in the current database, roughly 3100 compounds were detected, including 952 chemically defined compounds. A total of 579 of the known compounds have been identified in AraCyc and many of them participate in metabolic pathways described in that database. The metabolite data, along with the metadata, as generated by this consortium, are stored in the PM database.

Experiment Annotation

The pipelines used by the analytical laboratories in this consortium are captured through metadata for distribution via the database. Each step of the process requires the collection of information to provide users with an understanding of collection, distribution and extraction of sample material, along with the instrumentation setup and data processing (Figure 3-2). Experimental metadata provides information about the growth and harvesting regimen, including the temperature, illumination, duration of growth, humidity and storage parameters, which were used to produce the *Arabidopsis* tissue samples for analysis. This metadata also includes information that pertains to the genetic background of the samples.

The analytical metadata is collected in three sections: extraction, chromatography, and mass spectrometry metadata. Each section includes details about the tissue sample extraction process and the instrumentation models, settings and parameters used for the chromatogram and spectrometer for each analytical platform.

Each laboratory individually processes the metabolomics data obtained from the specific platform used. Metabolite identification is based on procedures developed in each individual laboratory that utilize comparisons of retention time, retention index and mass fragmentation patterns compared to those of authentic standard compounds (when available) included in both private and public mass spectrometry libraries. Metabolite peaks that cannot be chemically annotated are given a unique identifier (Bino et al., 2004). The raw data is processed and normalized based on each laboratory's instrument detection limit and analytical procedure. Specific processing procedures for each platform are available on the database on the protocols page. The processed and normalized data from each laboratory has been collected for each experiment and loaded into the database. The public can access the data online or download it for use in other applications.

Design of experiments:

The plant metabolome responds to both environmental (E) and genetic (G) factors during plant growth. Environmental parameters such as temperature, light intensity, growing medium, humidity and all other abiotic and biotic factors that affect plant growth and development are defined and stored. The genetic parameter is defined by the integrated expression of the alleles that encompass the organism's genome. The design of experiments conducted by the above consortium took both the genetic and environmental parameters into consideration. Genetic parameters were manipulated by using *Arabidopsis* stocks that contained T-DNA insertions in either a gene of known function (GKF) or gene of unknown function (GUF) obtained from the *Arabidopsis Biological Resource Center* (ABRC; Columbus, OH). The stocks were selected

based on availability and current knowledge and on gene predictions from sequence analysis and association networks (He, Lee, Walk and Rhee, manuscript in preparation). All mutant stocks were visibly screened for phenotypes that resembled wild-type seedlings. Pictures of each mutant line at 17 days after sowing are available within PM.

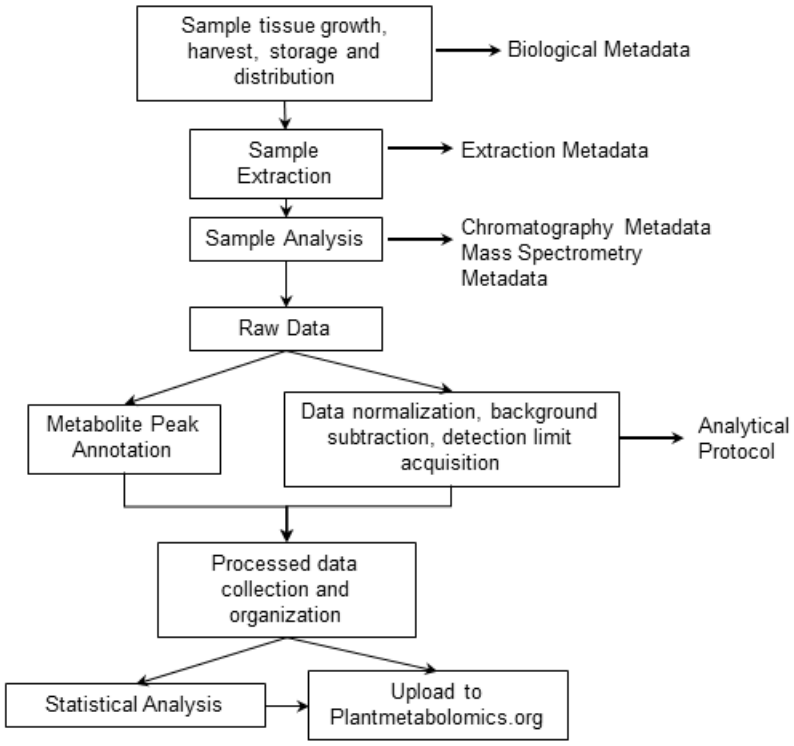

Figure 3-2 Schematic representation of the process used in generation of metabolite data

PM currently contains metabolomics data from two types of experimental designs that varied the G*E interactions (Table 3-1). The first setup used a combination of G*E variations, where the genetic parameter was comprised of two different genotypes (wild-type and one mutant stock) and the environmental parameter was changed in a single abiotic or biotic manner. The experiments that fall into this first category provide information on the overall effect that both the environment and genetic parameters have on the metabolome of *Arabidopsis* seedlings. The

second setup, which encompasses a large percentage of the data, varied only the genetic parameter and kept the environmental conditions constant during the growth period and across all experiments. Each experiment contained wild-type seedlings along with 8-15 seedlings representing *Arabidopsis* stocks carrying T-DNA mutant alleles. Holding the environment constant and varying the genetic parameter provides metabolomic data that is a consequence of the genetic change and can therefore provide information about the consequence of mutating a specific gene within each stock.

Experimental Data

Metabolite abundance data can be downloaded along with the metadata for each experiment contained in PM. There are three options for downloading metabolite abundance data. The first option allows the users to select and download data from specific experiments and/or analytical platforms. The downloaded file contains the user-chosen results compiled in a comma separated values (csv) format. This option also allows users to download the correlation coefficients between the various replicates along with the data. Once they download the csv file they can look at the correlation coefficients and determine if they want to exclude part of the data if the correlation among the replicates is low. PM does not exclude any data based on the data quality issues but equips the users with the analysis tools and measurements so that the users can make informed decisions. The second option allows users to download the compiled Excel workbooks for individual experiments that contain a single sheet for each analytical platform. The workbooks contain the original data as submitted by the respective labs. This option provides the data in an easy to use format that can be manipulated by the user for their own analyses. The third option comes in the form of a compressed file (.zip) that provides partially processed data for each of the mutant experiments. This download includes: 1. The further scaled metabolite abundance data which reduces the variation within biological

replicates; 2. Scatter plots and correlation coefficient values between biological replicate data that provide additional information about the consistency of the replicates; 3. All the metabolites with false discovery rate (FDR) adjusted t-test p-value and ratios between mutant and wild type, so that users can define the significantly altered metabolites by combining the p-value and ratio (fold-change); 4. MVA-plots that show changes in metabolite abundance by displaying the log concentration ratio vs. the average log concentration across replicates. Detailed information about the data processing can be found in the readme.doc file included in the datasets.

Table 3-1 Experimental set-ups used to generate metabolomics data contained in PM

	Factors Varied		
Experiment Name	Environmental (E)	Genetic (G)	Total (GxE)
EIE2	7	2	14
Fatb Induction	2	2	4
Elo1 Induction	3	2	6
ME1	1	9	9
ME2	1	11	11
ME3	1	16	16
ME4	1	14	14
ME5	1	11	11

.

Tutorials

The database contains tutorial information about the methodologies for the metabolomics studies developed by the consortium. These methodologies reflect metabolomics' utility in functional genomics and the current state of the technology. Metabolomics is not yet a widely utilized technology and it is thus important to train researchers in the methodologies, technologies, and standards in metabolite profiling. This ensures laboratory-to-laboratory reproducibility and facilitates meta-analyses across multiple experiments.

Three video tutorials demonstrate the methods used for tissue harvesting and distribution. The experimental metadata describes in detail the process used for harvesting tissue material and the "Tissue Harvest" video tutorial provides a visual guide for this process. This tutorial details the process used to open Petri dishes containing sample material, to collect the tissue and to immediately halt metabolism by submerging the tissue in liquid nitrogen. This process is completed within <2 minutes as seen by the elapsed time on the timer. Each laboratory requires specific amounts of tissue for each analytical platform. Collected samples must be weighed and sorted for shipment to the analytical laboratories. The two video tutorials "Sample weighing and Sorting of tissue samples" and "Sample Weighing (Closer View)" provide a visualization of this process. The three video tutorials provide an experimentalist with an additional tool to understand and repeat the process used to generate tissue samples required for metabolomics experiments.

The web portal also provides tutorials on how to browse, download and visualize the available data. These tutorials are provided as help buttons on all the main function pages as well as under the main "Tutorials" menu option. Many examples and screen shots of resulting pages are provided in the tutorials.

Data Analysis and Visualization of Experimental Data

The data analysis and visualization tools permit an analysis of data quality and hypothesis generation with dynamic graphs, which can be automatically generated with easy to use graphical user interfaces (GUIs).

Users may compare metabolite levels under different experimental conditions through the generation of dynamic ratio plots of the metabolites using the GUI. Users can choose any two experimental factors to compare and select specific analytical platforms to include in the analysis. The resulting ratio plot shows the abundance data (Figure 3-3). The x-axis shows the logarithm (base 2) ratio of the relative abundance of each metabolite between the mutant and wild type samples selected (see Materials and Methods). The metabolites that have a relatively low fold change between the two factors are close to the central vertical y-axis and the metabolites that have a relatively high-fold change are distant from the central vertical y-axis. The metabolites with one or more replicates with missing values are shown with different colored marks for quick inspection of data quality.

A summary of the metabolite abundance data is generated along with the ratio plot. This summary contains metabolite names that can be ordered according to the number of missing values (1, 2, 3 or more, or all null values). Detailed metabolite information is available by clicking on its name on the list or on the glyph on the ratio plot.

The error plot shows the change in the metabolite abundance level among the replicates. This helps the users to see if the significant change in the metabolite abundance is similar in replicate samples. The metabolite abundance data can also be visualized in a bar chart (Figure 3-4) where metabolite abundance under different experimental parameters is shown for each replicate.

The metabolite details page also provides links to other databases to give users access to more details about the metabolites. These links include metabolic pathway information from Aracyc and MetNetDB and compound information from PubChem, CAS, KEGG, and ChEBI. The names of all the pathways that contain a metabolite are shown on the metabolite annotation details page. Links to experimental data for all the other metabolites that participate in that pathway are also provided.

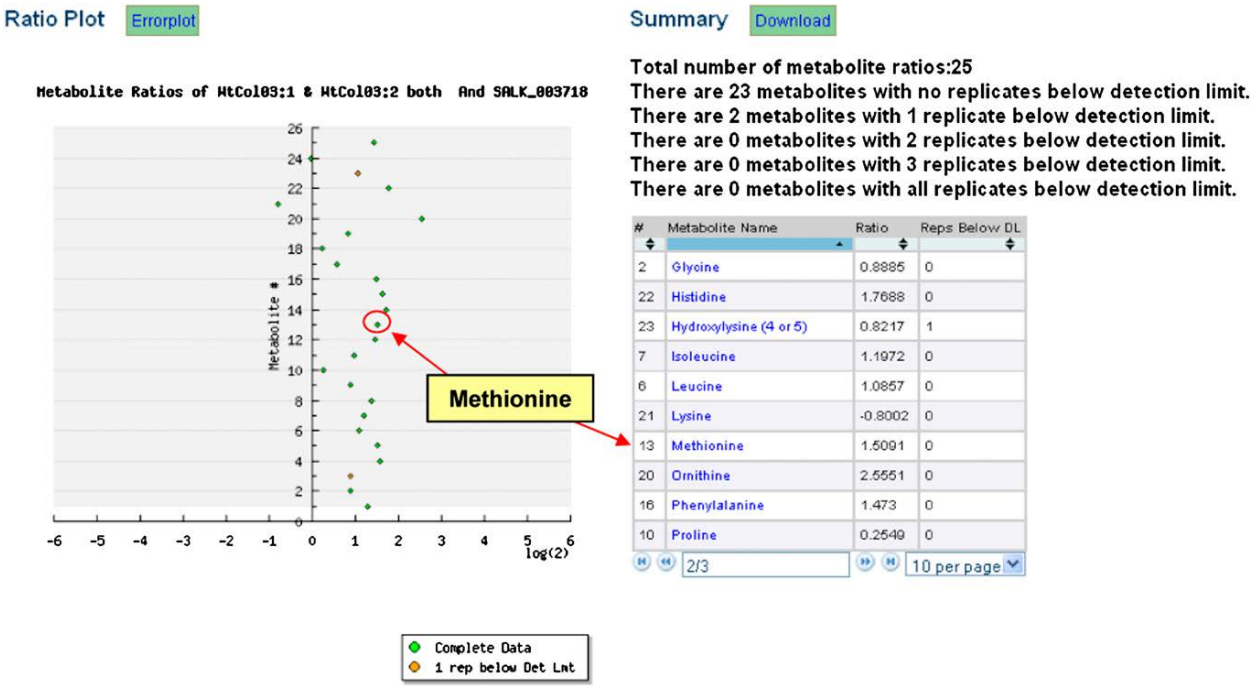

Figure 3-3 Ratio Plot

Compound Common Name:	L-methionine
Compound Synonyms:	M,met,L-methionine
Molecular Weight:	149.207
Chemical Formula:	(C 5)(H 11)(N 1)(O 2)(S 1)
Smiles Notation:	C(=O)(O)C(N)CCSC
INCHI:	InChI=1/C5H11NO2S/c1-9-3-2-4(6)5(7)8/h4H,2-3,6H2,1H3,(H,7,8)/t4-/m1/s1/f/h7H
CHEBI:	16643
CAS:	63-68-3
KEGG:	C00073
PUBCHEM:	84815 5255805
Aracyc Link:	MET
EC:	1.3.99.22 1.8.4.11 2.1.1.14 2.1.1.13 2.1.1.10 2.1.1.12 2.1.1.- 2.5.1.6
Aracyc Pathways:	lipoate biosynthesis and incorporation List SAM cycle List ethylene biosynthesis from methionine List adenosylmethionine biosynthesis List methionine biosynthesis List S-methylmethionine cycle List methionine degradation List tRNA charging pathway List
Metnet Pathways:	Search Metnet

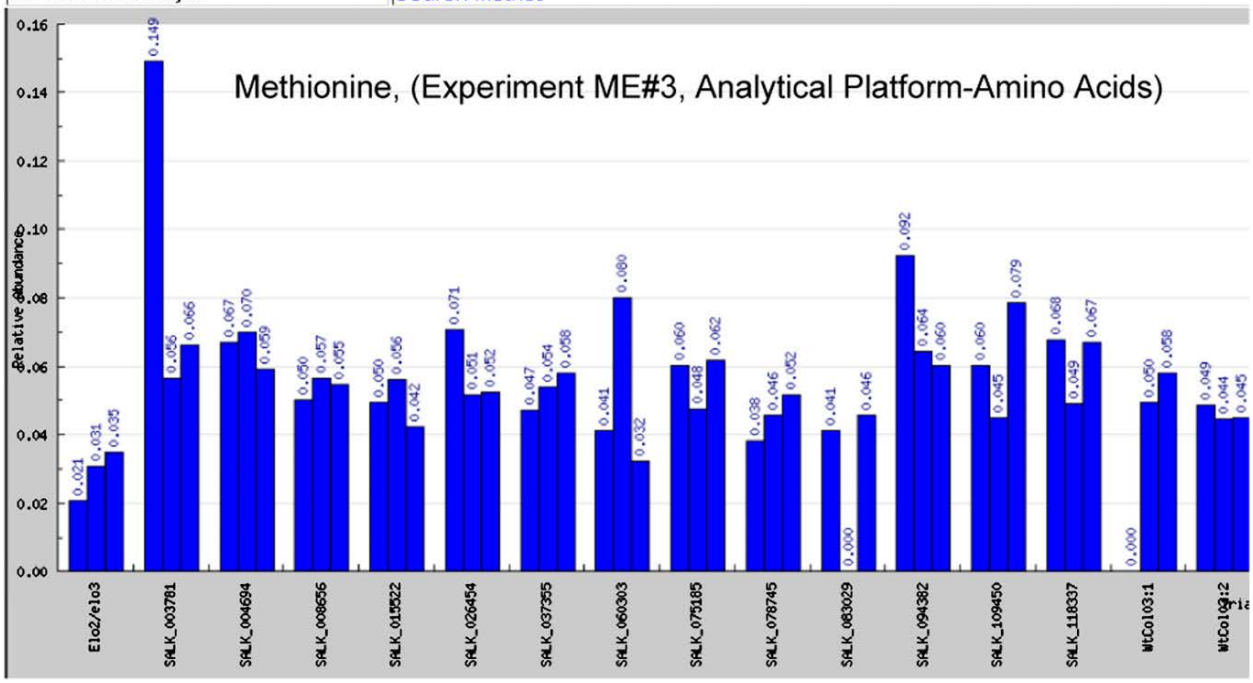

Figure 3-4 Metabolite details of Methionine

Data Quality Checks:

PM provides many options where a user can check the variability between different replicates (Figure 3-5) or see if some of the replicates are below the detection limit (Figure 3-3). The ultimate decision to exclude the data is left with the users. The data quality plots are provided along with the data. The ratio plot discussed in the previous section also provides instant access to replicates that are below the detection limit by showing them in different colors. A summary list provided with the ratio plots groups metabolites according to the number of missing values. The list can be ordered by the metabolite names to find if the same metabolite is detected by several platforms.

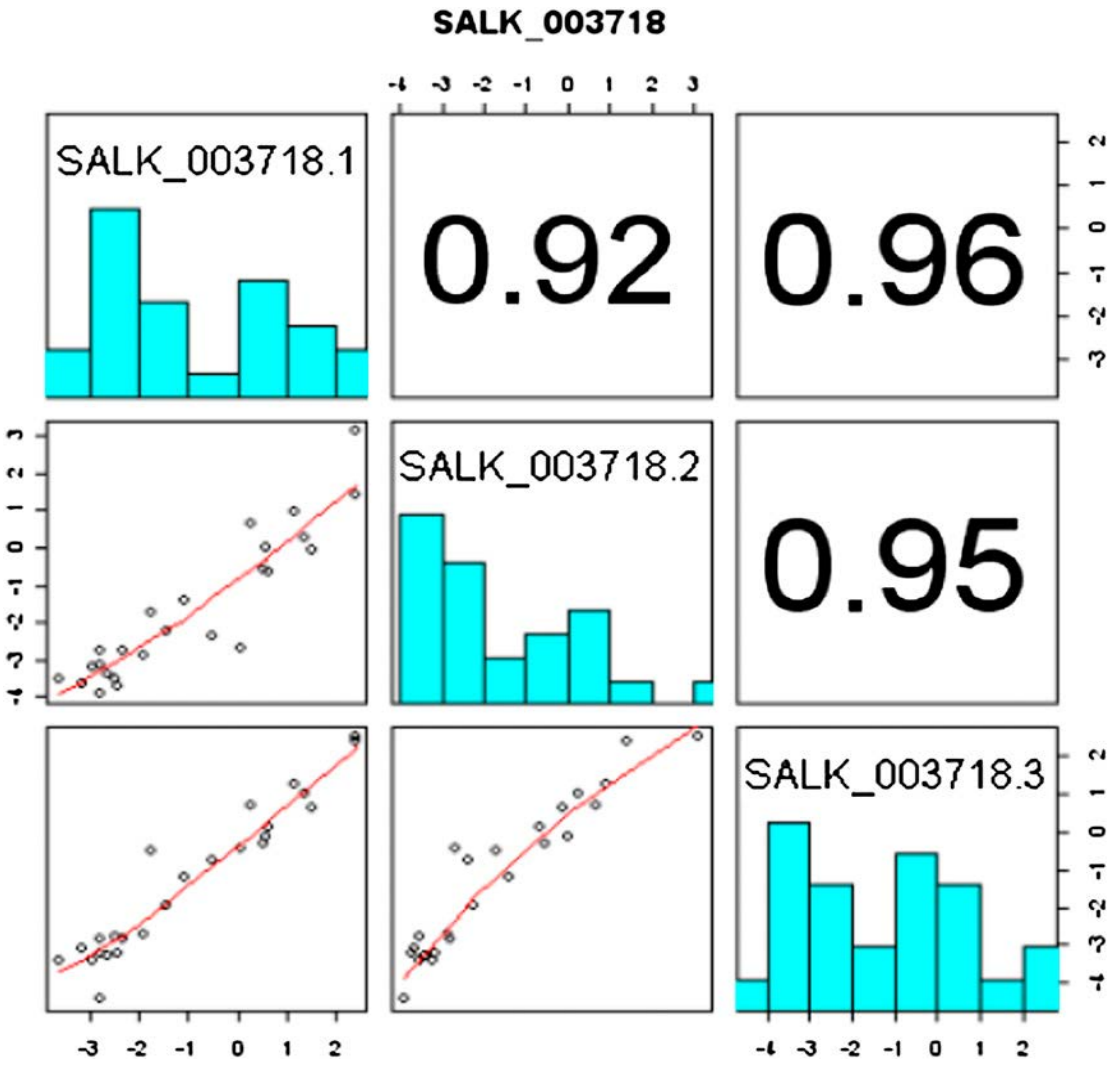

Figure 3-5 Replicate quality check

Query Capabilities

The database can be queried by individual metabolite or pathway names. This feature helps an investigator search for a particular metabolite across different experimental conditions. Once the metabolite is located in an experiment, the investigator can identify the pathways in AraCyc and MetNetDB in which this metabolite participates. The pathway search option finds all metabolites in the PM database that are part of the same pathway. The portal contains a local copy of AraCyc synonyms for metabolites along with the web links to AraCyc. This helps in

searching for metabolites by either the names by which they are stored in the PM database or any of the synonyms that can be resolved by AraCyc. The local copy is updated every 4 months.

Conclusions

One of the most important functions of any metabolomics database is to present collected data in a way that allows it to be used for comparison among different experiments and platforms. This requires that all metadata of standard operating procedures for sample preparation, data acquisition, and data processing be made available along with the corresponding results. Although there are some commercially available databases (Ridley et al., 2004), to our knowledge, PM is one of the first metabolomics databases available in the fundamental plant research arena. The database combines the results from many different platforms that were used in parallel to analyze the same biological material. At the end of the two-year pilot study, we have been able to provide data from 50 mutant lines and to capture baseline variations in metabolism in response to environmental condition variations during plant growth and tissue harvest. Web-based visualization tools in the portal make it easy for a non-statistician to do initial exploration of the data, perform quality checks and generate hypotheses. This platform not only provides the metabolomics data and the analysis tools, it also promotes the field of metabolomics by providing educational tutorials on performing the metabolomics experiments and implementing the MSI standards. We acknowledge that some of the data from the pilot project has low reproducibility between the replicates therefore the users are guided to carefully evaluate the data quality with easy to use visualization tools and tutorials so they can make educated decisions about exclusion of data from their analyses.

The metabolomics consortium expects to profile additional *Arabidopsis* mutant lines and upload the data to PM as it becomes available. We plan to enhance the resource by adding the derived spectral peak location, mass-spectra, and metabolite peak integration data as well as

make the actual chromatograms available for download in community accepted formats such as NetCDF and mzML. We plan to add more analysis and visualization tools to make this portal a better aid for generating hypotheses and promoting the field of metabolomics within the community. The web portal is also ready to accept MSI-compliant metabolomics data from other MS based metabolomics platforms for *Arabidopsis* and other plants.

Materials and Methods

Normalization and data processing

Metabolomics data generated are normalized and processed according to each specific laboratory's protocol. This process is detailed for each individual analytical platform and laboratory in the standard operating procedure protocols contained within PM.

Missing Values:

The detection limits for every run are typically experimentally determined by the corresponding labs and are reported along with the metabolite data. Missing values or below-detection limit measurements are replaced by ½ of the estimated detection limit if the detection limit is reported for that run; otherwise the missing values are replaced by ½ of the lowest value for that run (Helsel, 2005).

Ratio Plot

The x-axis ordinate is the logarithm (base 2) of the ratio of the relative abundance of each metabolite in the wild type vs. mutant plant:

$$x-axis = \log_2\left(\frac{\mu_{mt}}{\mu_{wt}}\right)$$

The values, μ_{mt} and μ_{wt}, calculated for each metabolite in each platform, are the sample means for the metabolite abundances of the replicates in the mutant and wildtype, respectively.

Error Plots

The standard error (SE) of the log-ratio was calculated using the delta method (or one-step Taylor-series) approximation,

$$SE = \frac{1}{\ln(2)} \sqrt{\left(\frac{SE_{mt}}{\mu_{mt}}\right)^2 + \left(\frac{SE_{wt}}{\mu_{wt}}\right)^2}$$

SE_{mt} and SE_{wt} are the standard errors of the average mutant and wild-type metabolite abundances calculated by: $SE = \sqrt{\frac{1}{N(N-1)} \sum_{i=1}^{N} (x_i - \mu_x)^2}$, where N is the number of replicates.

Compound Curation in AraCyc

The experimentally verified *Arabidopsis* compounds identified in the PM project were added into a broader metabolic pathway framework in AraCyc by first matching the names of PM compounds to existing compounds in the database to link the two resources. Named compounds that were not found in AraCyc were investigated using several resources such as MetaCyc (Caspi et al., 2008), SciFinder Scholar (Wagner, 2006), Chemical Entities of Biological Interest (Degtyarenko et al., 2008)(ChEBI, EMBL-EBI), PubChem (NCBI), and KEGG(Kanehisa et al., 2004) to find chemical structures and synonyms. These compounds were entered into AraCyc and linked to PM. Compound names that describe multiple structures that cannot be conclusively distinguished in the metabolomics experiments were entered as “classes”. These

contain the chemical formula of the identified compound, a text description, and, if possible, a partial structure using “R groups” to denote structural ambiguities. To place these compounds into the appropriate metabolic context, we searched the scientific literature and the databases used for compound identification. In addition, specific reactions between identified compounds were made based on generic reactions present in AraCyc.

Data Curation

Data is sent to the administrators using the sample spreadsheets. The spreadsheets are verified for format and then uploaded in the database by the administrators. The collaborators cannot upload the data themselves.

Supplemental Data

The following materials are available with this article.

Data Base Schema

The main structure and data organization of the PlantMetabolomics database are attached in Appendix–Supplementary Documents S1.

Website Map

The website map of PlantMetabolomics.org is attached in Appendix–Supplementary Documents S2.

How to use the web portal

A case study is provided in Appendix– Supplementary Documents S3.

Acknowledgments

This work was performed at the Virtual Reality Application Center, Iowa State University with the assistance of the Bioinformatics Laboratory of the Bioinformatics and Computational Biology program.

Supplementary Document S1: Data Base Schema

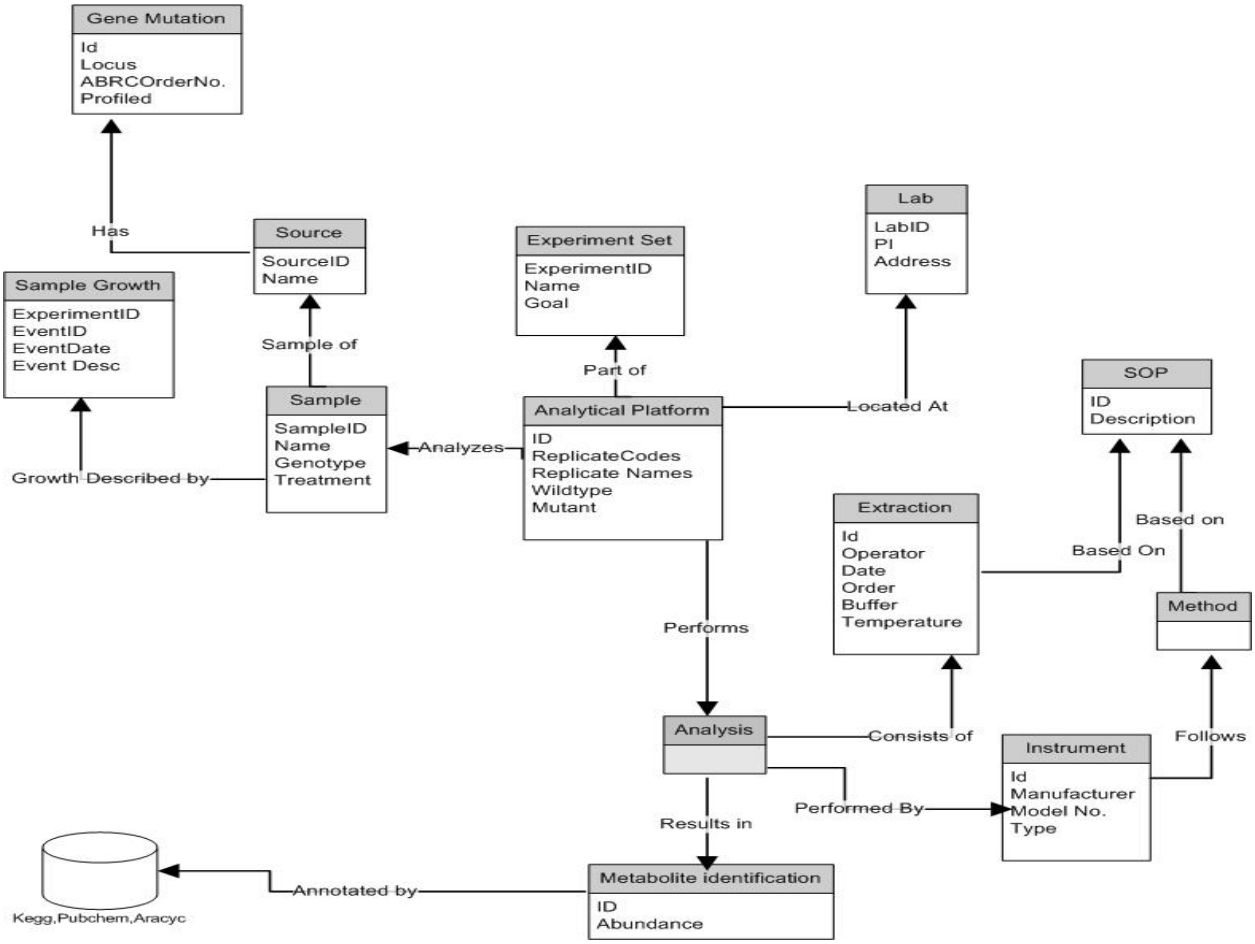

Figure 3-6 Database Schema

Supplementary Document S2: PM Website Map

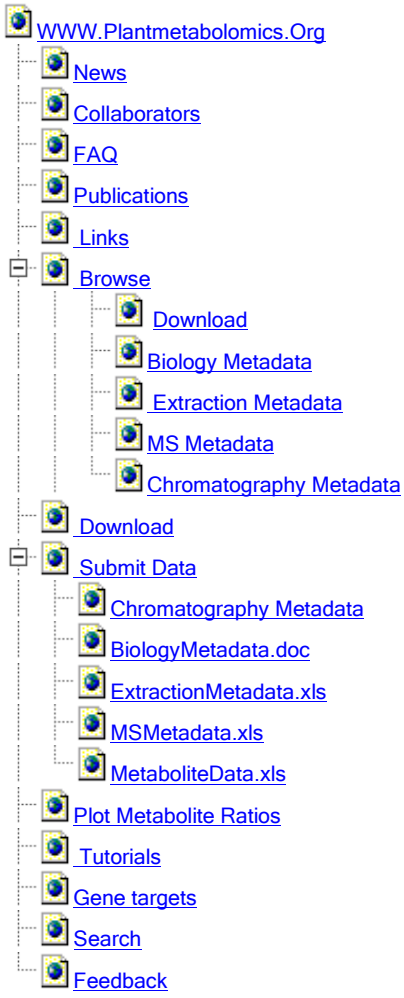

Figure 3-7 Website Map

Supplementary Document S3: Applications of PM and Availability of Metabolomics Data: Case Study

Case study: Visualize the difference in metabolite levels under two different stress conditions

Scenario: An investigator is interested in how the metabolome changes when comparing mutant and wild-type samples. Specifically, the investigator wants to know which metabolites show a significant change and which biochemical pathways they are involved in. The investigator is also interested in obtaining detailed information about specific metabolites from other web sources, other metabolites in the relevant biological pathways, and all the metadata associated with the selected mutant sample.

For example, the investigator is interested in the metabolome changes for mutant SALK_003718, which has a T-DNA mutation in the *Arabidopsis* gene At3g16950 which encodes a plastid lipoamide dehydrogenase, compared to the combined wild-type samples, WtCol01 and WtCol02, in Mutant Experiment #3 (ME#3). Below is a detailed description of a possible analysis path. Please note that analyses do not need to be done in this order. Help icons are located throughout the database to aid users in understanding the tools available at PM.

In the main web page, www.plantmetabolomics.org, the investigator first clicks on the "Tools" menu followed by "Plot Metabolite Ratio" on the menu option. To generate a ratio plot, data must be selected from individual experiments, for this example clicking on the drop down box by the experiment name "Mutant Experiment #3" allows the investigator to select the control (WtCol01) and the mutant (SALK_003718). By default, all metabolite profiling platforms for this experiment are selected, but the investigator can exclude/specify platforms. By clicking submit, the ratio plot is generated. Before submitting, the investigator can view the metadata for a

specific experiment by clicking on the experiment name. In this example, only the platform for amino acids is selected.

The ratio plot (Figure 3-3) contains all metabolites analyzed within the selected platforms (in this case, the amino acid platform in ME#3).Metabolite names appear by moving the cursor over the plotted points. Contained within this page is a summary of the data quality. Missing or below detection limit values are depicted by different colored marks on the plot. The right side of the page shows summary information in a list form. This data can be downloaded as a text file by clicking on the download button on top. This list groups the metabolites according to number of replicates that are below detection limit. This list also shows the fold changes between the mutant and the control and be easily sorted. Each point in the ratio plot or the metabolite name in the summary list is clickable and advances the investigator to the metabolite details page (Figure 3-4). For this example, the investigator clicks on the metabolite methionine, which is at the 13 point on the y-axis and has a 1.5 fold change between the samples SALK_003718 and WtCol01. The investigator is advanced to the metabolite detail page for methionine (Fig. 3.4). This page contains information about the molecular weight, chemical formula, CAS registry number, SMILES (Simplified Molecular Input Line Entry System) notation and pathway information for this specific metabolite. It also provides links to other databases including AraCyc and the LIGAND database from KEGG. The bar chart at the bottom of this page graphs the abundance level of methionine for each replicate of all the samples profiled in ME#3. Since a metabolite can be detected by multiple analytical platforms, the investigator can search for that metabolite using the “Search” functionality. The resulting page provides links to the metabolite details page where the data can be visualized for each analytical platform in bar charts as described above. The downloaded summary list is also helpful in finding a metabolite that is detected by multiple platforms. Since the different platforms use different extraction procedures, the amount of fold change can be different in different analytical platforms. In our

example, the downloaded summary list shows that methionine is detected by 3 analytical platforms in ME#1 and ME#2. The investigator can do a quick quality check of the platform at this time by generating the “Scatterplot of replicates” for SALK_003718 and WtCol01 lines in the 3 analytical platforms that detect it and make a decision to exclude any replicates that do not have a good reliability.

From the pathway information for methionine found on metabolite detail page, the investigator can search for other profiled metabolites involved in the “tRNA charging pathway”. The resulting page shows all the other metabolites in the PM database that belong to the given pathway. The investigator can download the CSV file of this data using the “Download” button provided at the top of the resulting page.

Metabolite abundance information can be downloaded by clicking the “Download” link on the top of the results or by clicking the “Download” menu option on the main page.

Following these steps provides a detailed analysis of a single mutant sample. A comparison of all the metabolites profiled by the consortium in a mutant sample to that of wild-type gives an overall view of the changes that are occurring in the metabolome of this mutant line. Biochemical mapping of the metabolites that are hyperaccumulating and hypoaccumulating compared to wild-type provides preliminary evidence of the biochemical pathway the target gene may be associated with, thus leading to an initial hypothesis about the function of that target gene.

CHAPTER 4. PLANTMETABOLOMICS.ORG:MASS SPECTROMETRY BASED ARABIDOPSIS METABOLOMICS-DATABASE AND TOOLS UPDATE

A paper published in Nucleic Acid Research 2012 database issue

Preeti Bais, Stephanie M. Quanbeck, Basil J. Nikolau, and Julie A. Dickerson

Abstract

The PlantMetabolomics (PM) database (http://www.plantmetabolomics.org) contains comprehensive targeted and untargeted mass spectrum metabolomics data for *Arabidopsis* mutants across a variety of metabolomics platforms. The database allows users to generate hypotheses about the changes in metabolism for mutants with genes of unknown function. Version 2.0 of PlantMetabolomics.org currently contains data for 140 mutant lines along with the morphological data. A web-based data analysis wizard allows researchers to select preprocessing and data-mining procedures to discover differences between mutants. This community resource enables researchers to formulate models of the metabolic network of *Arabidopsis* and enhances the research community's ability to formulate testable hypotheses concerning gene functions. PM features new web-based tools for data-mining analysis, visualization tools and enhanced cross links to other databases. The database is publicly available. PM aims to provide a hypothesis building platform for the researchers interested in any of the mutant lines or metabolites.

Introduction

PlantMetabolomics.org stores the data from an NSF-funded multi-institutional consortium that is developing metabolomics as a functional genomics tool for elucidating the functions of *Arabidopsis* genes without visible phenotype. The consortium has established mass

spectrometry based metabolomics platforms that detect approximately 2000 metabolites, of which ~1000 are chemically defined [1]. The consortium generates the *Arabidopsis* biological material at a single location followed by distribution to the analytical laboratories for targeted and untargeted analyses. Phase 1 focused on investigating the robustness of the *Arabidopsis* metabolome, and defining the conditions that minimize the environmental and developmental effects. Subsequently, the consortium profiled the metabolome of specific T-DNA knockout alleles for these targeted genes [2]. These MSI-compliant metabolomics data [3, 4] are integrated with phenotypic data and data concerning protein function, transcription and other studies to help users generate hypotheses concerning the functions of the targeted genes.

The updated PlantMetabolomics.org database features new datasets and morphological information for the plant community along with new web-based analysis tools. These tools include clustering and classification tools to distinguish between different mutants as well as determining which metabolites best differentiate the mutant. New visualization tools include ratio plots of metabolites and CytoscapeWeb [5] pathway visualization of metabolites on the AraCyc pathways [6].

Database Contents

PlantMetabolomics.org contains mass spectrometry based metabolomics concentration data for 140 novel single-knockout gene mutant lines in *Arabidopsis.* 53 the lines are novel since the last release and 35 were repeated to increase the number of replications. Approximately 998 known metabolites and 2020 unknown metabolites were detected using 7 different MS-based platforms for each of these mutant lines. The number of replicates for each line was also increased from 3 replicates to 6 replicates.

The database has also added morphological image data including features of the mutants' leaves, cotyledons and roots at 16 days after imbibitions (DAI) and mature seeds using an

Olympus stereomicroscope with reflected and transmitted light sources and a high-resolution digital color image and scanning electron microscope. Digital camera images of the roots of all the *Arabidopsis thaliana* tissue were collected at 6, 9, 13 and 16 days after imbibitions (DAI) in pixels and these were converted from pixels to root length measurements using Image J software [7]. A user can select a gene and compare its morphological images with the images from the wild type samples using a side by side image analysis tool in the database which is accessible from the when the user searches for a gene of interest from the home page or uses the search functionality to search for a gene.

New annotation links to LipidMaps [8] have been added for metabolites. Structurally known metabolites have been annotated with metabolic pathway information from the AraCyc database (version 8.0) [6]. This annotation helps users understand how changes in a metabolite might affect the metabolism of the entire organism. Figure 4-1 shows an example of the new annotation and the images.

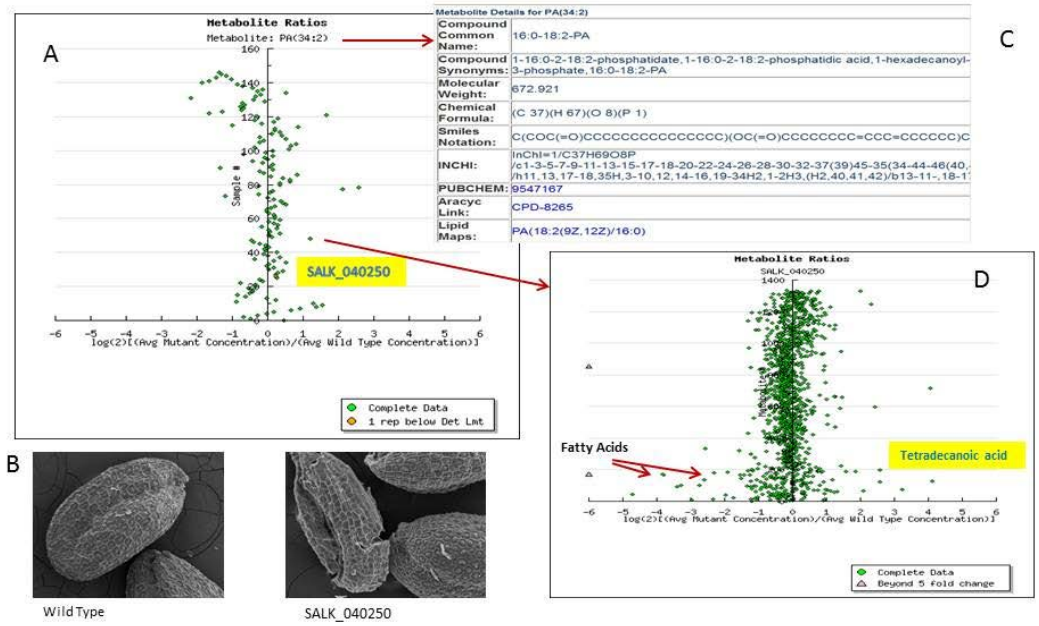

Figure 4-1 **Visualization tools**

Analysis Tools for Metabolomics

PlantMetabolomics.org includes new web based data analysis tools to aid a researcher in generating hypothesis about the metabolomics signature of a mutation. The data analysis wizard provides various options to normalize and preprocess data along with many choices of multivariate data analysis methods along with step by step guidance on the analysis pipeline. Default choices are provided at each step and the downstream analyses are made available only after the necessary preprocessing steps have been successfully performed. All the analysis results and figures are made available for download at the end of the analysis. The data analysis tool is developed with PHP and the R programming environment [9].

Data Preprocessing: The data preprocessing steps involve missing value imputation and normalization. For missing value imputation, the user selects a threshold to eliminate

metabolites that have a higher percentage of missing values than the threshold (e.g., for a threshold of 50%, a metabolite with 4 or more missing values out of 6 will be removed from further computation). For cases where there are fewer missing values, the missing values will be imputed by mean of the concentration for that metabolite over the remaining values. The next step is data normalization. Data normalization weights the metabolites to emphasize different attributes of the data. Common choices described in [10], Range Scaling, Pareto Scaling, and Auto Scaling, help weight metabolites equally regardless of overall abundance. Log Transformation is used to correct for heteroscedascity and make multiplicative effects additive. The equations and a discussion of each method are accessible from the “?” icon in the data analysis wizard. After the preprocessing and normalization steps, a user can choose one or more of the analysis tools to analyze the data. Examples have been provided at each data mining step to help users interpret their results.

Clustering Analysis: Biologists can generate hierarchical clustering plots to see which mutants are statistically close to each other and have similar metabolic profiles. Multiple choices for distance measure (Euclidean and Manhattan) and for the linkage method (Ward, complete, single, average, median, and centroid) are available. The goal is to group or segment a collection of samples (mutants) into subsets or "clusters", such that those within each cluster are more closely related to one another than objects assigned to different clusters. The result of clustering is presented as a dendrogram that a user can download from the PM website. Figure 4-2 A shows an example of a dendrogram using hierarchical clustering analysis tool with average linkage and Euclidean distance parameters.

Multi-Dimensional Scaling (MDS): A MDS plot is a commonly used multivariate exploratory data analysis tool. MDS is an exploratory multivariate data analysis method that is used in visualizing the structure of relations between entities by providing a geometrical representation of these relations in a lower dimensional space[11]. An MDS plot shows the similarities or

dissimilarities in data in two dimensions. In this case, the MDS plot shows statistical distances among samples based on their metabolomes' signatures (Figure 4-2 D). Commonly used distance measures (Euclidean and Manhattan) are provided for this tool as well.

Principal Component Analysis (PCA): PCA is one of the most commonly used methods used in high dimensional data analysis [12]. PCA provides a low dimension view of the multidimensional data by mathematically transforming a number of correlated variables into a smaller set of uncorrelated variables which are called Principal Components (PCs). A user can generate PCA plot against the first two principal components and also the scree plot that show the percentage of variability explained by subsequent principal components. The PCs are orthogonal and are ordered according to the variance explained. Therefore the first PC explains the maximum variance. If the variance in the data reflects the true biological difference then plotting first PC against the second can be used to visualize the separation in the different classes. The original variables that contribute the most to the first few PCs are considered to be the most important. The PCs can be downloaded for further analysis. Figure 4-2 B shows an example of PCA loadings plot for the first two PCs.

Random Forest Classifier: Random Forests are used in metabolomics for classifying mutants into different classes [13]. A Random Forest Classifier is an ensemble of classification trees [14]. Random Forests work well for classification when the number of features is much greater than the number of observations and they have good predictive performance even when most input variables are noisy[15]. Of importance to biologists is that the output is easy to understand because it does not transform the metabolite data and the output ranks variables that are responsible for classification.

The classification trees are built using a bootstrap sample of the data generated by using 2/3rd of the data for sample generation and keeping the remaining 1/3rd of the data for testing. A

small subset of the variables is used in building a tree. The random Forest R package provides classification analysis between two or more types of samples (e.g., Wild Type and a Mutant line) [16] and generates the variable importance score plots of the key metabolites (Figure 4-2 C).The list of top 30 key metabolites is also made available along with the annotations for the metabolites. One can click on a metabolite name on this list and see its annotation from various external databases such as KEGG, AraCyc and Lipid Maps. The automatically generated ratio plot shows the metabolite's behavior in the other mutants as compared to wild type samples. The complete list can be downloaded by clicking at the download file link and used in other applications. The random forest classifier can also be downloaded along with the number of correctly classified and misclassified samples in each class.

Download Results: At the end of analysis, the user can download all the results along with comma separated data files and as well as the R code used at each step of the analysis. Examples are also provided at each step to help the users with the interpretation of their results.

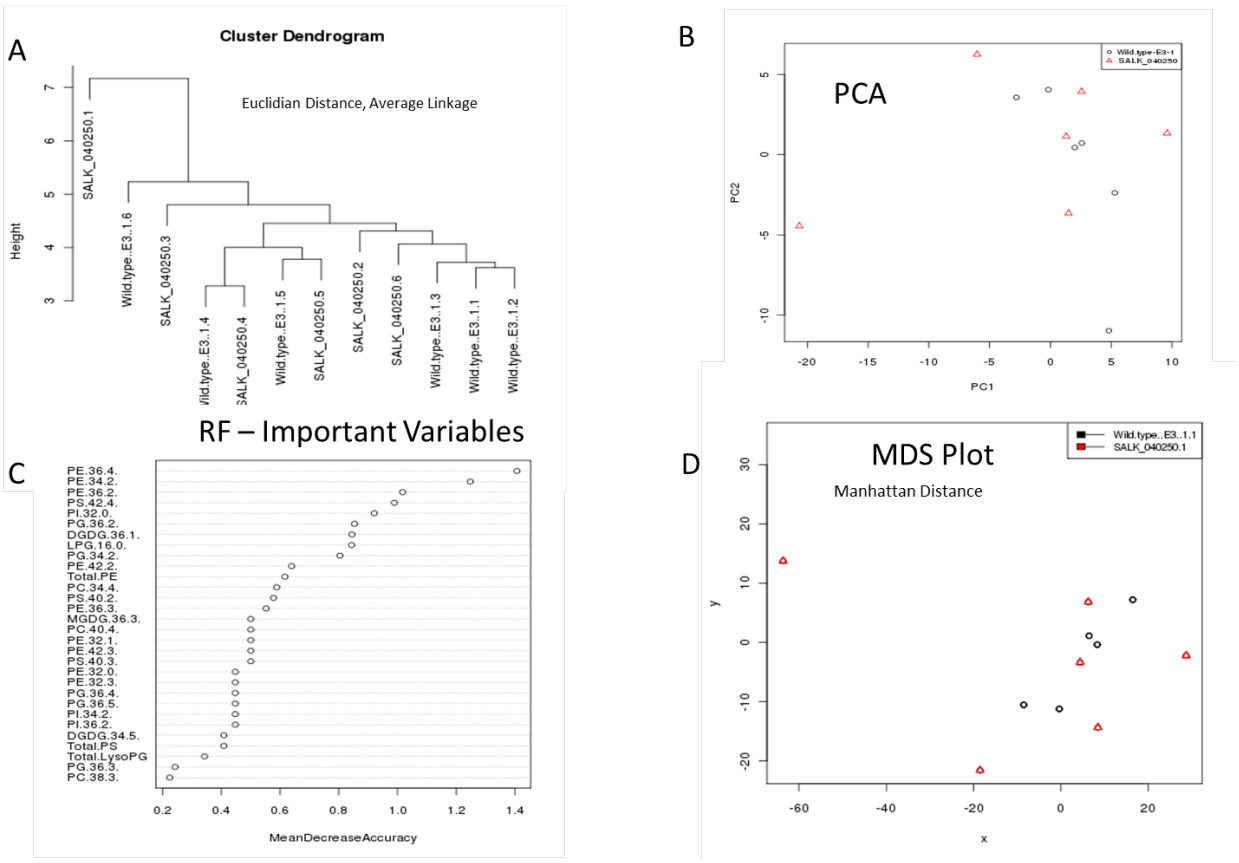

Figure 4-2 **Data analysis tools**

Visualization Tools for Metabolomics

New data visualization plots were added so a user can select a metabolite and see its behavior in 140 different mutations in a single plot (as a ratio of mutant and wild type samples). Similarly, a user can select a gene and see the behavior of all the metabolites (as compared to the wild type samples). After selecting a gene of interest, a user is taken to gene details page where they are shown the morphological data along with a log-ratio plot of the data. In the log-ratio plot for a gene, each point shows the log-ratio (to base-2) of a metabolite's abundance in the (mutant sample):(wild-type sample). The points are color coded according to the number of missing values for each metabolite and provide an instant data quality check. Clicking on a point in the log-ratio plot takes the user to a page where annotation of that metabolite with the information about its participation in pathways and links to other databases like KEGG [17]),

LipidMaps [8], and PUBCHEM [18] are shown. The metabolites are annotated with a local copy of the AraCyc database [19] which was updated to the latest release of version 8.0 of AraCyc.

Single metabolic pathways from AraCyc can also be viewed using CytoscapeWeb [5] and PathwayAccess tools[20]. From the annotation page, a user can select a pathway that contains their metabolite of interest and view the pathway with their metabolomics data superimposed for any of the experiments in the database.

Conclusions and Future Developments

This updated version of PlantMetabolomics.org provides metabolomics mass spectrometry-based metabolomics data from multiple analytical platforms. A user can analyze this data using our web based data visualization and mining tools and generate the hypothesis about the functions of gene of their interest. A user can also perform a comparative analysis on a metabolite or metabolic pathway of interest and see their behavior under different mutations. We plan to enhance our coverage mutant lines to 203 novel lines.

The next steps for this database are to create a viewer for extracting the spectra of the measured metabolite from the different platforms and replicates. This will create a valuable resource for mass spectra across many different platforms and gather information on measurement variability. This capability may allow PlantMetabolomics.org to link to the spectral data in the LC-MS *Arabidopsis* database, AtMetExpress [5] and the GC-MS Golm Metabolomics Database [23]. The flexibility of the pathway viewer will also be enhanced to give the user more ways to combine pathways into networks and select data.

Availability

The PlantMetabolomics.org database is available online and free to all without restriction at: http://www.plantmetabolomics.org/.

Funding

This work was supported by the National Science Foundation [MCB 08200823].

Acknowledgements

The following labs generated the metabolomics data in PlantMetabolomics.org: Oliver Fiehn (UC Davis), B. M. Lange(Washington State University), Lloyd Sumner (Noble Foundation), Ruth Welti (Kansas State University) , and Vladimir Shulaev (Virginia Bioinformatics Institute) as part of the Arabidopsis Metabolomics Consortium. The stereomicroscopic images were generated by Hilal Ilarslan and Jennifer Robinson of Iowa State. The annotations and links to AraCyc were provided by Kate Dreher and Sue Rhee of the Plant Metabolic Network Project and The *Arabidopsis* Information Resource (TAIR). NSF Research Experience for Undergraduate students, William Van Walbeek and William Petersen developed the Cytoscape Web pathway viewer tool.

References

1. Bais P, Moon SM, He K, Leitao R, Dreher K, Walk T, Sucaet Y, Barkan L, Wohlgemuth G, Roth MR et al.**PlantMetabolomics.org: a web portal for plant metabolomics experiments.** *Plant Physiol* 2010, **152:**1807-1816.
2. Alonso JM, Stepanova AN, Leisse TJ, Kim CJ, Chen H, Shinn P, Stevenson DK, Zimmerman J, Barajas P, Cheuk Ret al: **Genome-wide insertional mutagenesis of Arabidopsis thaliana.** *Science* 2003, **301:**653-657.
3. Fiehn O, Wohlgemuth G, Scholz M, Kind T, Lee do Y, Lu Y, Moon S, Nikolau B: **Quality control for plant metabolomics: reporting MSI-compliant studies.** *Plant J* 2008, **53:**691-704.
4. Fiehn O, Sumner LW, Rhee SY, Ward J, Dickerson J, Lange BM, Lane G, Roessner U, Last R, Nikolau B: **Minimum reporting standards for plant biology context information in metabolomics studies.** *Metabolomics* 2007, **3:**195-201.
5. Matsuda F, Hirai M, Sasaki E, Akiyama K, Yonekura-Sakakibara K, Provart N, Sakurai T, Shimada Y, Saito K: **AtMetExpress development: a phytochemical atlas of Arabidopsis development.** *Plant Physiol* 2010, **152:**566-578.
6. Matsuda F, Nakabayashi R, Sawada Y, Suzuki M, Hirai MY, Kanaya S, Saito K: **Mass spectra-based framework for automated structural elucidation of metabolome data to explore phytochemical diversity.** *Frontiers in Plant Science* 2011, **2**.
7. Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD: **Cytoscape Web: an interactive web-based network browser.** Bioinformatics. 2010;26(18): 2347-2348 8. Zhang P, Dreher K, Karthikeyan A, Chi A, Pujar A, Caspi R, Karp P, Kirkup V, Latendresse M, Lee C et al.: **Creation of a Genome-Wide Metabolic Pathway Database for Populus trichocarpa Using a New Approach for Reconstruction and Curation of Metabolic Pathways for Plants.** *Plant Physiology* 2010, **153:**1479-1491.
9. http://rsbweb.nih.gov/ij/index.html
10. http://www.lipidmaps.org
11. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J et al.**Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5:**R80.
12. van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ: **Centering, scaling, and transformations: improving the biological information content of metabolomics data.** *BMC Genomics* 2006, **7:**142.
13. Seber GAF: *Multivariate Observations.* Hoboken, NJ: John Wiley & Sons; 1984.
14. Spearman C: **The proof and measurement of association between two things.** *American Journal of Psychology 1904,Vol*. 100, No. 3-4. (r 1987), pp. 441-471.
15. Scott IM, Vermeer CP, Liakata M, Corol DI, Ward JL, Lin W, Johnson HE, Whitehead L, Kular B, Baker JM et al.**Enhancement of Plant Metabolite Fingerprinting by Machine Learning.** *Plant Physiology* 2010, **153:**1506-1520.
16. Breiman L: **Random Forests.** *Mach Learn* 2001, **45:**5-32.
17. Díaz-Uriarte R, Alvarez de Andrés S: **Gene selection and classification of microarray data using random forest.** BMC Bioinformatics, 2006, 7:3.
18. Liaw A, Wiener M: **Classification and Regression by randomForest.** *R News* 2002. R News 2(3), 18--22.
19. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: **KEGG for representation and analysis of molecular networks involving diseases and drugs.** *Nucleic Acids Research* 2010, **38:**D355-D360.
20. **PubChem Compound Database** [http://pubchem.ncbi.nlm.nih.gov]

21. Zhang P, Foerster H, Tissier CP, Mueller L, Paley S, Karp PD, Rhee SY: **MetaCyc and AraCyc. Metabolic pathway databases for plant research.** *Plant Physiol* 2005, **138:**27-37.
22. Van Hemert JL, Dickerson JA: **PathwayAccess: CellDesigner Plugins for Pathway Databases.** *Bioinformatics* 2010 , *26 (18): 2345-2346*.
23. Hummel J, Selbig J, Walther D, Kopka J: **The Golm Metabolome Database: a database for GC-MS based metabolite profiling.** In *Metabolomics.* Edited by Nielsen J, Jewett M. Berlin, Heidelberg, New York, : Springer-Verlag; 2007: 75-96

List of Figures

Figure 4-1: Visualization tools: (A) Log-ratio plot of a metabolite (PA 34:2) where each point shows the ratio of the concentration of the given metabolite in mutant samples vs. the wild type samples. The highlighted mutant line (SALK_040250) looks interesting as it is away from the central vertical axis and thus depicts difference between mutant samples and the wild type samples. (B) The user can instantly access the stereomicroscopic images for this mutant and compare them with wild type samples. Seed images at 250 X zoom of mutant's seeds look a little distorted as compared to the wild type seeds (Seed image courtesy of Jennifer Robinson). (C) The user can also access the details of the metabolites including cross links to other databases. (D) Clicking on any of the points in the log-ratio plot in (A) shows the log-ratio plot of all the metabolites for that mutant. For example, some fatty acids including tetradecanoic acid look interesting for this mutant as they are away from the central vertical axis and show large fold change between the wild type and mutant samples.

Figure 4-2: Data analysis tools: (A) Hierarchical clustering of lipidomics data from the Welti Lab compares SALK_040250 (At1g61720) mutant line with wild type samples using Euclidean distance and average linkage method. (B) PCA loadings plot of the first two PCs shows that the wild type and mutant are not linearly separable. (C) Important metabolites for the classification between wild type and the mutant line using the Random Forest tool shows that the most

important variables are glycerophospholipids with chain lengths of 34 and 36. (D) MDS plot of the mutant and wild type samples using the Manhattan distance measure which shows that the mutant and wild type are not separable and that there is an outlier in the data.

CHAPTER 5. DATA ANALYSIS PIPELINE IN FUNCTIONAL GENOMICS USING METABOLOMICS AND MACHINE LEARNING

A paper to be submitted to Plant Physiology

Preeti Bais, Basil Nikolau, David J. Oliver, Julie A. Dickerson

Abstract:

This analysis presents a biomarker discovery pipeline that uses machine learning and metabolomics to discover biochemical changes in a cell due to a single gene knock-out mutation in Arabidopsis. Since a single metabolomics technique cannot cover the whole metabolome, multiple mass spectrometry based metabolomics platforms are integrated together to get biomarkers of a mutation across a wide range of metabolite families. The use of different metabolomics platforms increases the coverage of the metabolome but multiple platforms present significant challenges on integrating data across the platforms. Different strategies for integrating the metabolomics abundance data from multiple platforms are compared to find the ideal method for biomarker discovery. The Random Forest machine learning algorithm is used for classification of mutant and wild type samples and to generate reproducible models with a small set of metabolites that are responsible for the classification. Unknown metabolites are a serious problem in any large scale metabolomics analysis as they do not provide any biological insight. Partial correlation networks are used in putatively identifying the unknown metabolites without the need for expensive and time consuming methods like NMR.

A proof-of-concept analysis on the oxoprolinase (*oxp1*) and gamma-glutamyl transpeptidase (*ggt1* and *ggt2*) single gene knock-out mutants in the glutathione degradation (GSH) pathway of the *Arabidopsis* confirms the known biology that OXP1 is responsible for conversion of 5-oxoproline (5-OP) to glutamic acid. In addition, *ggt1/ggt2*

analysis supports the hypothesis that the GGT genes may not be major contributors for the 5-OP production. Also, the *ggt2* mutation does not appear to alter the biochemical profile of the cells in comparison to the wild type samples, supporting the previous studies that it may have low level expression in the leaf tissues.

This data analysis pipeline is implemented in a web based metabolomics analysis and visualization suite of tools at www.plantmetabolomics.org.

Keywords: Metabolite Profiling, Arabidopsis, Machine Learning, GC-MS, LC-MS, Random Forests, Oxoprolinase, Correlation network

Introduction

Metabolomics is the science of measuring the pool sizes of metabolites (small molecules of molecular weight <1,000 Da), which collectively define the metabolome of a biological sample (Fiehn et al. 2000; Hall et al. 2002). Under stable environmental conditions, comparing the metabolome of a wild-type sample to that of a sample altered by a mutation at a target gene can provide clues as to the function of that gene (Bino et al. 2004). Metabolomics aims to capture the final outcome of the genes at the biochemical level and the metabolomics based biomarkers can provide an understanding of the biochemical networks involved in a cellular process. Since a single analytical technique can not cover all the metabolites of a biological system, multiple mass spectrometry (MS)-based metabolomics analytical and separation techniques were used on identical plant material to understand the gene functions (Bais et al. 2010). Different strategies of platform integration are compared to find the ideal method that not only classifies the mutants from the wild type samples with most accuracy but also provides the most biologically meaningful subset of metabolites for the classification.

Random Forest (RF) classifiers are used in classifying the mutant samples from the wild type samples and finding the key metabolites for the difference (Beckmann et al. 2007). RF classifiers have been shown to create usable models using metabolomics data (Enot et al., 2006, Scott et al. 2010). RF classifiers are non-linear classifiers which keep the features (i.e., metabolites) distinct and provide an importance ranking for the effectiveness of each feature. Finally, putative identifications for key unknown metabolites are provided by using a large scale partial correlation analysis across multiple mutation lines and manual inspection of the mass spectra. This helps in incorporating the unknown metabolites in understanding the biological significance of the biochemical difference between the mutants and the wild type samples. Other methods of structure determination of metabolites include MS-MS analysis or NMR which

are expensive and time consuming. Using the partial correlation networks across many different mutant lines to find if the key unknown metabolites are closely related to any known metabolites helps in hypothesizing the biological role of the unknown metabolites with the existing data.

Materials and methods

Plant Materials

The *oxp1* mutant (SALK_078745), the *ggt1* mutant (SALK_004694) and the *ggt2* mutant (SAIL_6_G02) have been described and characterized earlier (Ohkama-Ohtsu et al. 2007a, 2008) and were a result of T-DNA insertion into col-0 (ecotype Columbia:col-0) (Alonso et al. 2003). The data is available at the project web site (www.plantmetabolomics.org) as part of Experiment E1 (*oxp1*) and Experiment 3 (*ggt1* and *ggt2*). Six mutant samples were compared with the two sets of wild type samples (six samples in each set) from the same experiment batch. The wild type sets from the same experiment batch were also compared with each other. All the metadata about the plant growth conditions, extraction protocols, mass spectrometry, instruments etc. is available for download at www.plantmetabolomics.org (Bais et al. 2010).The partial correlation analysis was done using 70 mutant lines along with the one wild type line from GC-TOF platform (Supplementary Document 5.2).

Metabolite Detection platforms

The Arabidopsis Metabolomics Consortium combined parallel analytical outputs from seven analytical platforms conducted on aliquots of the identical plant material to generate abundance data on 1042 peaks. Approximately 60% of this data was obtained from non-targeted GC-TOF-MS and LC-MS platforms. The targeted analytical platforms were fatty acids, cuticular wax extraction, lipidomics, phytoesterols and isoprenoids extraction platforms. Currently 498 of the detected metabolite peaks are chemically defined by the analytical labs and 554 peaks have unknown structures. The complete data generation pipeline including each platform's extraction and analytical protocol is available at www.plantmetabolomics.org (Bais et al. 2010). Table 5.1

shows the number of metabolites detected by each platform along with the number of structurally known and unknown compounds. There was a six percent overlap of known metabolites between the platforms where the same metabolite was detected by more than one lab.

Table 5-1 Platform Summary

Platform	Known Metabolites	Unknown Metabolites	Total Metabolites
Fatty Acids (FA)	37	79	116
Cuticle Wax (CW)	37	25	62
Phytoesterols (PHY)	11	17	28
Isoprenoids (ISO)	6	3	9
Lipidomics (LPD)	171	0	171
GC-TOF-MS (GC-TOF)	195	420	615
LC-MS (LCMS)	41	0	41
Total	498	554	1042

Metabolomics Analysis Pipeline

The complete metabolomics analysis pipeline in Figure 5-1 is described in the following paragraphs.

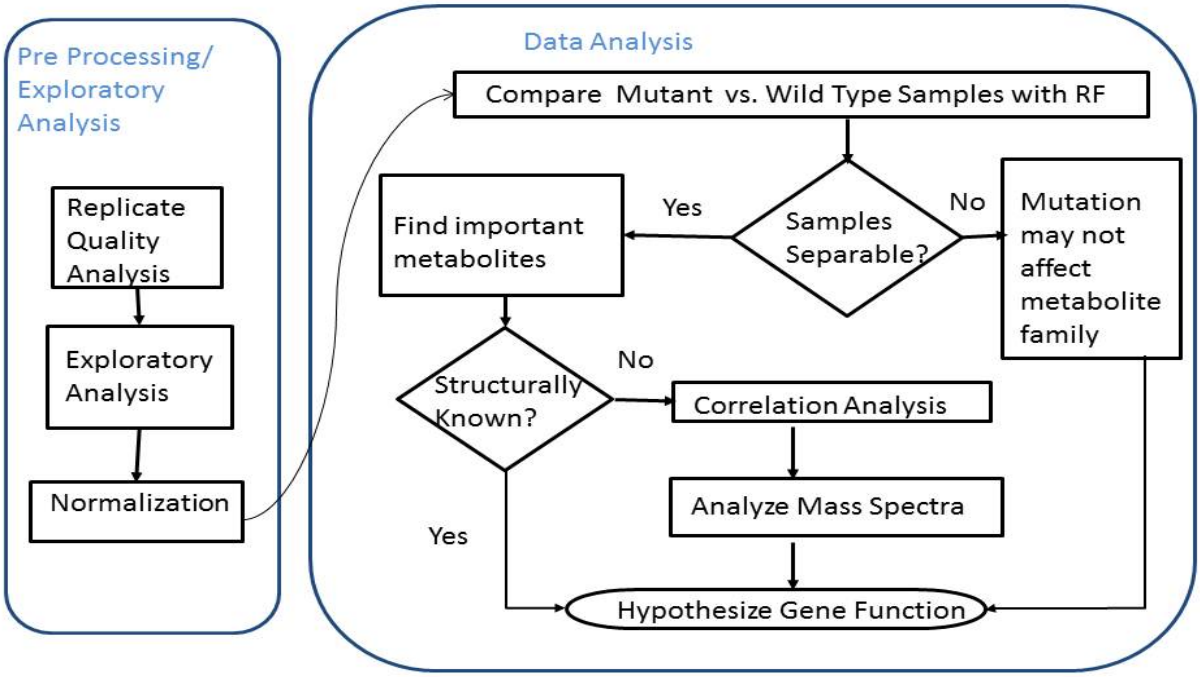

Figure 5-1 **Data Analysis Pipeline**

Data Preprocessing

Repeatability analysis was performed within the six replicates of any single genotype samples (wild type or mutant) for each of the seven platforms separately using the non-parametric Spearman's correlation (Spearman 1904) on log (base 2) transformed data. This analysis showed more than 50% Spearman's correlation coefficients between all pairs of replicates from a single genotype. For the *ggt1/ggt2* study, phytoesterols data was completely discarded due to lack of repeatability. Please see supplementary document 5.1 for the detailed discussion on repeatability analysis.

Exploratory Data Analysis

Log2 ratios of the average concentration of a metabolite from mutant samples and the average concentration from the wild type samples were calculated and plotted for all metabolites for each of the mutants. All the metabolomics platforms were combined together to generate a single ratio plot. Each point in the ratio plot showed the ratio of average abundance in mutant samples vs. the average abundance on the wild type samples for a metabolite on logarithmic scale. These plots visualize the overall trends in metabolite concentrations under the two genotypes by comparing the relative abundances of metabolites between mutant and the wild type samples on logarithmic scale and show which metabolites changed the most between the two types of samples. The points (metabolites) that were far from the central vertical axis had changed the most between the two conditions. The ratio plots were generated before any other normalization or scaling to see absolute changes in the metabolite levels.

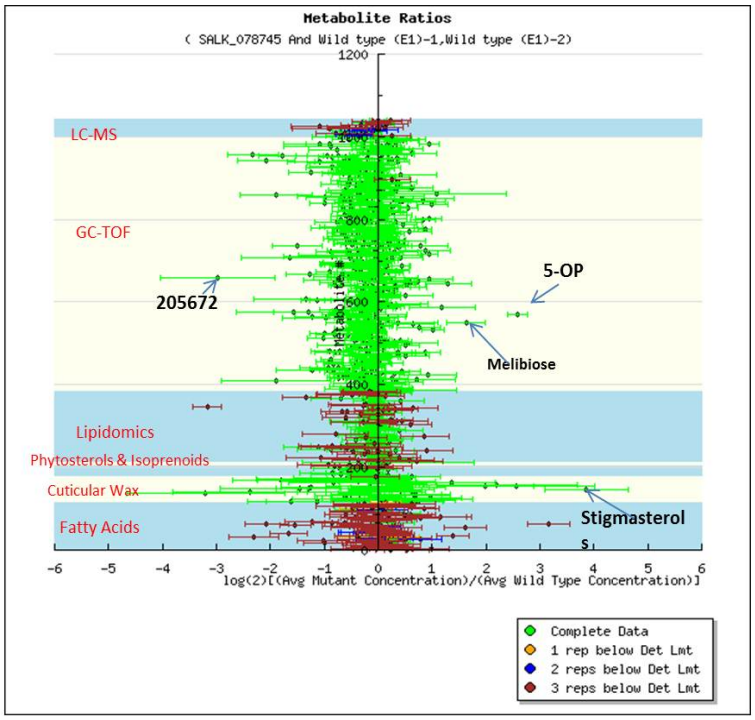

Figure 5-2 **Log2 ratio plot between *oxp1* mutant and wild type samples**

Data Preprocessing /Normalization

A metabolite was discarded if it had more than 50% missing values otherwise the missing values were imputed by the average value of the metabolite abundance. Range scaling was used in normalizing the data to a range of minimum abundance value for a metabolite and maximum abundance value for the same metabolite across all the samples to give equal importance to high and low abundance metabolites and to remove the instrumental response factor from the data for the integration of the platforms (van den Berg et al. 2006, Smilde 2005).

Data Integration across multiple platforms

Two methods for data integration were compared to find the optimal way to integrate the data from multiple platforms. In the batch integration method, all the metabolites from the seven platforms were concatenated side by side. All the metabolites were marked with the platform identification number to treat the common metabolites between two or more platforms as separate variables. In the hierarchical platform integration method, separate classifiers were built for each platform.

Random Forest (RF) Analysis

RF analysis (Breiman 2001) has been shown as a suitable method for classifying high dimensional data when the number of features is much higher than the number of samples (Lanz et al. 2009) . Reproducibility and biological significance are the major goals of any biomarker analysis. In this study, we used the RF's classification ability to determine if the mutant samples were metabolically different from the wild type samples using the .632+ bootstrap method (Efron et al. 1997). The .632+ bootstrap method is a widely used variation of the bootstrap resampling method and has been shown to perform well when the signal to noise ratio is small as in case of metabolomics data (Molinaro et al. 2005). In the basic bootstrap method, n samples are drawn with replacement for learning the model and the samples that are left out serve as the test set. The bootstrap estimate is the average error made on the left-out

samples. The basic bootstrap method tends to be high-biased because the number of samples in the learning set has .632n unique observations on average. The .632+ bootstrap method corrects for this by taking a weighted average of the bootstrap estimate.

The stability of a metabolite to be selected as a biomarker in the original samples was evaluated by its frequency of getting selected in the 100 bootstrap runs. High frequency for a metabolite meant more stability of that metabolite as a biomarker. Margins were calculated by subtracting mean class membership probability for the wrong class from the mean class membership probability of the right class in the 100 bootstrap runs. Higher margins showed more confidence in the class membership. The R package "randomForest" and "varSelRF" were used in the RF analysis (Liaw et al. 2002, Diaz-Uriate 2007).

Incorporating the structurally unknown compounds

Structurally unknown metabolites comprise of about half of our data. Many of the key metabolites from RF analysis are also structurally unknown and do not provide any biological insight even when they are selected as potential biomarkers from a classification method (e.g. RF) described in the above paragraph. We harnessed our in-house data of 70 different mutant lines to build partial correlation networks among the metabolites in a separate global analysis. This analysis highlighted the strong correlations among the pair of metabolites that survived 70 different mutations and thus helped us in forming the hypothesis about the structurally unknown metabolites. Previous attempts on generating linear correlation networks based only on Pearson's pairwise-correlation between the metabolites have not been very successful and do not match well with the actual known biochemical networks as they are unable to distinguish between direct and indirect reactions (Steuer et al. 2003a and 2003b, de la Fuente et al. 2004). Recent studies (Krumsiek et al. 2011) have suggested using partial correlation on the Pearson's correlation matrix to remove the indirect relations between the metabolites. Using the simulated and actual human lipid metabolomics data, the authors have shown that the low order partial

correlation networks are very similar to the actual biochemical networks. However, even the low order partial correlation analysis requires large number of samples and small number of features. We addressed this problem by analyzing a small subset of structurally unknown metabolites that were potential biomarkers from RF analysis along with all known metabolites using the data from 70 different mutant lines (503 samples)(Supplementary document 8.5).

First order GGMs were built using the pairwise Pearson correlation coefficients between two metabolites that were conditioned against the correlation with all other metabolites. Top pairwise correlations were chosen using a q value of 0.05. Once the correlation network was built, all the known metabolites that were directly connected by a single edge to an important unknown metabolite were highlighted. R package “GeneNet” (Schafer et al. 2005 a, 2005 b) and R package “igraph” (Csardi et al. 2006) were used in analyzing the graphs.

The mass spectrum of the unknown metabolite was then analyzed using the BinBase library (Fiehn et al. 2005) to find if the queried unknown compound’s mass spectrum closely matched with any known compounds. BinBase library shows 10 most similar mass spectrums to the queried compound using similarity criteria developed by Stein et al. (Stein et al. 1994). BinBase library comparison is available for the 75% of our unknown compounds from our database that are detected by the GC-TOF platform. This combined analysis gave more insight in hypothesizing biological role of unknown metabolites without more expensive structure determination methods.

Results and Discussion

The GSH degradation pathway was used to test the proposed methods for biomarker selection across platforms. In the first example, key biomarkers from the *oxp1* mutant were found which agree with existing results and the identity and functions of some key unknown metabolites

were hypothesized. In the second example, the key biomarkers helped identify the effects of the two GGT mutants (*ggt1* and *ggt2*).

Figure 5.3 shows the GSH degradation pathway with literature verified parts as solid lines and hypothesized reactions as dashed lines. In mammals, the γ-glutamyl cycle functions to recycle the amino acids in extracellular glutathione (GSH) before they are lost to the animal's excretory system. A sequence of reaction initiated by the extracellular γ-glutamyl transpeptidase (GGT) removes the glutamic acid (Glu) from GSH and thus initiates the import of the component amino acids back into the cell. The removal of Glu from GSH can either be a hydrolysis reaction or a transferase where the Glu is transferred to an acceptor amino acid resulting in a γ-glutamyl amino acid dipeptide. Both reactions are catalyzed by GGT (part 1A in pathway block diagram). Once the γ-glutamyl amino acid is returned to the cytosol it is converted to 5-OP and the free amino acid by the enzyme γ-glutamyl cyclotransferase (GGC) (part 1B). Oxoprolinase (OXP1) catalyzes the ATP-dependent conversion of 5-OP to glutamic acid (part 1C) (Van der Werf et al. 1971).The metabolism of GSH in plants is quite different. Instead of a single GGT, Arabidopsis has three or four such proteins. GGT1 and GGT2 are apoplastic with strong GGT1 expression throughout the plant and GGT2 predominately found in siliques (Ohkama-Ohtsu et al., 2007a).

In mammals, the extracellular GGT reactions appear to be the major route of GSH turnover. GGT1 knockout mutants in *Arabidopsis* show no change in GSH levels with respect to Wild Type, although the extracellular oxidized form of GSH (GSSG) levels are significantly elevated either causing or resulting from increased oxidative stress in this mutant in previous studies (Ohkama-Ohtsu et al., 2007b, 2008). The major route of GSH turnover in *Arabidopsis* appears to be via a cytosolic γ-glutamyl cyclotransferase (part2B-i, part 2B-ii) (Ohkama-Ohtsu et al., 2008). Kinetic estimates suggest that about 90% of GSH in Arabidopsis is metabolized by this route. The product of the cyclotransferase reaction is 5-OP which is converted Glu by 5-oxoprolinase (part 2C). There is a single copy of the gene (OXP1) and the knockout mutant for

this gene (OXP1) appears to be unable to metabolize 5-OP which accumulates to high levels (Ohkama-Ohtsu et al., 2008). In this study, the known part (part 2C) of GSH pathway confirms the results of biomarker detection methods described in this study and the unknown part of the pathway (part 2A) is used in hypothesizing the functions of the GGT1 and GGT2 genes.

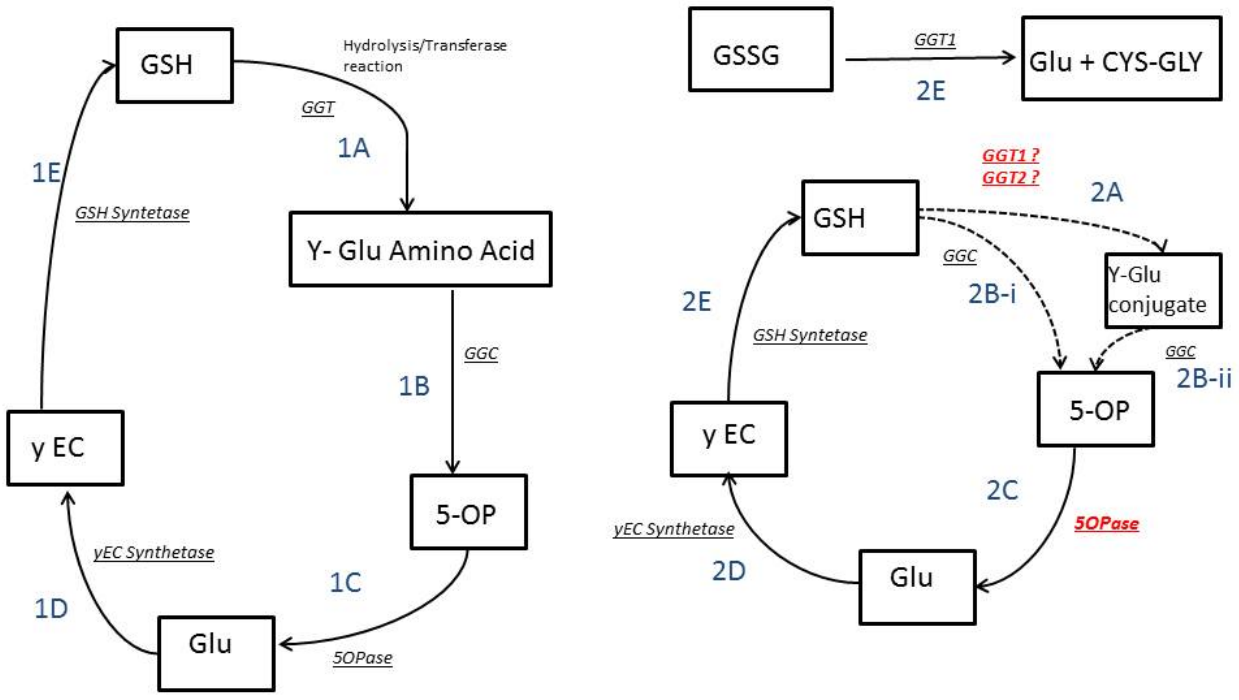

Figure 5-3 GSH degradation Pathway

Metabolomics platform integration

Table 5.2 compares the two data fusion strategies where samples from the *oxp1* mutant were compared against the two sets of wild type samples and the wild type (WT) samples are compared against each other. The WT1 vs. WT2 comparison showed negative margins and high bootstrap prediction error rates using 100 samples. The metabolic profiles of the two wild type samples were expected to be similar and thus the negative margins and high error rate predictions were as expected. Both fusion strategies were able to select 5-OP as a potential biomarker in the *oxp1*mutation with more than 10% frequency in 100 bootstrap runs. This measure asses how often a given metabolite, selected when running the variable selection procedure in the original sample, is selected when running the procedure on bootstrap samples and thus provides a measure of stability of a potential biomarker. The hierarchical fusion method had an ensemble of seven models that were built for each platform separately. It showed low error rates and high margins for the GC-TOF and LC-MS platforms but high error rates and low margins for all the other platforms. The hierarchical method provided more biologically meaningful results than the batch integration method (Table 5.2, Table 5.3). The *oxp1* mutation has been shown to work between 5-OP and glutamic acid in the literature (Ohkama-Ohtsu et al., 2008). The weak classifiers in fatty acids, cuticular wax, phytosterols, isoprenoids, and lipidomics platforms suggested that this mutation did not affect the metabolites detected by those platforms as they could not be used to differentiate between the mutants and wild type samples. The hierarchical fusion method was able to select glutamic acid as a potential biomarker in 15% of the bootstrap runs from the LC-MS platform. Although the classification accuracies for both the methods were at par for this mutant, the goal of a biomarker discovery is study is not only to classify the samples accurately but also select a small set of metabolites that can be repeatedly used in future studies when the class memberships are unknown. The hierarchical method also provides accuracy criteria for the

individual metabolomics detection platforms which can be used by the future studies to focus the attention on the platforms that provide strong models for a mutation under study.

Table 5-2 Comparison of Platform integration methods – *oxp1* vs. wild type

***oxp1* Study**				
	Low Level Fusion		**High Level Fusion**	
	***oxp1*-WT**	**WT-WT**	***oxp1*-WT**	**WT-WT**
Bootstrap estimate of prediction error (100 Iterations)	0.13	0.56	0.49 (FA), 0.42 (CW) 0.49 (PHY) 0.53 (ISO) 0.48 (LPD) 0.14 (GC-TOF) 0.24 (LC-MS)	0.63 (FA) 0.62 (CW) 0.35 (PHY) 0.55 (ISO) 0.59 (LPD) 0.53 (GC-TOF) 0.62 (LC-MS)
Average Margin	0.48	-0.15	-0.03 (FA) 0.13 (CW) 0.04 (PHY) -0.22 (ISO) -0.12 (LPD) 0.52 (GC-TOF) 0.27 (LC-MS)	-0.31 (FA) -0.27 (CW) 0.17 (PHY) -0.15 (ISO) -0.27 (LPD) -0.07 (GC-TOF) -0.31 (LC-MS)
# Of metabolites	952 (after missing values imputations)		62(FA) ,62(CW),28(PHY) 9(ISO),147(LPD),615(GC-TOF) 29(LC-MS)	
# of metabolites with >10% frequency in 100 bootstrap runs	7	NA	21 (After removing weak models FA,ISO,PHY,LPD)	NA

Table 5-3 Potential Biomarkers for *oxp1* mutation

#	Metabolite	U/D	Frequency in 100 Bootstrap Models		Correlated Compound	Similar Binbase Compounds (similarity score)
			HL	LL		
1	isothreonic acid	U	0.19	0.22		
2	GABA	D	0.18	0.17		
3	213179	U	0.17	0.21	5-OP	Glutamine, N-acetyl-glutamic acid (*)
4	succinic acid	D	0.13	0.11		
5	oxoproline	U	0.11	0.11		
6	303992	U	0.1	0.11	5-OP, tocopherol Y , galactose	Catechin (698.52)
7	200489	U	0.09	0.1	5-OP , homoglutamine	N-acetyl-D-mannosamine 3 (766.22), galactitol (751.66), glucose 2 (746.55)
8	melibiose	U	0.08	0.05		
9	205672	D	0.07	0.06	mannonic acid NIST	NA
10	alpha ketoglutaric acid	D	0.07	0.04		
11	4 hydroybutyric acid	D	0.05	0.04		
12	glycerol	D	0.05	0.02		
13	serine	D	0.05	0.02		
14	L Aspartic acid	D	0.59	0.01		
15	GABA	D	0.33	0.04		
16	L Isoleucine	D	0.33	NA		
17	L Tryptophan	D	0.26	NA		
18	L Citrulline	D	0.17	NA		
19	L Glutamic Acid	D	0.15	NA		
20	L Histidine	D	0.12	0.01		
21	L Proline	D	0.11	0.01		
22	L Threonine	D	0.1	NA		
23	L Tyrosine	D	0.09	NA		
24	L Phenylalanine	D	0.06	NA		

Incorporating the Unknown Metabolites

An unknown metabolite with ID "213179" which is up regulated in *oxp1* mutant samples appears with high frequency in the 100 bootstrap runs of RF classifiers in both methods. The partial correlation analysis shows that this metabolite is highly correlated with 5-OP across the 70 mutation lines. Chromatogram analysis shows that this metabolite has high similarity with N-acetyl-glutamic acid and glutamine. Our hypothesis about this metabolite is that it is a derivative of glutamic acid and is being up regulated in the mutant samples because 5-OP is not getting converted to glutamic acid due to the *oxp1* mutation.

Biological Confirmations and Discoveries

oxp1 ***Mutant vs. Wild Type***

The potential biomarker list consists of up regulated 5-OP and down regulated glutamic acid along with many amino acids that are also down regulated. The results support the literature evidence that OXP1 works between 5-OP and glutamic acid. The *oxp1* mutation down-regulates all the related amino acids as glutamic acid is a central molecule in amino acid metabolism in higher plants and the α-amino group of glutamic acid is transferred to all other amino acids via assimilation and dissimilation of ammonia. Both the carbon skeleton and α-amino group are involved in the synthesis of γ-aminobutyric acid (GABA), arginine, and proline (Forde et al. 2007). Sugars and the unknown metabolites putatively identified as sugars and flavonoid catechin are up regulated in the mutant samples. The lack of differentiation in the models created from the targeted metabolomics platforms also showed that the *oxp1* mutation may not affect lipids, fatty acid, cuticular wax, and isoprenoid or phytoesterol pathways.

ggt1 ***Mutant vs. Wild Type***

The best classifier for this mutation was generated by the LC-MS platform in the hierarchical clustering method. The list of potential biomarkers consisted of many amino acids which were up regulated in the mutant samples (Table 5.4). Notably, the potential biomarkers did not

include 5-OP. The results from this analysis confirm the hypothesis that GGT1 is not part of any major pathways for the 5-OP production and thus the *ggt1* mutation should not affect 5-OP concentration (Ohkama-Ohtsu et al., 2008).

Table 5-4 RF classification results for *ggt1* mutant using two platform integration methods

***ggt1* Study**				
	Batch Integration		**Hierarchical Integration**	
	***ggt1*-WT**	**WT-WT**	***ggt1*-WT**	**WT-WT**
Bootstrap estimate of prediction error (100 Iterations)	0.39	0.51	0.45 (FA) 0.45 (CW) 0.41 (ISO) 0.49 (LPD) 0.41 (GC-TOF) **0.17 (LC-MS)**	0.54 (FA) 0.59 (CW) 0.51 (ISO) 0.46 (LPD) 0.51 (GC-TOF) 0.58 (LC-MS)
Average Margin	0.15	-0.07	0.07 (FA) 0.04 (CW) 0.13 (ISO) -0.02 (LPD) 0.11 (GCT-OF) **0.46 (LC-MS)**	-0.14 (FA) -0.24 (CW) -0.06 (ISO) 0.08 (LPD) -0.01 (GCT-OF) -0.19 (LC-MS)
# Of metabolites	1043 (959 after missing values imputations)		91(FA) ,62(CW), 9(ISO),157(LPD),614(GC-TOF) 32(LC-MS)	
# of metabolites with >10% frequency in 100 bootstrap runs	12	NA	24	NA

Table 5-5 Potential Biomarkers for *ggt1* mutation

	Metabolite Name	Up or Down?	Frequency in 100 Bootstrap iterations	
			Hierarchical Integration	Batch integration
1	L Proline	U	0.73	0.21
2	L Histidine	U	0.7	0.14
3	L Citrulline	U	0.29	0.05
4	L Asparagine	U	0.28	#N/A
5	L alpha Amino n butyric acid	U	0.23	0.06
6	Beta Alanine	U	0.2	#N/A
7	L Glutamine	U	0.13	#N/A
8	delta Hydroxylysine	U	0.09	#N/A
9	L Homocystine	D	0.08	#N/A
10	Ethanolamine	D	0.07	#N/A
11	L Valine	U	0.06	#N/A

ggt2 *Mutant vs. Wild Type*

Both methods of data integration were unable to separate *ggt2* and wild type samples as evident from the high prediction error rates and low margins from models built on all seven platforms (Table 5.6). The high error rate and low margins were at par when the two sets of wild type samples were compared with each other. This suggests that the *ggt2* mutation may not have affected the biochemical profile of the cell, which supports the evidence that *ggt2* has low expression levels in leaf tissues. Also, 5-OP was not significantly changed in this mutation as well, suggesting that GGT2 is also not involved in 5-OP production confirming the hypothesis from the previous studies (Ohkama-Ohtsu et al., 2008).

Table 5-6 RF classification results for *ggt2* mutant

ggt2 Study				
	Batch Integration		**Hierarchical Integration**	
	***ggt2*-WT**	**WT-WT**	**ggt2-WT**	**WT-WT**
Bootstrap estimate of prediction error (100 Iterations)	0.48	0.51	0.49 (FA) 0.45 (CW) 0.46 (ISO) 0.45 (LPD) 0.46 (GC-TOF) 0.51 (LCMS)	0.55(FA) 0.57 (CW) 0.52 (ISO) 0.43 (LPD) 0.53 (GC-TOF) 0.56 (LCMS)
Average Margin	-0.05	-0.02	-0.03 (FA) 0.05 (CW) 0.03 (ISO) 0.04 (LPD) -0.01 (GC-TOF) -0.05 (LCMS)	-0.11 (FA) -0.21 (WT) -0.05 (ISO) -0.01 (LPD) -0.05 (GC-TOF) -0.18 (LCMS)
# Of metabolites	966 (after missing values imputations)		95 (FA), 62 (CW) 9 (ISO) 153 (LPD) 615 (GC-TOF) 32 (LCMS)	

Conclusions

This study shows that metabolomics platforms can be integrated effectively to form hypotheses about the functions of a gene. For example, in the *oxp1* mutation study, both the reactant and the product of a hypothesized reaction (5-OP and glutamic acid) were detected by different metabolomics platforms (GC-TOF and LC-MS) to generate the biomarker profile of this mutation. The analysis confirms that OXP1 works between 5-OP and glutamic acid in the gamma-glutamyl pathway and may not affect pathways involving fatty acids, lipids, isoprenoids, sterols and cuticular waxes. Lack of any metabolic change in the *ggt2* mutant samples supports the previous evidence that GGT2 may have low expression levels in the leaf tissues. We have explored a cost effective way of putatively identifying structurally unknown metabolites using partial correlation networks across many mutant lines.

List of Tables

1. Table 5.1: Summary of number of metabolites detected by each platform along with the number of structurally known and unknown compounds. Total 1042 metabolites are detected by seven mass spectrometry (MS) based platforms. 498 metabolites have known structures and 544 are unknown metabolites.
2. Table 5.2: Comparison of two data integration methods for *oxp1* mutant. High margins between class votes and low prediction error show that the model was strong and was able to classify the two genotypes. The wild type vs. wild type comparisons show weak models in both integration methods as expected. Similarly, *oxp1* vs. wild type models for all the platforms other than GC-TOF and LCMS also have weak models suggesting that there was no significant effect of this mutation on the metabolites detected by these platforms.
3. Table 5.3: List of biomarkers from the *oxp1* analysis. The first 13 compounds are detected by the GC-TOF platform and the next 11 are detected by LC-MS platform. The 3rd column shows if the metabolite is up or down regulated in the *oxp1* mutant samples. The two right columns show highly correlated compounds to the unknown compound and the BinBase matching compounds and similarity scores from the BinBase database. 5-OP and glutamate are both shortlisted as potential biomarkers using the hierarchical data integration method along with amino acids that are down regulated in the mutant samples. Glutamic acid is missed in batch integration method.
4. Table 5.4: List of biomarkers from the *oxp1* analysis. The first 13 compounds are detected by the GC-TOF platform and the next 11 are detected by LC-MS platform. The 3rd column shows if the metabolite is up or down regulated in the *oxp1* mutant samples. The two right columns show highly correlated compounds to the unknown compound and the BinBase matching compounds and similarity scores from the BinBase database. 5-OP and glutamate are both shortlisted as potential biomarkers using the hierarchical data integration method along with amino acids that are down regulated in the mutant samples. Glutamic acid is missed in batch integration method.
5. Table 5.5: List of biomarkers for *ggt1* mutant from the best performing LC-MS platform. 5-OP does not appear as a potential biomarker in any of the data integration method.
6. Table 5.6: Classification of *ggt2* mutant and wild type samples using batch and hierarchical data integration methods. The error rate and margins are at par with wild type vs. wild type

classification models in both data integration methods, suggesting that this mutation does not cause any biochemical changes in the cell and GGT2 may be a redundant gene.

List of Figures

1. Figure 5.1: Flowchart of data analysis pipeline.
2. Figure 5.2: Log2 ratio plot between *oxp1* mutant and wild type samples (WT1 and WT2) for all the seven platforms generated at www.plantmetabolomics.org. Metabolomics platforms are shown with alternating color bands. The x-axis shows the average log2 ratio between the mutant and wild type samples for each metabolite. The metabolites that are far from the central Y axis have more changes than the ones near the axis. 5-OP is up regulated in mutants along with significant changes in some other metabolites (e.g. 205672, melibiose) that are investigated further.
3. Figure 5.3: GSH degradation pathway in animals and Arabidopsis. Solid lines represent literature supported part of the pathway and the dashed lines represent the hypothesized parts of the pathway. In mammals, GSH is degraded by the sequential reaction of GGT, GGC and OXP1 to yield glutamate as shown in the left figure (part 1A, 1B and 1C). 5-OP to glutamate conversion is verified in Arabidopsis using OXP1 mutant (part 2C) but the action of GGT1 and GGT2 appears to be different as these mutations do not cause any significant changes in 5-OP levels.

Supplementary Documents

1. Appendix– Supplementary Documents: Supplementary Document 8.1 Replicate quality analysis of six replicates of mutant samples of GC-TOF platform and summary of replicate quality analysis for all the other platforms.
2. Appendix – Supplementary Documents: Supplementary Document 8.2: List of Genotypes used for Partial correlation Analysis

Acknowledgments

This project was supported by NSF grant #08200823. The following labs generated the metabolomics data in PlantMetabolomics.org: Oliver Fiehn (UC Davis), B. M. Lange

(Washington State University), Lloyd Sumner (Noble Foundation), Ruth Welti (Kansas State University) , Vladimir Shulaev (Virginia Bioinformatics Institute), and Basil Nikolau (Iowa State University) as part of the Arabidopsis Metabolomics Consortium.

References

Alonso, J. M., A. N. Stepanova, et al. (2003). "Genome-wide insertional mutagenesis of Arabidopsis thaliana." Science **301**(5633): 653-657.

Bais, P., S. M. Moon, et al. (2010). "PlantMetabolomics.org: a web portal for plant metabolomics experiments." Plant Physiol **152**(4): 1807-1816.

Beale, M. H., J. L. Ward, et al. (2009). "Establishing substantial equivalence: metabolomics." Methods Mol Biol **478**: 289-303.

Beckmann, M., D. P. Enot, et al. (2007). "Representation, comparison, and interpretation of metabolome fingerprint data for total composition analysis and quality trait investigation in potato cultivars." J Agric Food Chem **55**(9): 3444-3451.

Bino, R. J., R. D. Hall, et al. (2004). "Potential of metabolomics as a functional genomics tool." Trends Plant Sci **9**(9): 418-425.

Catala, R., E. Santos, et al. (2003). "Mutations in the Ca2+/H+ transporter CAX1 increase CBF/DREB1 expression and the cold-acclimation response in Arabidopsis." Plant Cell **15**(12): 2940-2951.

de la Fuente, A., N. Bing, et al. (2004). "Discovery of meaningful associations in genomic data using partial correlation coefficients." Bioinformatics **20**(18): 3565-3574.

Efron B , T. R. (1997). "Improvements on cross-validation: the .632+ bootstrap method." J American Statistical Association(92:548-560).

Enot, D. P., M. Beckmann, et al. (2006). "Predicting interpretability of metabolome models based on behavior, putative identity, and biological relevance of explanatory signals." Proc Natl Acad Sci U S A **103**(40): 14865-14870.

Fiehn, O., Robertson D, Griffin J, van der Werf M, Nikolau B, Morrison N, Sumner LW, Goodacre R, Hardy NW, Taylor C, Fostel J, Kristal B, Kaddurah-Daouk R, Mendes P, van Ommen B, Lindon JC, Sansone S-A (2007). "The metabolomics standards initiative (MSI)." Metabolomics **3**: 175-178.

Fiehn, O., Sumner LW, Rhee SY, Ward J, Dickerson J, Lange BM, Lane G, Roessner U, Last R, Nikolau B (2007). "Minimum reporting standards for plant biology context information in metabolomics studies." Metabolomics **3**: 195-201.

Fiehn, O. (2008). "Extending the breadth of metabolite profiling by gas chromatography coupled to mass spectrometry." Trends Analyt Chem **27**(3): 261-269.

Fiehn, O., J. Kopka, et al. (2000). "Metabolite profiling for plant functional genomics." Nat Biotechnol **18**(11): 1157-1161.

Fiehn, O., G. Wohlgemuth, et al. (2005). "Setup and Annotation of Metabolomic

Experiments by Intergrating Biological and Mass Spectrometeric Metadata." Proceedings Lecture Notes Bioinformatics **3615**: 224-239.

Fiehn, O., G. Wohlgemuth, et al. (2008). "Quality control for plant metabolomics: reporting MSI-compliant studies." Plant J **53**(4): 691-704.

Forde, B. G. and P. J. Lea (2007). "Glutamate in plants: metabolism, regulation, and signalling." J Exp Bot **58**(9): 2339-2358.

Hall, R., M. Beale, et al. (2002). "Plant metabolomics: the missing link in functional genomics strategies." Plant Cell **14**(7): 1437-1440.

Johnson, C. H., A. D. Patterson, et al. (2011). "Radiation Metabolomics. 4. UPLC-ESI-QTOFMS-Based Metabolomics for Urinary Biomarker Discovery in Gamma-Irradiated Rats." Radiat Res.

Kanehisa, M., S. Goto, et al. (2004). "The KEGG resource for deciphering the genome." Nucl. Acids Res. **32**(90001): D277-280.

Krumsiek, J., K. Suhre, et al. (2011). "Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data." BMC Syst Biol **5**: 21.

Lanz, C., A. D. Patterson, et al. (2009). "Radiation metabolomics. 3. Biomarker discovery in the urine of gamma-irradiated rats using a simplified metabolomics protocol of gas chromatography-mass spectrometry combined with random forests machine learning algorithm." Radiat Res **172**(2): 198-212.

Lee do, Y. and O. Fiehn (2008). "High quality metabolomic data for Chlamydomonas reinhardtii." Plant Methods **4**: 7.

Lin, P., J. Li, et al. (2008). "A missense mutation in SLC33A1, which encodes the acetyl-CoA transporter, causes autosomal-dominant spastic paraplegia (SPG42)." Am J Hum Genet **83**(6): 752-759.

Ling, Q. D., F. C. Chang, et al. (2006). "Synthesis and dynamic random access memory behavior of a functional polyimide." J Am Chem Soc **128**(27): 8732-8733.

Mackey, D., Y. Belkhadir, et al. (2003). "Arabidopsis RIN4 is a target of the type III virulence effector AvrRpt2 and modulates RPS2-mediated resistance." Cell **112**(3): 379-389.

Molinaro, A. M., R. Simon, et al. (2005). "Prediction error estimation: a comparison of resampling methods." Bioinformatics **21**(15): 3301-3307.

Monte, E., J. M. Alonso, et al. (2003). "Isolation and characterization of phyC mutants in Arabidopsis reveals complex crosstalk between phytochrome signaling pathways." Plant Cell **15**(9): 1962-1980.

Mueller, L. A., P. Zhang, et al. (2003). "AraCyc: a biochemical pathway database for

Arabidopsis." Plant Physiol **132**(2): 453-460.

Ohkama-Ohtsu, N., A. Oikawa, et al. (2008). "A gamma-glutamyl transpeptidase-independent pathway of glutathione catabolism to glutamate via 5-oxoproline in Arabidopsis." Plant Physiol **148**(3): 1603-1613.

Ohkama-Ohtsu, N., S. Radwan, et al. (2007). "Characterization of the extracellular gamma-glutamyl transpeptidases, GGT1 and GGT2, in Arabidopsis." Plant J **49**(5): 865-877.

Ohkama-Ohtsu, N., Y. Sasaki-Sekimoto, et al. (2011). "12-oxo-phytodienoic acid-glutathione conjugate is transported into the vacuole in Arabidopsis." Plant Cell Physiol **52**(1): 205-209.

Ohkama-Ohtsu, N. and J. Wasaki (2010). "Recent progress in plant nutrition research: cross-talk between nutrients, plant physiology and soil microorganisms." Plant Cell Physiol **51**(8): 1255-1264.

Ohkama-Ohtsu, N., P. Zhao, et al. (2007). "Glutathione conjugates in the vacuole are degraded by gamma-glutamyl transpeptidase GGT3 in Arabidopsis." Plant J **49**(5): 878-888.

Patterson, A. D., C. Lanz, et al. (2010). "The role of mass spectrometry-based metabolomics in medical countermeasures against radiation." Mass Spectrom Rev **29**(3): 503-521.

Schafer, J. and K. Strimmer (2005). "An empirical Bayes approach to inferring large-scale gene association networks." Bioinformatics **21**(6): 754-764.

Schafer, J. and K. Strimmer (2005). "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics." Stat Appl Genet Mol Biol **4**: Article32.

Scott, I. M., C. P. Vermeer, et al. (2010). "Enhancement of plant metabolite fingerprinting by machine learning." Plant Physiol **153**(4): 1506-1520.

Shannon, P., A. Markiel, et al. (2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks." Genome Res **13**(11): 2498-2504.

Smilde, A. K., M. J. van der Werf, et al. (2005). "Fusion of mass spectrometry-based metabolomics data." Anal Chem **77**(20): 6729-6736.

Stephen E. Stein , D. R. S. (1994). "Optimization and testing of mass spectral library search algorithms for compound identification." Journal of the American Society for Mass Spectrometry **5**(9).

Steuer, R., J. Kurths, et al. (2003). "Interpreting correlations in metabolomic networks." Biochem Soc Trans **31**(Pt 6): 1476-1478.

Steuer, R., J. Kurths, et al. (2003). "Observing and interpreting correlations in metabolomic networks." Bioinformatics **19**(8): 1019-1026.

Svetnik, V., A. Liaw, et al. (2003). "Random forest: a classification and regression tool for compound classification and QSAR modeling." J Chem Inf Comput Sci **43**(6): 1947-1958.

Ullah, H., J. G. Chen, et al. (2003). "The beta-subunit of the Arabidopsis G protein negatively regulates auxin-induced cell division and affects multiple developmental processes." Plant Cell **15**(2): 393-409.

van den Berg, R. A., H. C. Hoefsloot, et al. (2006). "Centering, scaling, and transformations: improving the biological information content of metabolomics data." BMC Genomics **7**: 142.

Van der Werf, P., M. Orlowski, et al. (1971). "Enzymatic conversion of 5-oxo-L-proline (L-pyrrolidone carboxylate) to L-glutamate coupled with cleavage of adenosine triphosphate to adenosine diphosphate, a reaction in the -glutamyl cycle." Proc Natl Acad Sci U S A **68**(12): 2982-2985.

Zhang, P., H. Foerster, et al. (2005). "MetaCyc and AraCyc. Metabolic pathway databases for plant research." Plant Physiol **138**(1): 27-37.

CHAPTER 6. PARTIAL CORRELATION NETWORKS TO PUTATIVELY INDENTIFY UNKNOWN METABOLITES IN NON-TARGETED METABOLOMICS

A paper to be submitted to Plant Methods

Preeti Bais, Basil Nikolau, Julie A. Dickerson

Abstract:

About a third of the total number of the genes in Arabidopsis cannot be functionally annotated using sequence genomics techniques alone. Comparing the biochemical signature of samples from single gene knock outs with the wild type samples can provide clues to the functions of that gene. We individually compared 70 single gene knock outs of Arabidopsis mutants with wild type samples to find the potential biomarkers of each mutant. The data was measured using non targeted gas chromatography mass spectrometry (GC-TOF) platform to get a global coverage of metabolite families. The Random Forest machine learning algorithm was used for classification of mutant samples from the wild type samples and selection of potential biomarkers for each mutant.

Structurally unknown metabolites comprise of a big portion of any larger scale non-targeted metabolomics analysis. These metabolites do not provide any biological information unless the structures are determined using expensive and time consuming methods like tandem mass spectrometry or NMR. We propose a new method of computationally determining the putative identification of structurally unknown metabolites using partial correlation networks across many genetic strains. A researcher can use our biomarker database to find the potential biomarkers of a gene (mutant) of interest, get putative identifications of all unknown metabolites, visualize the impacted pathways and find other mutants that have similar biomarker profiles. They can also

use integrated data mining tools to see the mutant's behavior using targeted platforms from our plantmetabolomics database for a more detailed analysis of the function of the gene along with the integrated morphological data to see the physical changes in plants and seeds.

Introduction

Metabolomics, which measures the concentration of small molecules (molecular weight < 1000 Da) can be used in finding the functions of genes when sequence genomics techniques alone are not adequate. Non-targeted metabolomics analysis performs a global analysis of all the metabolites in an organism without any previous knowledge on the chemical and physical properties of the metabolites. In a targeted metabolomics, small subset of known metabolites is enriched prior to the analysis increasing the sensitivity of analysis. Since the non-targeted analysis is de novo in nature, there is large number of detected metabolites where chemical and physical properties cannot be determined without using more expensive and time consuming methods like NMR or tandem mass spectrometry. These metabolites cannot be used in finding the biological relevance of the factor under study without the structure determination.

We propose a new method of putatively identifying the key unknown metabolites using partial correlation analysis across a large number of genetic strains. Unlike transcriptomic and proteomic correlation analysis, the observation of correlations in metabolites from the previous studies has not been able to identify known biochemical topologies using the standard correlation analysis with Pearson's and Spearman's correlations. Previous studies (Camacho et al. 2005, Steur R. 2006) have suggested various factors that contribute to the high correlations including (1) chemical equilibrium where metabolites reaching equilibrium show high positive correlations, (2) mass conservation – in a moiety conserved cycle at least one metabolite is negatively correlated to the rest of the group, (3) asymmetric control distribution – where intrinsic fluctuations a parameter that controls two metabolites causes high correlations among the metabolites (e.g. diurnal cycle) , and (4) unusually high variance in the expression of a

single gene. For example, a single enzyme that carries a high variance induces negative correlations between its substrate and product metabolites. A review by Steuer R. (Steuer R. 2006) pointed out several issues with the previous correlation studies including the use a predefined correlation threshold and tendency of the nodes that have a common neighbor to be identified as connected in a Pearson's correlation analysis. The author suggested using a partial correlation analysis on larger dataset across different genetic strains (or experimental conditions) to systematically identify the preserved correlations and thus detect the stable features or topologies of the underlying biochemical system.

In a recent study on lipid metabolomics (Krumsiek J. et al. 2011), the authors have shown that partial correlation networks are similar to the actual biochemical while zero order person's correlation networks are not. However, high order correlation networks have not been applied as often in metabolomics studies before because the number of features is much higher than the number of samples. In the present study, we took advantage of our in house data set which was generated using 194 structurally known metabolites across 503 samples. Using only the known metabolite, we generated first order correlation network algorithm GGM and show that the resulting networks are much sparser and match the actual biochemical pathways very well as the indirect correlations are removed. After showing that the highly correlated compounds from the partial correlation analysis are also neighbors in the actual biochemical networks, we use this strategy to find the neighbors of unknown metabolites .If those neighbor metabolites have known structures, we hypothesize that the unknown metabolite may belong to the same biochemical pathways as the known metabolites to have survived their correlations across so many mutations. This hypothesis can be verified with manually checking the raw chromatograms with our links to the GC-TOF library Binbase (Fiehn O. 2004). This method thus provides putative identifications of key unknown metabolites with the existing data and without the need for more expensive and time consuming methods.

Materials and methods

Plant Materials

Genetic parameters were manipulated by using Arabidopsis stocks that contained T-DNA insertions in a single gene and the environmental parameters were kept constant (Bais P. et al. 2010). A total 0f 70 different gene mutants (six samples of each mutant) were compared with six different sets of wild type plants (with 6 samples in each set). The list of mutants is shown in table 6.1.

Non targeted GC-TOF analysis

The metabolites were detected using gas chromatography-mass spectrometry (GC-MS) platform and a total of 614 metabolites were detected for each genotype. 194 metabolites were structurally known and the rest were unknown metabolites. All the metadata about the Plant growth conditions, extraction protocols, mass spectrometry, instruments are available along with the data at the www.plantmetabolomics.org (Bais et al. 2010). The total number of samples in the dataset was 503.

Data Analysis

Data Normalization

In a biological system, most of the metabolites are found in low abundance and only a small number of metabolites are in high abundance. Mean centering and range scaling methods were used to normalize metabolomics data. Range scaling uses biological range as the scaling factor (Dieterle et al. 2006, van den Berg et al. 2006).

Random Forest Classification

Random Forest (RF) machine learning method (Breiman 2001) has been used in transcriptomics and metabolomics studies for classification and biomarker discovery in recent years as it performs well when the number of samples are much lower than the number of features (Enot, 2006, Diaz-Uriate 2007). The selected features are not transformed as in PCA analysis and can be used directly in the biological interpretation and hypothesis generation. One of the main goals of a biomarker study is to find biomarkers that are not only able to separate the two classes of samples under study clearly but is also reproducible. 25 bootstrap runs using .632+ bootstrap method (Efron B. et al. 1997) were used in classification procedure to generate reproducible results. A base line significance of the models was generated using six sets of wild type samples (with six replicates in each set) with each other and then comparing the wild type samples across the batches. Previous studies on RF (Enot D. et al. 2006) have suggested using class margins as criteria for accessing model quality along with the normally used classification accuracy measure. Bigger margins in class votes show high confidence in the votes for the right class and thus clear separation between the classes. Margins were calculated by subtracting the votes to the wrong class from the votes to the right class.

Correlation Network Analysis

Gaussian graphical models (GGMs), remove the indirect associations by conditioning the pairwise correlation among two variables on the correlations with all the other variables. A GGM is an undirected graph in which each edge represents the pairwise correlation between two variables conditioned against the correlations with all other variables. The GGM networks were constructed in a three step process by first constructing the Pearson's correlation matrix, then calculating the partial correlations. Finally false discovery rate calculations were employed to remove the insignificant correlations. The metabolites were represented as vertices and the non-zero correlations between them were represented as edges in a graph.

R package “GeneNet” (Schafer et al. 2005 a, 2005 b) was used in analyzing the graphs in the GGM method. Top pairwise correlations were chosen using a q value of 0.05. The network analysis was performed using R package igraph (Csardi G. et al. 2006).

Results and Discussion

RF classification and biomarker selection

RF classification results of the same genotype (e.g. wild type VS. wild type model) had low average margins and high bootstrap (.632+) estimate of prediction error as expected because the metabolic signatures from the same genotype were expected to be very similar to each other under stable environment conditions. The error rate and average margins for the pairwise mutant vs. wild type RF classifications are shown in figure 6.1. This figure helps in finding out which mutant cause more biochemical changes and which mutants cause insignificant changes. We have recently shown that *ggt2* mutant may have low expression levels in leaf tissues and as shown in figure 6.1, this mutant’s classification with wild type samples is at par with wild type vs. wild type comparisons. Classification results from another mutant BCCP2 show high margins and low error rate in prediction and are investigated further.

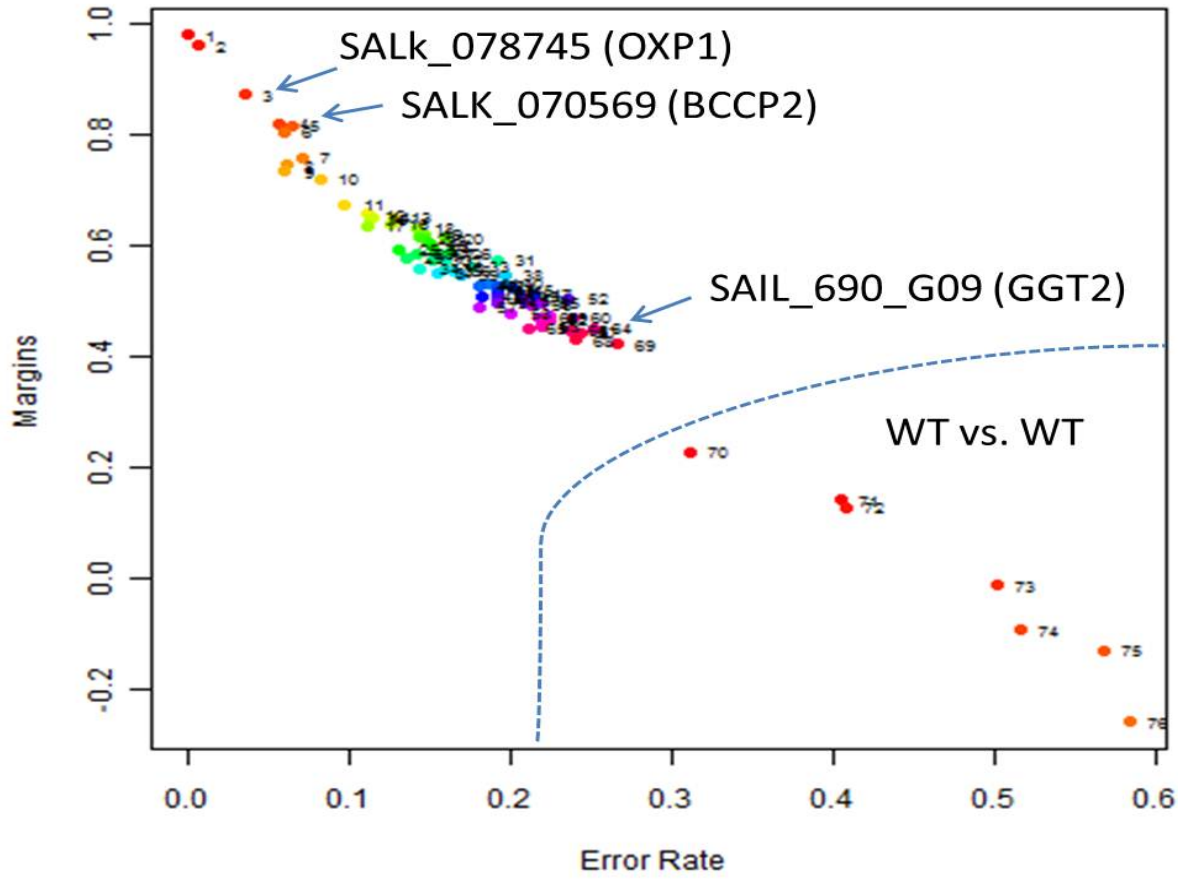

Figure 6-1 **RF classification of 70 mutant lines of Arabidopsis with wild type samples**

GGM Networks

In the first round, GGM networks were built using the 194 known metabolites alone from the 70 mutation lines using 503 samples to show that the correlation networks are similar to actual biochemical networks. The top 20 metabolite pairs with highest partial correlations are shown in table 6.1. Almost all the highest correlated metabolites are either from the same Aracyc pathway or share a common structure which can explain the high correlation coefficient between them. For example, mannonic acid NIST and gluconic acid which have similar structure in CHEBI database show a high correlation. Similarly, ornithine and N-acetylornithine which belong to the same Aracyc pathway “arginine biosynthesis II (acetyl cycle)” also show high correlation.

In the second round of correlation analysis, top biomarkers with unknown structures from the RF study described above for the entire 70 mutant vs. wild type classification analyses were picked for further investigation. The unknown metabolites that appeared in at least 75% bootstrap runs for any of the 70 mutants were selected along with all the 194 known metabolites for the GGM analysis because the goal was to find a highly correlated known metabolite for a key unknown metabolite.

Table 6-1 Top 20 highly correlated compounds from GGM networks

#	Metabolite 1	Metabolite 2	pcor	Comment
1	GABA	phosphoric acid	-0.77	
2	lignoceric acid	palatinose	0.74	
3	methionine sulfoxide	xylonolactone	0.69	
4	ornithine	N-acetylornithine	0.52	arginine biosynthesis II (acetyl cycle)
5	xylose	arabitol	0.52	Sugar and Sugar Alcohol
6	melibiose	digalacturonic acid	0.51	Sugar and Sugar Acid
7	malic acid	2-hydroxyglutaric acid	0.49	
8	stearic acid	heptadecanoic acid	0.46	Fatty acid family
9	GABA	oxoproline	-0.44	γ-glutamyl cycle :
10	octadecanol	1-hexadecanol	0.43	Both Alcohol
11	ornithine	citrulline	0.43	arginine biosynthesis II (acetyl cycle)
12	oxoproline	phosphoric acid	-0.42	γ-glutamyl cycle :
13	mannonic acid NIST	gluconic acid	0.41	Similar structure in CHEBI
14	phosphoethanolamine	adipic acid	-0.41	
15	glycine	aminomalonic acid	0.40	aminomalonic acid as an intermediate in the metabolic change of serine to glycine
16	phosphoethanolamine	3-ureidopropionate	0.39	
17	pelargonic acid	benzoic acid	0.39	Both are carboxylic acids
18	cis-sinapinic acid	cysteine-glycine	0.39	http://www.ncbi.nlm.nih.gov/pubmed/20188588
19	GABA	galactose	-0.38	glutamate degradation IV-Pyruvate-Glycolysis
20	conduritol-beta-epoxide	inositol allo-	0.38	EC 3.2.1.21 - beta-glucosidase bound as an ester of (+) chiro-inositol to aspartic acid

Table 6-2 RF classification results of SALK_070569 vs. wild type from 25 Bootstrap runs

Potential Biomarker	Frequency in bootstrap models	Up or Down in Mutant?	Top 3 Metabolites in partial correlation network	Top 3 Similar compounds from Binbase database	Other mutants with same biomarkers
ribitol	0.84	Up			SALK_022584C SALK_022584C SALK_151779C SALK_126891C ELO 2/4 SALK_040250 SALK_062847C SALK_110264
201060	0.28	Up	erythritol glycine ribitol threitol	erythritol, ribitol, xylitol 563	SALK_022584C SALK_022584C ELO 2/3/4 SALK_151779C SALK_022971C SALK_075185 ELO 2/3 SALK_053394 SALK_040250 SALK_009522 SALK_094382
211896	0.24	Up	glucoheptulose	cellobiotol	SALK_094382 SALK_110264
xylitol	0.12	Up			SALK_110264
arabinose	0.08	Up			SALK_110264 SALK_008656 SALK_008505 SALK_053394 SALK_040250
ornithine	0.08	Up			SALK_040250 SALK_021437 SALK_097354C ELO 2/4
212373	0.08	Up	allantoin.5TMS arabitol cysteine octadecanol ribitol	leucrose, glycerol-3-galactoside	ELO 2/4 SALK_040250
228911	0.08	Up	arabitol N.acetyl.D.hexosamine ribonic.acid	fructose 1 (sorbose 1) tagatose 1 fructose 2 (sorbose 2)	SALK_040250

Case Study and examples

A mutant SALK_070569 (biotin carboxyl-carrier protein ,BCCP2, cac1b-2 allele) has been shown to have no perceptible effect on plant growth, development, and fatty acid accumulation using morphological data and reverse genetic approaches in a recent study (Xu L et al. 2011). Our analysis for this mutant showed high class margins and low error rates as shown on top left portion of figure 6.1 (Point # 4). The potential biomarkers of this mutant did not contain any fatty acids or lipids confirming the literature evidence (Xu L et al. 2011). Table 6.2 shows the biomarker results for this mutant along with the top correlated compounds from the partial correlation analysis for the unknown metabolites. The results from table 6.2 show that the alcohols or the unknown compounds that are highly correlated with alcohols may be potential biomarkers for this mutation. The chromatogram analysis from Binbase database provides further evidence that the unknown metabolites may be alcohols and not fatty acids. The partial correlation analysis provides putative identification for the unknown compounds which match the observations from Binbase database.

The right most column of this table shows other mutants with similar potential biomarkers. An examination of the list shows that BCCP2 shares 4 biomarkers with mutant SALK_110264, which is homomeric acetyl-CoA carboxylase gene (ACC2) as per the literature evidence (Babiychuk E. et al. 2011). This information is relevant because both the studies point out functional redundancy in malonyl-CoA biosynthesis with these two mutants which are also supported by our results. The bottom half of the page shows all the affected Aracyc pathways for BCCP2 and includes many amino acid biosynthesis pathways.

Conclusions

We have created a database of metabolomics based potential biomarkers for 70 mutants in Arabidopsis that have not been annotated using sequence genomics techniques till date. We have also shown that the most correlated pairs of metabolites from the GGM analysis are either part of the same pathway or share structure similarity among each other and propose a new way to utilize this knowledge for annotating the unknown metabolites. A researcher can use the biomarker database to find the potential biomarkers for a gene of interest. They can then find putative identifications for the unknown metabolites in the biomarker list. All the other mutants that share those biomarkers are also made available along with the potentially impacted pathways that may be affected with the known metabolites and the using the putative identification of the unknown metabolites from the biomarkers list. Combining these analyses will help the users in finding the biological impact of a mutation and thus lead to hypothesis generation about the function of a gene. The users can also use the integrated morphological database to see the morphological changes in the plants and use data mining tools at the integrated plantmetabolomics database (www.plantmetabolomics.org) to analyze data from other targeted platforms like fatty acid, lipidomics, phytoesterols, isoprenoids , and cuticular wax from our consortium to further analyze the impact of the mutants. Detailed instructions and a case study with step by step process on using this tool and database is provided at the supplimentary document 8.6.

Availability of database

The biomarker database is integrated with our existing plant metabolomics database at www. Plantmetabolomics.org. A user can search for a mutant (or gene) from the home page and then follow the easy directions to see the biomarker models for the give mutant. A detailed use case is also provided in the Appendix – Supplementary Documents 8.6.

List of Tables

1. Table 6.1: List of 20 most highly correlated metabolite pairs from the GGM analysis. The two metabolites in each pair are either part of the same pathway or have structural similarity with each other.
2. Table 6.2: Case study of SALK_070569 (BCCP2) mutant. The frequency of a metabolite in 25 bootstrap runs is shown in the second column. The putative identifications for the unknown metabolites using partial correlation analysis are shown in the third column. The second column link s to the Binbase database and shows ten compounds that have most similar chromatograms for the queried metabolite. The last column lists other mutants with similar biomarker profiles.

List of Figures

1. Figure 6.1: Classification of 70 single gene knock out mutants are shown using the margin and predictive error from running the 25 bootstrap runs for each mutant. The wild type samples from within a batch and different batches show high error rates and negative margins. The mutants that have more impact on the biochemical profile have lower error rate and higher class margins and the mutants that have little impact have higher error rate and smaller class margins as shown in this graph.

Supplementary Documents

1. Supplementary Document 8.2: List of Genotypes used for Partial correlation Analysis
2. Supplementary Document 8.3 : Use case study on how to use the biomarker database

Acknowledgments

This project was supported by NSF grant #08200823.

References

Alonso, J.M., Stepanova, A.N., Leisse, T.J. et al. (2003) Genomewide insertional mutagenesis of Arabidopsis thaliana. Science, 301, 653–657.

Bais, P., S. M. Moon, et al. (2010). "PlantMetabolomics.org: a web portal for plant metabolomics experiments." Plant Physiol 152(4): 1807-1816.

Babiychuk E, Vandepoele K, Wissing J, Garcia-Diaz M, De Rycke R, Akbari H, Joubès J, Beeckman T, Jänsch L, Frentzen M, Van Montagu MC, Kushnir S. "Plastid gene expression and plant development require a plastidic protein of the mitochondrial transcription termination factor family" Proc Natl Acad Sci U S A. 2011 Apr 19;108(16):6674-9. Epub 2011 Apr 4.

Camacho D, de la Fuente A, Mendes P. The origin of correlations in metabolomics data. Metabolomics 2005;1:53–63.

Csardi G, Nepusz T (2006): "The igraph software package for complex network research", InterJournal, Complex Systems 1695. 2006. http://igraph.sf.net

Efron B, Tibshirani RJ: Improvements on cross-validation: the .632+ bootstrap method. *J American Statistical Association* 1997, **92:**548-560.

Fiehn, O., J. Kopka, et al. (2000). "Metabolite profiling for plant functional genomics." Nat Biotechnol **18**(11): 1157-1161.

Fiehn, O., G. Wohlgemuth, et al. (2005). "Setup and Annotation of Metabolomic Experiments by Intergrating Biological and Mass Spectrometeric Metadata." Proceedings Lecture Notes Bioinformatics **3615**: 224-239.

Gao, J., V. G. Tarcea, et al. (2010). "Metscape: a Cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks." Bioinformatics **26**(7): 971-973.

Ghassemian, M., J. Lutes, et al. (2006). "Integrative analysis of transcript and metabolite profiling data sets to evaluate the regulation of biochemical pathways during photomorphogenesis." Arch Biochem Biophys **448**(1-2): 45-59.

Hall, R., M. Beale, et al. (2002). "Plant metabolomics: the missing link in functional genomics strategies." Plant Cell **14**(7): 1437-1440.

Krumsiek J. et al. (2011) "Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data"BMC Systems Biology 2011, 5:21 doi:10.1186/1752-0509-5-21

Ohkama-Ohtsu, N., A. Oikawa, et al. (2008). "A gamma-glutamyl transpeptidase-independent pathway of glutathione catabolism to glutamate via 5-oxoproline in Arabidopsis." Plant Physiol **148**(3): 1603-1613.

Pearson, K. (1901). "On Lines and Planes of Closest Fit to Systems of Points in Space" (PDF).

Philosophical Magazine 2 (6): 559–572. http://stat.smmu.edu.cn/history/pearson1901.pdf

Reverter,A. and Chan,E.K.F. (2008) Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks.

Bioinformatics, 24, 2491–2497.

Shannon, P., A. Markiel, et al. (2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks." Genome Res 13(11): 2498-2504.

Spearman C.(1904)."The proof and measurement of association between two things" Amer. J. Psychol. 15 pp. 72–101

Steuer R, Kurths J, Fiehn O, Weckwerth W. Observing and interpreting correlations in

metabolomic networks.Bioinformatics 2003;19:1019–26.

Steuer R, Kurths J, Fiehn O, Weckwerth W. Interpreting correlations in metabolomic

networks. Biochem Soc Trans 2003;31:1476–8.

Van der Werf, P., M. Orlowski, et al. (1971). "Enzymatic conversion of 5-oxo-L-proline (L-pyrrolidone carboxylate) to L-glutamate coupled with cleavage of adenosine triphosphate to adenosine diphosphate, a reaction in the -glutamyl cycle." Proc Natl Acad Sci U S A 68(12): 2982-2985..

Xu L, Ilarslan H, Qian H-R, Li L, Che P, Wurtele ES, Nikolau BJ. 2011. Reverse genetic analysis of the two biotin-containing subunit genes of the heteromeric acetyl-CoA carboxylase indicates a unidirectional functional redundancy. ***Plant Physiology*** 155: 293-314.

CHAPTER 7. CONCLUSIONS

We have shown that metabolomics can be a very effective functional genomics tool. The database contains metabolomics concentration and morphological data from 140 novel Arabidopsis mutant lines and 1400 metabolites. A researcher interested in any of these mutant lines or metabolites can easily use our web based tools to visualize and perform data mining and generate lists of biomarkers or find if two or more genes have similar metabolic profiles and thus form hypothesis about the functions of those genes. They can also investigate which family of metabolites are affected most or not affected at all due to a mutation and concentrate their future efforts accordingly.

One of the biggest challenges in any metabolomics experiments is to incorporate structurally unknown metabolites and it is often very expensive to find the structures of all the unknown metabolites using NMR and other techniques. In this thesis, we have presented a novel way to incorporate more information from the unknown metabolites by using novel computational methods on the existing data. These methods not only prioritize the unknown metabolites for the future structure determination studies but also provide more insight into their possible structures.

We show that our methods not only confirm the known biology on GSH degradation pathway from Arabidopsis but this pathway in plants may be different than mammals which can be investigated further. Finally we present correlation networks of Arabidopsis that are built across 70 different mutant lines and show stable relationships between the metabolites from non-targeted GC-TOF platform. Many of the strong relationships from our analysis match very well with the known biochemistry and the some of the novel findings from our analysis can be investigated further to see if the strong relationships can guide us to novel pathways between those metabolites.

We have presented a web based relational database that can be easily adapted to mass spectrometry (MS) based plant metabolomics data from other plant species.

In summary, we hope that this work can provide the research community with new tools to incorporate more knowledge from mass spectrometry based metabolomics and help establish metabolomics as a functional genomics tool.

CHAPTER 8. APPENDIX – SUPPLEMENTARY DOCUMENTS

Supplementary Document 8.1: Data Quality analysis

Replicate analysis for GC-TOF-MS Data for oxp1 Mutant

This figure is used in replicate reliability analysis of an experiment. The bottom left corner of the figure shows the scatterplot matrix for different replicates. For an experiment with n replicates, n of these scatter plots will be drawn. Scatterplot on the ith row and the jth column corresponds to the replicates i and j. Each point in a scatterplot depicts a metabolite and its x and y coordinates are based on its concentration under two corresponding replicates. The experiments which have most of the points around the central diagonal in scatter plot are considered to be good because it shows that the experiment had similar results in all the runs. The numbers in the upper right corner show Spearman's correlation between two replicates. High numbers in this area indicate higher correlation between the replicates and thus high reliability between the replicates. The central figures show the distribution of metabolic concentrations in a replicate. Bell shaped figures depict that the metabolite concentrations are normally distributed and are considered to be good.

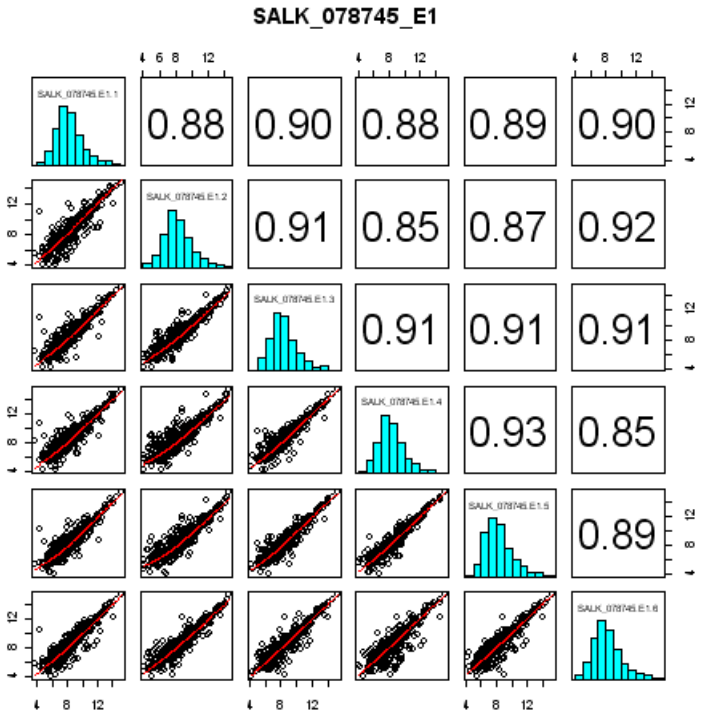

Figure 8-1 **Replicate quality analysis**

Table 8-1 Summary of Replicate quality analysis of platforms

Analytical Platform	*Spearman's Correlation Range*						
	oxp1	**WT1**	**WT2**	***ggt1***	***ggt2***	**WT1**	**WT2**
Fatty Acids	52 - 85	57-91	23-85	71-90	69-91	58-87	53-94
Cuticle Wax	61-79	56-86	57-85	41-90	72-88	62-89	52-83
Phytosterols	90-98*	89-98*	70-98*	38-94**	23-99**	22-99**	87-98**
Isoprenoids	85-1.00	71-97	95-100	95-100	93-100	95-100	97-100
Lipidomics	94-97	93-98	93-98	98-99	96-98	98-99	97-99
GC-TOF-MS	85-91	84-94	81-94	86-94	86-91	86-91	86-93
UPLC-Q-TOF	85-88	85-89	84-88	87-91	84-88	87-90	86-90

Range of Spearman's correlation between any two replicates of a single genotype for each analytical platform. Higher correlation number depicts high repeatability of experiments. Since different sets of WT1 and WT2 samples were used for comparison with the *ggt1* and *ggt2* than for comparison with the *oxp1* mutants, we have shown 4 different WT samples here.

* Metabolite "BML-PS2-14.180-472.5" was removed before the analysis because it had two negative values.

** 3 metabolites namely "BML-PS2-16.643-407.4", "BML-PS2-14.180-472.5" and "phytol" from the Phytosterols platform were removed because some of the data contained negative values.

Note: Phytosterol platform had some metabolites with negative abundance values because the intensity of a signature ion was lower in a sample than in the corresponding blank run and since

the platform protocol required subtracting the background at any given retention time and any given signature ion from the corresponding value of the sample (Personal communication with Dr. B. Lange), this resulted in negative values.

Supplementary Document 8.2: List Of Genotypes used for Partial Correlation Analysis

Table 8-2 List of Genotypes used for Partial correlation Analysis

#	Name	Genotype	No. of Samples	#	Name	Genotype	No. of Samples
1	SALK_105337C	A1g71697	6	36	SALK_009731	At5g16370	6
2	SALK_021437	At1g10670	6	37	SALK_074697	At5g19070	6
3	SALK_011304C	At1g12880	6	38	SALK_078745	At5g37830	6
4	SALK_137479C	At1g17160	6	39	SALK_112932	At5g37930	6
5	SALK_072982	At1g34350	6	40	SALK_130472C	At5g38460	6
6	SALK_110264	At1g36180	6	41	SALK_026454	At5g40010	6
7	SALK_133954C	At1g49820	6	42	SALK_093747C	At5g43380	6
8	SALK_013989	At1g50460	6	43	SALK_000892	At5g45300	6
9	SALK_126003C	At1g60230	6	44	SALK_008505	At5g47720	6
10	GABI_751B10	At1g65890	6	45	SALK_078710	At5g51420	6
11	SALK_092412	At1g71890	6	46	SALK_064795C	At5g53580	6
12	SALK_083029	At1g75000	6	47	SALK_025686C	At5g66550	6
13	SALK_008656	At1g76130	5	48	SAIL_690_G09	At1g03260	6
14	SALK_044892	At2g01170	6	49	SALK_040250	At1g07990	6
15	SALK_121515C	At2g03510	6	50	SALK_067448	At1g09430	6
16	SALK_032878	At2g17650	6	51	SALK_022584C	At1g22430	6
17	SALK_143417C	At2g27490	6	52	SALK_067463C	At1g35710	6
18	SALK_090658C	At2g35020	6	53	SALK_097354C	At1g58030	6
19	SALK_022971C	At2g39670	6	54	SALK_009522	At1g75960	6
20	SALK_063167	At2g42560	6	55	ELO 2/4	At3g06470/At1g75000	6

21	SALK_014489C	At2g46180	6	56	ELO 2/3/4	At3g06470/At4g36830/At1g75000	6
22	SALK_109405	At3g06470	6	57	ELO 2/3	At3g06470/At4g36830	6
23	SALK_003718	At3g16950	6	58	SALK_062847C	At4g11100	6
24	SALK_112040C	At3g19630	6	59	SALK_016312C	At4g14930	6
25	SALK_062081C	At3g49310	6	60	SALK_024747C	At4g22890	6
26	SALK_151779C	At3g52750	6	61	SALK_094382	At4g25000	6
27	SALK_000817	At3g56130	6	62	ELO 3/4	At4g36830/At1g75000	6
28	SALK_126891C	At4g08350	6	63	SALK_130673C	At4g39520	6
29	SALK_073183	At4g22880	6	64	SALK_137317	At5g01300	6
30	SALK_092408	At4g29540	6	65	SALK_053394	At5g07990	6
31	SALK_075185	At4g36830	6	66	SALK_020583	At5g13930	6
32	SALK_004694	At4g39640	6	67	SALK_114456C	At5g20080	6
33	SAIL_6_G02	At4g39650	6	68	aae17/8	At5g23050/At1g55320	6
34	SALK_090101C	At5g08120	6	69	SALK_083600C	At5g61790	6
35	SALK_070569	At5g15530	6	70	*SALK_021108*	At1g52670	36
71	Wild type	Col-0	54				
Total	**263**	**Grand Total**	**503**	**Total**			**240**
				Grand Total			**503**

Supplementary Document 8.3: Applications of biomarker database and putative identification of unknown metabolites: Case Study

Case study: Find the biomarkers of a mutation of interest

Scenario: An investigator is interested in how the metabolome changes when comparing mutant and wild-type samples for a mutant of interest. Specifically, the investigator wants to know which metabolites show a significant change and which biochemical pathways they are involved in. The investigator is also interested in obtaining detailed information about specific metabolites from other web sources, other metabolites in the relevant biological pathways, and all the metadata associated with the selected mutant sample.

For example, the investigator is interested in the metabolome changes for mutant "SALK_053394", which has a T-DNA mutation in the *Arabidopsis* gene "At5g099970".Below is a detailed description of a possible analysis path. Please note that analyses do not need to be done in this order. Help icons are located throughout the database to aid users in understanding the tools available at PM.

In the main web page, www.plantmetabolomics.org, the investigator selects the gene of interest from a drop down menu and clicks on "Go" button. The resulting page shows a ratio plot of that mutant compared with all the wild type samples from the same experimental batch. The investigator can visualize which metabolites have more significant fold changes from the wild type. Metabolite names appear by moving the cursor over the plotted points. Missing or below detection limit values are depicted by different colored marks on the plot. After the exploratory analysis, the investigator can click on the "Biomarker DB" on top of the plot button to see potential biomarkers for this mutant which were generated using 25 bootstrap runs of random forest algorithm to compare wild type samples with the samples from the mutant of interest.

The resulting page shows the statistics of the analysis along with the other entire mutant vs. wild type comparisons in a clickable chart. High margins and low error rate depict that a mutant causes significant biochemical changes in the cell. The user is also provided a list of potential biomarkers for the queried mutant. The user can click on the metabolite name to see details of a metabolite (e.g. structure information, pathway information etc.). The third column on this table shows 3 highly correlated metabolites to each metabolite in the first column from the partial correlation analysis. For example the first potential biomarker for the given mutant is "201060" which is an unknown metabolite. The second column shows that it is highly correlated with ribitol. The second column links the metabolite to the Binbase database and shows the chromatogram of the queried metabolite along with 10 other metabolites whose chromatograms are similar to this compound. Some of the matching compounds include, "xylitol 563" and "ribitol". The investigator can hypothesize that the unknown biomarker is a closely related compound to the alcohols from this analysis. The last column shows all the mutants in the database that share the biomarkers of the mutant of interest. This information can be used in finding if two mutants act in similar ways. Finally , the bottom portion of the page shows all the impacted Aracyc pathways with the mutation. All the known metabolites and putative identifications for the unknown metabolites are used in getting the pathways.

The investigator can also use the integrated morphological images to see the physiological changes in the plant under the mutation and compare them with the images from the wild type samples. The integrated data mining tools can also be used at this time to analyze data from the integrated targeted platforms to get a more comprehensive view of the changes occurred due to the mutation. The combination of the results from this database along with the plantmetabolomics database and data visualization tools will help in forming hypothesis about the functions of genes.