COMPUTATIONAL AND INTEGRATIVE OMICS APPROACHES TO STUDY THE EFFECT OF PERTURBATIONS ON METABOLIC PHENOTYPES

UMASHANKAR SHIVSHANKAR

NATIONAL UNIVERSITY OF SINGAPORE

2014

COMPUTATIONAL APPROACHES TO STUDY METABOLIC PHENOTYPES UMASHANKAR SHIVSHANKAR 2014

COMPUTATIONAL AND INTEGRATIVE OMICS APPROACHES TO STUDY THE EFFECT OF PERTURBATIONS ON METABOLIC PHENOTYPES

UMASHANKAR SHIVSHANKAR

(B.Tech. (Bioinformatics), Sathyabama University)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF BIOLOGICAL SCIENCES

NATIONAL UNIVERSITY OF SINGAPORE

2014

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Umashankar Shivshankar

21 August 2014

Acknowledgements

"*Enn enba eenai ezhuthu enba avereundum kanpol vazhum uyiruku*" (Couplet number 392, Thirukkural) a couplet in Tamil, which roughly translates as "Numbers and letters are windows to the world". As a bioinformatician, I see this couplet as meaning "Sequences and quantitative measurements as windows to the magnificent world of biology". Understanding the mysteries of biological systems through computers was the most exciting quest in my PhD.

The motivation and support provided by my supervisor Prof. Sanjay Swarup was much more than what I had wished for when I started my undergraduate final year project with him. The array of interesting projects and the exceptional guidance provided by him formed the roots of this thesis. Each project was an interesting challenge that helped me learn and enhance my knowledge. I express my immense gratitude and thanks for him being a fantastic supervisor.

I express my sincere thanks to Dr. Rohan Williams, for being my co-supervisor and provide great guidance. He has been the Statistics guru who introduced R to me, which formed the basis for statistical methods in my thesis. Working with him helped me understand the nuances of data analysis, and cultivating the attitude to never trust the data without "cold hard statistics" showing that the hypothesis holds good.

I would also like to thank my colleagues in Metabolite Biology Lab for their patience, support and for providing a great environment to work. In specific, close friends from the lab: Gourvendu Saxena, Amit Rai and Maria Yung Pui Yi for sound advice and discussions on everything under the sun, including research. My mentors from the undergraduate final year projects and now great friends Ambarish Biswas,

Sheela Reuben, Raghuraj Rao and Kalyan Chakravarthy for their guidance through all these years. I would also like to thank Ms. Reena Samynadan, Ms. Priscilla Li from DBS and Ms. Elaine Tay Teng Teng from NERI for assistance and timely administrative support.

I could not have come this far without my better half and wife-Ashwini Ravi Shankar. The person who has been by my side since high school and helped me dream and believe that we can conquer great heights together. This PhD would not have been possible without her continuous support and encouragement.

Family, provides the much needed inspiration and there can be no dearth of it from my family. Starting with my parents Umashankar and Prabha, then Raghav, Geetha, Srikanth, Vaidyanathan and Nagalakshmi, who have shown that with hard work, dedication and sincerity you can reach your dreams.

Nothing ever happens without a bunch of great friends, Raghav Shankar, Satish Rengarajan, Satishkumar, Arun Mahadevan, Sivashankaran, Deepan and the whole gang from 'The gumbal' in Singapore who made me feel at home in Singapore. I am also indebted to the vibrant online community at stackoverflow, Cross Validated, Simply Statistics, R-bloggers and Phenomena for invigorating discussions on science and statistics.

Finally, it's been a privilege to be born in a great country- India, which immensely helped me to shape my thoughts and provided me with a fantastic environment to grow up. I would also like to express my thanks towards the financial support provided by NERI, NUS and Singapore, that enabled high quality research to be performed with the state of the art facilities.

Publications

Journal

2015 **Shivshankar Umashankar**, Sanjay Swarup, Rohan Williams. *"Statistical methods for identification and removal of non-biological sources of variation in metabolomics data"*. (Under review in **Frontiers in Bioengineering and Biotechnology**, January 2015)

2015 **Shivshankar Umashankar**, Vejeysri Vello, Vinay Kumar, Peter Benke, Fook Tim Chew, Phang Siew Moi, Rohan Williams, Sanjay Swarup. "*Environmental and biochemical determinants of metabolic resource partioning in natural variants of microalgae- Chlorella*". (Manuscript in preparation)

2015 Amit Rai*, **Shivshankar Umashankar*,** Megha Rai, Lim Boon Kiat, Sanjay Swarup. "*Glycosylation regulatory factor affects innate immunity in Arabidopsis by co-regulating jasmonate biosynthesis and flavonoid glycosylation*". (* equal first authors, to be submitted to **PNAS**, February 2015)

2012 Wen Cai Zhang, Ng Shyh-Chang, He Yang, Amit Rai, **Shivshankar Umashankar**, Siming Ma, Boon Seng Soh, Li Li Sun, Bee Choo Tai, Min En Nga, Kishore Kumar Bhakoo, Senthil Raja Jayapal, Massimo Nichane, Qiang Yu, Dokeu A Ahmed, Christie Tan, Wong Poo Sing, John Tam, Agasthian Thirugananam, Monireh Soroush Noghabi, Yin Huei Pang, Haw Siang Ang, Wayne Mitchell, Paul Robson, Philipp Kaldis, Ross Andrew Soo, Sanjay Swarup, Elaine Hsuen Lim, Bing Lim. "*Glycine decarboxylase activity drives non-small cell lung cancer tumor-initiating cells and tumorigenesis*". **Cell** 148, 259-272 (2012)

2011 Ambarish Biswas, Raghuraj Rao, **Shivshankar Umashankar**, Kalyan C Mynampati, Sheela Reuben, Gauri Parab, Sanjay Swarup. "*datPAV—an online processing, analysis and visualization tool for exploratory investigation of experimental data*". **Bioinformatics** 27, 1585-1586 (2011)

Book chapter

2013 Amit Rai, **Shivshankar Umashankar**, Sanjay Swarup. *"Plant Metabolomics: From Experimental Design to Knowledge Extraction"*. Legume Genomics Vol. 1069, **Methods in Molecular Biology** (ed Ray J. Rose) Ch. 19, 279-312 (Humana Press, 2013)

Conferences

2014 **Shivshankar Umashankar**, Rohan Williams, Sanjay Swarup. *"Computational methods to study the effect of perturbations on metabolic phenotypes"*. Merlion Metabolomics Workshop 2014, National University of Singapore, Singapore, 19-21 November 2014

2014 **Shivshankar Umashankar**, Rasmus Kirkegaard, Sanjay Swarup, Rohan Williams. *"Systematic approaches for identification and removal of non-biological sources of variation in metabolomics data"*.

10th Anniversary of the International Conference of the Metabolomics Society (Metabolomics 2014), Tsuruoka, Japan, 23-26 June 2014

2014 Sanjay Swarup, **Shivshankar Umashankar**, Vejeysri Vello, Vinay Kumar, Peter Benke, Fook Tim Chew, Phang Siew Moi, Rohan Williams. *"Metabolomics approach to understand the resource partitioning in Chlorella during growth for enhanced biofuel production"*. 10th Anniversary of the International Conference of the Metabolomics Society (Metabolomics 2014), Tsuruoka, Japan, 23-26 June 2014

2014 **Shivshankar Umashankar**, Rohan Williams, Sanjay Swarup. *"Computational methods for metabolic networks: An integrative omics approach"*. EMBO Practical Course on Metabolomics Bioinformatics for Life Scientists, European Molecular Biology Laboratory (EMBL)-European Bioinformatics Institute (EBI), Cambridge, United Kingdom, 17-21 March 2014

***Publications not included in thesis**: The statistical techniques developed during my PhD were used in the projects listed below:*

Journal

2015 Gourvendu Saxena, Ezequiel M Marzinelli, Nyi Nyi Naing, Zhili He, Yuting Liang, Yvette M. Piceno, Suparna Mitra, Han Ping, Umid Man Joshi, Sheela Reuben, Kalyan Chakravarthy Mynampati, Shailendra Mishra, **Shivshankar Umashankar**, Ji-Zhong Zhou, Gary L. Andersen, Staffan Kjelleberg, Sanjay Swarup. *"Ecogenomics reveals land-use pressures on microbial communities in the waterways of a megacity"*. (in press, **Environmental Science & Technology**, January 2015)

Conferences

2014 Nicole Chua, Pui Yi Yung, **Shivshankar Umashankar**, I Sudiana, Sanjay Swarup. *"Harnessing microbial communities in tropical peatlands as resources for novel biocatalysts for plant biomass deconstruction"*. 16th European Congress on Biotechnology, Edinburgh, Scotland. 13-16 July 2014

2014 Pui Yi Maria Yung, Wei Ling Ng, Yong Jian Lee, Shao Bing Johanan Aow, Boon Kiat Lennon Lim, Nicole Chua, **Shivshankar Umashankar**, Sanjay Swarup. *"Functional and genomic analysis of the plant growth promoting bacterium I"*. 16th European Congress on Biotechnology, Edinburgh, Scotland. 13-16 July 2014

2013 Pui Yi Yung, **Shivshankar Umashankar**, Peter Benke, Shailendra Mishra, Stephan Schuster, Sanjay Swarup. *"Metagenomic analysis of microbial community in tropical lowland peatlands revealed specialized functional capabilities in freshwater environment"*. First EMBO Conference on Aquatic Microbial Ecology – SAME13, Stressa, Italy, 8-13 September 2013

2013 Rasmus Kirkegaard, Anisa Cokro, Chee Wai Liew, **Shivshankar Umashankar**, Victor Nesati, Sanjay Swarup, Per H. Nielsen, Rohan Williams, Stefan Wuertz. *"An Untargeted Metabolomics Survey from a Perturbation Model of Nitrogen Transformation in a Tropical Wastewater Community"*. Microbial Ecology and Water Engineering 2013 (MEWE 2013), Ann Arbor, United States of America, 7-10 July 2013

Contents

Summary

Organisms respond to genetic or environmental perturbations by modulating their cellular metabolism. Changes to these metabolic processes are orchestrated through the regulation of multiple biological processes, such as gene expression, protein synthesis or enzymatic reactions, among others. Metabolites that are intermediates or end-products of these regulatory processes, can be regarded as the ultimate biochemical phenotype of a cellular system. Traditionally, regulatory molecules and their mechanisms have been studied using a reductionist approach, targeting only the specific pathway and its intermediates. In order to understand the systems level regulatory interactions that determine the physiological state of a cell, statistical analysis of metabolomics data in combination with other omics data can be used. However, there are still large gaps in our understanding of how to systematically: (i) estimate and remove non-biological sources of variation in high-throughput datasets; (ii) characterize the influence of natural variation or genetic perturbation on the metabolome, and (iii) derive accurate and biologically informed identification of the regulatory control of metabolic networks of the cell.

Motivated by these unresolved challenges, this thesis aims to understand how an organism's metabolite profile is influenced by (i) unwanted non-biological artefacts; (ii) natural variation; (iii) induced genetic perturbation, with an aim to provide important insights on the regulatory and molecular mechanisms involved. By addressing these specific questions, the following knowledge was gained:

- To derive biologically meaningful information from high-density datasets, the data should be free from unwanted variation such as batch effects. To this end, multivariate statistical techniques were used to identify batch effects in an

untargeted metabolome survey and a filtering procedure based on the singular value decomposition was developed to remove these batch effects. This technique removed unwanted variation while permitting recovery of signals of biological origin (Chapter 3).

- To understand the influence of natural variation in the metabolic profiles, data generated from an untargeted metabolite survey of oleaginous algae-*Chlorella* species was used (Chapter 4). Statistical analysis of the metabolic profiles revealed (i) discordance between ribosomal-based phylogenetic classification and metabolic phenotypes; (ii) metabolic diversity between strains to be growth-stage dependent and influenced by habitat-specific variations; (ii) strain-specific associations with physicochemical traits. The top performing strains were enriched in metabolites belonging to isoprenoid and energy metabolism.
- To understand the regulation of biochemical processes that generate metabolite diversity, genetic perturbation-based approaches were used in *Arabidopsis* (Chapter 5). Analysis of multi-omics datasets using an integrative omics approach revealed (i) shared regulatory mechanisms between glycosylation of primary and secondary metabolites; (ii) coordinated regulation of processes associated with metabolite glycosylation and phytohormone biosynthesis; (iii) dependence of plant defence strategies on mechanisms that increase metabolite diversity.

Taken together, the approaches developed in this thesis, integrate environmental factors and metabolic network components with metabolomics data using statistical methods to provide insights into the functioning of complex cellular phenotypes.

List of tables

List of figures

List of abbreviations

Abbreviation	*Explanation*
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
MS	Mass spectrometer
m/z	Mass-by-charge ratio
rt	Retention time
PC	Principal component
PCA	Principal component analysis
SVD	Singular value decomposition
Feature	General term used to represent a measurement unit in high-throughput technologies. In this study, feature, refers to the peaks detected in the mass spectrometer. Each feature has a unique mass-to-charge (m/z) ratio, retention time, and includes the ion's abundance in each sample.
R	Statistical programming language
RunDay	Term used in this thesis to represent the sample processing days in the mass spectrometer

1. Introduction

"In God we trust; all others must bring data."

… W. Edwards Deming

1.1. Era of big data and integrative omics

Technological advances in the fields of physical sciences, computing and engineering have herald in drastic changes to the way data is generated, stored and analysed. Ease of data generation along with the availability of high computational power has led to increasing use of statistics in analysing networks, be it social or biological (such as transcriptional or metabolic networks). Biologists already familiar with handling big datasets from the time of microarray, have now embarked on designing larger experiments producing high density datasets.

This confluence of statistics, computing and biology has created an environment conducive for mining and analysing large biological datasets (Marx, 2013). In this study, we have focused on developing computational solutions for analysing metabolic phenotypes by integrating omics datasets to discover novel biological relationships.

1.2. Metabolomics and transcriptomics

Metabolites are substrates and end-products of enzymatic reactions regulated through dynamic biochemical and gene expression changes in the cell (Fiehn, 2002). Metabolomics attempts to study the role of metabolites in the physiological and developmental state of cells, tissues, organisms and their responses to perturbations. Measurement technologies such as mass spectrometry (MS) or nuclear magnetic resonance spectroscopy (NMR) (described in Section 2.1.5) are commonly used for profiling metabolite levels. Such measurements of metabolite concentrations, along with their metabolic pathway information are used for deriving biological

interpretations. By providing a real-time measure of the metabolite signals in various metabolic pathways, metabolomics approaches provide an accurate snapshot of the specific biochemical phenotype (Katajamaa and Oresic, 2007; Raamsdonk et al., 2001). Metabolomics approaches can be used to provide a critical assessment of complex phenotypes, as well as identify biomarkers related to diseases sub-types, physiological responses and the like (Baker, 2011).

The increasing availability of genome-wide datasets (Suhre et al., 2011; Wen et al., 2014), providing interactions between genes and metabolites (Hirai et al., 2005; Kresnowati et al., 2006) has now made it possible to establish that cellular metabolic phenotypes are directly affected by changes to gene expression levels. Cell growth and maintenance is orchestrated via enzyme mediated regulation of metabolite levels, that mostly occur via transcriptional and/or translational changes, post-translational modifications, binding of small molecules in response to genetic or environmental factors (Saito et al., 2010; Zelezniak et al., 2014). Knowledge of such regulatory mechanisms increasingly depend upon understanding the genetic structure and gene expression levels (Fiehn, 2002). In order to characterize the changes in the genome-wide RNA expression patterns, global transcriptome analysis is used.

1.3. Coordination of metabolic networks – the key to an organism's response to change

Metabolic pathways are co-ordinately regulated at multiple levels and are organized in the form of metabolic networks in a cell. These networks are scale-free, contain metabolites as nodes and provide basic biochemical building blocks, enable growth and maintenance of biological systems. Metabolic networks allow organisms to respond to perturbation and environmental factors by modulating metabolic reactions in these networks. For example, plants being sessile, modulate biochemical reactions to tolerate various abiotic stresses such as light, nutrient, temperature and

biotic stresses such as pathogens, and herbivores among others (Obata and Fernie, 2012).

1.3.1. Regulation of metabolic networks

Knowledge of regulatory mechanisms governing metabolic processes is essential for understanding the changes in the biochemistry and physiology of the system in response to perturbation. Variation in metabolite levels serve as functional indicators of regulatory processes influencing the physiological state of a system. Variation in the levels of metabolites can broadly be due to (i) non biological sources- such as instrument, sampling or experimental errors, (ii) natural variation or environmental factors, referring to macroscopic natural fluctuations such as changes to growth conditions, temperature etc., (iii) external perturbations, such as those that occur through the regulation of gene expression or that modify natural growth conditions.

Furthermore, the intricate network of interactions between metabolites are likely to regulate a number of biological processes, rather than affecting only a specific pathway. However, the feedback mechanisms that regulate metabolite levels via translational control, signal transduction pathways, or allosteric regulation are poorly understood, while protein abundances or enzymatic activities are also difficult to measure. Therefore, in order to understand an organism's response to perturbation, their genetic potential and genotype-phenotype relationships, a clear conceptual framework with multi-level measurements of biological entities is required.

Integrating metabolomics and transcriptomics datasets can provide new insights into biochemical processes by linking lower level biological entities (such as DNA, RNA and metabolites) and higher organizational levels (such as physiological and phenotypic response) (Bino et al., 2004; Hendriks et al., 2011). Furthermore,

metabolite profiling can also be used to identify novel functions for genes and aid in genome annotation (Prosser et al., 2014).

1.3.2. Metabolomic profiling- measuring metabolite levels

Metabolites possess increasingly diverse physicochemical properties with varying concentration levels that typically range from picomolar to millimolar (Bedair and Sumner, 2008; Boccard and Rudaz, 2014). Thus, analytical tools, which can cover this vast chemical space and also provide unbiased and accurate quantitative measurements of the concentration levels are required. With these criteria, mass spectrometry-based metabolomics provide the best platform technology for obtaining non-targeted, high-throughput metabolite profiles of the complete metabolome (Bedair and Sumner, 2008; Lei et al., 2011; Werner et al., 2008b). The high accuracy of mass spectrometers enable the detection of exact masses of metabolites that can serve as putative indicators of molecular formula or structures.

These non-targeted mass spectrometry-based metabolomics experiments which provide extended coverage of analytes along with measurements on interacting data layers, generate massive data structures. Furthermore, MS-based metabolomics experiments, being extremely sensitive to sample and analytical conditions, produce datasets wherein the actual biological variation is highly confounded with non-biological sources of variation. Additionally, genetic or environmental perturbations also alter metabolite abundances by reprogramming metabolic pathways. Thus, appropriate statistical procedures for handling such datasets should factor in the different sources of variation affecting the metabolite profiles.

Understanding and characterizing these complex datasets pose numerous analytical challenges at different stages of the experimental lifecycle, ranging from data extraction through to biological interpretation. Though a number of methods,

techniques and tools have been designed to address these challenges, unique problems associated with the application of metabolite profiling remain.

1.4. Motivation for current research

The basic tenets of scientific research are directed towards increasing knowledge and providing key avenues for further research in any domain. Thus, being in the increasingly challenging and emerging field of metabolomics, identifying metabolic strategies and networks in various biological systems is a key objective. Understanding the properties of such biological networks, including their architecture, regulatory processes, and robustness to evolutionary, environmental and genetic changes is key to predicting and engineering desired responses. Progress in understanding such networks is predicated on generating accurate descriptions, typically by measuring qualitative and quantitative relationships among the biological entities in different layers (e.g., transcriptome, metabolome, proteome) in response to perturbations.

Studies using metabolomics approaches have excellent capabilities to estimate the effect of different treatment conditions, and provide insights as to how genetic information and environmental factors can influence cellular metabolic responses and phenotypic characteristics (Hendriks et al., 2011). However, to fully exploit the wealth of information in such datasets, multivariate statistical approaches and integrative omics strategies that can elucidate complex biological interactions, taking into account the influence of other non-biological sources of variations, needs to be developed. The work presented in this study specifically attempts to contribute to the understanding of the influence of natural variation or perturbation on the metabolic phenotype of a cell, by emphasizing on issues related to metabolomics data analysis and knowledge discovery.

1.4.1. Organization of the thesis

A detailed review of the current state of metabolomics challenges, platform technologies, and data analysis techniques are provided in **Chapter 2**. Following are the specific questions addressed in the different results based Chapters of this study:

Chapter 3: How does non-biological variation affect the metabolomics profile?

Mass spectrometry-based metabolomics experiments are extremely sensitive to the non-biological sources of variation, such as sample extraction or analytical conditions. Furthermore, such experiments generate large and complex data that are confounded by multiple sources of technical and biological artefacts. Combining data from such experiments performed over long time periods of time (weeks/months) or assayed in different batches present numerous challenges. These are often overlooked in current pipelines, and may lead to systematic errors, such as batch effects, that could be misinterpreted as being of biological origin (Leek et al., 2010).

We have developed approaches for identifying and removing sample and assay-related (batch) effects in untargeted metabolome data using multivariate statistical techniques. We demonstrate the use of a filtering procedure based on the singular value decomposition (SVD) on untargeted metabolite profile data from oleaginous algae- *Chlorella*, to remove structure in data related to day of sample assay. The batch effect corrected data is then analysed in an integrated manner to derive meaningful biological information in Chapter 4.

Chapter 4: How do environmental and biochemical factors affect metabolic resource partitioning strategies?

Cellular metabolism of organisms is tightly regulated in response to environmental pressures. Understanding the impact of habitat and biochemical factors on an organism's growth and physiology is extremely important, especially in

biological systems, whose bio-products are extremely sensitive to such factors. Increasing energy demands has led to the search for alternative sources of energy, with biofuel from algae being one of the most promising (Brennan and Owende, 2010).

The yield of bio-energy products in high cell densities depends on the metabolic resource partitioning strategies employed by the organism. Therefore, in this study, we used untargeted high-resolution mass spectrometry along with biochemical profiling to understand the metabolic differences at exponential and stationary growth stages of 22 naturally varying *Chlorella* strains isolated in Malaysia by our collaborators from University of Malaya (UMA).

Chapter 5: What is the impact of genetic perturbations on the metabolic network?

Metabolites have important functional and ecological roles, such as regulating defence, growth, providing stress tolerance, and are highly valuable as pharmaceuticals. These diverse functions are orchestrated through intricate metabolic networks, for example in plants, these networks involve almost 200,000 secondary metabolites (Wink, 2010). The diversity of metabolites mainly arise through biochemical processes such as conjugation (e.g., glycosylation). While the individual enzymes and metabolites involved in these processes are known, there are large gaps in the field about (i) how the different molecular entities that are involved in conjugation processes, function in coordinated networks; and (ii) how metabolite conjugation is regulated in response to other processes, such as development and defence.

We used a putative glycosylation regulatory mutant- *tt8*, in the model plant *Arabidopsis thaliana*, to understand the regulatory processes governing metabolite conjugation and its impact on the gene-metabolite relationships in the metabolic

network. To discover novel glycosylation targets of TT8 and its regulatory network, we developed and used an integrative omics approach.

Finally, **Chapter 6**, summarizes the key findings and contributions of the thesis, and provides recommendations for future work. An outline of the data analysis tools developed/implemented in two interdisciplinary collaborative projects during the present work is also provided in the Appendix. Important aspects of this study, along with the overall organization of the topics analysing perturbational effects on metabolic networks, are depicted in Figure 1.1.

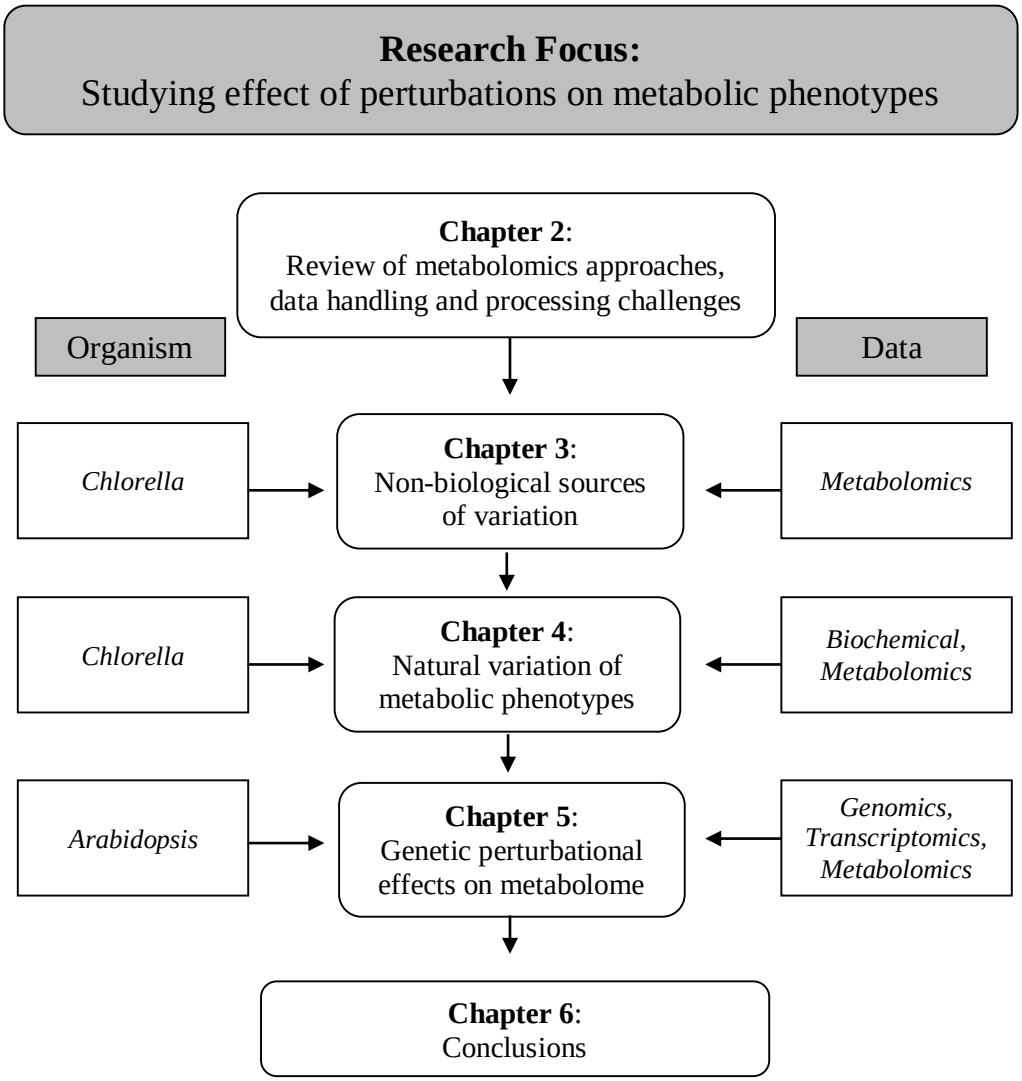


Figure 1.1. Scope of the present work- research depth, breadth and width

2. Literature review

"What was vital was overlaid and hidden by what was irrelevant.
Of all the facts which were presented to us, we had to pick just those
which we deemed to be essential"
... Sherlock Homes in 'The Adventure of the Naval Treaty'

The literature reviewed here has been organized into two parts. The first part of the review provides an overview of metabolome, metabolomics approaches and data analysis strategies. In the second part, metabolic networks, multi-omics approaches and challenges are discussed.

2.1. Metabolome

The term "metabolome" refers to the complete collection of metabolites synthesized by a biological system. Metabolites have important structural and functional roles, and are low molecular weight intermediates or end products of biochemical reactions occurring within cells (Bhalla et al., 2005; Fiehn, 2002). Real-time measurement of metabolite levels through metabolite profiling techniques help determine the active biochemical processes in an biological system (at the time of measurement), and provide an accurate biochemical phenotype (Fiehn, 2002). Measuring metabolite level fluctuations provides important insights into the interactions between the genotype and the environment, and also the various sub-cellular modifications that are a part of homeostasis. This information can be used to assess the cellular response to environmental changes (Bhalla et al., 2005) and in functional genomics strategies (Bino et al., 2004; Fiehn, 2002).

Parts of this review is published as a book chapter- *Rai A, Umashankar S and Swarup S,* "Plant Metabolomics: From Experimental Design to Knowledge Extraction". Legume Genomics Vol. 1069, Methods in Molecular Biology.

The metabolic profile of a biological system isolated from a specific location, developmental stage or environment represents a unique metabolite signature for that particular system observed at that instant of time and in that highly specific physiological state.

2.1.1. Metabolite classes and diversity

Metabolites possess an enormous range of physicochemical diversity, for example, in plant kingdom alone there are nearly 200,000 different types of metabolites (Wink, 2010). The activity and sub-cellular specificity of metabolites to organs, tissues and biochemical pathways arise mainly due this diversity. Metabolites are classified as either primary metabolites such as sugars, amino acids, and organic acids or secondary metabolites such as phenylpropanoids, terpenoids and alkaloids. This classification is based on their functional roles, with primary metabolites playing an active role during growth, development and central energy conversion cycles, while, secondary metabolites, are mostly involved in specialized functions, such as coordinating cellular response to environmental perturbations and in signalling (Hartmann, 2007). These functions of secondary metabolites, require them to be synthesized and localized in specialized cells, tissues, or organs.

Secondary metabolism pathways are generally specialized to cell or tissue type during initial differentiation stages (Rhodes, 1994). Such specialization of metabolic pathways also exist between different compartments within the cell. For example, recent reports have emphasized on the dynamic and highly specific phyto-metabolome by assessing the specificity of foliar metabolic responses in plants to fungus (Schweiger et al., 2014).

The diversity in structures and functions of secondary metabolites arise mainly through biochemical processes, such as conjugation, that change the basic chemical properties of a small number of core metabolites (Hartmann, 1996). Such changes to the chemical structure of the metabolites confer new functional properties such as

altered bioactivity, subcellular mobility, compartmentalization among others. For example, these processes help mediate inactivation of toxic forms of metabolites (Winkel-Shirley, 2001) and de novo biosynthesis of new compounds among others (Sakakibara, 2006).

A number of conjugation processes such as glycosylation, sulfation, acetylation, methylation, amino acid conjugation, gluthatione conjugation and lipophilic conjugation exist in nature (LeBlanc, 2007). Specialized enzymes such as carbohydrate active enzymes (CAZy) (Lombard et al., 2014) are involved in conjugation of metabolites and are responsible for generating the secondary metabolite diversity. The importance of these processes in secondary metabolism can easily be judged by the fact that nearly 4% of the genome of higher plants encodes CAZy.

Advances in metabolomics approaches have led to a better understanding of the classification and roles of primary and secondary metabolites. With many secondary metabolites also identified to have important roles in growth and development, most scientists now consider the differentiation between primary and secondary metabolites as obsolete.

2.1.2. Metabolomics

Metabolomics is the comprehensive analysis of the metabolome by profiling metabolite levels (Fiehn, 2002) using various measurement technologies (described in Section 2.4). The practice of using 'metabolomics' to describe analytical and quantitative measurement of metabolite began with Oliver et al (Oliver et al., 1998). However, it is only in the last decade where technological developments have provided a breakthrough to the increasing use of metabolomics in research, as witnessed in the exponential growth in the number of articles returned by PubMed search for the term 'metabolom*' (* is a wild card, includes metabolome, metabolomic and metabolomics, accessed June 2014) (Figure 2.1).

Metabolomics complements other omics approaches and provides unique advantages that help understand the relationship between mechanistic biochemistry and cellular phenotype (Gary et al., 2012; Goodacre et al., 2004).

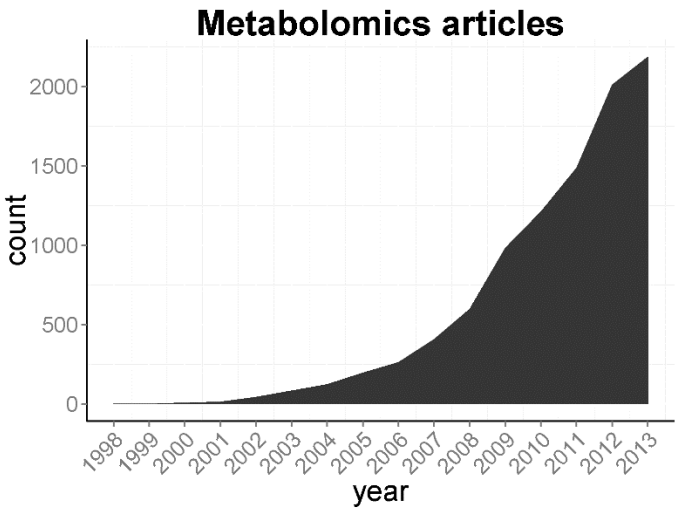


Figure 2.1. Number of articles in PubMed

Firstly, due to the complex regulatory mechanism in cells, changes to gene or protein expression levels might not directly result in a change in the morphological or biochemical phenotype. Unlike genes or proteins that might be subjected to post-translational or epigenetic regulations, metabolites whose structures are determined via metabolite profiling techniques serve as direct signatures of biochemical activity, and can be used to detect unexpected pleiotropic effects. Thus, metabolomics approaches can be used to determine biochemical processes that are activated in a particular phenotype. Such analysis in combination with the transcriptomic information can be used to understand regulatory networks controlling these phenotypes, thereby providing important clues to understand the genotype-to-phenotype relationship. For example, comparing metabolite profiles using differential metabolite analysis helped uncover the effects of a silent mutation in yeast (Raamsdonk et al., 2001) and in potato (Weckwerth et al., 2004).

Secondly, with advances in computational and analytical technologies, metabolomics has attained the technical robustness to provide alternate and

complementary measure of phenotypes. Using the concept of Metabolic Control Analysis (Kell and Mendes, 2000; Teusink et al., 1998) which states that metabolite levels serve as direct substitute for physiological measurements, metabolomics experiments can be used to measure changes to biological parameters such as gene or protein expression levels, without knowing anything about an organisms genetic makeup or its regulatory networks. Multi-omics approaches provide quantitative descriptions of cellular regulation and identify early metabolic biomarkers in disease progression. For example, by analysing metabolite fluxes together with transcriptome profiles, metabolic reprogramming strategies leading to tumorigenesis have been identified (Sreekumar et al., 2009; Zhang et al., 2012).

Systematic analyses of metabolic snapshots of environmental and microbiome samples offers great potential to identify hitherto unknown novel bio-active compounds and pathways (Medema et al., 2011; Steen et al., 2010). Understanding such novel biological designs will lead to better metabolic engineering strategies in synthetic biology applications, for example, to produce or consume key metabolites in response to environmental cues. Furthermore, with most natural products from secondary metabolism being induced under conditions associated with their habitat and lifestyle, exploring these natural variants can aid in the identification of natural products with important commercial and pharmaceutical values (Clardy and Walsh, 2004). This strategy has been used to understand the metabolic markers of biotechnological traits useful for biofuel production in Chapter 3.

2.1.3. Analysing the metabolome

At any given time, the metabolic state of a cell is maintained through the regulation of dynamic biochemical processes in response. By coordinating metabolite levels, these processes provide the basic building blocks of cell metabolism ensuring a thermodynamically favourable environment for growth and development. The regulation of metabolite levels and metabolite network connectivity can be studied

using metabolomics approaches. This provides a unique metabolic fingerprint for each system, where the specific metabolic response can be correlated to different sources of variation. Studying perturbations which alter gene and metabolite levels, and thereby regulating metabolic activity provide a mechanistic view of the regulatory networks (Jansen, 2003). In particular, it will be of interest to understand the degree to which gene expression changes affect metabolite concentration. There are two possible scenarios in which metabolomics approaches can provide valuable insights:

2.1.3.1. Natural intrinsic variations

This is an intriguing feature of cell metabolism, wherein genetic variation between organisms, or diverse environmental factors create distinct metabolic phenotypes. These metabolic signatures are indicative of the natural variation present in such organisms, and can be used to understand how an organism's genetic makeup complement's its environment and enable it to survive in a unique ecosystem. The advent of multi omics technologies has facilitated an ecosystems biology approach, wherein, genome wide association studies of an entire population are sampled and the inherent genetic and metabolic variation are analysed. These approaches are also used for identifying novel bio-products in their natural state and can provide the data for unravelling complex interplay between genes and environment. For example, understanding ecological principles governing growth and production of desired compounds in microbiomes, can help design efficient metabolic engineering strategies (Bouslimani et al., 2014; Nah et al., 2013; Yen et al., 2013). Environmental samples have higher levels of inherent variation and require robust experimental design [discussed in Section 2.5].

In this thesis, the variation in the metabolite profiles due to (i) non-biological sources (Chapter 3); (ii) natural variation (Chapter 4); and (iii) genetic perturbations (Chapter 5) are discussed.

2.1.3.2. Deliberate and controlled perturbations

The genetic basis for metabolite regulation, diversity, concentration, and the interrelationships between metabolic pathways, regulatory networks can be studied using perturbation-based approaches. These studies are designed to link genotype with the corresponding phenotype. Such perturbational strategies can be:

2.1.3.2.1. Localized

The change in metabolite levels resulting from a localized intervention, such as targeting genes at specific steps in a pathway are used to understand the cause and effect relationship. Typically loss-of-function (gene silencing, mutagenesis), gain-of function (transgenesis), chemical elicitors or inhibitors, RNAi or amiRNA approaches are used in this functional genomics approach (Jansen, 2003). Furthermore the effect of the localized intervention can be used to estimate the causal relationships arising out of pleiotropic effects in the global metabolic network.

2.1.3.2.2. Global

Biological systems thrive even in harsh environmental conditions by reconfiguring their metabolic networks to ensure that metabolic homeostasis is maintained. Measuring metabolic responses during treatments applied to the entire system such as biotic or abiotic stressors, transient or diurnal time series in natural conditions, other environmental changes (stress, nutrients etc.) can provide clues to understand the changes to the organism's biochemical processes as a response to perturbations. In this case, changes to the metabolic network might be induced at multiple branch points, thereby affecting a large number of metabolites simultaneously. These network-wide perturbations can also be used to guide a more localized analysis. For example, identifying that GLDC enzyme was correlated with tumorigenesis, led to a more targeted study using a perturbational models [described in Appendix 1] (Zhang et al., 2012).

2.1.4. Qualitative Vs Quantitative approaches in metabolomics

Metabolomics approaches can be designed to provide either a qualitative description, such as nature, physicochemical properties of metabolites and/or quantitative analysis where metabolites and their levels are analysed. For example, to differentiate classes of metabolites selectively enriched between legumes grown in a fertile soil with those grown in drought conditions, qualitative analysis can be performed. The results from such analysis can be used for identifying biomarkers indicating the physiological state of an organism in response to its environment, as well as to compare wild-type and genetically engineered systems (Alisdair et al., 2004).

Accurate absolute or relative metabolite concentrations can be quantitatively measured to analyse flux changes, metabolic reprogramming strategies, differential activation of pathways when comparing different genotypes, treatment conditions, environmental perturbations etc. An unbiased approach to profile as many metabolite as possible is used. These analyses help derive specific biological question, wherein either a targeted approach or a non-targeted approach can be used.

2.1.4.1. Targeted metabolomics

This approach is generally used when testing a specific hypothesis, wherein the possible metabolic targets or pathways affected are known. For such approaches, good knowledge of the biological problem is desired as this will ensure that the sampling strategy captures the maximum change in metabolite levels. Furthermore, sound knowledge of extraction chemistry is also required to design the extraction and sample preparation steps for isolating specific classes of metabolites (Halabalaki et al., 2014; Kim and Verpoorte, 2010; Parab et al., 2009). The extracted metabolites are then quantified using tandem MS or NMR spectroscopy-based approaches. Using such techniques, new classes of metabolites and novel connections in metabolic networks can be discovered. For example, these approaches can be effective in analysing the levels of desired compounds, such as in the field of nutritional metabolomics (Jones et

al., 2012), food safety/quality (Cevallos-Cevallos et al., 2009), environmental chemistry and toxicology (Viant and Sommer, 2013), identification of biomarkers of diseases (Kaddurah-Daouk et al., 2008), or effects of genetic modifications on a specific enzyme (Fiehn, 2002).

2.1.4.2. Non-targeted metabolomics

To obtain a comprehensive, unbiased coverage of the entire metabolome, a non-targeted metabolomics approach is used. The non-targeted nature of these approaches provide a methodological starting point generating data-driven hypotheses. These approaches can be used to identify novel bio-products (Bouslimani et al., 2014), new connections between metabolic pathways, and uncovering biochemical phenotypes of novel biological systems such as microbiomes (Segata et al., 2013). These new links between cellular pathways and biological mechanisms aid in a better understanding of cell biology, physiology and can be used to engineer novel products.

A combination of first non-targeted, and then targeted approach is suggested to perform an unbiased characterization of the metabotype and to subsequently identify novel metabolites. For example, non-targeted approach can be used to identify the most affected pathway, then a targeted analysis can be performed to accurately obtain the concentration levels of metabolites from that pathway.

2.1.5. Metabolomics platform technologies: Choice of metabolomics hardware based on experimental approach

A number of technical and analytical challenges exist in performing metabolomics experiments. Firstly, there is wide physicochemical heterogeneity between metabolites and a broad dynamic range of abundance (Wink, 2010). These require multiple extraction strategies coupled with combination of analytical techniques to achieve adequate metabolite coverage (De Vos et al., 2007; Fernie, 2007; Patti, 2011). Secondly, analytical instruments with ultra-high resolution, high scan speeds are required to ensure that the chemical space of a broad range of metabolites is

covered thoroughly. These in turn produce large amounts of data, which needs advanced statistical analysis to obtain meaningful information (Werner et al., 2008a). Lastly, the biological question, sample type and experimental design should be the basis for choosing an analytical instrument. With the above criteria, MS and NMR instruments that facilitate metabolomics experiments to be performed with high specificity, reproducibility and in both qualitative and quantitative manner should be selected. A brief overview of the characteristics of these instruments is provided below.

2.1.5.1. Nuclear magnetic resonance

NMR provides the option to have a highly selective and non-destructive approach, thus, it is widely used for structure elucidation, confirmation and quantification of both known and novel metabolites (Kim et al., 2011; Wishart, 2008b). Unlike MS, NMR can be used to analyse samples existing in both solutions or as solid-state samples. However, NMR has relatively low sensitivity, thus limiting the metabolic coverage. This makes NMR to be preferred mainly for targeted approaches, such as flux analysis using 1-D or 2-D NMR. Nevertheless, NMR has great potential in quality control measurements, and in chemotaxonomy to classify and characterize biological systems based on their distinct metabolic signatures. (Wishart, 2008b) has discussed NMR-based metabolomics in depth.

2.1.5.2. Mass spectrometry

Mass spectrometry-based metabolomics is widely popular in both targeted and non-targeted approaches as it provides high resolution, sensitivity and coverage required for identification and quantification of metabolites. The ratio of mass-to-charge observed of ions is measured in this technique. These observations provide specific chemical information which can directly be related to the chemical structure and formula. For example, accurate mass, isotope distribution patterns and characteristic daughter ions are all produced using MS. These results are used for fragmentation-based structure elucidation or identification via spectral matching to

compound databases such as HMDB (Wishart et al., 2013), KEGG (Okuda et al., 2008) or MetaCyc (Zhang et al., 2005).

The high sensitivity of MS is used to detect metabolites even at picomolar or femtomolar levels. High mass accuracy (detection differences less than 2ppm) and high resolution, make the next generation instruments, such as Fourier transform-ion cyclotron resonance MS (FT-ICR-MS) and Orbitrap MS to be used derive semi-quantitative measurements of metabolite concentrations. For absolute quantitative measurements of metabolite levels, quadrupole ion trap mass spectrometers or triple quadrupole mass spectrometers (for tandem MS/MS) are used. The high sensitivity and resolution, comes with its caveats, requiring robust data analysis strategies for obtaining biological information. A detailed discussion on various stages of mass spectrometry-based metabolomics experiments is provided (Dettmer et al., 2007).

2.1.5.2.1. *Hyphenated mass spectrometry techniques*

The two most common approaches for MS-based metabolic profiling are either direct injection of the sample into MS or using chromatography techniques such as gas chromatography (GC), high-performance or ultra-performance liquid chromatography (LC) and capillary electrophoresis (CE) in conjunction with mass spectrometry (hyphenated mass spectrometry) to provide better separation and resolution of metabolite profiles.

Direct injection-based approaches are faster compared to others as time spent in the chromatographic run is saved. Thus, for screening large number of samples, such as those from a population-based research or a large environmental study these approaches are used. However, direct injection of compounds are prone to matrix effects such as ion suppression or enhancement and lead to inaccurate quantification of metabolites. There are also challenges in resolving adducts during metabolite identification.

MS with chromatography-based separation enables the separation of ions based on its physicochemical properties such as size and charge, and thus, avoid matrix effects. These advantages makes chromatography coupled MS to be one of the most widely used approach in analysing complex samples. Furthermore, additional information on metabolites detected, mainly in the form of retention times of those metabolites is obtained. Such information is especially useful while performing database-dependent metabolite identification.

GC-MS-based approaches require the analytes (thermo-labile) to be in gas phase, and are suitable for both volatile and non-volatile compounds following derivatization. This approach is widely used in targeted metabolomics where the chemical properties of metabolites are known. For example, GC-MS is popular in plant metabolomics to detect volatile metabolite contributing to aroma (Shuman et al., 2011). Furthermore, GC-MS has comprehensive robust metabolite libraries that can be used for metabolite identification (Hummel et al., 2007). The major limitation of GC-MS is the extensive derivatization steps, and restrictions based on the chemical properties of metabolite classes. Thus, liquid phase-based methods such as LC-MS have found favour among metabolomics researchers.

LC-MS is generally less restrictive than GC-MS. For example, in LC-MS, samples can be mildly heated during ionization. This property of LC-MS makes it ideal for non-targeted metabolomics approaches that are performed to detect both thermo-labile and thermo-stable metabolites. An additional analysis strategy is the flexibility to use a number of columns based on reverse phase, ion exchange and hydrophobic interactions principles (Allwood and Goodacre, 2010). Ultra-high performance liquid chromatography (UHPLC) provides a fast and efficient way to increase chromatographic resolution and detection range, while decreasing the analysis time compared to HPLC. The major challenge in LC-MS-based approaches is the bottleneck in metabolite databases. With a number of variations in separation columns, a

comprehensive database facilitating cross-comparisons between different approaches has been difficult to construct. Thus, putative metabolites are usually validated via NMR or tandem MS/MS and MRM (Multiple Reaction Monitoring) methods. These methods provide a detailed description of the chemical structures and aid in validation of the metabolite detected. LC-MS-based metabolomics has been reviewed in this article (Bin et al., 2012).

Separation of charged metabolites can be performed using CE-MS. However, its sensitivity is low, hence metabolites have to be enriched before being used in CE-MS. Furthermore, it also lacks comprehensive reference libraries. Depending on the objectives of the experiment, nature and range of metabolites to be detected, a number of ionization techniques, such as, electron ionization (EI), electrospray ionization (ESI), atmospheric pressure chemical ionization (APCI), chemical ionization (CI), MALDI, desorption ESI (DESI) and extractive ESI (EESI) can be used. Detailed reviews (Ernst et al., 2014; Lei et al., 2011; Rai et al., 2013) discuss strategies used in analytical platforms to data acquisition techniques for mass spectrometry-based metabolomics.

2.1.6. Design of Experiments

It is important to note that metabolomics data can be observed at different scales, such as tissues, organs, organism or communities. Each of these systems have their own complexities that affect, the rate at which metabolite concentration changes, time delays between gene-metabolite responses among others. Therefore, it is important to develop an experimental design that can provide insights into biological responses without confounding effects (Leek et al., 2010). For example, pilot studies should be performed in order to optimize various extraction and analytical procedures. The nature and class of the metabolites specific to the experiment should be examined, before conducting a large-scale study.

The experimental design can be tuned to the biological question of interest, after primarily addressing the data acquisitions challenges. Robust study designs minimizing nuisance variations (sample handling and analytical) and maintaining sample integrity should be used (Hendriks et al., 2011; Leek et al., 2010). Furthermore, good experiment designs should not only aim to reduce analytical-measurement variations, but, should also ensure that the experiment has considerable statistical power to answer the biological question. For these reasons, a pilot study along with thorough literature survey should be performed. These results can then be used to ascertain the levels of biological and technical variation in the samples. Adequate number of biological and technical replicates should be selected after accounting for the predicted variation among replicates. Such careful analysis can ensure that the actual experiment meets the coverage, reliability and reproducibility criteria to provide sound biological information. To ensure minimal confounding effects due to instrument or analytical variations, factorial or randomized study designs can be considered. Minor influences due to non-biological sources of information can be detected by performing exploratory data analysis during data pre-processing.

If careful statistical considerations are taken into account at the experimental design phase of a multi-omics project, then there is an opportunity to build rigorous systems-level statistical models that fully take advantage of the interdependent workings of biological molecules. A number of reviews (Fernie et al., 2011; Gibon and Rolin, 2012; Goodacre et al., 2007) have provided detailed recommendations for performing well-designed, robust metabolomics experiments.

2.1.7. Data analysis strategies in metabolomics

High-throughput metabolomics experiments enable the detection of large number of signals, such as, measuring 10,000 features (ions) across different conditions. These measurement technologies typically produce gigabytes of data having intricate patterns hidden in their data structures. Such datasets, require enhanced data standards and

strategies involving robust, high quality statistical procedures to obtain biological knowledge. The data analysis procedures depend on the (i) choice of analytical platform, (ii) experiment design and biological question, and (iii) inherent properties of the data (Boccard et al., 2010)

Data extraction, handling and treatment procedures which greatly influence the ability to identify and quantify metabolites of interest, have a direct role in the biological interpretation (Katajamaa and Oresic, 2007). Extracting the relevant information from the overwhelming amount of data generated by these high throughput techniques is an important objective for knowledge discovery in this field (Boccard et al., 2010; Goodacre et al., 2004). Development of bioinformatics techniques, specifically for data storage and management, raw data extraction, pre-processing and statistical analysis, integration with other omics datasets, metabolite identification and metabolic modelling is crucial for future progress of metabolomics and systems biology (Shulaev, 2006; Wishart, 2009).

To extract valuable information, irrespective of the analytical technique used, metabolomics data is analysed in the following step-wise manner: pre-processing, pre-treatment, data analysis, validation and interpretation (Eliasson et al., 2011; Goodacre et al., 2007; Hendriks et al., 2011; Katajamaa and Oresic, 2007). Pre-processing methods transform raw signals into a representation facilitating robust statistical comparisons. They typically include filtering, peak alignment, noise removal, baseline correction, normalization and scaling (Boccard et al., 2010; van den Berg et al., 2006). Pre-processing is usually followed by a quality control strategy, where exploratory data analysis is performed to check data quality and identify any issues in sample processing, analytical or technical errors, batch effects etc.

Data analysis techniques try to reduce the multi-dimensional datasets into smaller components, thereby enabling the researchers to identify differential metabolites, interesting patterns in the data structure, and visualize the dynamic information using

both univariate and multivariate analyses (Saccenti et al., 2014). Most analysis strategies utilize both supervised methods (uses prior information to guide the classification) such as ANOVA, partial least squares (PLS) and discriminant function analysis (DFA), and unsupervised (to describe the overall pattern or data structure) methods such as hierarchical clustering, principal component analysis (PCA) and sel*F*-organizing maps (Broadhurst and Kell, 2006). To aid in biological interpretation the differential features identified using the above techniques are then mapped onto metabolic pathways. Robust data analysis strategies not only provide interesting biological interpretations, but also help design better experiments, optimize protocols and reduce experimental errors (Parab et al., 2009).

In a typical metabolomics dataset, the biological differences between samples are hidden under intricate patterns in the data, confounded with obscuring sources of variability introduced at various stages of sample generation or analysis, such as systematic errors during experiments, misplaced samples or instrument errors. Thus, separating out the relatively small but important patterns in metabolite concentrations related to genetic variation or multitude of environmental changes is not straightforward. Naïve analysis of such datasets can lead to serious misinformation (Leek et al., 2010). Multivariate statistical techniques for identification and removal of non-biological sources of variation are discussed in depth in Chapter 3.

The vast volume of data has facilitated the development of a number of application-specific software pipelines and advanced statistical techniques for data handling, data processing and mining, and visualization aiming at disentangling the complex regulatory processes in biological systems (Biswas et al., 2010; Pluskal et al., 2010; Smith et al., 2006; Xia et al., 2012). (Sugimoto et al., 2012) provide a comprehensive review on the bioinformatics tools and techniques available for mass spectrometry-based metabolomics data analysis.

The functional annotation along with data mining and extraction of knowledge from the wealth of information obtained is one of the grand challenges of metabolomics. The metabolomics research community, functioning under the umbrella of Metabolomics Society (http://metabolomicssociety.org/), has developed a set of guidelines ensuring minimal reporting standards for experiments, thus, providing the much needed benchmarks for data analysis, exchange and comparison of metabolomics experiments (Fiehn et al., 2008; Goodacre et al., 2007; Members et al., 2007). These require researchers to diligently record metadata such as temperature, growth conditions (minimal set of data of reporting standards and general guidelines) that aid in biological interpretation and reduce experimental errors.

2.2. Metabolic networks

2.2.1. What are metabolic networks?

Biological systems derive their important characteristics like adaptability, stability and resilience through the regulation of highly interconnected chains of metabolic pathways that encompass heterogeneous biological entities including DNA (genes), mRNA, proteins and metabolites. The complex interactions between various components in a biological system is best characterized as networks. A number of biological networks, each donning a different functional role such as protein-protein interaction (PPI) networks, metabolic networks, transcriptional regulation networks, signal transduction networks interact with each other. These biological networks being dynamic and selectable, respond to environmental pressures by regulating various biochemical processes in the cell. Thus, these networks are a reflection of biological processes such as metabolism, transcription and translation that take place in the cell, (Ideker et al., 2001), and, provide an overview of the regulation by proteins, activation or inactivation of enzymes by posttranslational processes and feedback loops

The collection of metabolites, their pathways and their inter-relationships, holding information about a series of biochemical events constitute an organism's

metabolic network. For example, using both computational predictions and experimental procedures, a metabolic pathway database, PlantCyc, containing more than 800 pathways from over 300 plant species has been developed (Zhang et al., 2005). These pathways are co-ordinately regulated at multiple levels, such as by feedback regulation of metabolic reactions and transcriptional regulation of sets of metabolic genes. Through metabolomics approaches it is now possible to sample thousands of unique ions, assign putative formula and structure, add new pathways, and even develop naïve metabolic networks (Dettmer et al., 2007). Metabolic networks act as scaffolds for metabolic models and can be used to predict cellular function and study of the role of individual reactions.

Robustness and modularity of metabolic networks, are the two major properties that dictate how metabolic networks function and respond to external stimulus. Robustness is the property through which the cellular metabolism tries to maintain homeostasis on encountering genetic or environmental perturbations (Smart et al., 2008). Robustness of metabolic networks is achieved mainly through the presence of multiple isozymes, therefore ensuring redundancy and a tight feedback control. Thus, the effect of blocking one enzyme or a pathway, often leads to the activation of an alternative route through its complementary isozymes. For example, the conversion of glucose-6-phosphate into glyceraldehyde-3-phosphate can be achieved using both glycolysis and pentose phosphate pathways. Such properties confer the ability to use multiple alternative pathways to synthesize the same metabolites, thus increasing the odds for an organism to survive in unfavourable conditions. Stable isotope-based metabolomics approaches wherein the movement of the targeted metabolite into different pathways or tissues is tracked, help in understanding the plasticity and the dynamic nature of metabolic networks.

This observed robustness of metabolic networks, can be explained using the concept of modularity. Modularity describes the potential of independent and self-

contained property of a system, i.e. the process through which a number of regulatory processes govern activation of specific modules to produce the desired metabolic response to perturbations. Modularity and plasticity of metabolic networks follow a power law distribution, with focus being on the few critical hub nodes, made of important genes and metabolites, in the dense highly interconnected community structure. These key genes or metabolites appear in many reactions, while most other genes or metabolites appear in only one or few reactions. Such key metabolites serve as common substrates at branch points of diverse metabolic pathways where a high level of coordinated gene expression exists (Huss and Holme, 2007). Identifying these critical branch point metabolites can provide insights into the master regulators and strategies for understanding metabolic response (Holme, 2011). For example, in *Arabidopsis*, different branches of isoprenoid metabolic network such as carotenoids and brassinosteroids metabolic pathways exist independently and are activated based on subcellular localization of metabolic pathways (Vranova et al., 2012). The specific activation of metabolic modules occur based on compartmentalization and localization of metabolic response, thus, providing the mechanism for targeted production of metabolites in a tissue-dependent manner as a desired response to environmental factors (Brown et al., 2003).

The analysis of these nonlinear, multivariate and multi-layered networks, identification of functional modules, along with the complexities in accurately detecting, quantifying and interpreting metabolomics datasets have led to the development of multiple techniques (Brohee et al., 2008; de Oliveira Dal'Molin et al., 2010; Hamilton and Reed, 2014; Junker et al., 2006; Lewis et al., 2012; Mithani et al., 2010; Ruppin et al., 2010; Yeang, 2009). The structure and properties of metabolic networks especially the topology and methods for the reconstruction of metabolic networks have been described in detail (Palsson, 2006). The tight regulation of

metabolic networks and effects of genetic perturbations on metabolic networks are analysed in detail in Chapter 5.

2.2.2. Integrative omics approaches

Systems Biology seeks to "*study the behavior of an in vivo biological process by systematically perturbing them and then monitoring the interactions between gene, protein, metabolite and informational pathways*" (Ideker et al., 2001). Biological networks are complex systems which are highly inter-connected, non-linear, and dynamic with interactions at multiple levels. Traditionally, genes or proteins involved in different processes have been studied in a reductionist manner, for example characterizing genes only affected in specific metabolic pathways involved in human disease or plant defence response. However, coordination of cellular processes involves interconnectivity via networks of various biological pathways and their control by signalling and regulatory networks.

Gene-regulatory motifs form the building blocks of functional modules by regulating the expression of genes and in turn metabolic pathways. These pathways then form functional modules which are highly interconnected and interact with different biological networks in the cell, thus forming a large-scale biological network requiring systems level approaches to understand biological processes in a holistic manner. Such integrative approaches to biological questions can yield important insights inaccessible to traditional reductionist methods. The main utility of systems approaches lies in the possibility to predict the results of experimental or natural perturbations. For example, Gal4p and Gcn4p lead to the breakthrough discovery showing how transcription factors work (Ptashne, 1988). However, to understand the transcriptional regulatory control orchestrated by these transcription factors, systematic analysis of data from different biological layers was needed.

Technological developments in analytical platforms along with the rapidly decreasing cost of multiple omics measurements has caused an important paradigm

shift in the field of systems biology by facilitating integrative omics approaches (Cai, 2012; Gary et al., 2012; Nielsen and Jewett, 2007; Segata et al., 2013; Wang et al., 2013b). By combining these heterogeneous biological information into a single systems level analysis, using techniques such as correlations, metabolic control analysis, information theories, and network/graph models, complex regulatory interactions can be better studied (Joyce and Palsson, 2006). Such analysis can reveal the dynamic interactions and connectivity in metabolic networks, permitting the discovery of new correlations and pathways among biological entities (Choi and Pavelka, 2012). However to connect these highly multivariate datasets in terms of biological networks, a clear conceptual framework with good experiment design, dedicated tools and statistically sound data analysis is required.

Statistical techniques capable of handling heterogeneous datasets, which might contain of binary, categorical or continuous data, as well as being able to accommodate missing data, and remove artificially induced systematic biases should be developed (Wang et al., 2012). Furthermore, these techniques should be able to provide sound biological interpretation and visualization of multiple layers of data. The overall strategy in any multi-omics analysis is

(i) identifying differential features for each dataset, such as genes, proteins or metabolites independently;

(ii) combining these results by mapping these features onto the biological networks using pathway information from databases such as KEGG (Kanehisa and Goto, 2000; Kanehisa et al., 2014), MetaCyc (Zhang et al., 2005), and Human Metabolome Database (Wishart et al., 2013) or by developing genome-scale metabolic models. This is then used for identifying patterns of correlation or co-regulation, For instance, transcriptomics and metabolomics data were integrated to identify clusters of genes and metabolites that were co-ordinately modulated in response to specific nutritional stresses in the model plant *Arabidopsis thaliana* (Hirai et al., 2004);

(iii) analyse network structure to identify enriched sub-networks, functional modules; and

(iv) develop in silico cellular models using models such as Flux Balance Analysis (FBA) to predict reaction rates and network activities that give rise to cellular phenotypes (Wang et al., 2012). Bayesian methods, which are modelled to avoid over-fitting datasets are also widely used. (Choi and Pavelka, 2012; Cline et al., 2007; Hirai et al., 2007; Kresnowati et al., 2006; Li et al., 2013; Segata et al., 2013; Takahashi et al., 2011) provide examples and detailed discussions on the integration of omics data.

The two most commonly used approaches for data integration are correlation analysis, techniques such as PCA and PLS belonging to the unsupervised approaches and Genome-scale metabolic models (GEMs). In the first approach, normalized gene and metabolite datasets are analysed to identify significant gene-metabolite pairwise correlations (Allen et al., 2010; Hirai et al., 2005), this reveals the presence of co-expressed connections. PLS can be used to model the metabolite abundances as a function of gene expression profiles (Pir et al., 2006). These are then visualized in the form of networks or used in enrichment analysis. Constraint-based approaches aim to develop a genome-scale metabolic model and incorporate the metabolic fluxes and reaction kinetics (Hamilton and Reed, 2014; Lewis et al., 2012; Price et al., 2003; Zelezniak et al., 2014). FBA uses GEMs to calculate how metabolites flow through the metabolic network. This enables researchers to predict the production rate of specific metabolites or the growth rate of an organism.

These multi-omics approach are used in diverse areas ranging from studying ecological networks, where biotic interactions between species are of focus, to modelling biochemical networks within a cell and in diverse domains such as, toxicology, metabolism and pharmacokinetics. They have also been successfully applied to study regulatory interactions (Hirai et al., 2007), functional genomics (Raamsdonk et al., 2001; Tohge et al., 2005) and to identify genome wide quantitative

trait loci (Riedelsheimer et al., 2012; Saito et al., 2010). Analysing the associations between genotypic and phenotypic characteristics has important ramifications in pathological studies for explaining disease pathways and identifying biomarkers for prognosis and diagnosis (Kaddurah-Daouk et al., 2008; Sreekumar et al., 2009; Zhang et al., 2012). Such integrated analyses provide important clues that help understand how genetic blueprints combined with non-genetic, environmental factors influence a biological system.

The challenges in integrating multiple datasets, such as lack of uniform and standardized databases, lab-to-lab variations, are expectedly the same as observed in metabolomics data analysis. MetaboLights (Haug et al., 2013; Salek et al., 2013b), a database for freely storing metabolomics data with detailed experimental protocols and meta information, is championed by the European Bioinformatics Institute, and promises to be an important tool to overcome these barriers.

2.2.3. Challenges in metabolomics

The last decade has witnessed great advances in statistical techniques and measurement technologies aiding in robust characterization of the complete metabolome of an organism. However, a number of challenges ranging from sample preparation to metabolite identification and biological interpretation hinder comprehensive utilization of metabolomics data (Hegeman, 2010; Vuckovic, 2012).

The high chemical and structural diversity of metabolites require specialized extraction protocols, taking the spatio-temporal location, genotype and compound classes into account. A suitable extraction procedure that provides both comprehensive coverage and specificity should be developed before using any metabolomics platform. Furthermore, to obtain the metabolite levels at the exact moment defined in experiment design, conditions that can possibly affect degradation or inter-conversion of metabolites should be prevented. This is usually performed by quenching the metabolism of the targeted biological system (Álvarez-Sánchez et al., 2010). For

example, intermediate metabolites such as those from Calvin cycle and nucleotides have a very short turn-over time and require immediate quenching (Fernie et al., 2011). The physicochemical characteristics of the metabolites in the biological system under consideration, should be the basis for developing any extraction buffer/method. (Kim and Verpoorte, 2010; Parab et al., 2009; Vuckovic, 2012; Want et al., 2013) provide a detailed reviews and protocols on sample preparation for metabolomics experiments.

The above articles describe a variety of protocols for targeting the metabolites of interest, and for optimizing instrument parameters. However, trying to identify novel metabolites or pathways where the nature of metabolites are largely unknown is extremely challenging. Furthermore, the variation in extraction protocols has made it difficult to have standardized metabolomics databases, especially for MS-based approaches. For example, different instruments have slightly varying fragmentation patterns for the same metabolite, thus, standardized libraries are very hard to develop. Data curation is also difficult as curators are having to comprehend the uncharacterized measurement noise associated with high-throughput measurements, and errors during metabolite identification (Salek et al., 2013a) . (Wishart, 2009) provides list of databases used for metabolite identification and pathway mapping.

These issues have resulted in a number of molecules that are detected by the instrument, but are not assigned to any metabolite and thus not included in the metabolite databases or repositories (Kind et al., 2009). A survey conducted by the American Society for Mass Spectrometry (ASMS) in 2009 revealed metabolite identification to the biggest bottleneck among users (Spectrometry, 2009). The current practices allow high confidence identification for only major primary metabolite such as sugars, sugar phosphates, amino acids, and organic acids and certain secondary metabolite classes such as phenylpropanoids, and alkaloids. This is because, in a typical mass spectrometry analysis, for a particular metabolite, number of features such as their isotopic forms, adducts and daughter ions are produced. These ambiguities further

complicate metabolite identification via direct matching of the m/z ratios to the databases. Therefore, to identify novel compounds with high confidence, researchers use the rather slow technique of structure elucidations using fragmentations patterns obtained using tandem MS and NMR approaches. NMR-based identification strategy relies on detecting and matching the characteristic and unique "chemical-shift" fingerprint for each metabolite (Moco et al., 2007). However, with the probable chemical space of around 600 million compounds- determined using the seven golden rules, and up to 8 billion chemical formulas (theoretically possible C, H, N, S, O, *P*-formulas for compounds up to 2000 Da), it is an enormous challenge identify novel compounds when the closest reference compounds are not available in the databases.

The time, effort and cost involved in experimentally determining characteristic properties for millions of molecules present in any given bio-system is a daunting challenge. Thus, a significant improvement in experiment design and data analysis tools is an urgent necessity to enhance systems biology-based knowledge discovery.

2.2.4. Metabolomics as a tool in biological research

The potential of metabolomics as a tool far outweighs the challenges, a delightful scenario to be in as it encourages and rewards technological advances. By providing the insights into biochemical regulations, metabolomics immediately adds a new dimension as an analytical technique. The areas in which metabolomics techniques have been applied are diverse, and new applications are continuously being explored, a selected few are discussed here.

- Functional genomics, systems biology and biotechnology (Alisdair et al., 2004; Hamilton and Reed, 2014; Nielsen and Jewett, 2007; Saito and Matsuda, 2010): Understanding regulatory networks in biological systems, developing genome-scale metabolic networks, studying cellular dynamics using mathematical modelling, effects of perturbations, metabolomics as a tool in

enhancing and developing compounds with useful traits in various biological systems.

- Plant biology, plant-microbe and plant pathogen interactions, agriculture (Bhalla et al., 2005; Dixon et al., 2006; Hall, 2006; Lee et al., 2013; Narasimhan et al., 2003; Okazaki and Saito, 2012; Rasmussen et al., 2012): Characterizing biochemical and genotype-phenotype relationship, cellular responses to different environments, identifying novel plant products and developing metabolic engineering strategies for producing compounds with important pharmaceutical and commercial values.
- Food science and nutrition (Dervilly-Pinel et al., 2012; Jones et al., 2012; Wishart, 2008a): For detecting contaminants, enhancing nutritional value of foods, optimizing fermentation and bioremediation processes, impact of fertilizers and pesticides on plants and environment, assessing substantial equivalence from genetically modified organisms compared to natural cultivars, among others.
- Human health and disease (Aboud and Weiss, 2013; Kaddurah-Daouk et al., 2008; Spratlin et al., 2009; Suhre et al., 2011; Wikoff et al., 2009): For identifying early biomarkers useful as disease and prognostic indicators, diagnosing pathologies, and in drug development and assessing therapeutic targets of disease, assessing associations between genetic variation and human disease phenotypes, to understand the complex interactions of host, diet and gut microflora in human health.
- Environmental metabolomics and natural products research (Bundy et al., 2009; Nguyen et al., 2012; Rochfort, 2005; Viant and Sommer, 2013; Zhang et al., 2010): Assessing ecotoxicology, microbiome structure and functions, components of ecosystems biology, interactions of organisms with environment, discovering natural products, studying evolutionary relationships

using phylogenomics and metabolic networks, designing new bio-parts using ecological principles in synthetic biology applications.

The following chapters of this thesis systematically address different issues related to metabolomics data processing and integration of multiple biological data layers.

3. Statistical methods for identification and removal of non-biological sources of variation

'Data does not equal information; information does not equal knowledge; and, most importantly of all, knowledge does not equal wisdom. We have oceans of data, rivers of information, small puddles of knowledge, and the odd drop of wisdom.'

… Henry Nix (1990) in 'A National Geographic Information System – An Achievable Objective?'

3.1. Background and introduction

Metabolomics technologies have now reached a stage of development where the primary concern is not about generating high quality data but rather about obtaining meaningful biological knowledge from gigabytes of information. This ability to generate large amounts of data will aid in the understanding of previously inaccessible domains of biology (Goodacre et al., 2004; Kell, 2004). Furthermore, robust data analysis of biochemical phenotypes can provide unique insights in the context of both hypothesis generating (exploratory) and hypothesis testing (confirmatory) phases of research (Jaeger and Halliday, 1998).

Designing systematic experiments, analysis protocols and extracting relevant knowledge from the wealth of data is critical to all omics applications. MS-based metabolomics experiments are extremely sensitive and provide unparalleled detection and coverage of metabolites. As these technologies not only increase the quantity of data, but also affect its properties (Godzien et al., 2013), successful application of these experiments depend on both the analytical system and the data mining strategies. Typically, in a large dataset the true biological responses detected are hidden under that façade of data confounded with unwanted variation (Leek et al., 2010). Such

experiments demand advanced statistical procedures to identify and remove unwanted non-biological sources of variation (Eliasson et al., 2011; Hendriks et al., 2011).

Metabolomics data processing typically includes exploratory data analysis to check the quality of data (such as presence of batch effects) before proceeding to pre-treatment, pre-processing and statistical analysis for identifying differential metabolites (Katajamaa and Oresic, 2007). Exploratory data analysis (EDA) should be conducted to identify any systematic errors before further experiments can be performed (Boccard et al., 2010).

A description of the web-based exploratory data analysis application- datPAV (Biswas et al., 2011) developed during the initial part of this research is provided in Appendix 1. datPAV provides various statistical and visualization options for exploratory data analysis. To enable quick examination of high-throughput omics data and cater to the needs of the wide range of omics studies, datPAV has been designed as an web-based tool for performing multi-omics exploratory data analysis.

3.1.1. The importance of batch effects in omics data

Limited by sample processing time (each sample requiring on average 20 minutes for LC MS-based analysis), large sample metabolomics experiments often need to be performed over weeks or in different batches. Even with robust protocols and optimization of sample preparation methods, each batch represents a unique analytical environment having its own time and place-dependent experimental nuances. For example, in a metabolomics experiment with two samples A and B, each with four biological replicates; there might be differences in the resultant outcome of one biological replicate, for example A1 due to batch, extraction or analytical errors. This difference can influence A1 to have a metabolite profile which is completely different from other replicates of A and might even influence A1 to be similar to B. If these errors and batch effects are not removed, then the statistical power to compare the variance between two groups in parametric (t-tests) or non-parametric tests is lost. This

is mainly due to the increased variation in group A as the result of the outlier replicates. In large blinded studies or for non-supervised analysis, A1 might also be grouped with B. If such systematic errors are not undetected, then they lead to confounding biological interpretation. A real example, where the outlier replicates of wild-type, were found to be similar to outlier replicates of mutant lines, is shown in Chapter 5.

Careful experiment designs can minimize batch effects, however it can only be eliminated if the whole study is conducted in a single batch. Thus, batch effects are almost an inevitable consequence of large experiments. For example, human error (differences during sample preparation), different analytical platforms, instrument variations during long periods of operation such as temperature changes or ionization efficiency, changes in chromatography such as column conditions can introduce unwanted experimental sources of variations in the metabolite profiles (Leek et al., 2010). There can also be differences due to unwanted biological variation such as differences in sample mass, concentration or cell number among others. These systematic experimental or analytical influences lead to qualitative and quantitative differences in the relative peak intensities of the metabolites, that are unrelated to the biochemical phenotype-the main focus of the study. The resulting peak intensity of each metabolic feature (counts) is a combination of both the biological signal, as well as unwanted (non-biological) variation.

Analysing data from such experiments presents numerous challenges due to the influence of these batch and sampling/measurement variables (Ernest et al., 2012). In order to distinguish biologically relevant signals from experiment noise, robust normalization procedures are required. Batch effects can affect subsets of metabolite features in different ways (Redestig et al., 2009). The standard pre-processing steps such as binning, alignments, normalization and scaling procedures can be used for adjusting technical variation due to abundances (van den Berg et al., 2006). These pre-processing steps assume general invariability for all metabolite features and do not include corrections for systematic variations due to batch effects (Leek et al., 2010).

Thus, potentially misinterpreting signals arising out of experimental artefacts as being of biological origin. This can lead to major problems and incorrect biological conclusions when such spurious differences due to batch effects are completely correlated with the biological question (outcome of interest). Furthermore, ignoring batch effects during data analysis can increase confidence intervals, therefore affecting robust identification of differential metabolites (Ernest et al., 2012).

3.1.2. Existing solutions for removing batch effects

For microarray-based expression profiling, sophisticated normalization methods that directly incorporate batch adjustments in statistical models such as ComBat (Johnson et al., 2007), Surrogate Variable Analysis (Leek and Storey, 2007), and Remove Unwanted Variation (Gagnon-Bartsch and Speed, 2012) have been developed. A thorough review on batch effects and their removal methods for microarray-based expression studies is provided in (Lazar et al., 2013). For metabolomics experiments, which are susceptible to an even higher amount of unwanted variation than microarray studies, the development of batch effect correction procedures have been relatively minimal (De Livera et al., 2012; Ernest et al., 2012; Wang et al., 2013a). However, batch effect removal procedures using these applications require a specific experimental design. Thus, these applications could not provide a ready solution for experiments having customized sampling strategies. The specific details pertaining to each statistical application mentioned above is discussed later in this Chapter.

Unsupervised methods such as PCA have been successfully used for capturing systematic variation due to latent variables (such as batch effects) in a large datasets (Alter et al., 2000; Leek et al., 2010; Leek and Storey, 2007). After normalization and scaling, PCA makes use of co-variances or correlations among the metabolite features to decompose the original data matrix onto a lower dimensional space. The reduction of the variation across thousands of features resulting from biological differences and

confounded with the influence of nuisance latent variables into orthogonal PCs, provides a robust statistical measure to quantitatively characterize the metabolite data structure. The PCs are actually linear combinations of the original data variables. Maximum variation and the most interesting phenomena is typically observed in the first few PC loadings, with subsequent components explaining decreasing amounts of variation (Ivosev et al., 2008; Liland, 2011). Visualization of the PCs that describe the maximum variation in the data structure can reveal the underlying relationships between the metabolite features as a manifestation of biological differences or batch effects. PCA shares a close mathematical relationship with singular value decomposition (SVD) (Alter et al., 2000; Shlens, 2014). Typically, SVD-based calculations are carried out within PCA.

In this Chapter, I developed a statistical approach for removing batch effects in the large-scale untargeted metabolomics data using SVD. I developed this approach using data obtained from a survey of natural variation in oleaginous algal species. Specifically, metabolite measurements were obtained from 22 *Chlorella* strains, compared over two growth phases and run in four batches, to remove structure in data related to day of sample assay (268 samples being run in 4 batches spread over a month). Here, the term batch refers to a collection of samples processed at a particular instance using the same instrument under identical conditions. In this study, batch refers to the day on which the samples were processed in the mass spectrometer (abbreviated as RunDay). Furthermore, batch is an all-encompassing term for both observed and unobserved variation affecting the samples processed in a particular day.

3.2. Materials and methods

3.2.1. Experimental design

In order to survey the natural variation in oleaginous microalgae and identify strains for efficient biofuel production, 22 algal strains were isolated by colleagues at the University of Malaya (Vello et al., 2014) from 7 different locations and a total of

16 different habitats in Malaysia (refer to Chapter 4 for detailed sampling strategy). These strains were then isolated and cultured in laboratory conditions (as described in Vello et al., 2014). We then performed untargeted metabolite profiling for these 22 strains at 2 growth phases, namely, exponential (between day 4 to 6) and stationary (at day 12). A total of 264 samples (262 samples, as some strains did not have all the replicates), *22 Strains* X *2 Growth Stages* X *3 Biological replicates* X *2 Technical replicates*, apart from blanks (for determining instrument or analytical noise/errors) and matrices (sample extraction matrix) were profiled. Blanks undergo the entire extraction process, but without the sample material. They were run after each set of 6 samples (3 biological replicates and 2 technical replicates from one strain). The large number of samples required metabolite profiling to be carried out over 4 different batches spread over 2 weeks (May 15 to May 23, 2013) (Figure 3.1).

Such experimental designs are prone to batch effects, as over long periods of time, instrument characteristics might change. Furthermore, samples that are collected at the same time but profiled in batches might also face issues with sample degradation. This can be minimized by storing the samples at -80°C without any freeze/thaw cycles. In cases where the sample numbers are large, experiments can only be run in batches. Such experiments should be carefully designed with adequate randomization procedures used both during extraction and MS analysis stage. Furthermore, to utilize statistical techniques that can handle and correct for batch effects, multiple internal standards, along with pooled biological samples should be run in a randomized order in each batch. The experimental design used in this study did not use (i) randomized extraction or MS run order and (ii) did not have pooled biological samples or internal standards, thus, the metabolite profiles were influenced by batch-specific non-biological sources of variation.

3.2.2. Metabolome profiling

3.2.2.1. Metabolite extraction

Cell disruption was achieved using bead beating of the cell pellets, specifically a Tomy micro smash MS 100 bead beater (Tomy Seiko Co., Tokyo, Japan) along with lysing matrix Y from MP Biomedical lysing kit (MP Biomedicals, Solon, OH). The screw-top micro centrifuge tubes containing Lysing matrix constituted 0.5 mm diameter Yttria-Stabilized Zirconium oxide beads. Lyophilized algae cells weighing 50 mg were added to lysing matrix along with 1 ml pre cooled 80% methanol. The sample was subjected to bead beating at 4000 rpm for 20 s and then thawed in ice for a minute. This procedure was repeated five times, the extract was then centrifuged twice at 11,000 x g for 10 min at 4°C. The supernatant was pipetted out and filtered through 0.2 µm syringe filters (Sartorius Stedim Biotech). The final extract was kept in -80°C refrigerator till the analysis. Metabolite extraction was performed by Ms. Vejeysri Vello (University of Malaya).

3.2.2.2. LC-MS analysis

Chromatography separations were carried out using a Zorbax Eclipse Plus-C18 (2.1x50 mm, 1.8-µ) reverse phase column on an Agilent Infinite 1290 UPLC system. The temperatures for column and auto sampler were 50°C and 7°C, respectively. The mobile phase consisted of de-ionized water with 0.1% formic acid (solvent A) and LCMS grade acetonitrile (ACN) with 0.1% formic acid (solvent B). A gradient elution was conducted for separations using the following method: isocratic elution with 5% B for 0.5 min, followed by a 10 min gradient to 98% B, which was kept for 2 min, then re-equilibrated at 5% B for 2.5 min. The flow rate was 0.3 mL/min and injection volume was 3 µL. The samples were subjected to an Agilent Q-TOF 6540 mass spectrometer after separation through liquid chromatography. The analysis was carried out in positive mode with ESI as source for ions with a mass range between 50 to 1200

m/z. The nebulizer pressure (psi), source gas temperature (°C), dry and sheet gas flow (L/hour), capillary voltage (V) and sheet gas temperatures (°C) were 40, 250, 12, 12, 4000 and 350, respectively. LC-MS profiling was performed by Dr. Peter Benke and Mr. Vinay Kumar from (Metabolites Biology Lab, NUS).

For the data-dependent MS/MS, UHPLC system with column was setup in-line with mass spectrometer, with a 14 min long separation method same as described above for untargeted metabolic profiling. Parameters used were: drying gas temperature at 250ºC with 12L/min (nitrogen) flow rate, nebulizer gas at 40 psi, sheath gas temperature at 350ºC with 12L/min (nitrogen) flow rate, capillary voltage at 4000V, nozzle voltage 1500V, skimmer voltage 65.0V, fragmentor voltage 100V and octopole RFPeak voltage 750V. Parameters for precursor selections were: fixed collision energy at 20eV and 40eV, max precursors per cycle at 10, threshold (absolute) at 100cps, active exclusion enabled with exclusion after 2 occurrences and release of active exclusion after 30 s. Data acquisition was performed in centroid mode at the resolution of 30,000 with MS scan rate set at 8 spectra/s and MS/MS scan rate set at 4 spectra/s. Metabolite identification for MS/MS data is ongoing, the results presented in this chapter use the MS1 data extracted from the MS/MS dataset for validating SVD-based approach.

3.2.3. Metabolomics data analysis

3.2.3.1. Data processing and analysis

Raw data files from Q-TOF (.d files) were converted into mzXML format using msconvert of the ProteoWizard suite (Chambers et al., 2012; Kessner et al., 2008). The parameters defined for Q-TOF were optimized for this dataset (method = ‘centWave’, ppm = 30, peakwidth = c(5,60), prefilter = c(0,0), snthresh=6, peak grouping: bw = 5, minsamp = 1, mzwid = 0.015; retention time correction algorithm: ‘obiwarp’) (Patti et al., 2012), a total of 67,467 features were extracted using XCMS package (version 1.38) (Smith et al., 2006) in statistical programming language R (version 3.01) (R Core

Team, 2014). This produces a data matrix where samples are represented in columns and the metabolite features in rows. In the resulting data matrix, each row (mass features) is characterized by a unique mass-by-charge ratio (m/z) and a retention time (rt; or the time taken for an ion to elute through the chromatography column). The columns of this data matrix provide the abundance or the counts (number representing concentration of that particular mass feature in the sample).

All statistical analysis were performed using R version 3.01. Exploratory data analysis was performed using R and datPAV. Version control has been implemented using Git in RStudio™. All the scripts will be uploaded onto GitHub upon publication.

Log transformation followed by centering and scaling was performed on the dataset during the data pre-processing stages. Raw Total Ion Chromatograms (TIC) were developed to obtain a visual representation of the reproducibility of metabolic profiles between biological and technical replicates. TICs plot the retention time on the x-axis, and the total ion current detected for all features/ions at that particular instance on the y-axis. The similarity between the TICs for the replicates indicate the similarity or differences in the metabolite profiles (including the abundance of each metabolite) between the replicates. Thus, visualizing TICs provides a quick measure of the relationship between the metabolite profiles of samples. Variation in the TIC profiles between replicates can indicate whether certain replicates of strains are potential outliers. An example is shown in Figure 3.2A, where the technical replicate-D4_104_b3_r002 (here D4 indicates growth stage, 104 is the strain number, b1 is biological replicate 1 and r002 is the second technical replicate of that biological replicate) shown as a pink line clearly has a different profile. The coefficient of variation (CV) was calculated for each replicate of a strain. Specifically, the CV of the abundances of all features in a particular replicate was calculated, and this was done for each replicate of a strain. In an ideal scenario, the replicates will have similar CVs as the metabolite abundances should be similar. However, replicates that had different TIC profiles had a markedly different coefficient of variation. For example, the

coefficients of variation for the replicates of strain D4_104, were 43.62, 41.49, 43.44, 42.61, 41.31 and 75.72 for D4_104_b1_r001, D4_104_b1_r002, D4_104_b2_r001, D4_104_b2_r002, D4_104_b3_r001 and D4_104_b3_r002, respectively. Based on the CVs, we can conclude that the variation in the abundances of features in D4_104_b3_r002, and therefore the metabolite profile, is different compared to the other replicates of strain D4_104_b3_r002.

Figure 3.2B shows the distribution of the abundances for all the features in each replicate as a boxplot. Furthermore, to investigate the cause of variation in the CVs and distribution in the abundances of features between replicates, the number of missing values (zeroes) in each replicate were analysed. Interestingly, the outlier replicate that showed a different profile had a 20% increase in the number of features whose abundances were zero (Figure 3.2C). We can conclude that TICs, CV and the number of zeroes are inherently related to each other for each replicate. This observation was as expected as data points analysed here are abundances of features, thus are non-negative in nature and have a minimum value of zero. Therefore, the abundances are likely to follow a mixture of Poisson-distribution (for non-negative abundance) and normal-distribution (for zeroes). This observation suggested a heuristic strategy for identifying and removing outliers.

To identify outliers among the CVs of replicates of a strain, boxplot statistics was used. Specifically, the CVs of replicates that were 1.5 times the interquartile range, above the upper quartile and below the lower quartile were deemed as outliers. Using this strategy, we identified the following strains to contain outliers (the number of outlier replicates in each strain is indicated in the brackets): D12_001 (1), D12_006 (1), D12_014 (1), D12_051 (1), D12_177 (1), D12_187 (1), D12_207 (2), D12_252 (2), D12_255 (1), D12_258 (1), D4_014 (1), D4_094 (1), D4_104 (1), D4_245 (1), D4_254 (1), D4_268 (1) and D4_325 (1). Thus, after removing 19 outlier samples, the final data matrix used for analysis contained a total of 243 samples (125 from day 4 and 118 from day 12).

Figure 3.1. Sample (Strain) allocation to different batches. The rows indicate the strain and columns indicate the batch in which they were processed. Each batch indicates a separate day in which the samples allocated to that batch were processed.

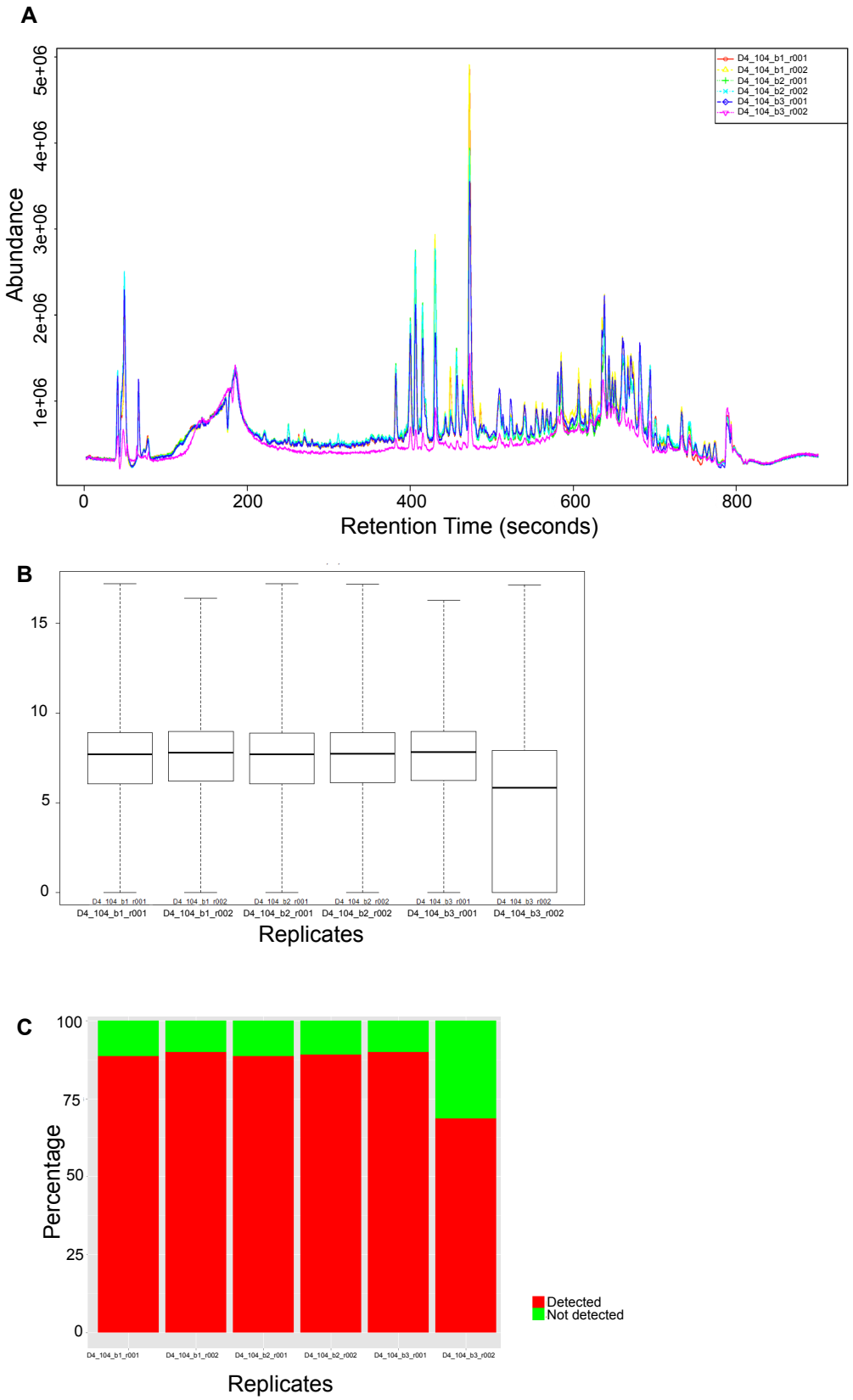


Figure 3.2. Identification of outliers within replicates (A) TICs of the biological and technical replicates of a strain; (B) Boxplot showing the distribution of metabolite abundances in each replicate; (C) For each sample, the percentage of features whose abundances were zero is indicated in green shade. Red shade indicates the number of features whose metabolite abundances were detected.

3.2.3.2. Missing values

Missing values arise when a feature is identified in some samples but is not detected in others. For the sample in which the feature is not detected, the abundances are marked as zeroes or 'NA' in the metabolite data matrix. Thus, due to biological or technical variation between replicates of a sample, some replicates might contain a high number of missing values. For example, the feature might be present but not be detected in a sample if the concentration is below instrument limits or if there were analytical errors (Karpievitch et al., 2012). These missing values are a key challenge in quantitative analysis as they influence the distribution and variance of the metabolite profiles. Furthermore, the reason behind the missing-ness often cannot be easily determined, and a number of approaches have been developed for imputation of missing values (Gromski et al., 2014).

In this study, to ensure that absence of certain features were not a result of software limitation, the actual chromatograms in the form of raw TICs were investigated to determine whether features really had missing values. These raw TIC plots were analysed for all the samples before initiating xcms-based data processing. Finally, in xcms, $fillpeaks$ function that identifies peak groups where the sample is not represented and then integrates the signal in the region of that peak from the raw data, was used as a data imputation step.

In typical metabolomics experiments, filters are set to select ions whose abundances are above a certain threshold such as 500 or 1000 counts. However, for our analysis, to obtain raw unfiltered dataset, no such filter was applied. Features that had missing values even after the imputation step were removed. Features present across all samples were selected mainly to avoid scenarios wherein the features not detected by the instrument, have their values artificially imputed during SVD filtering approach. Thus, for this study, we used a Filter by Flag (Flagged for presence or absence of a metabolite) approach, wherein metabolite features detected and having an abundance

(intensity) value among all the samples at either exponential phase (day 4) or stationary phase (day 12) were used. This greatly reduced the data matrix from 67,467 features to 13, 443 features in exponential phase (day 4) and 10,687 in stationary phase (day 12). Figure 3.3 shows the number of zeroes (missing values) in the replicates before the Filter by Flag approach. As expected in Figure 3.3A the blanks towards the left (indicated by the red box) and the matrix towards the right (indicated by the green box) had the maximum number of missing values. This plot also provides an indication of the outliers, i.e. replicates of strains having an unusually high number of zeroes. Figures 3.3B and 3.3C provide a comparison of number of zeroes in each strain (sum of zeroes in the replicates) for each feature.

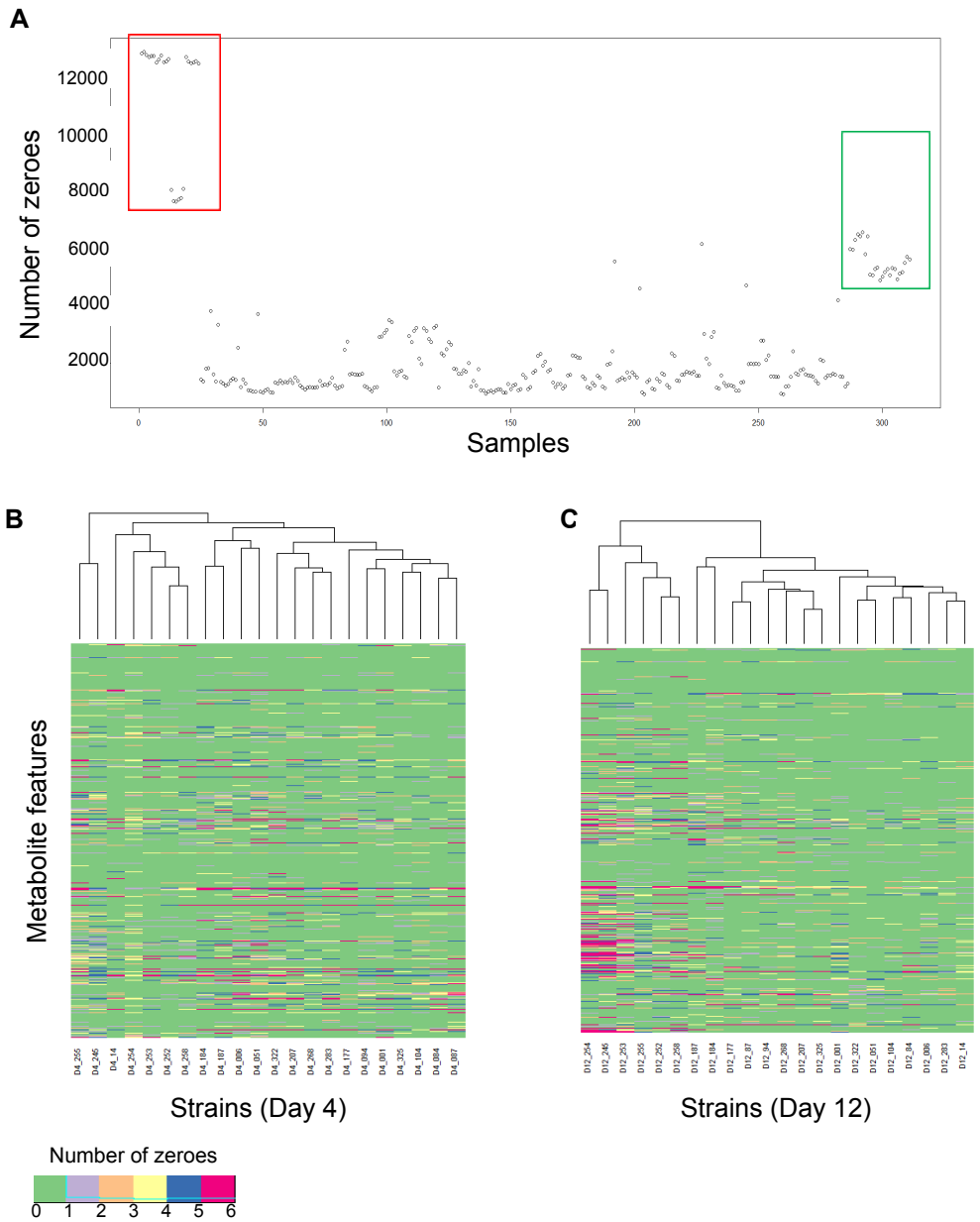


Figure 3.3. Identification of missing values (A) the number of zeroes in each sample is shown here. The red and green boxes indicate samples from blanks and matrix respectively. As blanks are without the biological sample and matrix is the sample buffer, the number of metabolites in each of these samples are very low compared to the metabolite profiles of the biological samples (strains). Thus, blanks and matrix contain a higher number of zeroes. (B) heatmap showing the number of zeroes for each strain and each feature at exponential phase-Day 4 and (C) stationary phase-Day 12.

3.3. Results and discussion

3.3.1. Identification of batch effects

"In high dimensional data, with far more measured variables than observations, it is almost always possible to find a satisfactory separation between two or more classes" (Saccenti et al., 2014).

The mass spectrometry-based metabolite profiling was performed in 4 batches. This could induce potential batch-specific variations in the metabolic profiles. An initial screening of the TICs of the blanks and matrix indicated that batch effects might influence the variation in metabolite profiles. Figure 3.4 shows the TICs of the blanks and matrix. Clearly there is a shift in the metabolic profiles in both blanks and matrix datasets. Furthermore, after pre-processing the data matrix, we used Bray-Curtis measure to create a dissimilarity matrix and visualize the separation using ordination plots (Figure 3.5). The points in the plots represent each blank or matrix sample. The points are coloured according to the RunDay. In experiments where batch effects do not influence the metabolic profiles, there would be no clear trends observed within either of the two subsets of control samples (blanks or matrix, respectively). However in this case, we observe RunDay to be a distinguishing factor between samples in both blank and matrix. This separation is indicative of instrument variations as the same sample preparation strategy was applied for all samples and in all batches.

We then attempted to study how the RunDay differences affected the samples. Exponential and stationary phase data sets were treated as two separate datasets mainly for the following reasons:

(i) Growth stage differences are themselves confounded within batches: From Figure 3.1, we observe that batch 4 did not have any samples from stationary phase. However, 14 out of 22 strains from the exponential phase were run in batch 4. Similarly 12 strains from stationary phase were processed in batch 2, while only 3 strains from the exponential phase were processed in the same batch. This experimental

design resulted in unequal number of samples in each batch, and batch differences were confounded with growth stage-specific differences, thus requiring complex modelling approaches. Furthermore, any modelling solution used here, not only had to factor in the nesting of strain within batch, but also the unequal distribution of samples from the two growth stages into different batches. Therefore, attributing the differences in the metabolite profiles to strain-specific, and growth stage-specific effects from the combined dataset would be complicated due to this experimental design.

(ii) Loss of information in the combined dataset: The main criterion for selecting mass features from the raw dataset for statistical analysis, was that a mass feature should be detected across all samples. When the datasets were treated separately, exponential phase had 13,443 features and stationary phase had 10,687 features that passed the above criterion. However in the combined dataset, only 9,421 features could be selected based on the above criterion. Thus, exponential phase and stationary phase when treated separately, had 4,021 (42%) and 1,266 (13%) features more, respectively. These differences could be attributed to the substantial and expected differences in the metabolic profiles of the strain during exponential and stationary growth stage. Therefore, treating the growth stages as a single experimental dataset would result in loss of growth stage-specific information, which were mainly determined from the features that were unique to each growth phase.

For these reasons, we decided to treat the metabolite profiles of exponential and stationary growth stages as separate datasets.

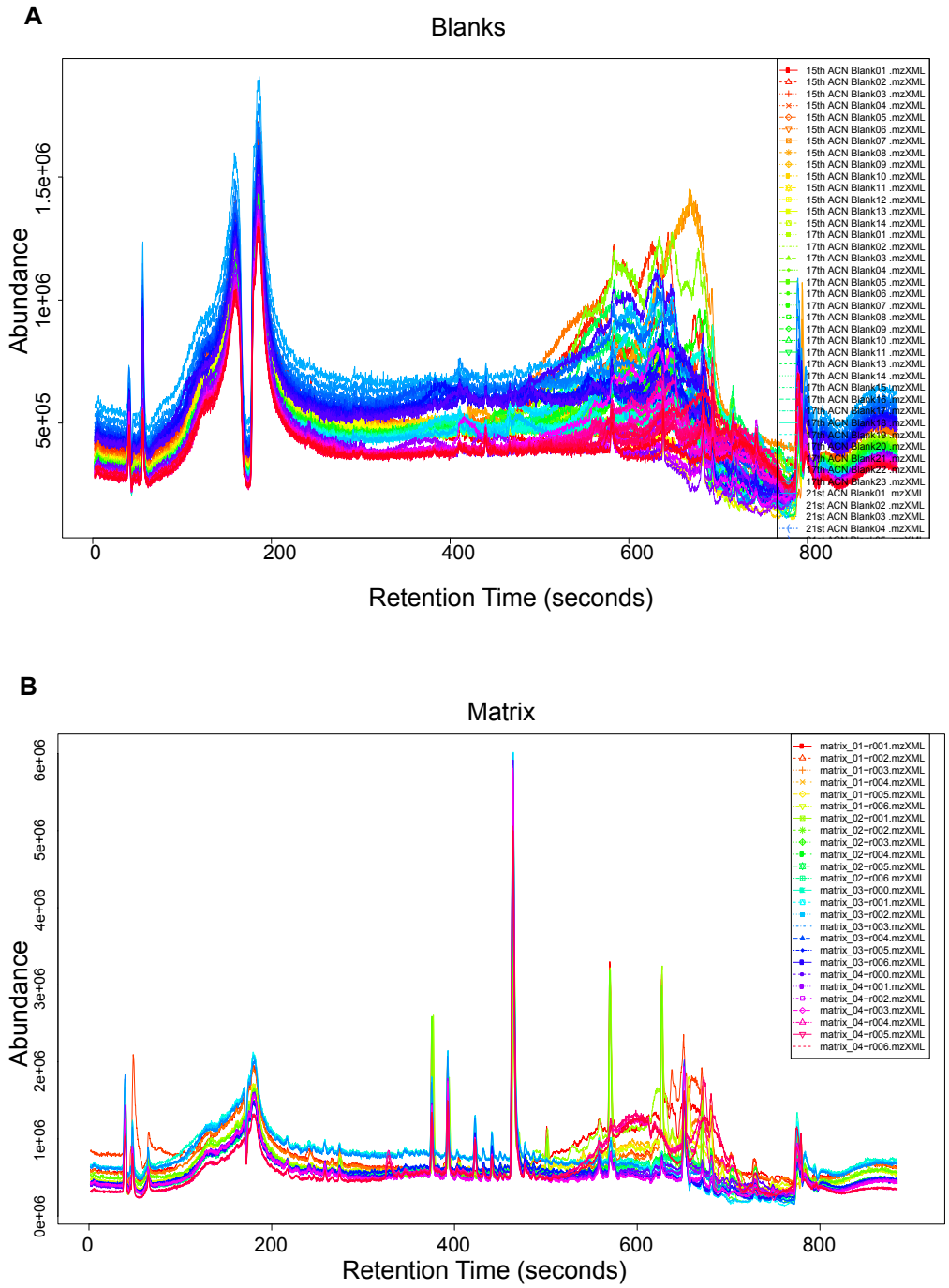


Figure 3.4. Total Ion Chromatograms of (A) Blanks and (B) Matrix. These TICs clearly indicate a shift in the metabolite profiles. In ideal conditions all the TICs should overlap with each other.

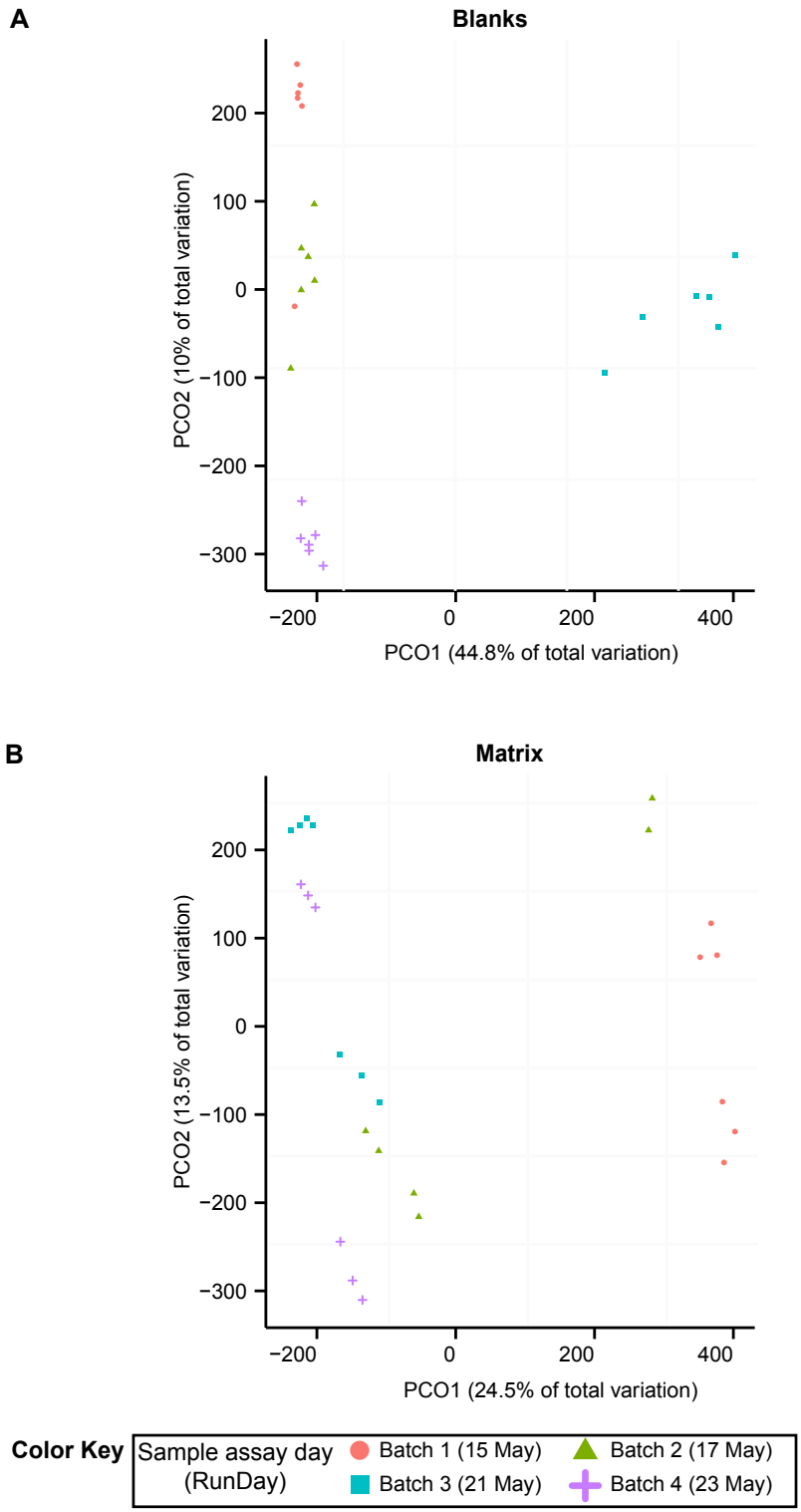


Figure 3.5. Principal coordinates analysis of (A) Blanks and (B) Matrix. Both the ordination plots show that blanks or matrix run in different batches- indicated by run day, have different metabolite profiles.

Using the same Bray-Curtis-based ordination analysis, we visualized the separation between strains at both exponential and stationary growth phases (Figure 3.6). As expected, we observed clear separation between the strains based on RunDay. Analysis of distance using the *adonis* function from 'vegan' package (Jari Oksanen, 2013) in R, was computed between the metabolite distance matrix and Strain or RunDay. We note that replicates cluster strongly within strains (for samples from exponential growth phase (day 4), the analysis of distance results using Bray-Curtis measure for calculating dissimilarities, indicated the R^2 values to be 0.564, *p*-values <0.001 and for stationary growth phase (day 12), R^2=0.634, *p*-values <0.001. Comparable results when Euclidean distance was used to define inter-sample distances). This rules out the possibility of the influence of carry over effects (in cases where the sample is not eluted completely during the run) or labelling errors (as the replicates within a strain clustered together). The points in the plots represent samples and are coloured based on RunDay to help visualize the relationships between strain-specific differences and batch effects. We also observed small but significant associations using Bray-Curtis measure with RunDay, with the coefficient of determination (R^2) values for day 4 being 0.190 and for day 12 it was 0.205, with *p*-values less than 0.001. Similarly results were observed when Euclidean distance was used. All calculations were performed with 999 permutations.

To quantitatively assess the associations between priori factors (RunDay and strain) and variation in the metabolite data structure, we used a linear regression model. First the feature level data matrix was decomposed into orthogonal PCs using *princomp* function from the 'stats' package in R. The covariance matrix was used for generating the eigen values and eigen vectors. To interpret the variation captured in each component in terms of strain-specific and RunDay effects, we calculated R^2 between the PC loadings and RunDay or strain as dependent factors. The significance of such correlations were tested using ANOVA.

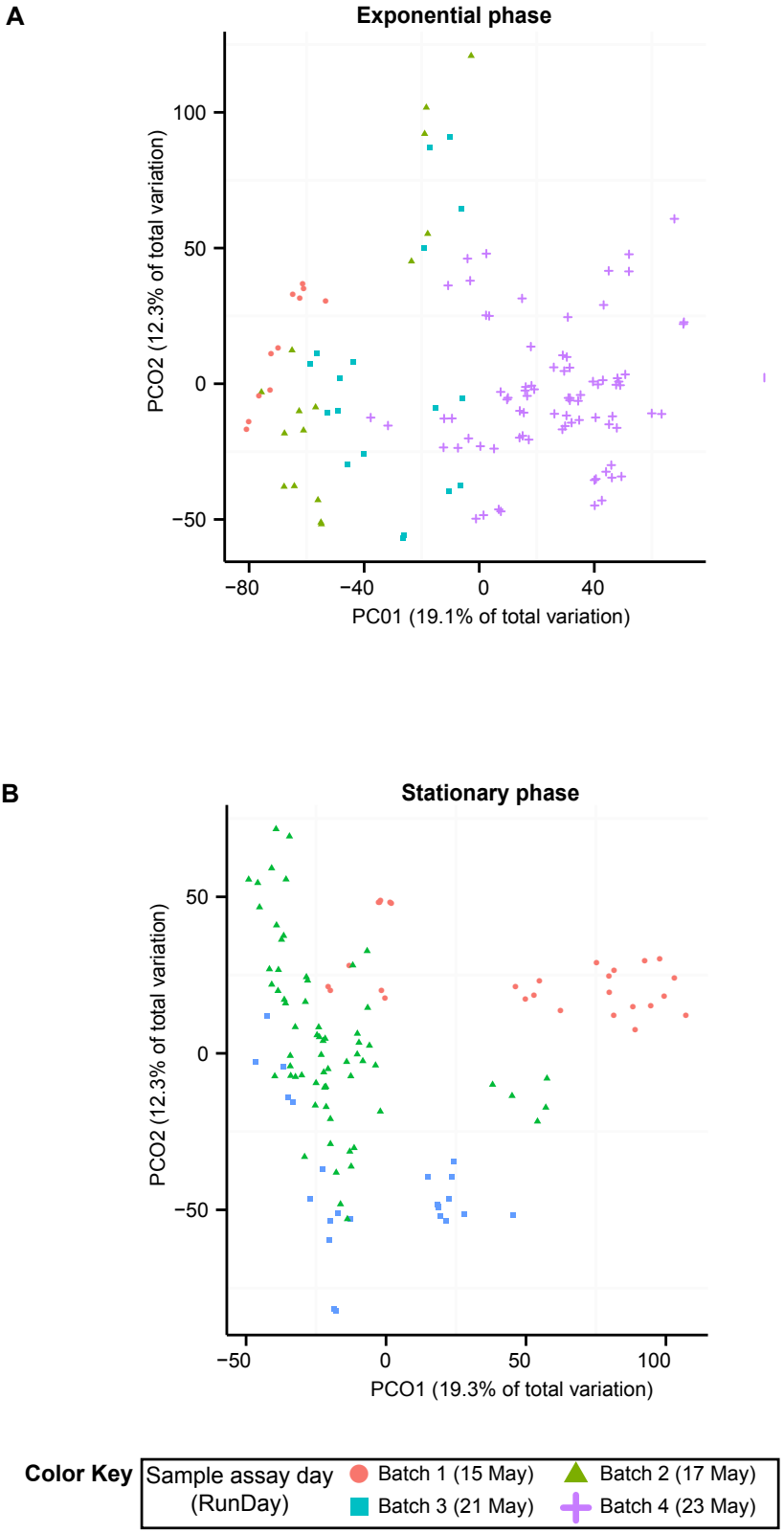


Figure 3.6. Principal coordinates analysis (plotted using PRIMER6) of strains (A) exponential phase-Day 4 and (B) stationary phase-Day 12. The run day (batch) in which the samples were run clearly influences the variation in both PC1 and PC2 for exponential and stationary phase.

It is important to note that both RunDay and strain-specific (biological) variation influences metabolic differences. The R^2 statistic and its associated *p*-values for both day 4 and day 12 are plotted in Figure 3.7 against their corresponding PC. We observe that both strain-specific association and RunDay effects are significantly associated with the leading PCs (blue box in Figure 3.7). Furthermore, strain-specific association survives on higher PCs but RunDay variation tends to disappear after the initial PC. This is shown in Figure 3.7 wherein the R^2 and *p*-values for RunDay (shaded green) become insignificant ($R^2 < 0.1$, *p*-values >0.1) after the top few PCs. Now that we have clearly established RunDay to be a significant batch effect, it is of critical importance to isolate and remove the effect of RunDay to observe natural variation or genotypic differences between the strains.

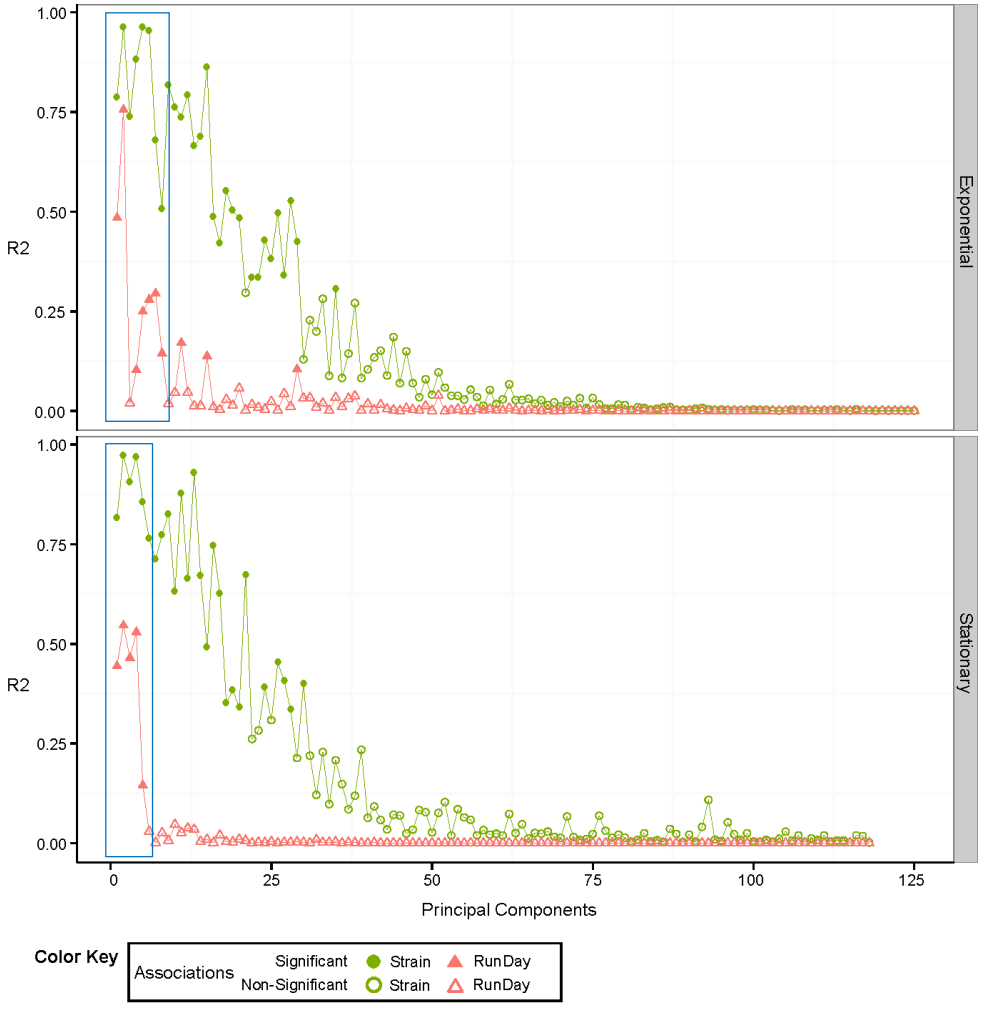


Figure 3.7. Associations between principal component loadings of metabolomics data with RunDay and strain at exponential-day 4 and stationary-day 12 phase. The blue boxes highlight the confounding effects between the batches- RunDay (shaded green) and biological differences (shaded red) on higher PCs. R2 values describing degree of association between each PC eigenvector and either RunDay or strain identity. Associations that are statistically significant are shown in closed symbols.

3.3.2. Current methods for removing unwanted variation

The current solutions for quantifying batch effects are based on (i) data normalization using internal standards or pooled quality controls, (ii) linear regression models, and (iii) advanced statistical techniques. We briefly discuss each of these approaches and why they were not suitable for this study.

3.3.2.1. Internal standard and quality control-based approaches

Internal standards (IS) are compounds or commercial standards that can be used to assess the reproducibility of the instrument. These compounds have chemical properties that enable them to be clearly distinguishable from the samples under investigation. For example, IS might contain unique peaks or elute at different retention times, thus making it easier to identify them from the peaks generated from the sample. Quality control (QC) samples on the other hand are pooled mixtures of the biological samples under investigation (Naz et al., 2014). In this approach, the idea is to capture all metabolites which have the potential to be detected by the instrument. In this manner, the instrument can sufficiently capture the entire spectrum of the metabolome under investigation. Dunn et al., 2012 provides an excellent review on the importance of QC's in metabolomics analysis(Dunn et al., 2012)(Dunn et al., 2012)(Dunn et al., 2012). Studies have shown that using quality control samples help mitigate batch effects better than using internal standards (Van Der Kloet et al., 2009).

On detecting batch effects, existing methods normalize the feature intensity to that of an internal standard. For example, the normalized log abundances for each non-IS or QC feature in each sample is obtained by subtracting from the log abundance of the IS or QC. This will introduce an additional bias if the features are not correlated

with the IS or QC samples. The other major drawback of using single internal standards is the assumption that unwanted variation occurs only after sample preparation and that every metabolite in a sample undergoes the same type of unwanted variation. This might not be true as the influence of batch effects might vary based on the chemical properties of the metabolite. Furthermore, selection of an appropriate IS also depends on knowing the chemical background of the samples under investigation. This is not possible for untargeted studies where the analytical properties of the samples are not known beforehand.

The unwanted variation detected using the IS might also be influenced by cross-contamination effects of the other metabolites. This has led to the development of methods such as cross-contribution compensating multiple standard normalization (CCMN) to identify and remove batch effects influenced by run order in GC-MS (Redestig et al., 2009). Such approaches provide an excellent solution for removing run order, provided adequate internal standards were chosen. If the unwanted variation was not due to the instrument, but resulted during sample extraction or collection periods, then using IS or QC will not remove these confounding factors. There are very limited strategies that can be used when even the QC samples or internal standards are affected by batch.

In this study, there were no dedicated QC or IS samples. Even then, as witnessed in Figure 3.5, the blanks and the matrix themselves showed significant batch effects. Thus, for such large-scale studies, multiple IS or advanced statistical techniques need to be utilized. For a smaller study, randomized sample and extraction order might have been feasible, however in this study it was not possible to partially extract samples from strains for each batch as it was not designed to facilitate a random order. In specific, in the absence of randomized design, retroactive randomization wherein samples are aliquoted each time from the same sample solution, in order to analyse the samples in different batches will lead to active degradation of the samples during the freeze and thaw cycles. Therefore, for experiments with large number of samples, a

clear randomized pattern for sample storage, extraction and MS analysis, with multiple pooled biological quality controls and internal standards should be included in the experimental design.

3.3.2.2. Linear regression

Confounding effects can be modelled as either fixed or random effects in a mixed linear model. If the latent variable (batch effect) is modelled as a fixed-effect then the resulting treatment groups means will include the latent variable. Thus, limiting the variance and confidence interval to be modelled only on the remaining residual errors and sample sizes. Furthermore, model inferences can only be estimated based on the latent variable. However, if the latent variable is modelled as a random effect, then this factor becomes a source of random variation and the experiment results will cover all probable scenarios under the influence of the latent variable. This approach is used in programs such as MetabR that implement linear mixed models to normalize metabolomics data based on fixed effect confounding variables (Ernest et al., 2012). MetabR requires the metabolite features to have a normal distribution, which is not feasible, as it is challenging to ensure normal distribution for environmental samples collected in an untargeted approach.

3.3.2.2.1. Linear model

We performed a naïve removal to RunDay effect and subtracted the influence of RunDay, by using the residuals from the above model to perform linear regression using strain as a factor. Our analysis showed that though batch effects were reduced, RunDay still had a significant association with metabolite features (Figure 3.8A). Using adjusted p-values, calculated using *mt.maxT* function from the multtest package in R, the number of significant features in exponential phase were determined to be 3,208 and stationary phase had 2,555 significant features (p-value < 0.05). Figure 3.8B shows the number of significant features associated with RunDay, strain, and common to both RunDay and strain.

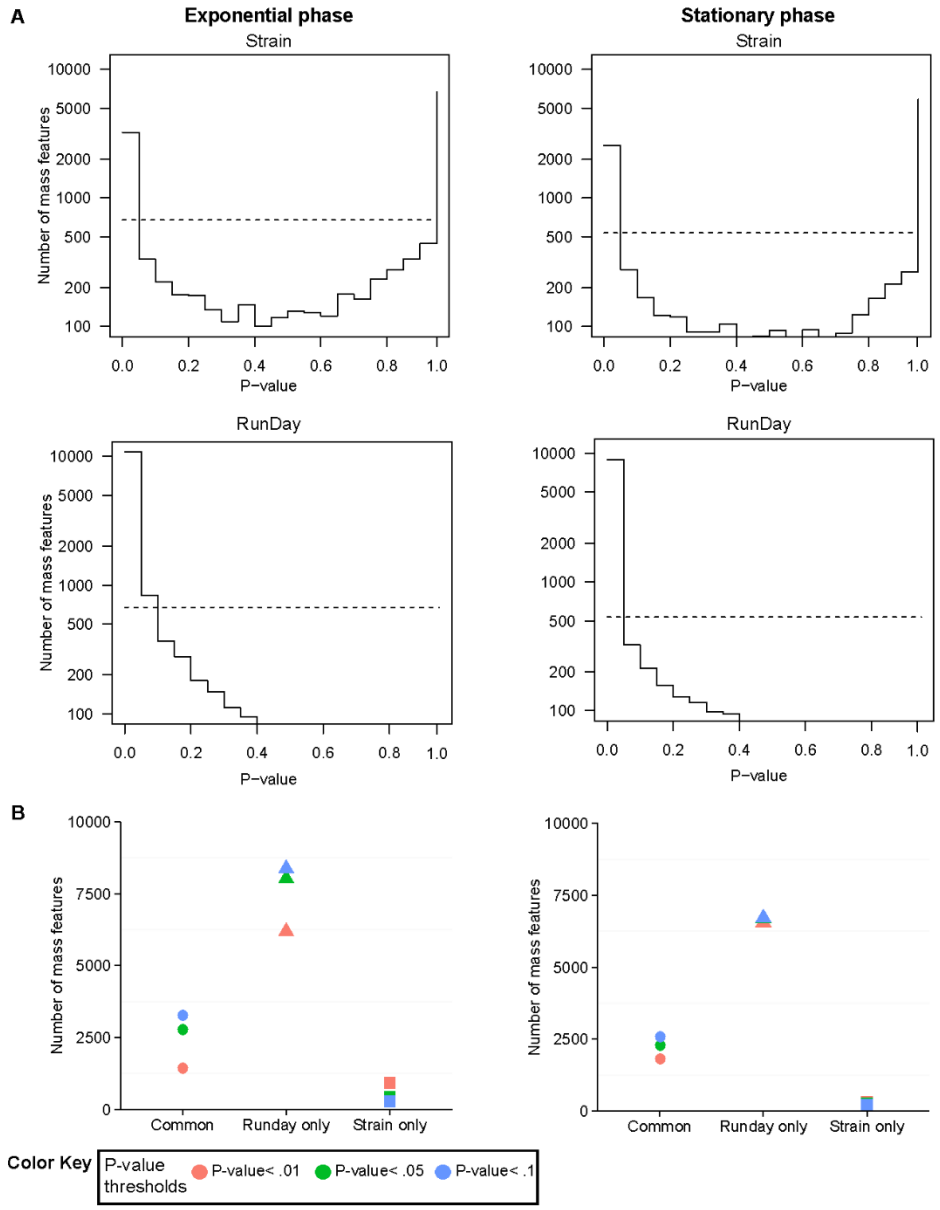


Figure 3.8. Naïve removal of RunDay effect using linear model. The left column shows the results obtained at exponential phase, whereas the right column shows the stationary phase. (A)The p-values are plotted as a histogram, with the x axis splitting the range of FDR adjusted p-values in bins of 0.05 and the y axis the corresponding number of features in that range. The dotted line indicates the number of significant features that are purely by chance alone ($p < 0.05$). The first row shows the distribution of p-values that associated with strain, whereas the second row shows the p-values associated with RunDay; (B) Significant features that are common between RunDay and strain, and unique to each RunDay or strain.

3.3.2.2.2. Nested linear model

A nested linear model was then used to test whether this approach can mitigate the batch effects and identify strain-specific significant features. The nested model, shown below, was used for each of exponential and stationary growth phase.

$$lm(x \sim as.factor(RunDayId) + as.factor(RunDayId)/as.factor(StrainId))$$

Strain and RunDay (batch) are treated as fixed effects as we are interested in differences in these specific days and strains. Furthermore, as strain is completely nested within each batch, it is fit within RunDay in a nested model. Residuals from the above model were calculated, and the residuals for each feature were tested against each of strain and RunDay (as factors). This analysis effectively tested for the presence of any feature in the residual for significant association to strain or RunDay, once the nested nature of these factors had been taken into account. The results from the above analysis did not detect any significant association of feature against either RunDay or strain, as the *p*-values of all such features from the residual dataset were above 0.1.

These results were as expected, as the main factors influencing variation in the metabolite profiles were strain-specific differences and batch effects. Therefore, modelling strain and RunDay as fixed effects showed that the residuals did not contain signal associated with strain or RunDay. The above model further supports the observation that strain-specific and RunDay-specific differences were the major factors that were driving the variation in the metabolite profiles.

Nested model does not reduce the influence of batch effects: The ideal scenario is, after fitting the nested model, we would want to detect a higher number of features that are associated with strain alone and minimum number of features associated with RunDay. We then estimated the number of significant features associated with (i) RunDay, and (ii) strain (Figure 3.9). We observed that the number of significant features associated with RunDay were 11,736 at exponential and 9,600 at stationary. Furthermore, 12,405 and 9,479 features were also significantly associated with strain. The distribution of *p*-values of these features are shown in Figure 3.9

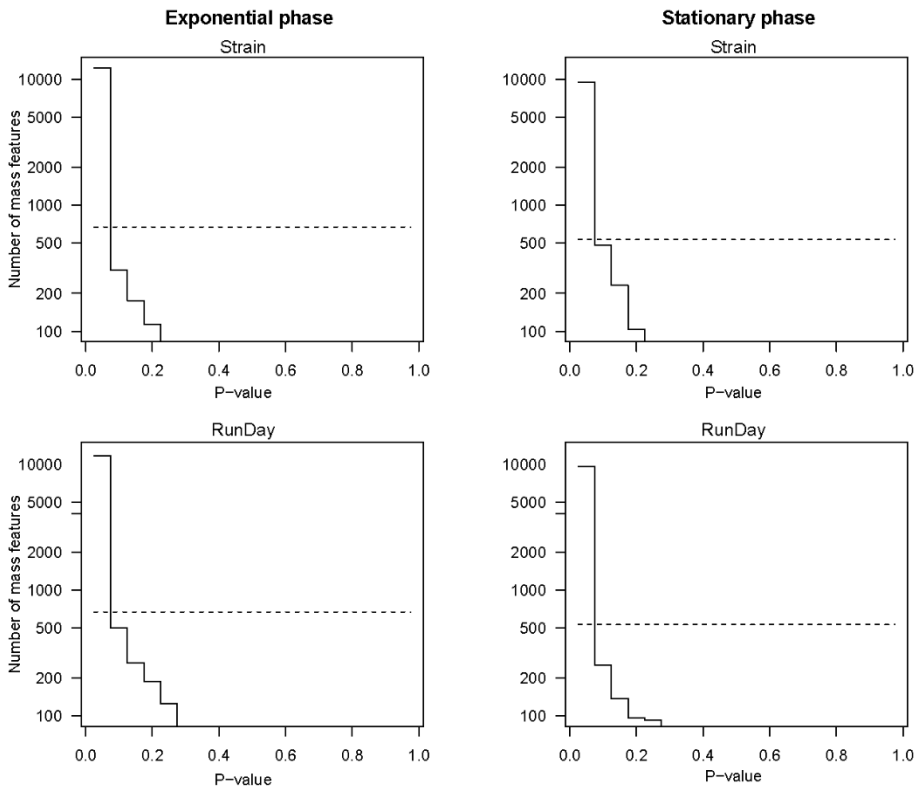


Figure 3.9. Significant features detected using a nested linear model. The first row shows the distribution of *p*-values that associated with strain, whereas the second row shows the *p*-values associated with RunDay. The dotted line indicates the number of significant features that are purely by chance alone (*p*-value < 0.05).

As we observed that the number of significant features that were associated with both RunDay and strain were quite similar, we then calculated the number of significant features that were (i) common to both (overlapping features of RunDay and strain), (ii) significant in RunDay only, (iii) significant in strain only. The results shown in Figure 3.10 indicate that a high number of features, that were detected as significant, were common to both RunDay and strain. Furthermore, the number of features that were detected as significant only in RunDay or strain is less than the number that are common to both RunDay and strain (Figure 3.10). These results further strengthen the observation, that for this experimental design, a nested linear model is unable to separate out the influence of batch effects from strain-specific differences.

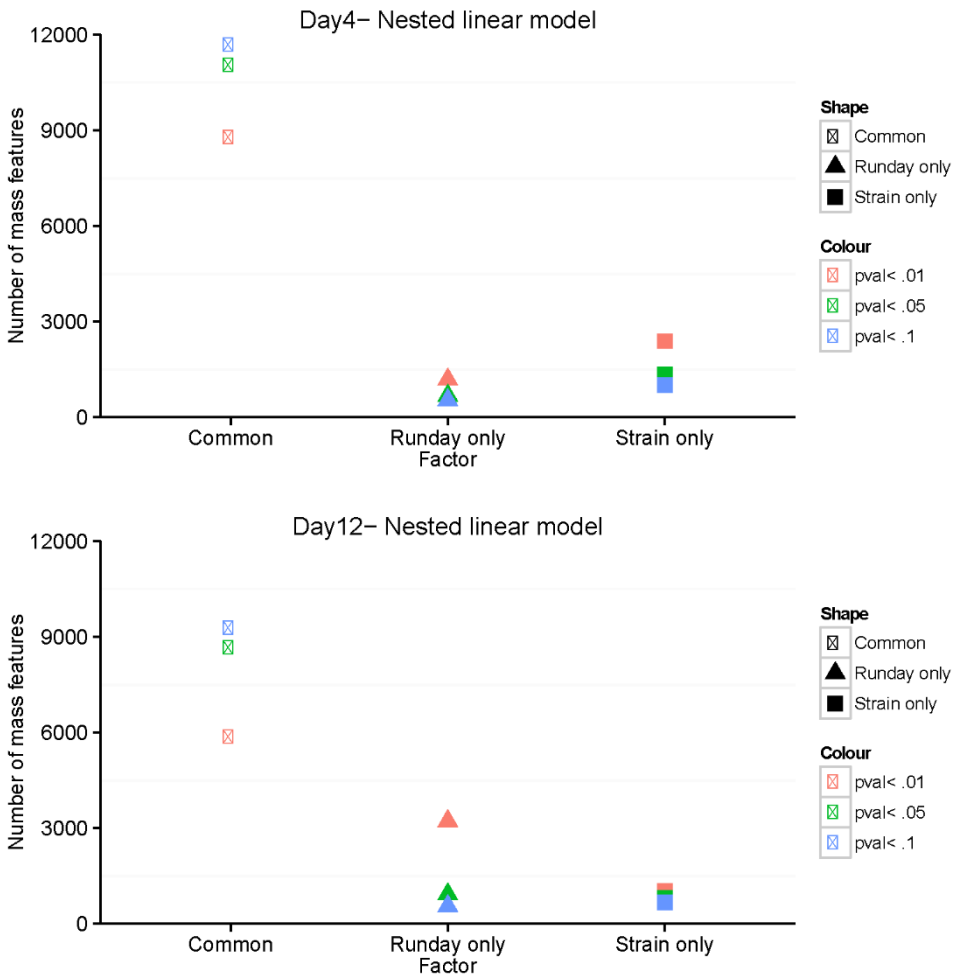


Figure 3.10. Overlaps between significant features detected using a nested linear model.

3.3.2.3. Advanced statistical models

These models have been designed to identify surrogate variables manifested in the form of batch effects. They identify specific parts of the data matrix that are affected by batch effects, and perform targeted removal of unwanted variation. ComBat (Johnson et al., 2007), uses an empirical Bayes framework to fit a linear model including both biological factors and batch covariates. ComBat works exceptionally well for small datasets when the latent variables are known (Chen et al., 2011). For large datasets, with potential unknown non-biological sources of variation factor analysis, methods such as surrogate variable analysis (SVA) (Leek et al., 2012; Leek and Storey, 2007) can be used. SVA combines SVD and linear model analysis to capture the unwanted variation due to multiple factors.

A number of approaches for removing unwanted variations or batch effects, use factor analysis. For example, Remove Unwanted Variation (RUV) (Gagnon-Bartsch and Speed, 2012) and for metabolomics RUV-2 (De Livera et al., 2012) use negative controls to detect the presence of unwanted variation and subsequently remove them using linear regression models. These negative controls are features that are not affected based on biological factors or experiment design. RUV and CCMN rely on the identification of specific compounds as negative controls. If these compounds are provided then it can identify compounds that are not affected by batch effects. Furthermore, in the absence of any non-changing features, RUV-2 can utilize internal standards or blanks as negative controls. The hypothesis is that any variation in the negative control set is only due to the influence of unwanted variation.

A recent report evaluating six batch effects correction methods for expression microarray data suggested ComBat to be the most effective (Chen et al., 2011). In the same review, the researchers mention that for cases when genuine biological variations are completely confounded with batch effects, none of the methods could effectively reduce batch effects.

In this study, RUV-2 and CCMN could not be applied as we did not have non-changing features in our dataset. Different strains being run on different days (Figure 3.1), led to strain being completely nested within RunDay and resulted in a singular design model matrix in ComBat. ComBat cannot invert design matrixes in a singular system. Using such an experimental design in ComBat, resulted in the following error:

Error in solve.default(t(design) %% design) : # Lapack routine dgesv: system is exactly singular: U[24,24] = 0*

SVA has been used for identifying surrogate variables, however, for this experiment, when SVA was used, the confounding effects of RunDay with strain-specific differences resulted in no surrogates being identified. A possible cause for this outcome is that the strains were completely nested within batches, therefore limiting the ability of SVA to detect surrogate variables. It will be an interesting future exercise

to test the exact reason that led SVA not to detect any surrogate variables. This can be tested by permuting the sample observations between strains and batches.

Motivation for an SVD-based approach:

From the Figures 3.6, 3.7, a clear progressive effect of batch differences could be detected. Figure 3.7 also showed that batch effect was associated with the leading PCs. These results suggested that a PCA-based approach could provide a solution to reduce the influence of batch effects. PCA-based approaches have been shown to effectively correct widespread batch effects (Leek and Storey, 2007; Pickrell et al., 2010). In a recent report by Goldinger et al, the authors compare the effectiveness of using PC-filtering approaches with approaches that use linear models. They put forward a caution that, PC-filtering might remove biologically relevant data. They state that the approach is suitable when linear models are not effective in removing batch effects. As shown above, for the experimental design used in this study, regression-based methods were ineffective in removing batch effects, and existing tools such as ComBat or SVA could not reconcile the confounding factors in the experimental design.

Confounding variation due to batch effects have been effectively mitigated by filtering out multiple PCs (Fehrmann et al., 2011; Goldinger et al., 2013; Price et al., 2006; Stranger et al., 2012). These reports highlighted the use of SVD-based filtering for mitigating batch effects. It is important to note that in an SVD-based approach when PC filtering is applied, there is a tendency to lose some information that is not associated with non-biological sources of variation. One of the arguments put forward by Goldinger et al as the limitation of PCA for mitigating batch effects is that the components of variation that contribute to each principal component are often unknown. However, in this study, we first computed the association between and the factors of interest, namely strain and RunDay (described in Section 3.3.1 and Figure 3.7). With this motivation, we attempted to investigate the effectiveness of using SVD-based filtering for mitigating batch effects from untargeted metabolomics data.

3.3.3. Removal of batch effects: Solution based on SVD

In this study, we utilize the data decomposition technique of SVD to develop a novel statistical framework for removing batch effects. Upon identifying PCs that are significantly associated with RunDay and strain, we employ a filtering procedure using SVD to minimize or remove batch effects.

This approach consists of 3 steps (Figure 3.11). The first step is to quantify the association between PC and the priori factors, in this case strains and RunDay (Figure 3.7) using analysis of variance models. This resulted in the identification of PCs that were most correlated with the surrogate and with the outcome. In the second step, the proposed algorithm decomposes the metabolite matrix into UDV^{T} components using SVD. Lastly, it nullifies a small number of singular values of the diagonal matrix D, based on their association with RunDay and finally re-computes the metabolite matrix preserving the overall strain-specific biological variation and removing confounding batch effects. The procedure below explains these steps in detail.

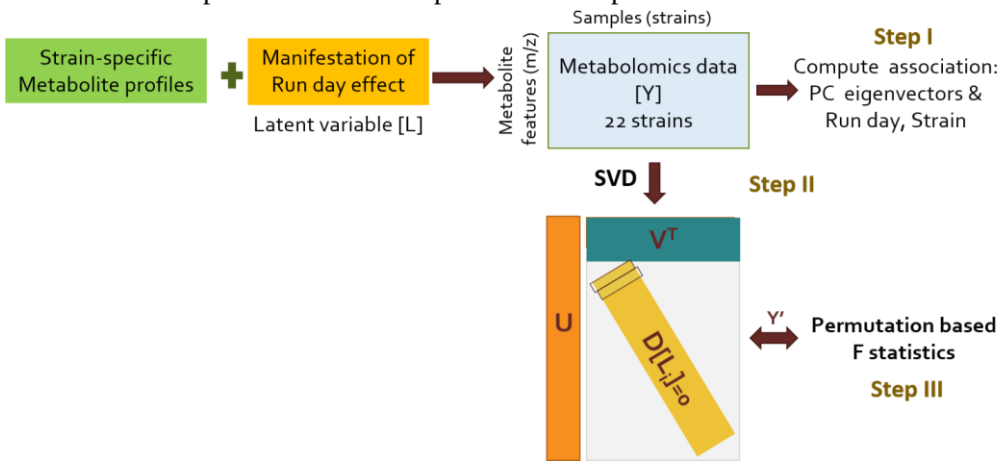


Figure 3.11. Procedure for batch effect removal using SVD

3.3.3.1. Approach to perform batch effect removal using SVD:

1. For a $m \times n$ scaled metabolite feature matrix Y, with m metabolite features observed over n samples and $m > n$, the variation in Y results from strain-specific metabolic differences confounded with systematic variation caused by RunDay effect. The RunDay effect is depicted here as a function of latent variable L.

2. To identify the influence of RunDay on the metabolite matrix, PCs of Y, which form the linear transformations of the original variables, are calculated using the *princomp* function on the covariance matrix in R.
3. Associations between PC (using loadings) of Y and influence of the RunDay latent variable L is calculated using the below given linear regression model Run day:$\mathrm{lm}(\mathrm{x}\sim \mathrm{as.factor}(\mathrm{RunDay}))$ and Strain:$\mathrm{lm}(\mathrm{x}\sim \mathrm{as.factor}(\mathrm{Strain}))$
4. A filtering procedure using SVD (*svd* function in R) is then performed on the scaled original metabolite matrix, yielding $Y = UDV^{\mathrm{T}}$. Here U is an $m \times n$ orthonormal matrix, D is a $n \times n$ diagonal matrix containing positive singular values, and V is an $n \times n$ orthonormal matrix. PCs are then rows of DV^{T}, where the i^{th} PC is found in the i^{th} row of DV^{T}. The columns of U are the loadings of their respective PCs. The above description is for cases when m > n. It should be noted that for real-valued data, U can be an $m \times n$ orthogonal matrix and D an $m \times n$ rectangular diagonal matrix.
5. The SVD algorithm was run on all components in the decomposed matrix. For each $\mathrm{PC_i}$ which was significantly associated (p-values <0.05) with run day (determined from the linear model), we replaced the same i^{th} singular values present in the diagonal matrix D with zero '0'. The diagonal matrix D with the replaced value was then used for recomputing the metabolite matrix, resulting in $Y^{*} = UD^{*}V^{\mathrm{T}}$.
6. By nullifying the i^{th} values of PCs that were significantly associated with RunDay and re-computing the new matrix, we effectively filtered out the variation influenced by RunDay. These steps are illustrated in Figure 3.12.
7. Furthermore, the re-computed matrixes were confirmed for efficient removal of RunDay effect by testing the number of differential metabolite features associated with RunDay and strain using permutation-based F-statistics. Each feature in the recomputed metabolite matrices was tested using the *mt.maxT*

function from the multtest package (Pollard et al., 2005) in R. This function provides permutation adjusted *p*-values for step-down multiple testing procedures. The null hypothesis corresponds to no differential metabolite features across samples when RunDay or strain was used as a factor. We performed 1000 permutations for each feature and iteratively tested the recomputed metabolite matrices to arrive at the stage where we could observe no significant features based on RunDay, whereas strain still had a significant effect. Due to the multiple testing, false discovery rate (FDR) adjusted *p*-values were generated using Benjamini-Hochberg procedure.

Test of effect size and inflation statistics: We tested for (i) effect size using correlation as function of test statistic, and (ii) inflation effects by observing the distribution of numerator and denominator of *F*-statistics.

Correlation as function of test statistic: Correlation as a function of test statistic was used as function to calculate the R^2 value associated with an *F*-statistic. The correlation coefficient can be used to measure of the strength of the effect rather than to test the significance of the effect (Rodgers and Nicewander, 1988). In the presence of multiple groups, Rodgers et al provided a measure to determine the relationship between the coefficient of determination R^2 and *F*-statistic through the formula $R2 = F(k - 1)/[F(k - 1) + (N - k)]$

In the above formula, F is the *F*-statistic, k is the number of groups (strain or RunDay), and N is the total number of samples. The R^2 value was calculated for each feature. Using the above relationship, the correlation statistic [effect] is computed for the *F*-statistics for each of batch and strain-specific effects. This relationship is plotted in the in the third rows of Figures 3.12 A, B and C. The features deemed as significant (FDR adjusted *p*-value < 0.05) after permutation-based tests are coloured in red. The significant features have a high R^2 and *F*-statistic, indicating that the model describes the variation well, and, the variation between group means are high. Expectedly,

features which did not have a significant FDR adjusted *p*-values, also had a low *F*-statistic value and low coefficient of determination.

We then tested for inflation of the *F*-statistics, related to the choice of whether shrinkage estimators are required. The numerator and the denominator of the *F*-statistics were obtained using the $mt.teststat.num.denum$ from the multtest package in R. The numerator assesses the variation of the means between groups. The denominator is an average of the sample variance estimated for each group. Artefactual inflation of test statistics may occur in cases of small differences (numerator) amplified by small within group variances, and a simple analysis was undertaken to explore this issue. In this analysis, the ratio of denominator to numerator is plotted in the last rows of Figure 3.12 A, B and C. These plots indicate the number features (in y axis) and ratio of the test statistic (in log scale in the x axis). In an ideal scenario the features that are deemed as significant should have low denominator to numerator ratio. Such ratios indicate that the between group means are largely different, and thus, the numerator has a high degree of variation. In Figures 3.12 A, B, C, the number of strain-specific significant features that have low denominator to numerator ratio is higher than those affected by batch effects (on the bottom left panel). As the PCs are removed, this trend increases for the strain-specific features, while decreasing for batch affected features. Taken together, these results help conclude that the SVD-based filtering is effectively removing only the variation associated with batch effects, as the strain-specific variation still exhibit significant ratios (with low values) between denominator-to-numerator of the *F*-statistic.

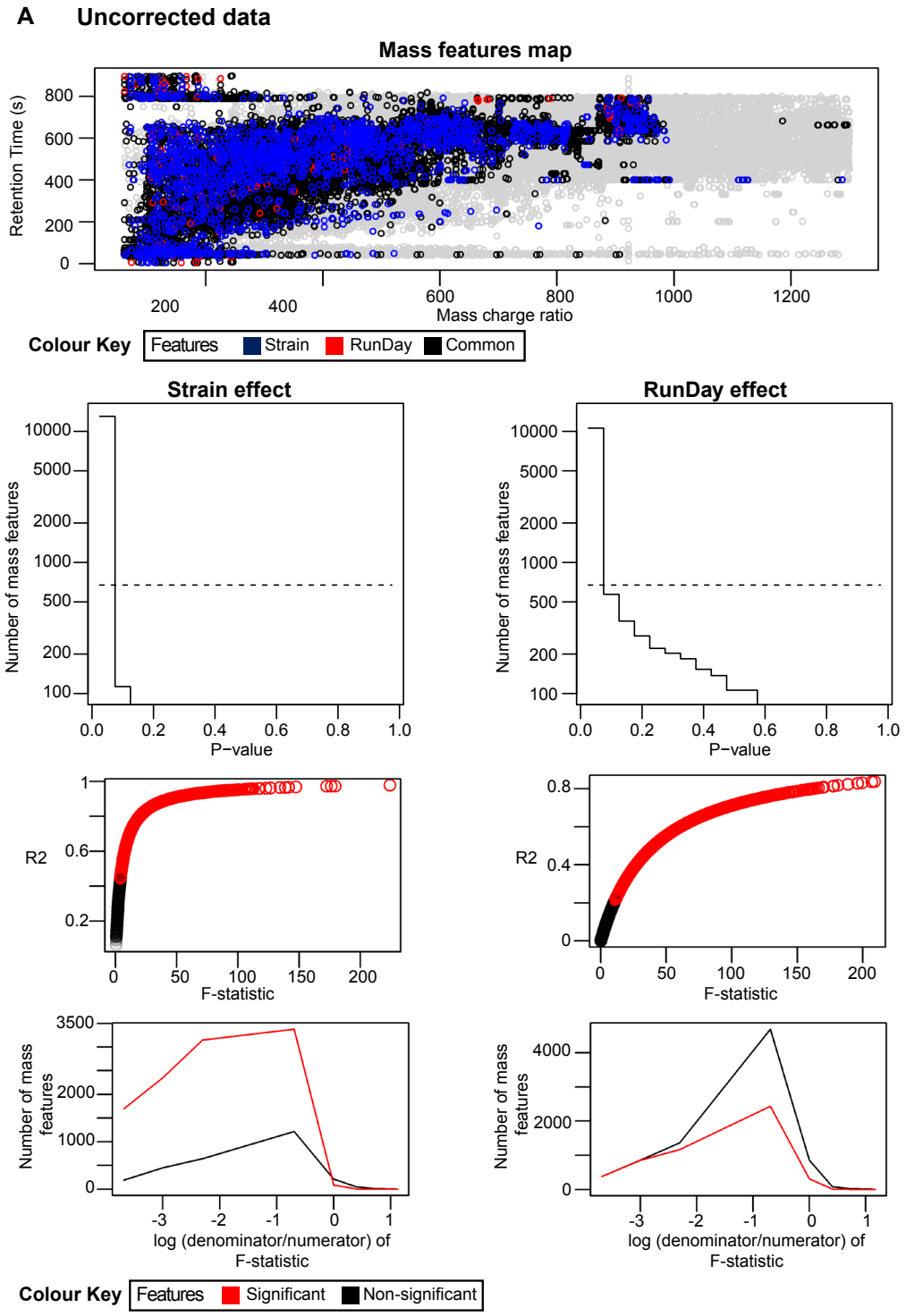


Figure 3.12. Illustration of batch effect removal using SVD. (A) The uncorrected data matrix is shown. The first row show displays the metabolite matrix in a mass-by-charge retention time plane. The grey shaded points represent the full data matrix. Points shaded in blue, red and black represent the significant features associated with strain, RunDay and those which are common, respectively. The second row shows the number of significant differential features (p-value < 0.05) associated with strain and RunDay. The left column shows the results obtained with strain as a factor, whereas the right column shows results with RunDay. The third row plots the relationship between R2 and F-statistic for each feature. The last row plots the distribution of the ratio of denominator to the numerator of the F-statistic.

B Removing top 5 PCs

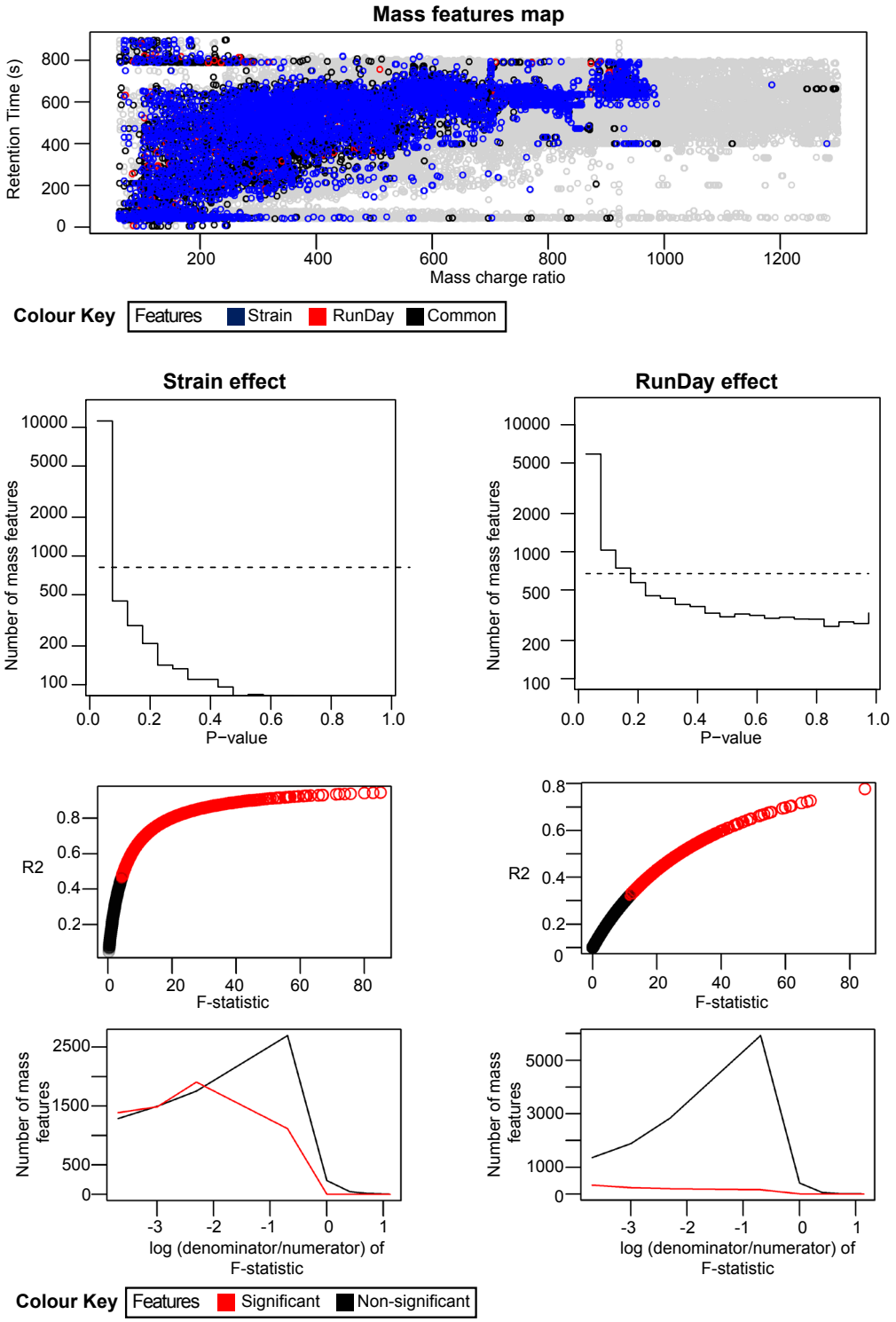


Figure 3.12B. The significant features associated with RunDay has reduced.

C Removing top 10 PCs

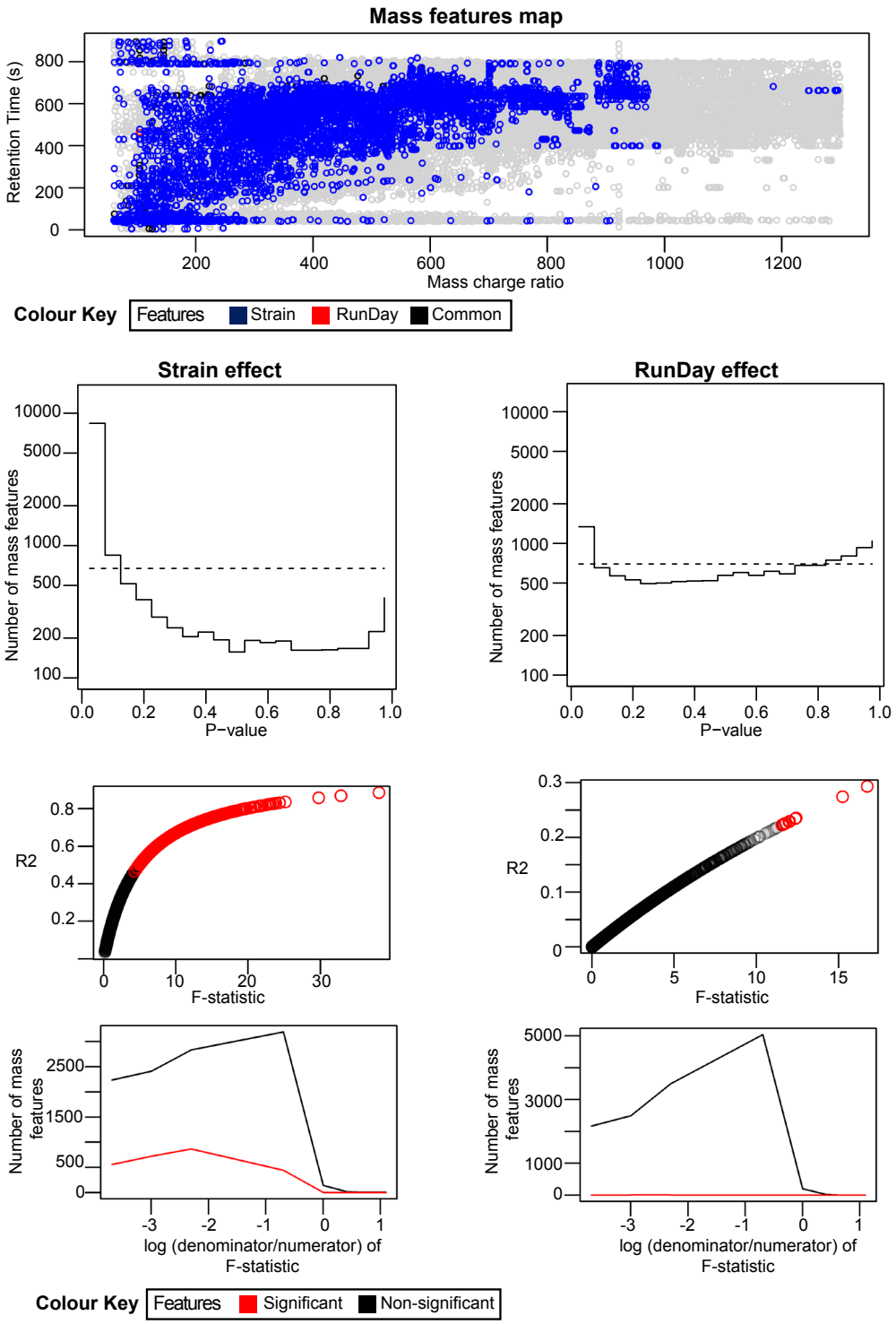


Figure 3.12C. The significant features associated with RunDay are negligible and batch effects are significantly reduced.

3.3.3.2. Results from the application of the batch effect removal algorithm

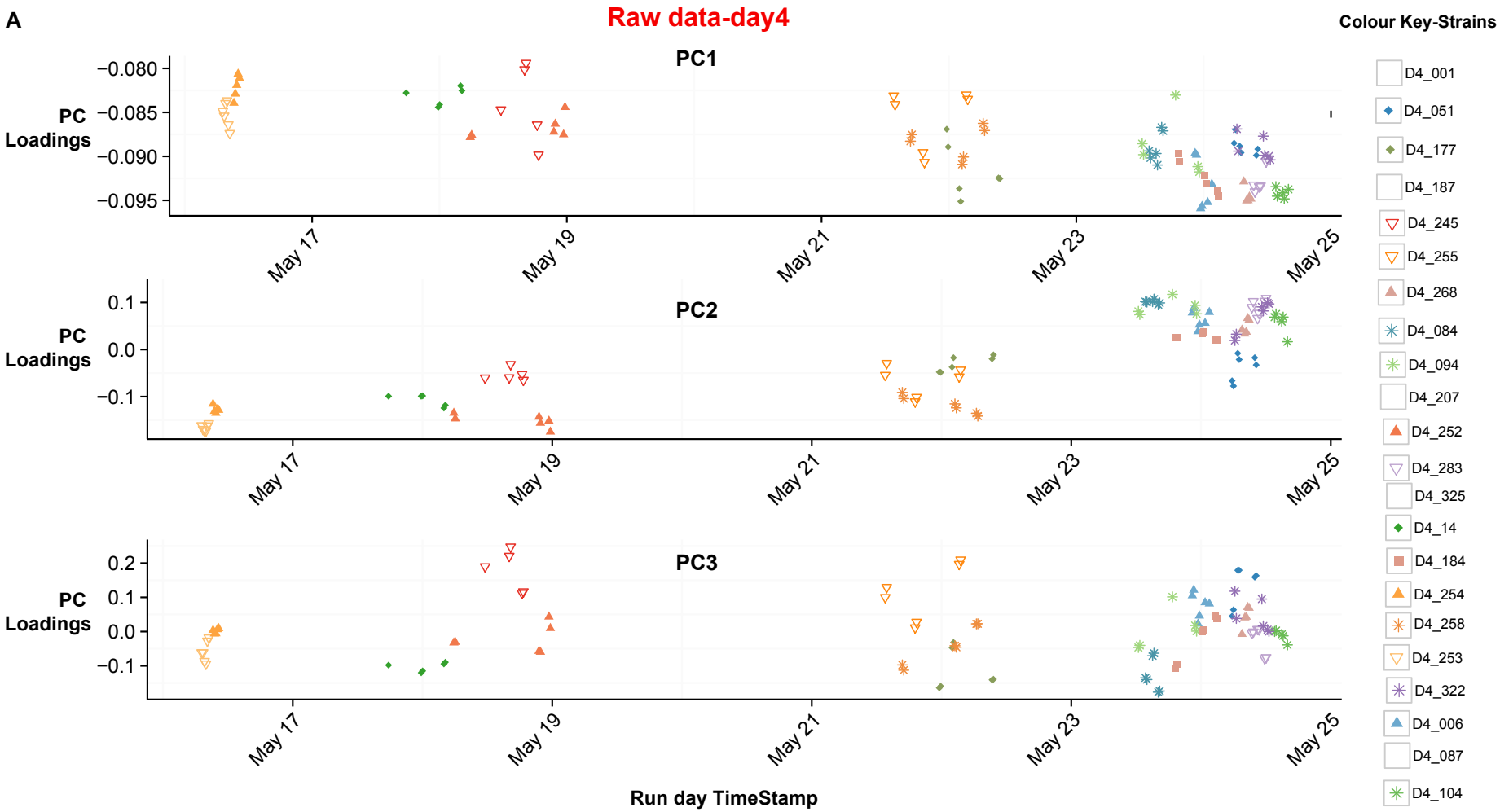
To test whether the proposed algorithm works well for removing unwanted variation, we would have to understand that the differences between strains are correlated with the unwanted variation as a result of batch differences. Thus, an effective solution will (i) reduce the number of features which are influenced by RunDay, (ii) identify features which are influenced only by strain-specific variation, and (iii) remove the influence of RunDay and preserve only the effect due to strains for features which are confounded with both RunDay and strain effects (Figure 3.12A, B, C).

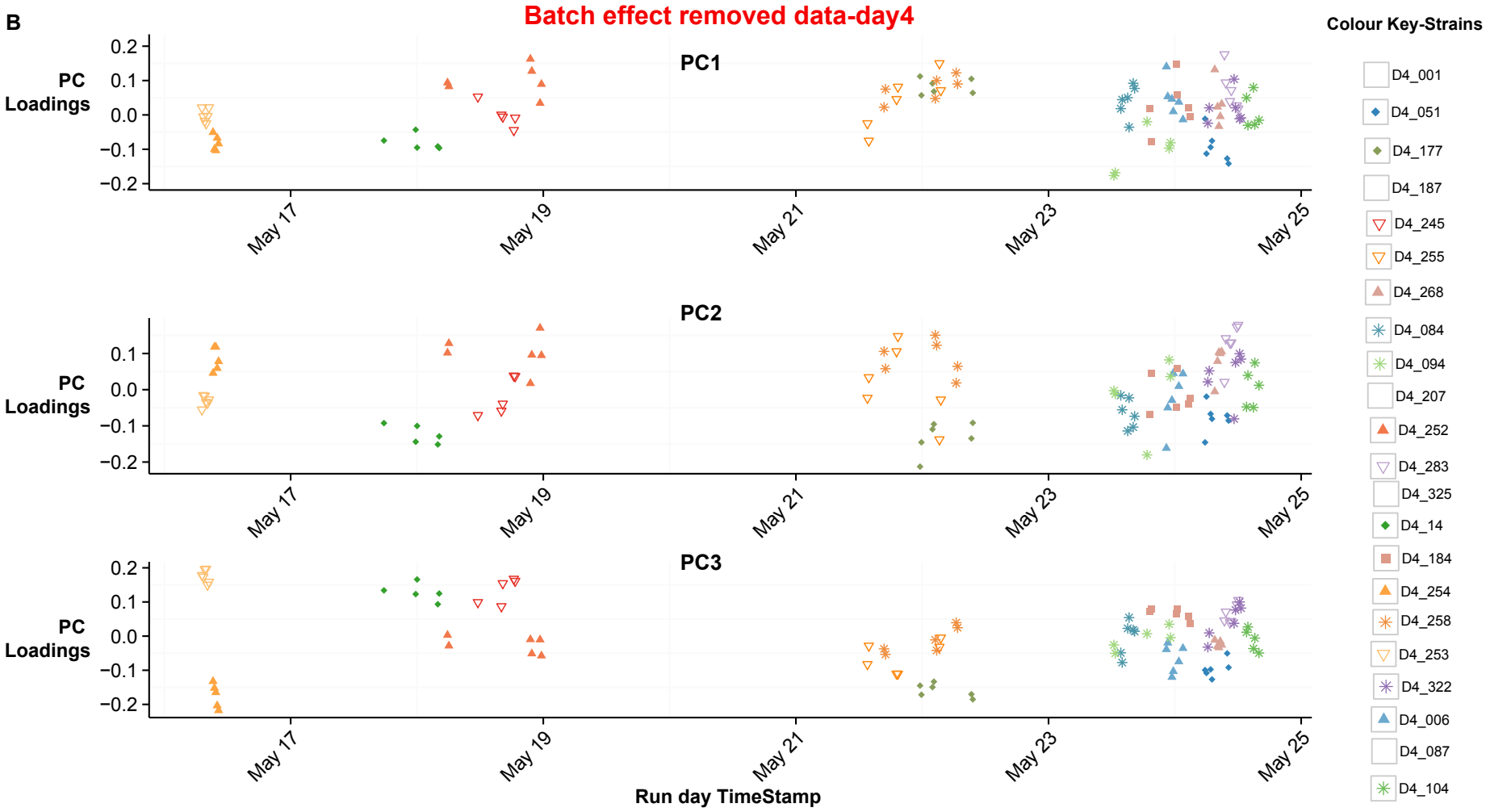
The above approach was used for removing batch effects from metabolite matrices of strains profiled at both exponential (day 4) and stationary phase (day 12). For both data matrices, significant association with RunDay was observed with mainly the top few PCs. Thus, the top PCs captured the systematic variation due to batch effects. Before batch correction, at exponential stage, there were 10,664 differential features between strains and 5,165 features showing significant differences based on RunDay (Table 3.1). Similarly for the day 12 matrix, we observed 9,044 and 5,116 features due to strain and RunDay differences, respectively. Using the above approach, batch effects were substantially minimized by nullifying the singular values of the first 4 PCs for stationary and first 7 PCs of exponential phase. For example, Figure 3.13 shows groupings of strains before and after batch correction. This illustration shows an increasing or decreasing trend in the raw (scaled) data due to RunDay in loadings of PCs 1, 2 and 3. The removal of RunDay associated variance resulted in the strains to cluster based on their biological differences. For stationary phase, we retained 18.99% of the residual variance of the original matrix containing 10,687 features. Out of this set, 5,878 features were significant due to strain-specific differences and 6 features significant due to RunDay effect. However the features affected by RunDay were different from those affect due to strain, hence were ignored during metabolite identification. For exponential phase, there were multiple sources influencing the

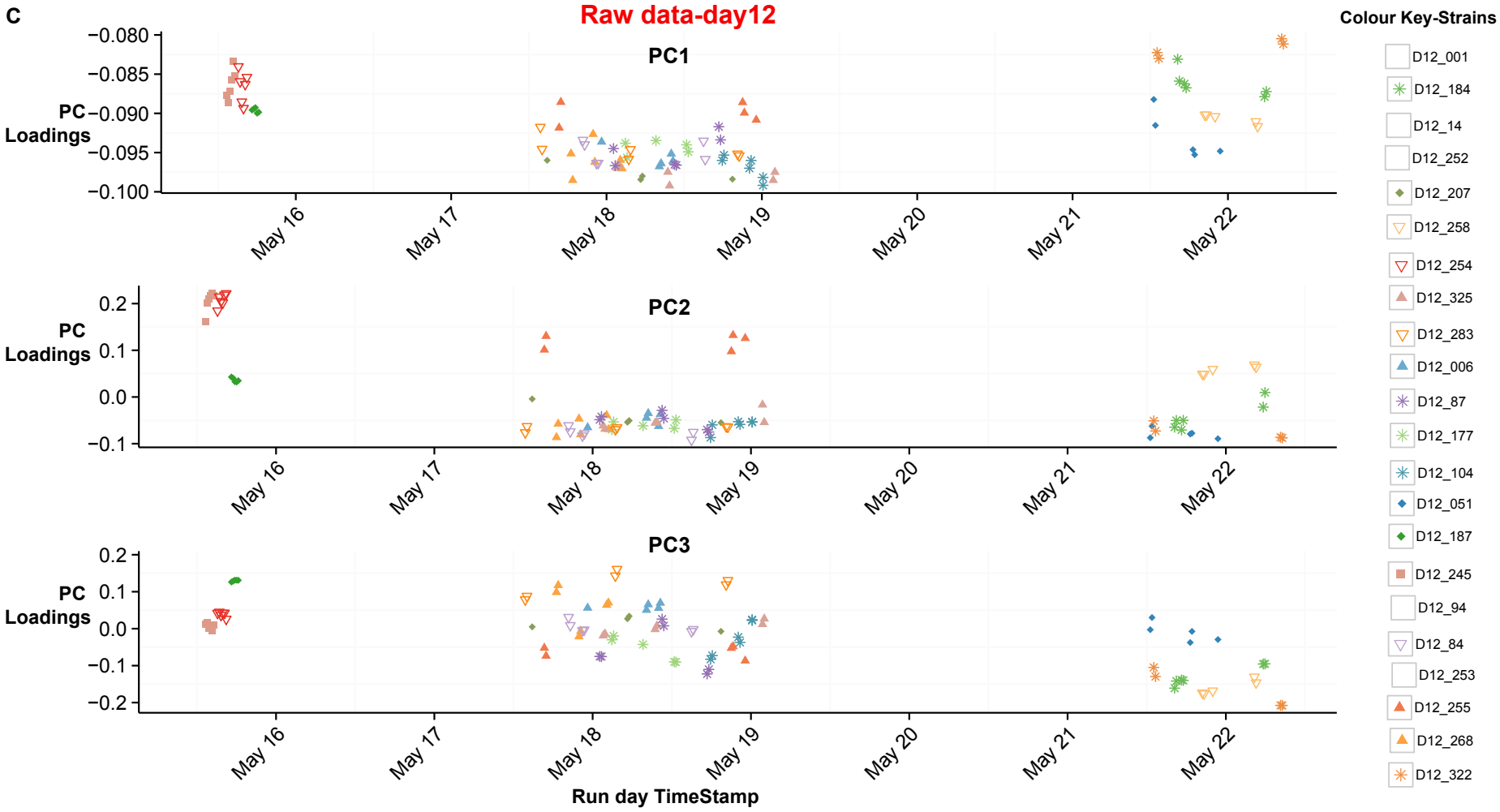
metabolite matrix, such as differences in growth rate between strains which were further confounded with batch effects. Thus for exponential phase, after batch effect correction, we retained 13.31% of the residual variance from the metabolite matrix encompassing 13,444 features. The batch correction resulted in identification of 3,979 features showing significant strain-specific variation and 138 features which were significantly affected by RunDay. Over all, the batch effect removal procedure greatly reduced the variation in metabolite profiles attributed to run day to almost 0% for stationary (day 12) and around 1% for exponential (day 4) phase. The features that were still associated with RunDay were artefacts or adducts and hence did not map onto any metabolite from with METLIN (Sana et al., 2008) or MetaCyc- constrained to the *Chlorella* metabolome (Zhang et al., 2005) databases.

Table 3.1. Significant metabolite features before and after batch effect removal

	Exponential phase	***Stationary phase***
	Raw data	
Features detected	67,468	67,468
Complete features	13,444	10,687
Residual variance	100%	100%
Strain-specific features	13,012	10,309
RunDay-specific features	10,537	8,910
	Batch effect removed	
Residual variance (PCs removed)	13.31% (7 PCs)	18.99% (4 PCs)
Metabolites identified	1,102	996
Strain-specific features	3,979	5,878
Strain-specific metabolites	466	655
RunDay-specific features	138	6
Metabolites influenced by RunDay	0	0


A
Raw data-day4
PC1
PC2
PC3
PC Loadings
−0.080
−0.085
−0.090
−0.095
0.1
0.0
−0.1
0.2
May 17
May 19
May 21
May 23
May 25
Run day TimeStamp
Colour Key-Strains
D4_001
D4_051
D4_177
D4_187
D4_245
D4_255
D4_268
D4_084
D4_094
D4_207
D4_252
D4_283
D4_325
D4_14
D4_184
D4_254
D4_258
D4_253
D4_322
D4_006
D4_087
D4_104


B
Batch effect removed data-day4
PC1
PC2
PC3
PC Loadings
0.2
0.1
0.0
−0.1
−0.2
May 17
May 19
May 21
May 23
May 25
Run day TimeStamp
Colour Key-Strains
D4_001
D4_051
D4_177
D4_187
D4_245
D4_255
D4_268
D4_084
D4_094
D4_207
D4_252
D4_283
D4_325
D4_14
D4_184
D4_254
D4_258
D4_253
D4_322
D4_006
D4_087
D4_104


C
Raw data-day12
PC1
PC2
PC3
PC Loadings
−0.080
−0.085
−0.090
−0.095
−0.100
0.2
0.1
0.0
−0.1
−0.2
May 16
May 17
May 18
May 19
May 20
May 21
May 22
Run day TimeStamp
Colour Key-Strains
D12_001
D12_184
D12_14
D12_252
D12_207
D12_258
D12_254
D12_325
D12_283
D12_006
D12_87
D12_177
D12_104
D12_051
D12_187
D12_245
D12_94
D12_84
D12_253
D12_255
D12_268
D12_322

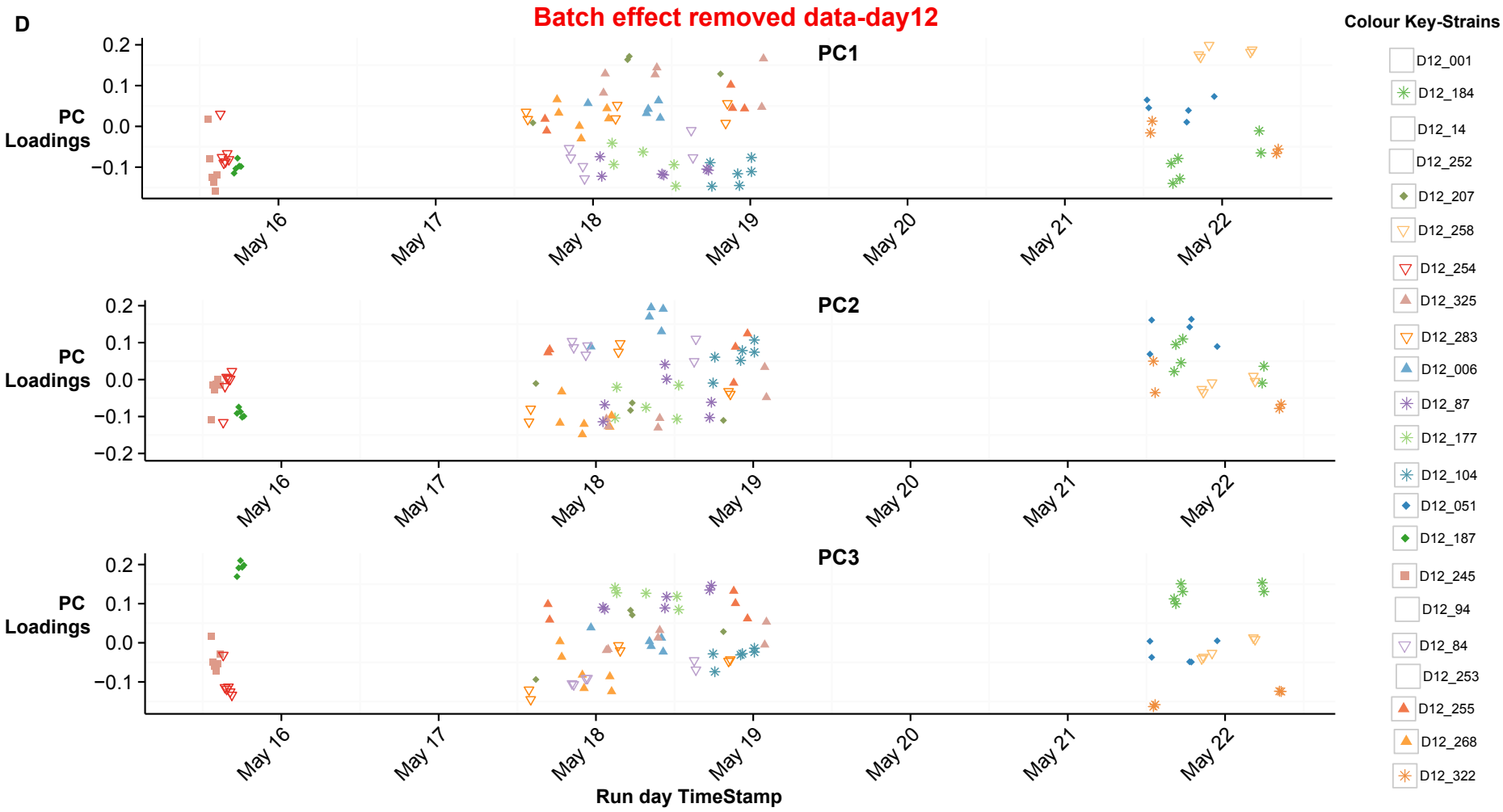

D
Batch effect removed data-day12
PC1
PC2
PC3
PC Loadings
0.2
0.1
0.0
−0.1
−0.2
May 16
May 17
May 18
May 19
May 20
May 21
May 22
Run day TimeStamp
Colour Key-Strains
D12_001
D12_184
D12_14
D12_252
D12_207
D12_258
D12_254
D12_325
D12_283
D12_006
D12_87
D12_177
D12_104
D12_051
D12_187
D12_245
D12_94
D12_84
D12_253
D12_255
D12_268
D12_322

Figure 3.13. The plots depict the first two PC loadings of PCA performed on strains before and after batch correction for (A) exponential phase- day 4, uncorrected data; (B) exponential phase- day 4, after batch effect correction; (C) stationary phase- day12, uncorrected data; (D) stationary phase- day12, after batch effect correction. The y axis in first, second and third row show PC 1, 2 and 3, respectively. The x-axis represent RunDay time stamps. The absence of within batch grouping based on run order indicate that there was no within batch variation and that only RunDay difference (between batches) was the major confounding factor. Interestingly, the strains achieve a better (random) spread after removing the batch effect from the data structure.

To assess the feasibility of using SVD filtering in the present case, the following analyses have been undertaken to show that SVD-based filtering approach used in this study retains some strain-specific biological information:

Case study 1: Strains that were run within a single batch in each growth stage: We designed an analysis to investigate the validity of the SVD procedure by only examining inter-strain differences within a given RunDay batch. Specifically, to estimate the percentage of variation explained by strain-specific differences before and after the application of the SVD batch effect correction on all samples, but only comparing samples within the same batch to their uncorrected counterparts. This was tested using analysis of distance measure ($adonis$ in vegan package) on 14 strains which were run in batch 4 from exponential phase and 12 strains from stationary phase run in batch 2 (Figure 3.1). Table 3.2 describes the results from $adonis$ which shows that the within batch strain-specific differences are preserved following the application of the SVD procedure. Interestingly, the variation explained (R2) by strains, seems to increase by 6% for exponential phase while decreasing by 3% for stationary phase after batch effect correction procedure.

From the same analysis the distribution of *F*-statistics for the strain-specific differences (based on 999 permutations) (Figure 3.14A) also indicate an overlap in the magnitude of the *F*-statistics between the uncorrected and batch effect corrected data. The y axis in the plot shows the density of *F*-statistics values. The topological ordering between the strains is represented by plotting the first two axes of the double centered distance matrix obtained from $cmdscale$ in Figure 3.14B. .

Table 3.2. Analysis of distance results for within batch comparison

Growth stage	***Dataset***	***Factor***	***DF***	***SS***	***MS***	***F.model***	***R^2***	***Pr(>F)***	
Exponential Phase	Raw data (batch23)	Strain	13	61300	4715.4	2.4771	0.32792	0.001	***
		Residual	66	125636	1903.6	NA	0.67208		
	Corrected data (batch23)	Strain	13	56657	4358.2	3.2337	0.3891	0.001	***
		Residual	66	88951	1347.7	NA	0.6109		
Stationary Phase	Raw data (batch17)	Strain	11	77775	7070.4	4.719	0.49013	0.001	***
		Residual	54	80907	1498.3	NA	0.50987		
	Corrected data (batch17)	Strain	11	60677	5516.1	4.2697	0.46517	0.001	***
		Residual	54	69762	1291.9	NA	0.53483		

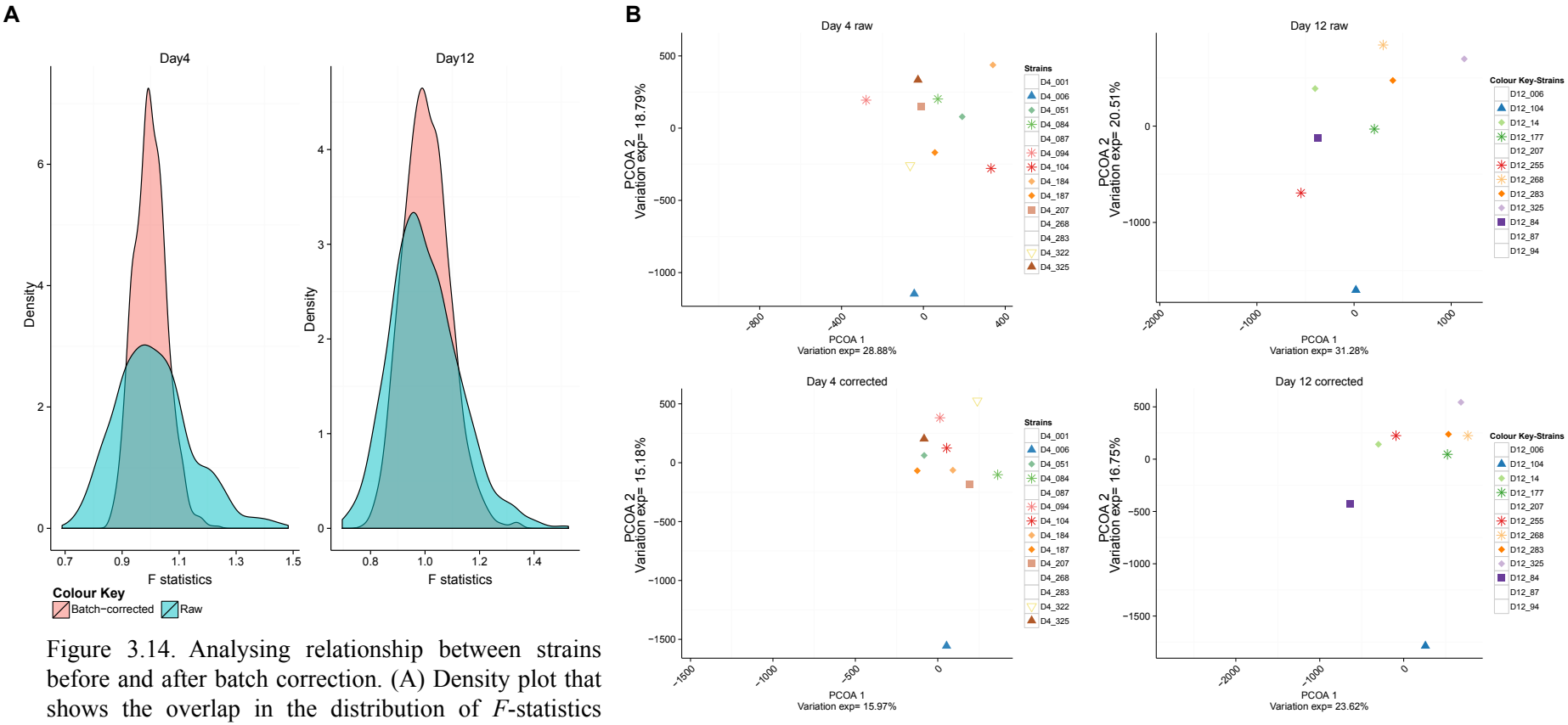


Figure 3.14. Analysing relationship between strains before and after batch correction. (A) Density plot that shows the overlap in the distribution of *F*-statistics between the uncorrected and batch corrected data; (B) Relationship between strains plotted using the first 2 axis of the distance matrix

These plots indicate that though the variation between strains seems to have decreased, the topological ordering of strains such as, (i) D4_001 and D4_006 at exponential stage, and (ii) D12_006 and D12_014 at stationary stage, appear to be preserved before and after batch effect correction.

By definition, the issue of whether it is appropriate to use SVD across batches cannot be assessed using this analysis, but these results clearly indicate there is preservation of biological signal following application of the SVD procedure. The issue of across-batch correction is addressed in the following sections.

Case study 2: Validation of biological interpretations and strain relationship using independent data

The biological relationships between strains used in this study, were tested using an independent experiment performed in August 2014. This experiment was a targeted tandem MS/MS analysis that was performed for validating the putative metabolites predicted from this study. We extracted MS1 data from this datset to compare the relationship between strains.

Targeted metabolomics was performed using tandem MS/MS on 6 strains that were shown to have the most diverging physicochemical traits. They were UMACC001, UMACC187, UMACC253, UMACC254, UMACC322, and UMACC051 profiled at both exponential and stationary phase. The MS1 spectrum was extracted from the tandem MS/MS data and processed using the same methods as described in the thesis. Importantly, these new data on selected strains were run in the one batch. Thus, they provide an ideal test case to understand whether batch effect correction procedure preserves the biological relationship between strains. For comparative analysis, the same 6 strains were selected from the original dataset and the batch corrected dataset, in both exponential and stationary phase. Analysis of distance measure was used to determine the variation explained by the strains (Table 3.3) and distribution of *F*-statistics were assessed.

Table 3.3. Analysis of distance comparing relationship among 6 strains

Growth stage	*Dataset*	*Factor*	*DF*	*SS*	*MS*	*F.model*	R^2	*Pr(>F)*	
Exponential Phase	Raw data	Strain	5	33367	6673.5	3.8113	0.39654	0.001	***
		Residual	29	50779	1751	NA	0.60346		
	Corrected data	Strain	5	21817	4363.4	3.2625	0.36	0.001	***
		Residual	29	38785	1337.4	NA	0.64		
	MS/MS data	Strain	5	5531.9	1106.38	3.9194	0.62022	0.001	***
		Residual	12	3387.4	282.28	NA	0.37978		
Stationary Phase	Raw data	Strain	5	39261	7852.2	4.4525	0.47104	0.001	***
		Residual	25	44088	1763.5	NA	0.52896		
	Corrected data	Strain	5	38042	7608.5	6.5802	0.56823	0.001	***
		Residual	25	28907	1156.3	NA	0.43177		
	MS/MS data	Strain	5	6821.8	1364.36	5.6202	0.70076	0.001	***
		Residual	12	2913.1	242.76	NA	0.29924		

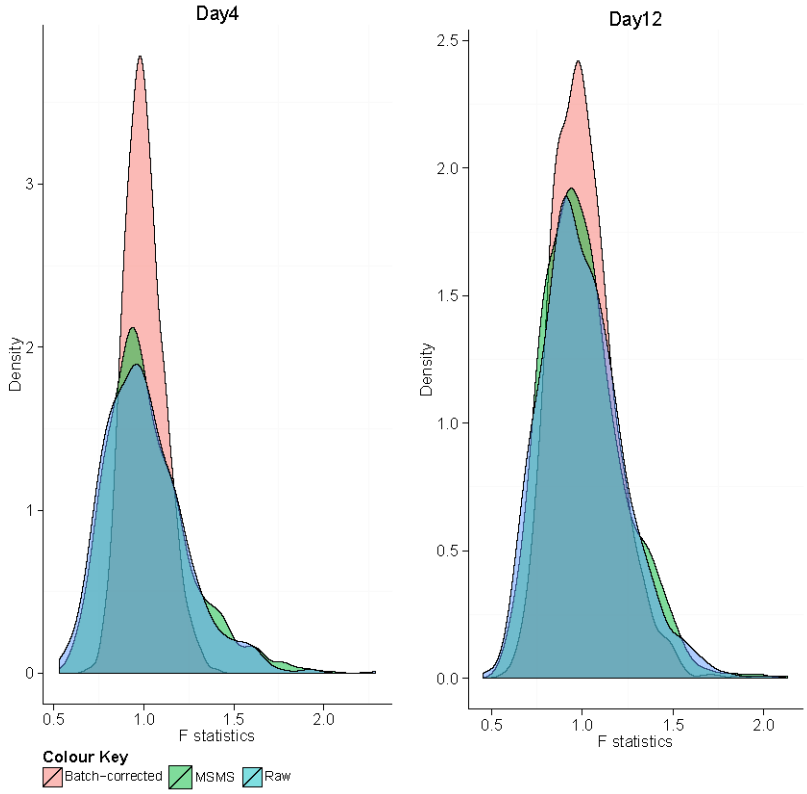


Figure 3.15. Density plots that depict the distribution of F values obtained from the permutation test for the raw, batch corrected and independent MS/MS data

The results again reflect the earlier trend, with a minor change, in this case, batch effect correction in stationary phase seems to explain 5% more variation than the uncorrected data. However for exponential phase, there seems to be a 3% decrease. The distribution of *F*-statistics show an overlap between the raw data, corrected data and MS/MS data from the new experiment (Figure 3.15). Furthermore the topological ordering of strains remains similar in all 3 datasets (Figure 3.16). From these results, we can conclude that relationship between strains is preserved. Importantly, it also supports the claim that SVD-based filtering effectively removes batch effects while still retaining strain-specific differences.

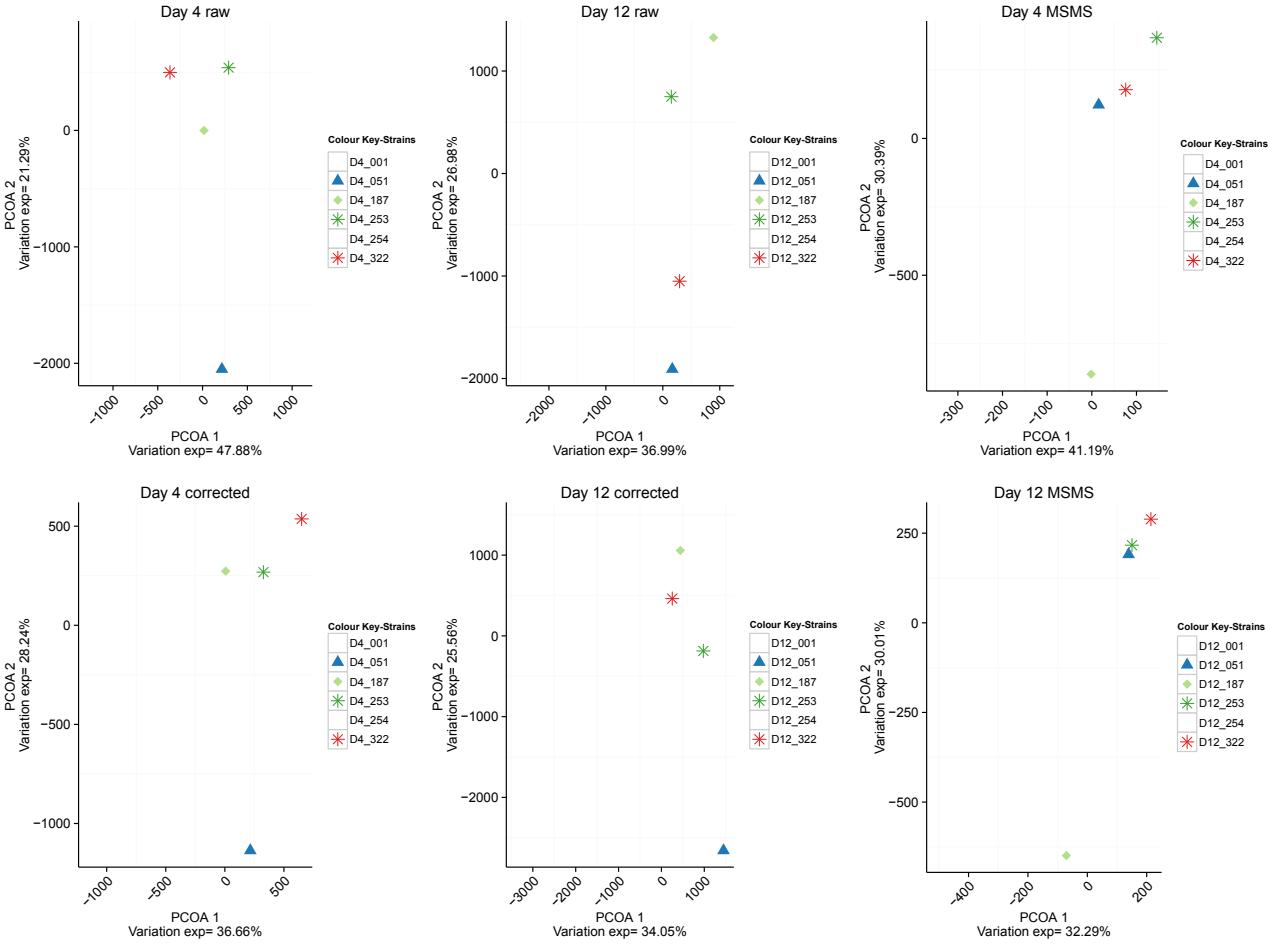


Figure 3.16. Relationship between 6 strains in the raw data, batch corrected data and in the new MS/MS dataset, plotted using the first 2 axis of the distance matrix

SVD-based approach preserves biological interpretation

Figure 3.17 shows the effect of SVD-based filtering of PC components. The y axis indicates the number of significant features (determined using permutation-based *F*-statistics) that were associated with Strain (coloured red) and RunDay (coloured green). The x axis indicates the number of PCs that have been removed.

As witnessed in the Figure 3.17, the number of significant features that are associated with RunDay drops of rapidly after removal of the top few PCs in both exponential (Day4) and stationary (Day12) phase. Furthermore, while RunDay associated features are removed after the top few PCs, there still seems to be a significant number of features that are associated with strain differences. Unlike, the nested linear model, these are strain-specific features alone and are not significant when RunDay is used as a factor.

A detailed representation of the effect size, inflation statistic and the relationship between strain-specific and run-day specific features are provided in Supplementary Figures 1 and 2 for exponential and stationary phase, respectively. Furthermore, as observed in Figure 3.14, the overall relationship between strains analysed in the same batch, before and after batch correction procedure, is similar. Therefore, SVD-based correction seems to effectively reduce non-biological sources of variation.

Metabolites in batch effect corrected data show significant associations with PC

A recent method 'Jackstraw' that computes the association between metabolites and principal components was tested on the raw and batch corrected data. It uses a resampling method to produce accurate significant measures of associations between the observed metabolites and their principal components.

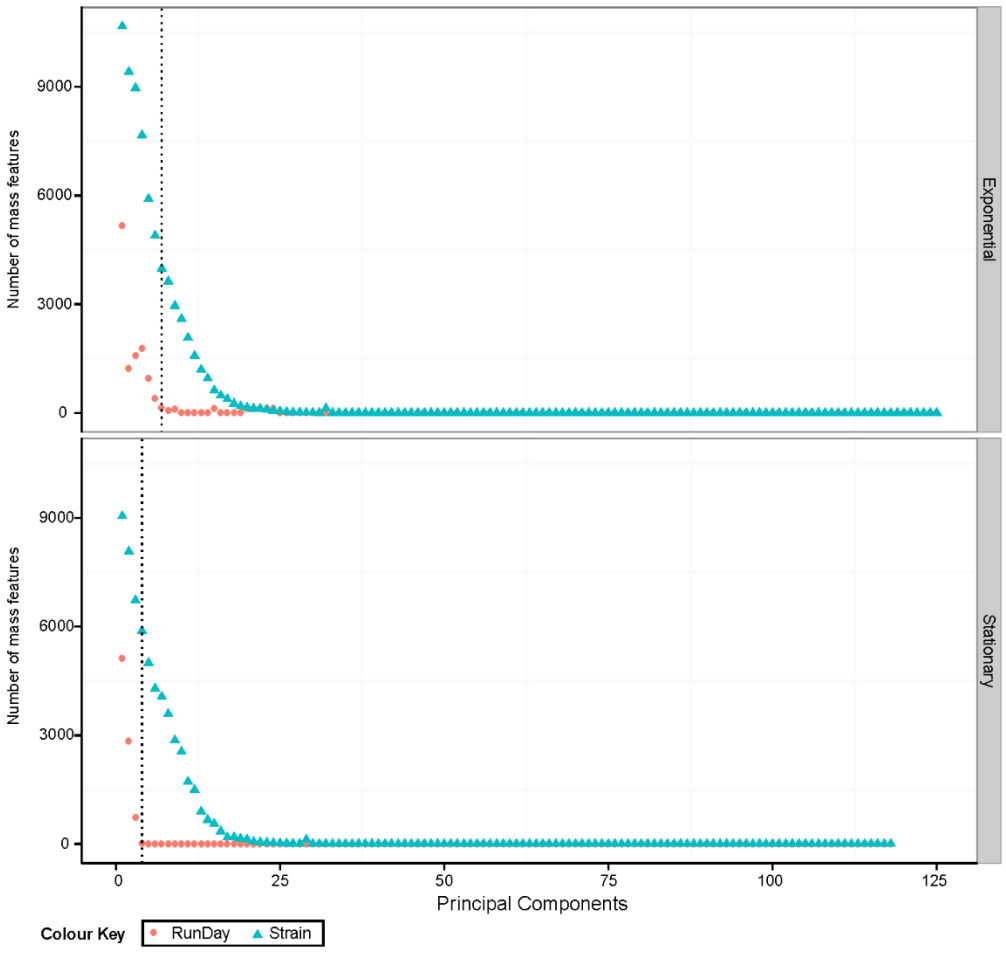


Figure 3.17. Number of significant features that are associated with RunDay and strain after removing each PC. The x axis indicates the PCs and the y axis indicates the number of significant features.

The over-fitting characteristics that result from computation of principal components from the same set of variables are also taken into account (Chung and Storey, 2014). The *permutationPA* function from the jackstraw package was used to estimate the number of significant principal components from both raw and batch corrected data at each growth stage. Interestingly, the test did not detect any significant principal components in the raw data. However, when the *permutationPA* function was used on the batch effects corrected dataset, it determined the first 26 components of exponential phase and first 22 components of stationary phase to have significant associations with metabolites (Figure 3.18). The x axis in Figure 3.18 represents the PCs and the y axis shows the percentage variation explained by each PC.

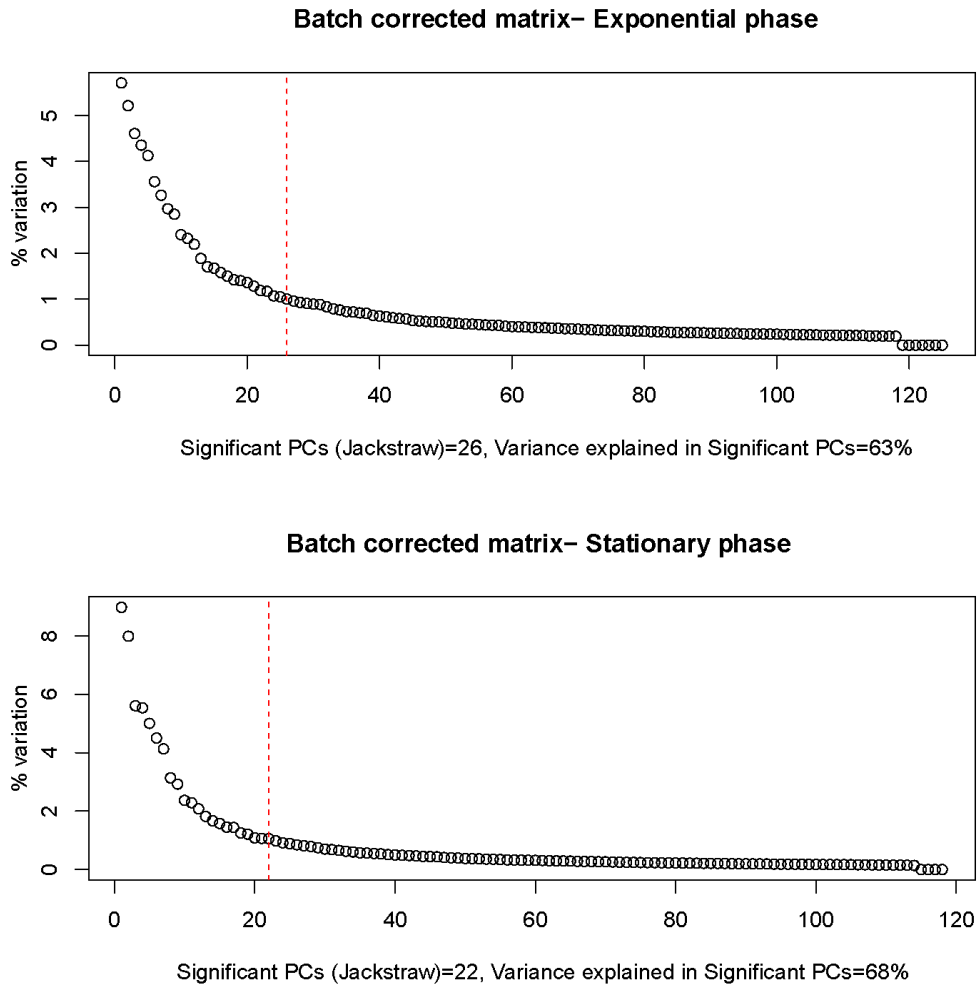


Figure 3.18. Significant associations between principal components and metabolites. The y axis indicates the percentage variation and the x axis indicates PCs.

These results suggest that the variation in the uncorrected dataset might be masked due to batch effects, thus, the metabolites did not show any significant association with the principal components. Furthermore, the above results further support the assumption that SVD-based filtering retained significant variation due to strain-specific effects. This observation is based on the detection of significant association between metabolites and principal components of the batch effects corrected dataset that explained 63% and 68% in exponential and stationary phase, respectively.

Comparison of SVD-based approaches with standard batch effect removal approaches

The effectiveness of using PCA-based approaches in this fashion has been subject to some debate, and in particular, the generality of this procedure is unclear, hinging on the fact that the influences of non-biological, batch-related variables are thought to manifest themselves in the first few principal components. In practice, this is probably extremely difficult to establish in general terms, due to the context-specific and diverse nature of such non-biological influences in large sample series. Goldinger et al. argue that such PCA filtering is less effective than linear modelling, and demonstrate the even principal components with small variance contributions can be associated with batch related variables (Goldinger et al., 2013)

Recently, Leek et al compared the effectiveness of using PCA (using SVD) and SVA for identifying and mitigating batch effects (Leek, 2014). In this study, using three published datasets and one simulated dataset, the author compared supervised SVA and svaseq for count data (similar in nature to metabolite abundances in the metabolomics data matrix), with standard methods for removing batch effects. The author uses SVD to perform PCA-based batch effect removal. For the simulated dataset, when batch effects had low correlation with group effect, SVA and RUV-based approaches performed better than PCA-based approach. However, the author shows that in datasets where there is moderate or high correlation between batch and group effects, then unsupervised SVA and PCA-based approaches perform better than RUV-based approaches.

In the present case, we have analysed the association between biological and non-biological variables, and eigenvector composition, and at least here, PCA effectively decouples the influence of these factors, despite their highly confounded nature. But the generality of this procedure is unclear and we emphasize, in common with others, that the influence of batch effects needs to be carefully investigated on a

case-by-case basis, and undue reliance should not be placed on prescriptive solutions. The extent to which PCA filtering should be used a general strategy for removing batch effects, potentially even in situations when these factors are either not available or not recorded, is unclear and more work is needed in this area.

3.4. Conclusions

We used significance tests and analysis of variance measures to evaluate the impact of batch effects on the metabolite features when both batch and biological differences were confounded. Given a set of linear combination of PCs estimated from the metabolomics data, we developed a (i) filtering procedure that minimizes such unwanted variations resulting in preservation of biological signals (ii) and does not require negative controls, QCs or standards for removing unwanted variation. We use the non-targeted algal metabolomics data to show that the proposed filtering procedure can be used to effectively remove nuisance variation caused by batch effects, while preserving the biological source of variation to serve as a direct indicator of biochemical phenotype. The analysis of this dataset is further examined in the next Chapter (discussed in Chapter 4).

We have demonstrated the need for incorporating batch effect correction methods as a standard protocol especially for high throughput datasets. This can largely simplify complex methods and provide meaningful biological interpretations. Furthermore, with increased sharing of metabolomics datasets among the scientific community through initiatives such as MetaboLights and COSMOS (Salek et al., 2013b), it is imperative that comprehensive meta information is recorded for removing batch differences. Similar to the case of expression microarrays, these data standards will facilitate increased meta-analysis and large-scale data mining providing further justification for implementing robust experimental design and data analysis strategies (Gibon and Rolin, 2012; Rai et al., 2013). It is also important to note that experimental

designs should incorporate the use of multiple internal standards and pooled standards. The samples should be extracted and processed in a randomized order. These steps are provide vital information that can be used to identify and mitigate batch effects.

The SVD-based approach described in this study, provides a suitable alternative for mitigating batch effects, when the experimental design issues render regression-based methods as ineffective. The methods described in this Chapter may also have a role to play in identifying and removing batch effects in any large-scale experiments such as Next Generation Sequencing (NGS) studies. Taken together, implementation of these techniques will lead to improved experimental designs and enhanced data standards in untargeted metabolome surveys from any setting

.

4. Environmental and biochemical determinants of metabolic resource partitioning in naturally varying microalgae- *Chlorella*

"This preservation of favorable variations and
the rejection of injurious variations,
I call Natural Selection."
… Charles Darwin (1859) in 'Origin of Species'

4.1. Background and Introduction

The increasing energy demands along with the depleting fossil fuels has resulted in finding alternative sources of energy critical for sustaining modern life (Georgianna and Mayfield, 2012). Biofuel production using microalgae has clear advantages over other renewable sources of energy due to the following characteristics:

(i) Microalgae produce and accumulate oil as nonpolar lipids, such as triglycerides (TAGs), from sunlight and carbon dioxide (Wijffels and Barbosa, 2010). These rich sources of TAGs can be converted to high quality biofuels.

(ii) Compared to other alternatives such as crop biomass, microalgae grow relatively rapidly (Wijffels and Barbosa, 2010) and are easy to metabolically engineer for producing the desired bio-products.

(iii) Microalgae can be farmed on non-arable land (Chisti, 2007) and using non-potable water (Phang, 1990), thus, minimizing wasteful diversion of resources that can be better utilized and crucial for growing food crops.

The vast natural diversity of algae along with their enormous chemical and physiological variability provides an environment conducive for identifying efficient strains for biofuel production (Stengel et al., 2011). Such diversity also highlights the ability of microalgae to thrive in diverse ecosystems (Radakovits et al., 2010). Unbiased high-throughput screening of the biochemical profiles of such populations,

especially inter-species comparisons, can provide insights into the genetic and environmental factors influencing production of valuable commercial products from oleaginous microalgae (Stengel et al., 2011). Furthermore, such studies can also provide insights into the factors that shaped evolutionary diversity in these species *i.e.* the role of environment in shaping evolutionary and regulatory divergence between species.

Microalgae exhibit enormous diversity in their lipid profiles, ability to synthesize energy, growth rates and biomass productivity, all of which determine the yield (Stengel et al., 2011). Thus, bioprospecting natural variants with the desired traits such as strains that can grow quickly having high biomass and lipid content, can drastically reduce the time required to optimize metabolic engineering strategies for large-scale production of biofuels, and have high economic benefits (Georgianna and Mayfield, 2012).

In this study, we focus on the green oleaginous microalgae- *Chlorella*, which are widely distributed in fresh water environments (Eckardt, 2010). Initially, a number of algal isolates were assigned to the genus *Chlorella*. However, this taxonomy classification was not reliable because of the lack of distinct morphological characteristics. With further molecular analysis, these isolates were then separated into two classes of chlorophytes, namely the Trebouxiophyceae (true *Chlorella*), and the Chlorophyceae (Blanc et al., 2010). Similar to other microalgae, *Chlorella*, has been the focus of interest mainly for (i) producing biofuels and high-value bioactives, (ii) sequestering carbon dioxide from the environment, (iii) and as biofertilizers or for bioremediation (Arbib et al., 2014). Furthermore, *Chlorella* has an inherent capacity to produce high amounts of lipids (Pribyl et al., 2012) and biomass (Doan et al., 2011). *Chlorella* has been studied for a number of years starting from 1969 (Fott and Nováková, 1969), with most of the efforts initially focusing on identifying and screening for specific bioactive algal compounds (Onofrejova et al., 2010; Schumacher

et al., 2011), with later emphasis on analysing the effect of nutrients, growth conditions (Xin et al., 2010) and determination of optimal conditions for producing high quality biofuel (Rodolfi et al., 2009).

However, the untapped potential of naturally varying microalgae as efficient producers of biofuel has never been comprehensively studied using non-targeted metabolomics approaches (Stengel et al., 2011). Previous efforts in screening tropical microalgae were largely focused on their use as food supplements or as fertilizers (briefly reviewed by (Vello et al., 2014)). Furthermore, there is growing evidence indicating that organisms vary in their ability to regulate both the levels and configurations of a given set of metabolic enzymes (Rhee et al., 2011), related to both variation in genetic and environmental factors. (Breunig et al., 2014; Chan et al., 2010; Wen et al., 2014). The development of advanced analytical measurement technologies (discussed in Chapter 2) combined with multivariate statistical techniques, have now provided opportunity to profile the diverse chemical space and richness of algal compounds. This approach also provides an unbiased characterization of the biochemical phenotype facilitating the characterization of effect of genetic and environmental (habitat) niche on the metabolic diversity.

A major challenge is to understand the complex factors influencing the allocation of cellular resources to various processes such as growth, lipid productivity and biomass in oleaginous algae. Understanding how microalgae can convert the single carbon compounds into bio-products of interest can be studied using metabolomics. For example, comparing naturally varying strains using non-targeted metabolomics profiling can generate accurate quantitative biochemical phenotypes that help understand the preferential utilization of metabolic pathways in the cellular resource partitioning strategies. Knowledge of such processes can help in strain prioritization (Krug and Muller, 2014) whereby efficient strains that facilitate easy manipulation of metabolic pathways can be identified. This approach can also aid in removing

bottlenecks such as cell-density limits during synthetically re-engineered microbial biofuel production. Additionally, mass spectrometry-based metabolomics can reveal the genomic potential and characterize the metabolite concentration changes influenced by environmental factors.

As inter-species comparison of strains isolated from different habitats are analogous to studying the environmental effects on metabolic phenotypes of oleaginous algae, we examined the natural variation of 22 *Chlorella* strains isolated from 7 different geographic locations in Malaysia. To characterize the metabolic diversity of these strains, non-targeted metabolic profiling was performed. In this Chapter, I have undertaken a systematic analysis of these to analyse the metabolic diversity between 22 *Chlorella* strains, with the specific aim of (i) understanding metabolic changes during growth; and (ii) identifying habitat-induced variation and biochemical determinants of metabolic phenotypes.

This study is the first report to profile natural variation in oleaginous microalgae using a combination of non-targeted metabolomics, phylogenetic analysis and physicochemical profiling. We assess the strain-specific metabolic reprogramming strategies and analyse associations between physicochemical and metabolic profiles to identify key metabolic correlates of biotechnology related traits. The overall objective of this study is to identify algal strains that have biochemical and metabolic characteristics suitable for biofuel production.

4.2. Materials and methods

4.2.1. Sampling strategy

To survey the natural variation in oleaginous microalgae, untargeted metabolomics profiling was performed on 22 strains of *Chlorella* obtained from University of Malaya Algae Culture Collection (UMACC). These 22 strains were isolated from 7 different geographic locations (Figure 4.1), comprising of 16 diverse habitats (Table 4.1) in Malaysia (Courtesy: Ms. Vejeysri Vello and Prof. Siew-Moi Phang, UMA). Colleagues from UMA had previously characterized the lipid productivity and the fatty acid composition of these strains to identify promising strains for biofuel production. For detailed procedures on collection, culturing and storage of samples, kindly refer to (Vello et al., 2014).

Figure 4.1. Sampling locations in Malaysia

Phylogenetic analysis of these 22 strains based on partial 18S rRNA sequences revealed that 15 strains belonged to the true Chlorella clade (within the class

Trebouxiophyceae) and 6 strains were from the *Parachlorella* clade, while 1 strain (UMACC 184) was not sequenced (Table 4.1)(Vello et al., 2014).

Table 4.1. Species and sampling site description of 22 *Chlorella* strains (adapted from (Vello et al., 2014))

Strain	***Species***	***Origin***
UMACC 001	Chlorella	Pond at IPSP Farm, University of Malaya
UMACC 006	Chlorella	Fish tank containing chicken manure, IPSP Farm, University of Malaya
UMACC 014	Chlorella	IPSP Farm, University of Malaya
UMACC 051	Chlorella	Aerobic pond for POME treatment, Tenamaran Palm Oil Mill, Selangor
UMACC 084	Chlorella	Digested POME, enriched with goat dung, IPSP Farm, University of Malaya
UMACC 087	Chlorella	Digested POME, enriched with goat dung, IPSP Farm, University of Malaya
UMACC 094	Chlorella	Tenamaran Palm Oil Factory, Selangor
UMACC 104	Chlorella	Muddy water of Sementa Mangrove, Selangor
UMACC 177	Chlorella	Plastic container, Kuantan Pahang
UMACC 184	Unidentified	NA
UMACC 187	Chlorella	Tin, Chinese graveyard, Kuantan Pahang
UMACC 207	Chlorella	Concrete tank, shop houses, Kedah
UMACC 268	Chlorella	Raw palm oil effluent pond, Labu Palm Oil Mill, Negeri Sembilan
UMACC 283	Chlorella	Anaerobic pond 3, Labu Palm Oil Mill, Negeri Sembilan
UMACC 322	Chlorella	Wastewater treatment pond at oil refinery, Negeri Sembilan
UMACC 325	Chlorella	Wastewater treatment pond at oil refinery, Negeri Sembilan
UMACC 245	Parachlorella	Seawater from Terengganu
UMACC 252	Parachlorella	Sea Bass Pond at Sepang, Selangor
UMACC 253	Parachlorella	Sea Bass Pond at Sepang, Selangor
UMACC 254	Parachlorella	Sea Bass Pond at Sepang, Selangor
UMACC 255	Parachlorella	Sea Bass Pond at Sepang, Selangor
UMACC 258	Parachlorella	Sea Bass Pond at Sepang, Selangor

It is important to note that though the strains are similar to those used in (Vello et al., 2014), the data for this study comes from a different batch. This study was specifically conducted to identify the differences in metabolic strategies during growth phases and to determine the associations between biochemical traits and metabolic profiles.

4.2.2. Experimental design

Experiments were designed to assess the influence of habitat and genotype on the metabolic phenotypes. Metabolite profiles of 22 strains at two growth phases- exponential (day 4) and stationary phase (day 12) (Figure 4.2) were obtained using Liquid Chromatography Mass Spectrometry (LC-MS) (described in Chapter 3). Biological replicates of each of these strains were developed in a manner permitting assessments of metabolic variation. Specifically, for each strain, we profiled 3 biological replicates and for each biological replicate 2 analytical replicates were used at each growth phase.

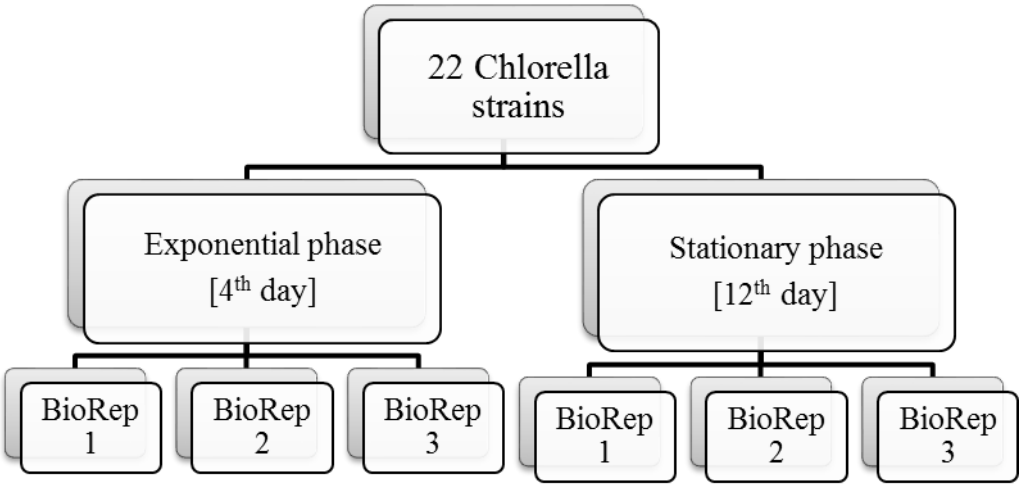


Figure 4.2. Experimental design for generating metabolome and biochemical profiles. The physicochemical measurements were collected for 3 biological replicates.

The physicochemical measurements collected by Ms. Vejeysri Vello (UMA) are:

- Specific growth rates: monitored based on OD_{620nm} and chlorophyll *a* (Chl*a*) concentration. Figure 4.3 provides an illustration of the growth rate measurements for these 22 strains. There is no growth rate associated with strains at stationary phase.
- Biomass (g/L)
- Biomass productivity (g/L/day): expressed as the dry biomass produced (in g L−1 day−1), at the end of exponential growth phase, specifically it is biomass density (g L−1)×specific growth rate (day−1)

- Lipid productivity (mg/L/day)
- Total lipid content (% dry weight)
- Total protein content (% dry weight)
- Total carbohydrate content (% dry weight)

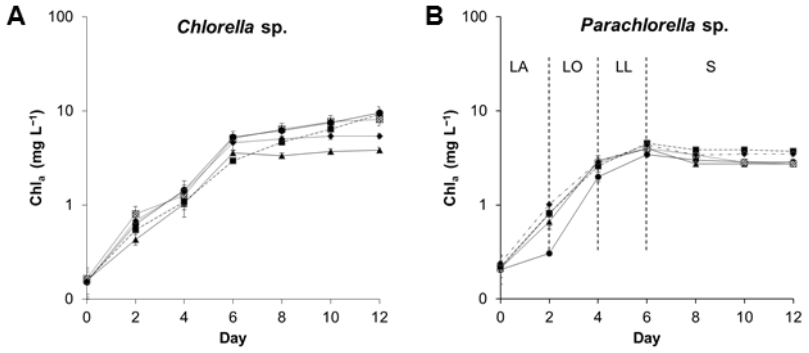


Figure 4.3. Representative growth rates for (A) *Chlorella* and (B) *Parachlorella* strains

4.2.3. Metabolite identification

We used the metabolomics datasets that were corrected for batch effects in this study. CAMERA package (Kuhl et al., 2012) in R (R Core Team, 2014) was used with the default parameters to remove isotopes and adducts before metabolite identification. Figure 4.4 shows the distribution of standard deviation of the abundance (in the y axis) for each m/z feature (in the x axis). As there was higher variation in the larger metabolite masses, a 10 ppm mass and retention time window of 5 s was set after manually inspecting the peak width in the extracted ion chromatograms. After deionization, metabolite features which elute within 5 s of each other and having m/z within 10 ppm were grouped and the median m/z for this was calculated. The m/z was then used in database search for predicting putative metabolite identifications. The m/z and retention time for all the features detected in each dataset is provided in a Supplementary dataset (Dataset 1).

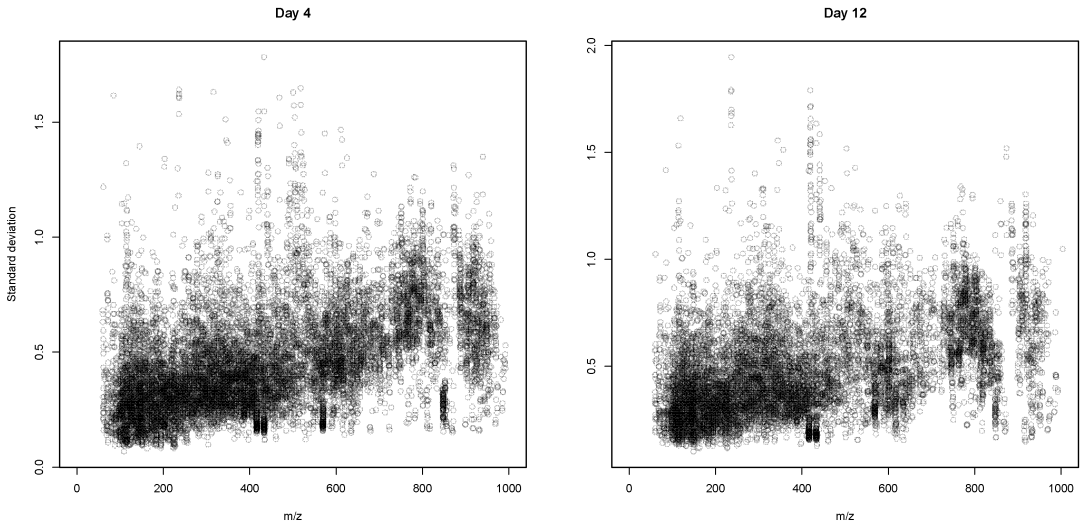


Figure 4.4. Standard deviation of metabolite abundances at exponential (day 4) and stationary (day 12) growth stage. The standard deviation increases linearly with increase in mass, indicating that higher masses are more variable, possibly due to the analytical limitations.

We then used the binned m/z feature list to predict putative metabolites by matching them against MetaCyc- constrained to the *Chlorella* metabolome (Zhang et al., 2005) and METLIN (Tautenhahn et al., 2012) databases using the PCDL manager (Agilent) with a search window set to 10 ppm in positive mode. Specifically, the database was first searched for exact matches to the given m/z, if there were no exact matches, then the search extends to 1pp window, then 2 and so on till 10 ppm. The best match is the one which has the minimum difference in ppm from the given m/z value. All the metabolites described in this Chapter were putative identifications based on database matches. Supplementary Dataset 1 lists all predicted metabolites with their ppm differences (within 10 ppm). The predicted metabolites can provide valuable clues to possible metabolites and their pathways, however, these need to be validated using standards and tandem MS/MS techniques for absolute confirmation. The total number of putative metabolites, at (i) exponential phase -1,102 were detected in the batch corrected dataset, while 466 metabolites were differential, and at (ii) stationary phase - 996 metabolites were detected using batch corrected dataset while 655 were differential metabolites.

All statistical analyses in this study were performed using R and on batch effect corrected data. In figures, strain labels starting with UMACC are replaced by 'D*X*_' for clarity. Here *X* refers to the growth stage, thus *X* is 4 for exponential phase (D4_) and 12 for stationary phase (D12_). However, when the strains are referred to in the context of their genetic sequences or genomes, UMACC label is still used as there are no differences in these labels based on growth stages. Using this dataset, we identified significantly different (FDR adjusted *p*-values < .05) features between the 22 strains, amounting to 5,878 features at stationary phase, and 3,979 features at exponential phase.

As a follow up to these initial surveys, data-dependent MS/MS is currently being performed using Agilent quadrupole time of flight mass spectrometry (Agilent Q-ToF 6540) with ESI probe in positive mode of ionization. The resulting fragments will be used to identify metabolites based on chemical structure similarity with standards or/and matching with MassBank database (Horai et al., 2010).

4.3. Results and discussion

4.3.1. Genetic divergence between algal strains

rRNAs being key elements of the translation mechanism in cells are extremely conservative and are not structurally affected by artefacts produced due to lateral transfers (Pace et al., 1986). Furthermore, the length of the SSU-18S rRNA sequences are adequate to provide statistically robust comparisons between species. These characteristics make them ideal entities to be used for deriving phylogenetic relationship between organisms. For this study, phylogenetic analysis using the 18S rRNA sequences (courtesy Ms. Vejeysri Vello, Prof. Phang Siew Moi, UMA) provided a measure of the algal diversity sampled, especially given that these were uncharacterized microbes assayed over diverse geographical locations and habitats.

With the phylogenetic relationships already mapped out by Vello et al., 2014, we set out to derive a measure for calculating the genetic distance among the 21 strains. While the phylogenetic relationship was analysed by Vello et al, the work in Figure 4.5A was mainly performed to assess genetic divergence within 21 strains that had metabolomics data. The phylogenetic analysis from Vello et al could not be used for the same purpose as (i) it used 29 strains, and (ii) Vello et al compared these 29 strains with 83 other taxa, mainly related to other Chlorella (from the GenBank database). In this Chapter, genetic differences were analysed between 21 strains (UMACC 184 did not have 18S data) in order to provide a measure of comparison between the metabolic distances for the same set. Specifically, it was used to analyse whether strain-specific metabolic traits could be related (at both growth stages) to the genetic differences between the strains. The genetic distance between the 21 strains was calculated using the $stringDist$ function (Levenshtein distance) from the 'Biostrings' package (Pages et al.) in R. Hierarchical clustering based on average linkage method was then used to visualize this distance matrix (Figure 4.5A).

From Figure 4.5A, we observed a clear separation between strains isolated from Seawater or Sea Bass Pond, namely UMACC 252, UMACC 245, UMACC 258, UMACC 253 and UMACC 254 and others. Interestingly, the above five strains are from the *Parachlorella* clade, thus exerting an influence on the genetic distance. The tight clustering within this group also suggest that these strains might have a similar genetic background. Furthermore, Strain UMACC 255 isolated from Sea Bass Pond and belonging to *Parachlorella,* formed a separate cluster with UMACC 322 which was isolated from a waste water treatment plant and belonged to the *Chlorella* clade. Surprisingly, the above two strains formed a unique subclade separate from the remaining 19 strains. This could possibly be due to the unexpectedly high similarity in the 18S rRNA sequences between these two strains. It will be interesting to perform complete sequencing of these strains to analyse their evolutionary and regulatory

relationships. These observations from genetic distance-based clustering led us to further characterize the metabolic relationships between strains.

4.3.2. Metabolic divergence between algal strains

Metabolites identified from exponential phase-1,102 (full data), 466 (differential) metabolites and from stationary phase we had 996 (full data), 655 (differential) metabolites were used for calculating the metabolic divergence among the 21 strains using $vegdist$ function (Euclidean distance) of 'vegan' (Jari Oksanen, 2013) package in R. In the above description, full data refers to all the metabolites which were detected in the metabolomics profile for that growth stage, whereas differential refers to all the metabolites which were significantly varying in abundance between the strains. The distance matrix produced was visualized using hierarchical clustering based on average linkage method (Figure 4.5 Exponential phase: B, Stationary phase: C).

There appeared to be substantial differences in the clustering patterns between genetic and metabolic distance matrices and even between growth stages in the metabolic profiles. For example, the clusters formed based on habitat or genotype in the genetic distance-based tree (Figure 4.5A) were conspicuous because of their absence in the metabolic distance trees. Furthermore, strains D4_322 (Figure 4.5B) (which had also shown significant genetic divergence, Figure 4.5A), and D12_001 (Figure 4.5C), showed the maximum metabolic divergence at exponential and stationary phase, respectively. These large metabolic profile based divergence indicate that these strains have a markedly different metabolome compared to the others. UMACC 001 and UMACC 014 which formed a cluster based on genetic distance, were also clustered together based on metabolic profiles at exponential phase. However, this trend was not repeated in the metabolite profiles at stationary phase. We observed that strains that were genetically divergent, were converging in the same cluster based on metabolic distances at exponential (D4_051 and D4_087, Figure 4.5B) and stationary

phase (D12_014 and D12_255, Figure 4.5C). Apart from these obvious groupings, there were no clear clustering patterns in the metabolic distance trees at both exponential and stationary phase.

We then performed a Mantel test (*mantel* function in 'vegan' package, with 999 permutations) computing Spearman's correlation coefficient to analyse the trends between metabolic and genetic distance matrices. Mantel test, a non-parametric statistical method, that tests the significance of correlations between two distance matrices using permutations of rows and columns. In this implementation of Mantel test, the Spearman's correlation coefficient 'r' will fall in the range of -1 to +1 depending on the correlation between the elemental (strain) distributions in both the genetic and metabolic matrices. A strong negative correlation brings 'r' closer to -1, whereas in case of positive correlation the correlation coefficient value 'r' will be closer to +1. An 'r' value of '0' indicates no correlation between the two matrices. We did not observe any significant correlations between the genetic and metabolic distances at both growth stages (Table 4.2). This result is in accordance with previous studies on closely related genotypes, where studies conducted on *Arabidopsis thaliana* analysed the natural variation in different accessions and also showed that there was only a minor or no correlation between its genetic and metabolic diversity (Chan et al., 2010; Houshyani et al., 2012). Furthermore, they observed that metabolite levels were largely influenced by environment. The absence of genotype-metabotype correlation indicates that there is no one-to-one relationship between the genome and the metabolome of an organism. Thus, an unbiased metabolomics survey might be a better analytical technique to capture novel metabolites or understand the genomic potential of uncharacterized strains. Furthermore, there was no correlation (based on Mantel test) between the metabolic profiles of exponential and stationary phase, either using full data (r = 0.079, *p*-value = 0.297) or differential metabolites (r = 0.021, *p*-value = 0.43).

The above results indicate that the metabolite strategies in strains differ between growth stages.

Table 4.2. Mantel test statistic (Spearman's correlation coefficient r) between genetic and metabolic distances

Growth stage	***Dataset***	***Vs Genetic distance***
Exponential phase	Full data	0.108 (*p*-value = 0.22)
	Differential metabolites	0.180 (*p*-value = 0.10)
Stationary phase	Full data	-0.076 (*p*-value = 0.74)
	Differential metabolites	-0.112 (*p*-value = 0.81)

4.3.2.1. Metabolic diversity between algal strains

To obtain a more intuitive feel for the growth stage-specific differences at metabolites, we generated a presence/absence matrix of the metabolites detected in each growth stage. The batch corrected dataset is used for putative identification of metabolites using a database search (described in Section 4.2.3). The metabolites identified in different growth stages number 1,102 in exponential and 996 at stationary phase. This number represents the total number of metabolites that were predicted based on mass similarity (within the given 10 ppm threshold) to existing metabolites. Using the m/z features that were differential between strains, we obtained 466 metabolites at exponential and 655 metabolites at stationary phase. A matrix representing all the detected metabolites in this experiment was created. This presence/absence matrix had 4 different columns, wherein the presence of a metabolite in that column category is scored as '1' and absence as '0'.

This is visualized as a heatmap (Figure 4.6), wherein the detection of the metabolite in that set is indicated by a grey shade. The majority of metabolites detected seems to be common between exponential (day 4) and stationary phase (day 12), when the complete profile is used. However the differential metabolites, i.e. metabolites that vary in their abundance among the 22 strains were mostly unique to either exponential or stationary phase.

The Venn diagram in Figure 4.7 reveals that out of 655 differential metabolites at stationary phase, 311 were unique to the stationary phase alone. In other words, the metabolic differences between strains at stationary phase were due to significant changes to the abundance levels in 311 metabolites. The concentration levels for these metabolites did not significantly change or were not detected at exponential phase. A similar analysis shows that 56 metabolites were varying in their concentration only at exponential phase and not stationary phase. Taken together, a total of 655 unique metabolites are accountable for the strain-specific metabolic differences among the 22 strains at exponential and stationary phases. Analysis of these metabolites will help understand the metabolic individuality of algal strains.

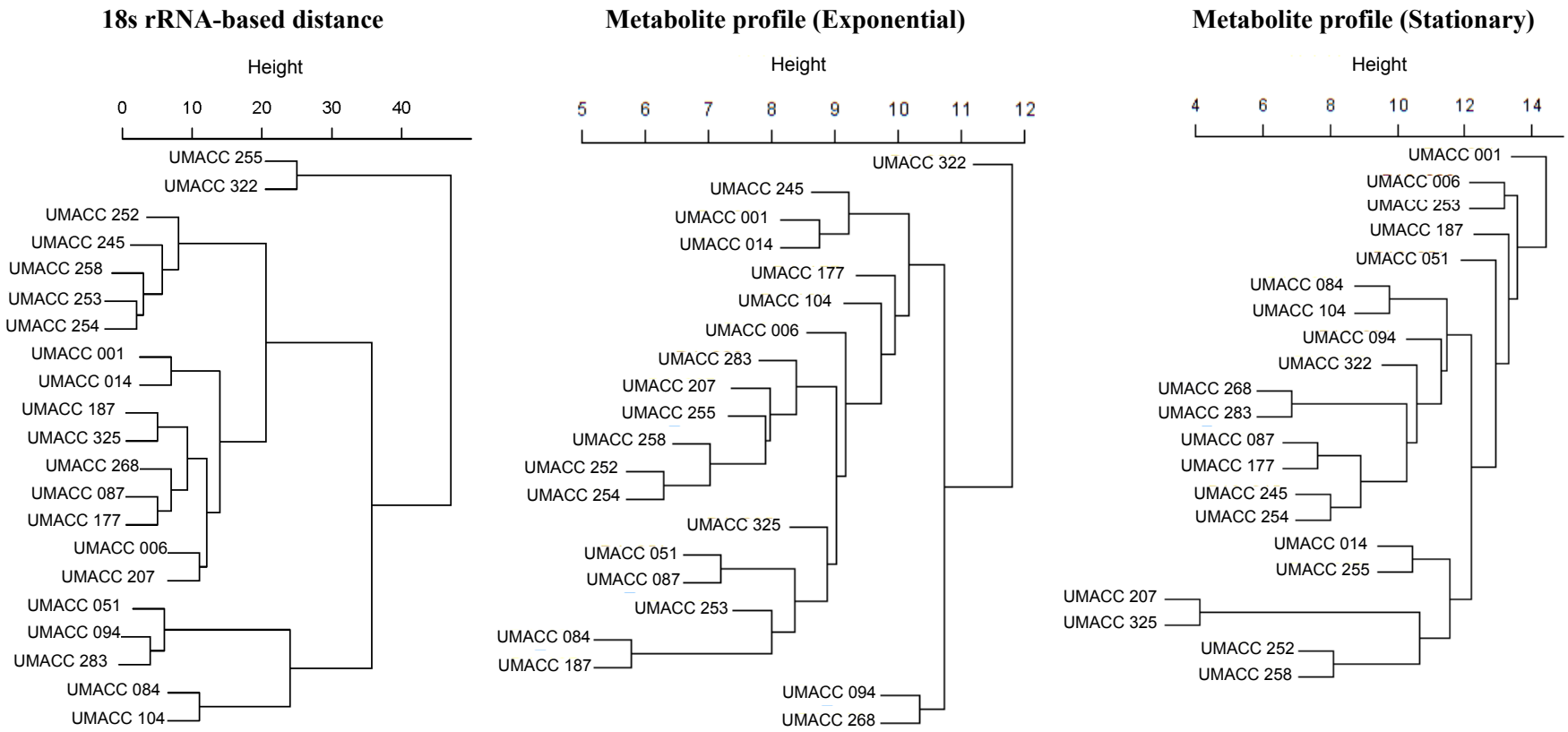


Figure 4.5. Distances between 21 strains (A) Genetic distance (Levenshtein distance), Metabolic distance (Euclidean distance using full dataset) at exponential phase (B) and stationary phase (C)

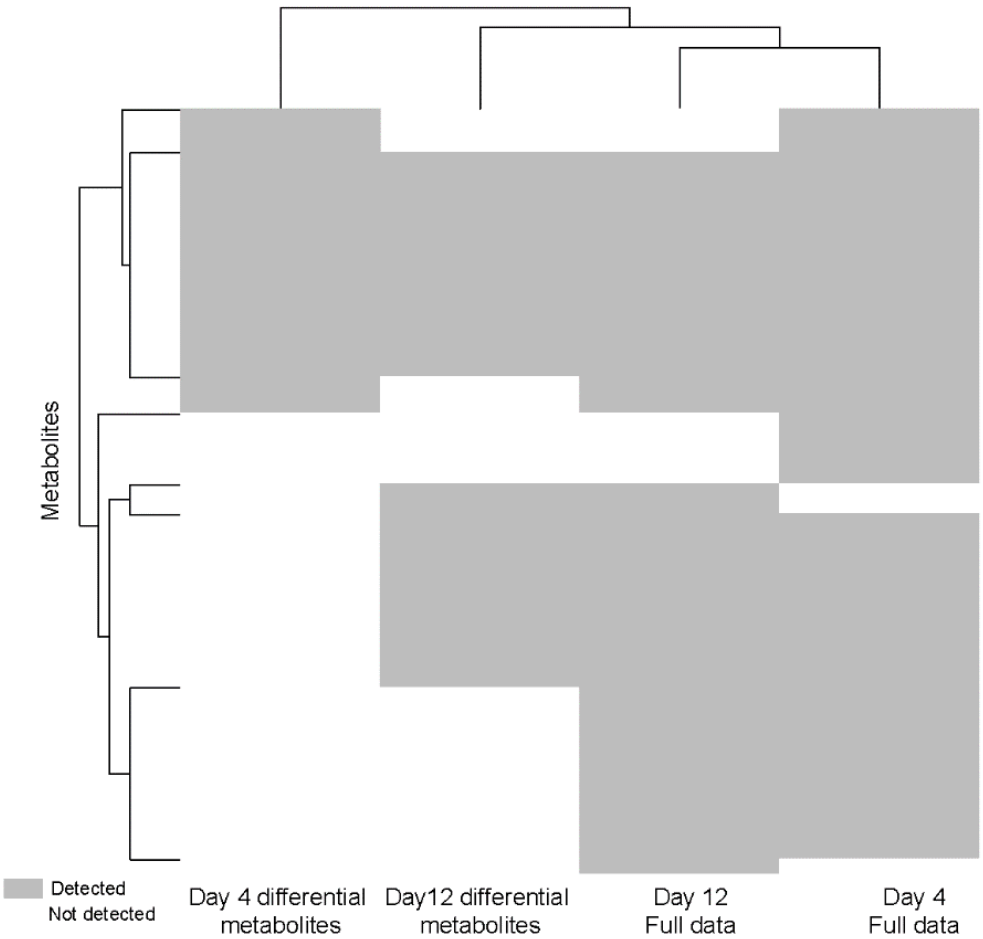


Figure 4.6. Heatmap (presence of a metabolite indicated in grey shade) showing differences in metabolites detected between growth stages. The rows represent metabolites, and columns are different datasets.

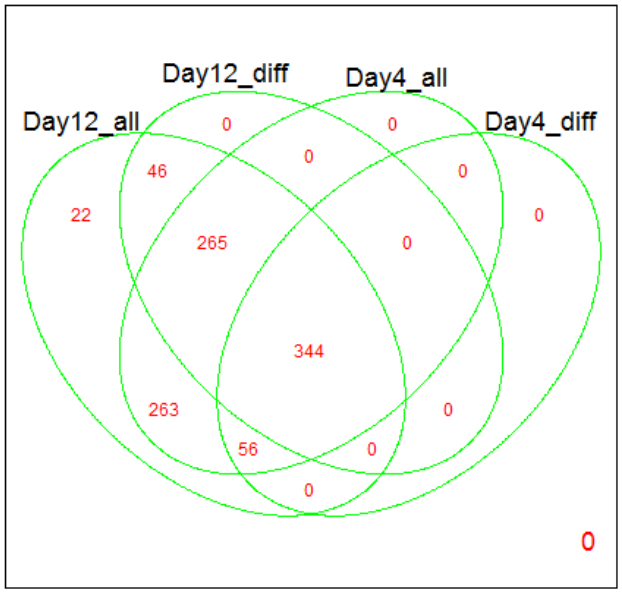


Figure 4.7. Venn diagram highlights the differences between the metabolites detected in each growth stage. *Abbreviations:* Day12_all- Total metabolites detected in stationary phase, Day12_dif*F*- Differential metabolites in stationary phase, Day4_all- Total metabolites detected in exponential phase, and Day4_dif*F*- Differential metabolites detected in exponential phase.

Strain-specific metabolic phenotypes arise from both regulatory (concentration) and structural (chemical structure) differences in metabolites. Strains of the sub cluster of *Parachlorella* strains D4_255, D4_258, D4_252 and D4_254 at the exponential phase (Figure 4.8A) were all isolated from the same location in Selangor. It is interesting to observe the similarity between the metabolic profiles of these strains and raises the hypothesis that habitat might have an influence on the metabolic processes at early growth phases (No such clustering was observed during the stationary phase (Figure 4.8B)). However, the above hypothesis, though interesting, is purely speculative and warrants more experimentation and increased sample size to achieve statistically robust results.

Interestingly, there were no differential metabolite features i.e. features with significant variation in their abundance when the strains were grouped according to their genotype (*Chlorella* Vs *Parachlorella*), at both exponential and stationary phase. Therefore, the major source of differences in the metabolite levels of these strains are due to strain-specific regulatory variations and/or environmental factors (such as habitat) influencing the biochemical phenotype.

This analysis, highlights the fact that the time at which the metabolome is profiled is indicative of the metabolic state of a system. This should especially be taken note of when trying to understand an uncharacterized organism's metabolic potential. In ideal scenarios metabolic profiling should be performed mostly over a time series.

To identify metabolic pathways that were detected in both growth stages, these metabolites were mapped on *Chlorella*-specific metabolic pathways in KEGG (Table 4.3 and Figure 4.9). Typically pathway or function enrichments in omics analysis are calculated by performing overrepresentation analysis using methods such as hypergeometric tests or Fisher's exact test. This produces in an enrichment score, typically in the form of *p*-values as an indicative measure of pathway enrichment. Such an analysis was also performed for this study, specifically $fisher.test$ was used to test

for overrepresentation of these pathways in the total dataset and in the features that were differential between strains at each growth stage. It should be noted that this test was performed by treating the number of metabolites detected at each stage in the full dataset as the total population size for that stage. In typical experiments, such analysis is performed by treating the total number of entities in the pathway as the population size for that pathways, and the total number of entities in all metabolic pathways as the total population size. However, in metabolomics experiments, it should be noted that the maximum number of entities in a metabolic pathway for an experiment, is limited by the detection method (extraction protocols and capacity of mass spectrometers) used and the database. Therefore, in order to apply a uniform and unbiased measure of enrichment, the total population size for each stage was calculated using the total number of metabolites detected in that stage. The resulting *p*-values from this analysis are shown within brackets for each pathway in Table 4.3. Only 2-Oxocarboxylic acid metabolism in stationary phase had a significant enrichment score.

The objective of this experiment was to observe the differences in the type of metabolites and their associated pathways. In order to capture differential enrichment of metabolic pathways, a targeted study with an extraction protocol optimized for such analysis and using instrumentation (for example tandem MS/MS) that can facilitate such comparisons is required. In scenarios where such experimentation is not possible, alternate complementary levels of information such as those measuring the transcriptome levels should be used to strengthen the statistical and biological interpretation (an example for such an application is described in Section 5.3.1).

However, biological interpretation based on the (i) presence and absence of metabolites, and (ii) overall coverage of the untargeted metabolic profiles when overlaid on the canonical *Chlorella*-specific pathways, can still be derived from the metabolomics analysis (without MS/MS data) conducted in this study. As the same technique was used for profiling strains in both growth stages, it is assumed that the

detection limits for metabolites in both growth stages will be similar. With this hypothesis, the presence or absence of metabolites clearly relates to the differences in the type of metabolites and their associated processes that are active in metabolic pathways. These shifts in metabolic strategies during growth stages as shown in Figure 4.9, can provide valuable clues to understand metabolic diversity of the 22 strains. In Figure 4.9, the total number of metabolites are represented as 100 percentage in the columns. The corresponding percentage of a pathway in the full data is shaded dark grey, whereas the number of differential metabolites identified onto that pathway is coloured orange (the percentage is calculated based on the total number of metabolites for that pathway).

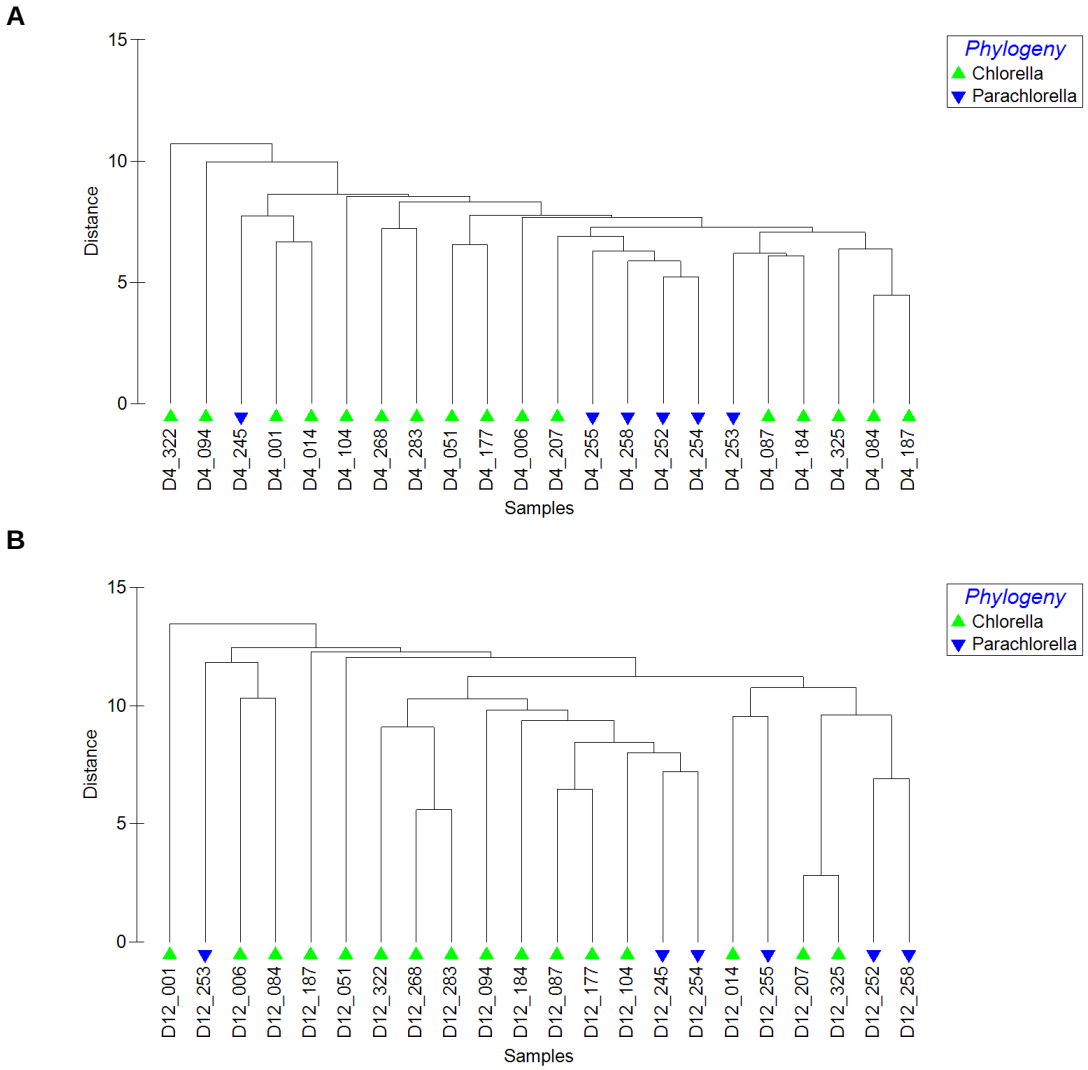


Figure 4.8. Hierarchical clustering of differential metabolites using Euclidean distance and average linkage method at (A) Exponential phase, (B) Stationary phase.

Expectedly, the top pathways are associated with secondary metabolism or energy (Table 4.3). The maximum number of differential metabolites between the strains at both exponential and stationary phase belonged to secondary metabolism pathways such as (i) Carotenoids, ubiquinone and other terpenoids biosynthesis; (ii) lipid metabolism; Interestingly for the above two pathway classes, the number of differential metabolites in stationary phase is almost double the number found in exponential phase. These numbers indicate that the diversity in lipid production between the 22 strains arise mainly during the stationary growth phase; (iii) alpha-linolenic acid (jasmonic acid) metabolism and (iv) phenylpropanoid metabolism. Similarly, differential metabolites were greater in stationary phase than in exponential phase for the above metabolic pathways. Linolenic acid is one of the major determinants of fatty acid content in algae, thus the differences in numbers indicate that the strains have the diverse metabolic strategies in alpha-linolenic acid metabolism that eventually result in varying fatty acid levels among these strains. It will also be interesting to perform a targeted analysis of the jasmonic acid pathway (discussed in Chapter 5) to determine whether these algal strains also have rudimentary mechanisms for phytohormone biosynthesis (Blanc et al., 2010). Phenylpropanoids include a large class of metabolites (discussed inn Chapter 5), with diverse functional roles, thus targeted analysis using MS/MS can further elucidate the implications of differences in this metabolite class.

Finally, differential metabolites which mapped onto pathways related to energy generation were mainly associated with (i) amino acid metabolism, (ii) 2-Oxocarboxylic acid metabolism and (iii) Aminoacyl-tRNA biosynthesis. The trends between the number of differential metabolites in stationary and exponential phase were mixed and require confirmation using MS/MS data to understand the differences in metabolic strategies (Table 4.3). This is also illustrated in Figure 4.9. Supplementary Dataset 2 provides metabolites identified in each pathway as separate lists.

Table 4.3. Number of metabolites detected in each pathway for each growth stage. Enrichment values are shown within the brackets.

Pathway name	***Total Metabolites***	***Day12-FullData***	***Day12-Differential***	***Day4-FullData***	***Day4-Differential***
cvr00906 Carotenoid biosynthesis	115	19	16 (0.17)	24	13 (0.19)
cvr01210 2-Oxocarboxylic acid metabolism	NA	13	12 (0.08)	13	8 (0.14)
cvr00260 Glycine, serine and threonine metabolism	51	13	10 (0.49)	13	5 (0.74)
cvr00564 Glycerophospholipid metabolism	52	10	10 (0.04)	11	7 (0.14)
cvr00592 alpha-Linolenic acid metabolism	40	11	10 (0.14)	11	6 (0.32)
cvr00130 Ubiquinone and other terpenoid-quinone biosynthesis	80	13	9 (0.72)	14	5 (0.81)
cvr02010 ABC transporters	122	14	9 (0.84)	15	7 (0.5)
cvr00330 Arginine and proline metabolism	90	10	8 (0.44)	9	5 (0.34)
cvr00970 Aminoacyl-tRNA biosynthesis	52	9	8 (0.23)	10	4 (0.7)
cvr00270 Cysteine and methionine metabolism	57	9	7 (0.52)	8	2 (0.93)
cvr00940 Phenylpropanoid biosynthesis	65	7	7 (0.1)	7	3 (0.66)
cvr00960 Tropane, piperidine and pyridine alkaloid biosynthesis	68	11	7 (0.84)	12	5 (0.66)
cvr00460 Cyanoamino acid metabolism	46	8	6 (0.61)	9	2 (0.96)
cvr00590 Arachidonic acid metabolism	75	6	6 (0.14)	6	5 (0.06)
cvr00860 Porphyrin and chlorophyll metabolism	124	10	6 (0.89)	11	4 (0.78)
cvr00950 Isoquinoline alkaloid biosynthesis	93	6	6 (0.14)	7	2 (0.89)
cvr01220 Degradation of aromatic compounds	NA	9	6 (0.78)	9	3 (0.83)
cvr00240 Pyrimidine metabolism	66	5	4 (0.57)	6	3 (0.53)
cvr00360 Phenylalanine metabolism	72	7	4 (0.9)	7	3 (0.66)
cvr00380 Tryptophan metabolism	82	8	4 (0.96)	9	3 (0.83)
cvr00591 Linoleic acid metabolism	28	4	4 (0.27)	3	3 (0.08)
cvr00600 Sphingolipid metabolism	25	4	4 (0.27)	4	3 (0.22)
cvr01040 Biosynthesis of unsaturated fatty acids	54	5	4 (0.57)	4	3 (0.22)
cvr00230 Purine metabolism	92	5	3 (0.87)	5	3 (0.38)
cvr00250 Alanine, aspartate and glutamate metabolism	24	4	3 (0.69)	4	3 (0.22)
cvr00310 Lysine degradation	52	5	3 (0.87)	6	2 (0.82)
cvr00400 Phenylalanine, tyrosine and tryptophan biosynthesis	35	5	3 (0.87)	5	0 (1)
cvr00760 Nicotinate and nicotinamide metabolism	47	4	3 (0.69)	4	2 (0.58)
cvr00770 Pantothenate and CoA biosynthesis	55	4	3 (0.69)	4	1 (0.9)
cvr01200 Carbon metabolism	NA	6	3 (0.94)	7	3 (0.18)

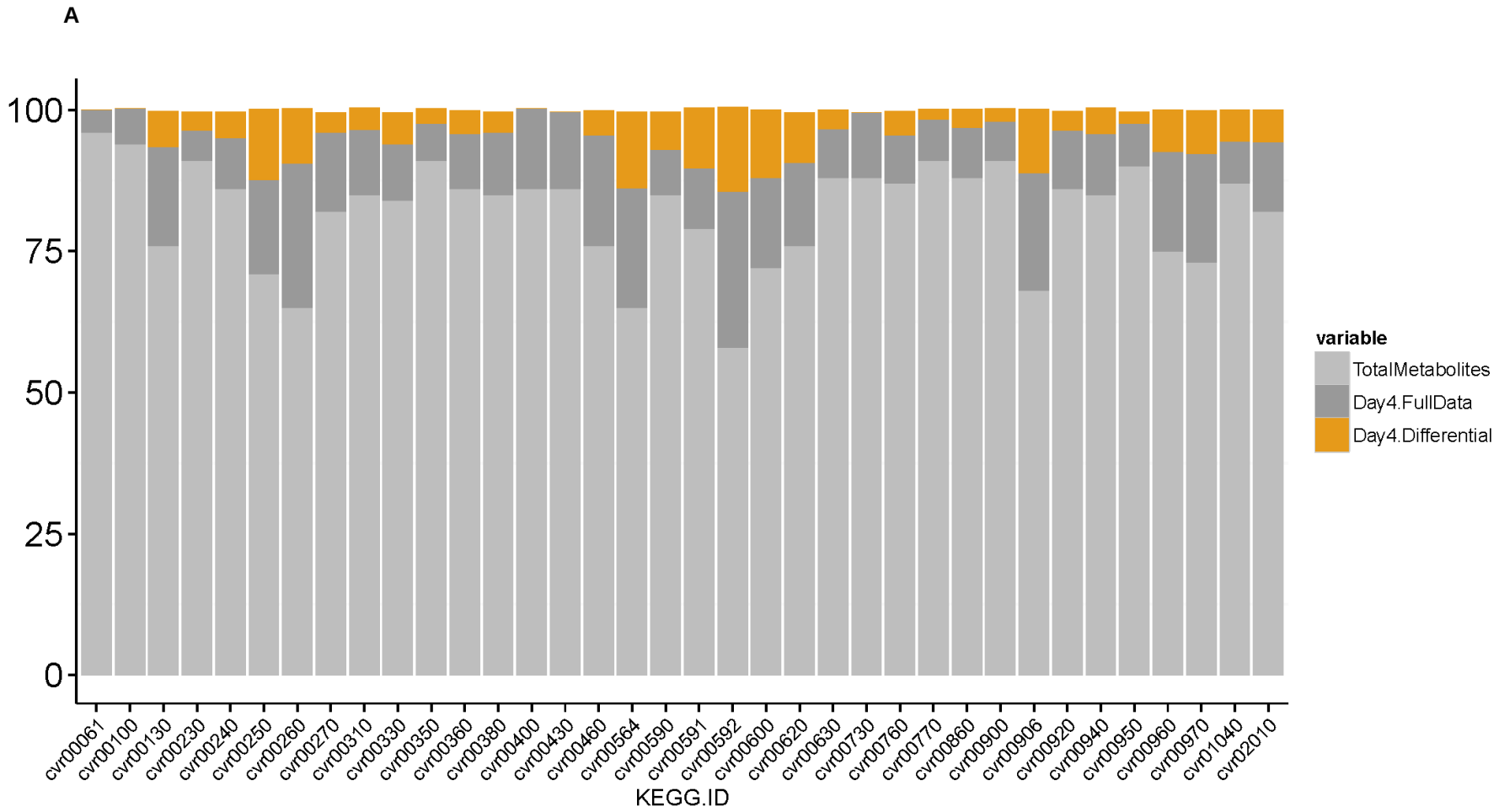


Figure 4.9A. Differential presence of metabolites in metabolic pathways at Exponential phase. *Abbreviations:* Day4.FullData- Total metabolites detected in exponential phase, Day4.Differntial- Differential metabolites in exponential phase.

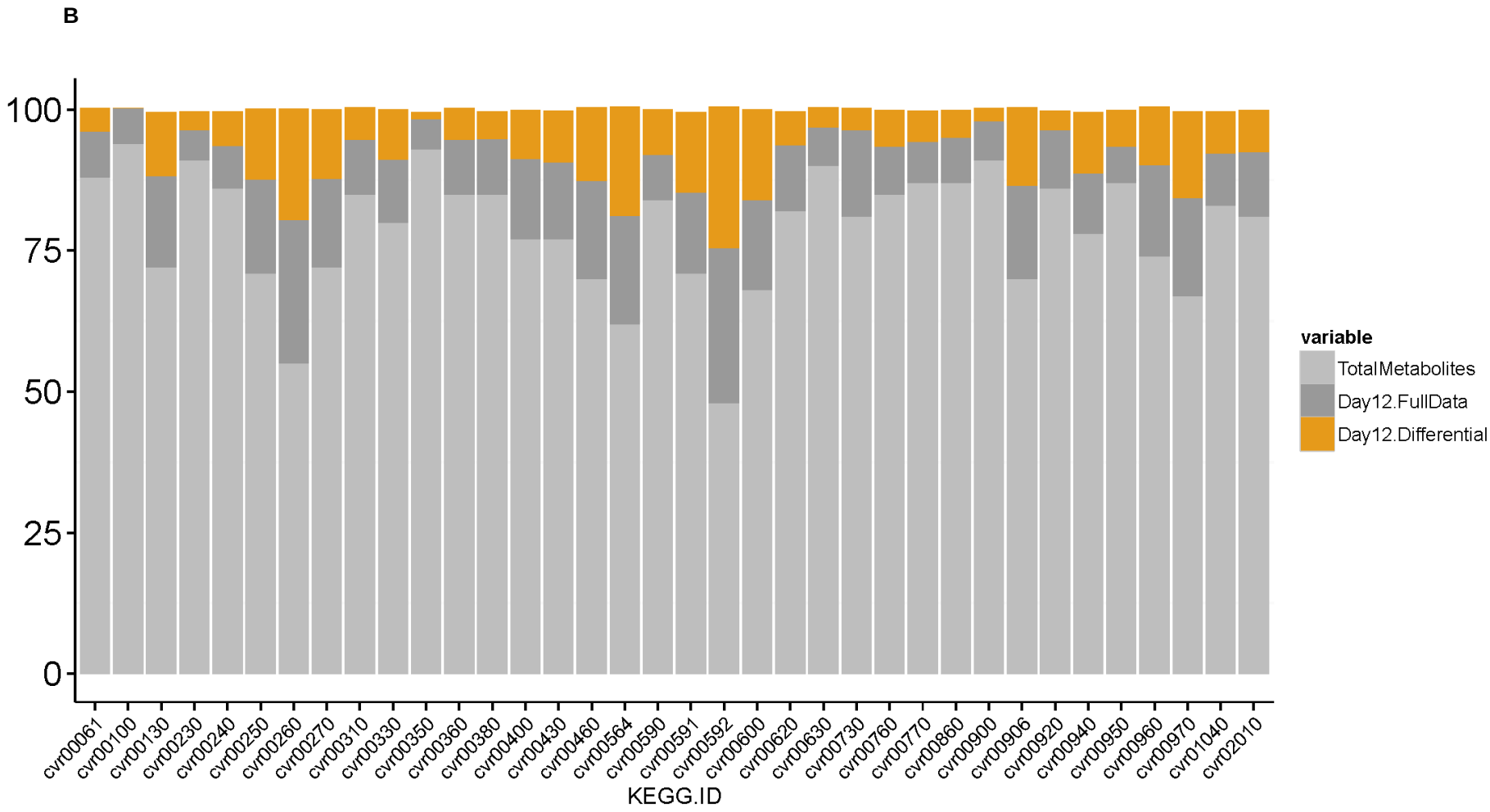


Figure 4.9B. Differential presence of metabolites in metabolic pathways at Stationary phase. *Abbreviations:* Day12.FullData- Total metabolites detected in stationary phase, Day12.Differntial- Differential metabolites in stationary phase.

4.3.3. Physicochemical profiles

The physicochemical profiles, totalling 7 different measurements (see Materials and methods) were characterized for the 22 strains. With the observation of strain-specific and growth stage-specific differences in the metabolic profiles, we analysed the physicochemical data to identify whether they exhibited a pattern similar to the metabolic profiles. We performed PCA on the centered and scaled physicochemical data using the *prcomp* function from the 'stats' package in R. The separation between strains at exponential and stationary growth stage was observed in the second PC axis which explained around 35% variation in the data. Interestingly, the variation in the first PC was largely associated to the influence of lipid or protein content in strains.

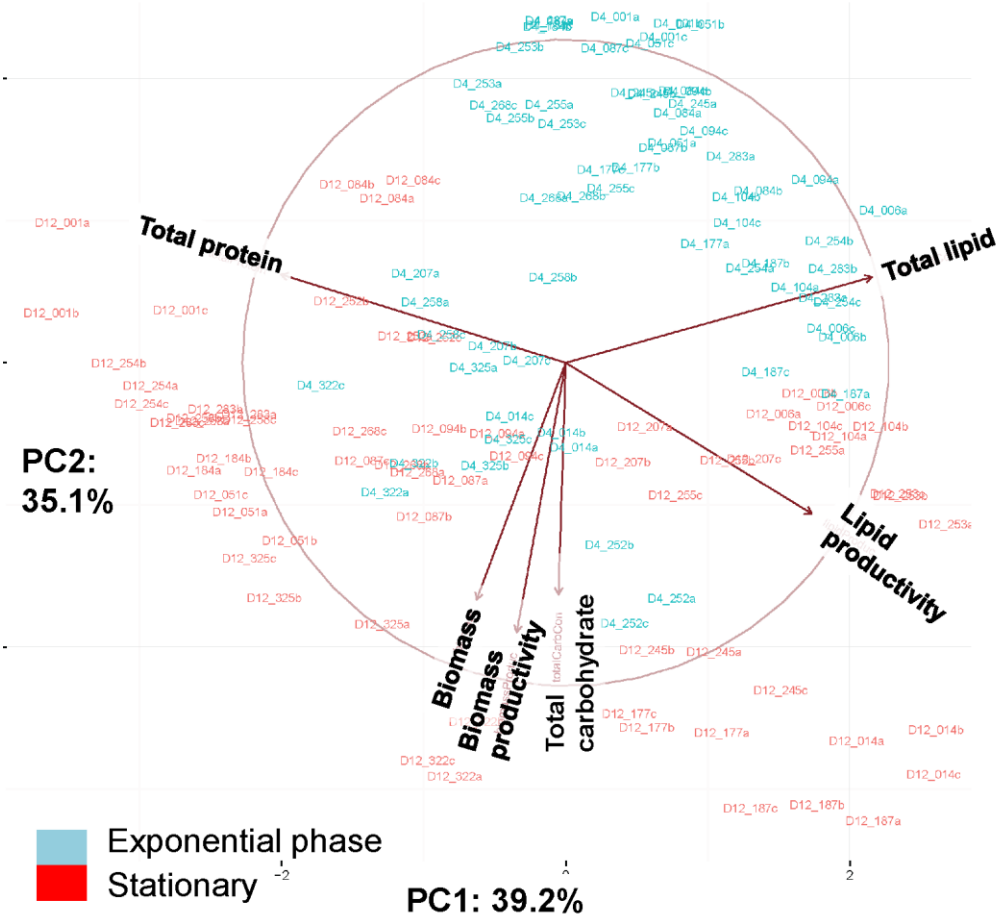


Figure 4.10. PCA of physicochemical profiles of 22 strains. Strains in exponential phase are coloured blue, whereas stationary phase is coloured red.

We observed that the total protein and lipid productivity were inversely related to each other. To analyse this relationship further, we performed Kendall tau-based correlation (Figure 4.11).

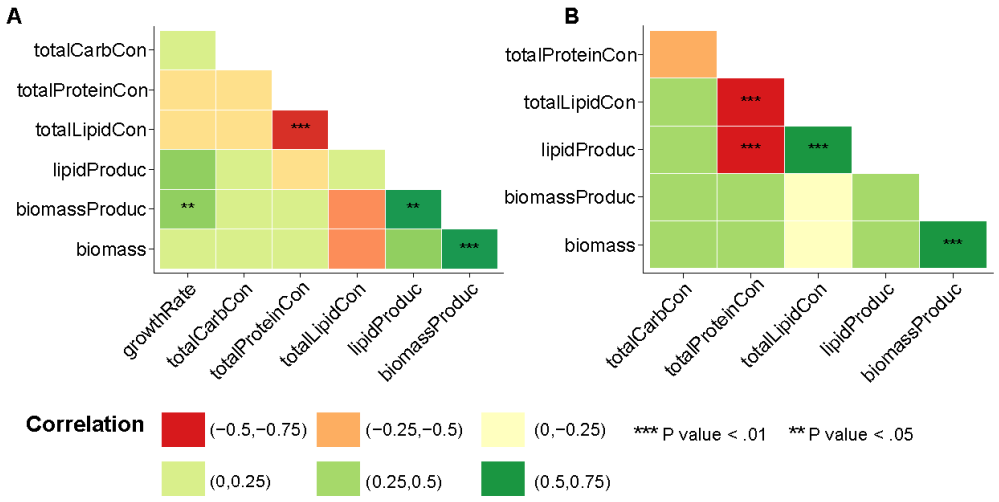


Figure 4.11. Correlation between physiochemical measures of 22 strains at (A) Exponential phase, (B) Stationary phase. *P*-values were adjusted using Bonferronni corrections. *Abbreviations:* totalCarbCon- Total Carbohydrate Content, totalProteinCon- Total Protein Content, totalLipidCon- Total Lipid Content, lipidProduc- Lipid Productivity, biomassProduc- Biomass Productivity, biomass- Biomass, growthRate- Growth Rate.

From Figure 4.11, we observed that total lipid concentration was negatively correlated with total protein at both exponential and stationary phase. At exponential phase, both biomass productivity and lipid productivity were significantly correlated. The lack of correlation between biomass and lipids during stationary phase is not surprising as the lipid accumulation is commonly observed at the start of stationary phase when nutrients become limited. As lipid productivity is often used as an indicator for biofuel production efficiency, its correlation with biomass productivity, substantiates its usefulness as a suitable physicochemical-biomarker (Huerlimann et al., 2010).

The significant negative correlation between total lipid and total protein indicates that likelihood of inherent lipogenic and proteogenic strains. Understanding the differences in molecular mechanisms between such strains can provide valuable insights into their resource partitioning strategies. Expectedly as the biochemical

profiles significantly vary between growth stages, the biochemical distance matrix of both exponential and stationary phase calculated using $vegdist$ function (Euclidean distance) in 'vegan, showed no correlation (r = -0.004, *p*-value = 0.48, calculated using Mantel test, $mantel$ function using Spearman's correlation coefficient in 'vegan' package, 999 permutations). We then used the Mantel test to analyse pattern similarity between biochemical profiles and metabolite distances at exponential and stationary phase (Table 4.4).

Table 4.4. Mantel test statistic (Spearman's correlation coefficient r) between biochemical and metabolic distances.

Growth stage	***Dataset***	***Vs Biochemical profiles***
Exponential phase	Full data	-0.011 (*p*-value = 0.55)
	Differential metabolites	0.024 (*p*-value = 0.39)
Stationary phase	Full data	0.263 (*p*-value = 0.013)
	Differential metabolites	0.248 (*p*-value = 0.016)

From the above table, we observe that there were small but significant correlations between the metabolic profiles of strains in the stationary phase and their biochemical profiles. Interestingly, there was no correlation between strains at exponential phase and their physiochemical profiles. The lack of a significant correlation between biochemical traits and metabolic profiles at exponential phase indicate that microbes decide their metabolic resource partitioning strategy late during their growth phase.

To analyse if there were significant associations between any biochemical trait of the strains and the observed metabolic variation, we performed PERMANOVA (Permutational Multivariate Analysis of Variance Using Distance Matrices) with 999 permutations using the $adonis$ function in 'vegan' package in R. Only biomass had a significant association with the metabolite profiles, both with differential metabolites (R2 = 0.084, *p*-value = 0.015) and full data set (R2 = 0.079, *p*-value = 0.019) at stationary phase. With biomass explaining around 8% of the variation in metabolite profiles, the results hint that the resource partitioning strategy of strains is reflected in

its end phenotype, thus, stressing the importance of analysing cellular metabolism in order to understand the molecular mechanisms that determine the efficiency of biofuel production. None of the other biochemical traits were significantly associated with metabolic profiles at stationary phases, whereas there were no correlation between any biochemical trait and the metabolic profiles at exponential phase.

4.3.3.1. Biochemical determinants of resource partitioning

The success of large-scale microalgae-based biofuel production depends on growth rate and oil content in strains. The high- and low-yielding groups are interesting candidates for the identification of naturally varying species-specific metabolic traits. Thus, we set out to determine if the best strains in each of the biochemical traits have markedly different metabolite profiles when compared to the worst strains (lowest producers, growers etc.) in each trait. Centered and scaled data were used to determine the best and worst strains. For each trait, strains in the top and bottom 10% of the range (values measured for each trait) were extracted (Figure 4.12) using the *quantile* function in R.

Traits	DAY4						Traits	DAY12					
	HIGH			LOW				HIGH			LOW		
growthRate	D4_245	D4_252	D4_254	D4_094	D4_184	D4_255	growthRate	NA	NA	NA	NA	NA	NA
biomass	D4_187	D4_255	D4_322	D4_001	D4_051	D4_253	biomass	D12_177	D12_187	D12_325	D12_006	D12_084	D12_253
biomassProd	D4_014	D4_252	D4_322	D4_001	D4_051	D4_253	biomassProd	D12_177	D12_187	D12_325	D12_006	D12_084	D12_253
lipidProd	D4_187	D4_252	D4_254	D4_001	D4_051	D4_253	lipidProd	D12_014	D12_104	D12_187	D12_001	D12_184	D12_254
totalLipidCon	D4_006	D4_094	D4_254	D4_207	D4_322	D4_325	totalLipidCon	D12_006	D12_104	D12_253	D12_001	D12_184	D12_254
totalProteinCon	D4_207	D4_252	D4_325	D4_006	D4_104	D4_283	totalProteinCon	D12_001	D12_254	D12_258	D12_014	D12_245	D12_253
totalCarbCon	D4_258	D4_268	D4_322	D4_184	D4_245	D4_254	totalCarbCon	D12_014	D12_245	D12_322	D12_001	D12_094	D12_104

Figure 4.12. Classification of strains based on their biochemical profiles. Strains which are consistently best performers are coloured dark blue; worst performers are coloured red; and strains which switched between lipid and protein productivity are coloured green. *Abbreviations:* growthRate- Growth Rate, biomass- Biomass, biomassProd- Biomass Productivity, lipidProduc- Lipid Productivity, totalLipidCon- Total Lipid Content, totalProteinCon- Total Protein Content, totalCarbCon- Total Carbohydrate Content.

From the above figure, we observe that strain UMACC 322 which had the maximum metabolic divergence at exponential phase (D4_322, Figure 4.5B) is the top

strain in among three different biochemical traits (Biomass, Biomass productivity and Total Carbohydrate Content) at exponential phase. Similarly, UMACC 001 which had the maximum divergence based on metabolic distance at stationary phase (D12_001, Figure 4.5C), is the worst performer in three different traits (Biomass, Biomass productivity and Lipid productivity).

We had earlier observed that D4_252 and D4_254 formed a separate sub cluster at exponential phase (Figure 4.5B). The possible divergence of their metabolic profiles emerges when comparing their biochemical traits. From Figure 4.12, we observe that strains UMAC 254 and UMACC 252 are one of the top lipid producers. However at stationary phase, UMAC 254 (D12_254) has switched to become a strain having one of the highest total protein content, thus, suggesting that early markers of resource partitioning strategy can be identified using an organism's metabolic profile. Similarly UMACC 187 (D12_187) which is not one the top strains at exponential phase, but has high biomass and biomass productivity and lipid productivity formed a separate sub cluster based on the metabolic profiles at stationary phase (Figure 4.5C).

To analyse whether the differences in the biochemical traits reflect in quantitative differences in their metabolic profiles, *F*-tests were performed on metabolite matrices using the $mt.maxT$ function (with 1000 permutations) from the 'multtest' package (Pollard et al., 2005) in R. In this test, the top 3 strains in each category were grouped into one category and their metabolite profiles were compared against the bottom 3 for the categories given below (Table 4.5). Importantly, we did not observe influence of growth media on the metabolic profile differences, thus ruling out batch affects during culture preparation.

Biomass which had a significant association with metabolite profiles at stationary phase seems to elicit the maximum difference (showing 795 features) between the metabolite profiles of strains at stationary phase. In other words, strains which differ based on biomass had divergent metabolite profiles. At exponential phase

biomass and lipid productivity (showing 196 and 116 features, respectively) which are dependent on the growth rate, along with lipid content (104 features) seem to elicit the maximum difference in metabolite profiles. The metabolic and biochemical relationships between strains are vastly different between growth stages. Strains having extreme biochemical profiles also have relatively divergent metabolic profiles compared to the others.

Table 4.5. Biochemical determinants of metabolic diversity among 22 *Chlorella* strains. The numbers represent the number of significant differential features for each comparison.

Comparisons	***Day 4***	***Day 12***
	Top 10% vs bottom 10%	**Top 10% vs bottom 10%**
Chlorella Vs Parachlorella	0	0
Prov-Seawater media and other media	0	0
Growth rate	8	0
Biomass	55	795
Biomass productivity	196	795
Total carbon content	25	265
Lipid productivity	116	128
Total lipid content	104	173
Total protein content	39	190

Thus, in order to explore the behaviour of strain-specific metabolite associations in determining resource partitioning strategies, we used Haygood's measure (Haygood et al., 2007). Haygood et al. described an approach for identifying tissue-specific gene expression patterns using maximal expression criteria: since few genes are expressed in a completely tissue-specific manner, even if that gene shows maximum expression in one tissue, there is a high probability that this expression level might also be witnessed in other tissues. This problem was recast into linear algebra by Haygood et al. Specifically, each gene is treated as a "vector" in the "tissue expression" space", thus, multiple tissues are treated as separate dimensions. In this approach,

tissue-specific measures are described by computing the square of the cosine angle between gene-vector and each of the tissue axes. However, this is equal to computing the σ_{gt} of each gene, with the number of genes g=1,..,G and tissues ranging from t=1,..,T

$$\sigma_{gt} = \frac{E_{gt}^2}{\sum_{k=1}^{T} E_{gk}^2}$$

In the above expression E_{gt} is measured as the expression level of gene g in tissue t. This equates to simply normalizing the expression of a gene to the total expression measured across all included tissues. The advantage of such an approach is that the sum of σ_{gt} is always going to be unity across all tissues. Furthermore, for any gene, the average of σ will always be 1/T. Thus, deviations from this value can be used to explore tissue-specificity. Furthermore, Haygood measure is independent of the overall expression level (application for metabolomics developed in discussion with Dr. Rohan Williams). In this study, the above concept had been modified to reflect strains as equivalent to tissues and metabolites abundance levels measured in the place of genes. This concept is particularly suited for metabolomics, as a higher or lower abundance of a metabolite in a particular strain hints at the possibility that the specific metabolite might play a critical role in determining the activation of its associated metabolite pathway. In our dataset, this translates into normalizing the abundance of each metabolite to the total abundance of the metabolite detected across all strains. Thus, the Haygood measure for a metabolite which does not have any specific abundance trend will be around 0.0454 (1/22 strains=0.0454). The above calculation assumes that a metabolite, which is within the normal abundance levels has an equal association with all the strains. Thus if the total abundance level is 1, this is split equally among all the strains resulting in 0.0454 being the Haygood measure for a non-strain-specific metabolite. This analysis was performed on the 466 differential metabolites from exponential phase and 655 differential metabolites from stationary phase (Figure

4.13). It is important to note that while the heatmap in Figure 4.6 shows the presence/absence of a list of metabolites across growth stages, the values represented for each metabolite in Figure 4.13 is the Haygood measure for that metabolite in that particular strain.

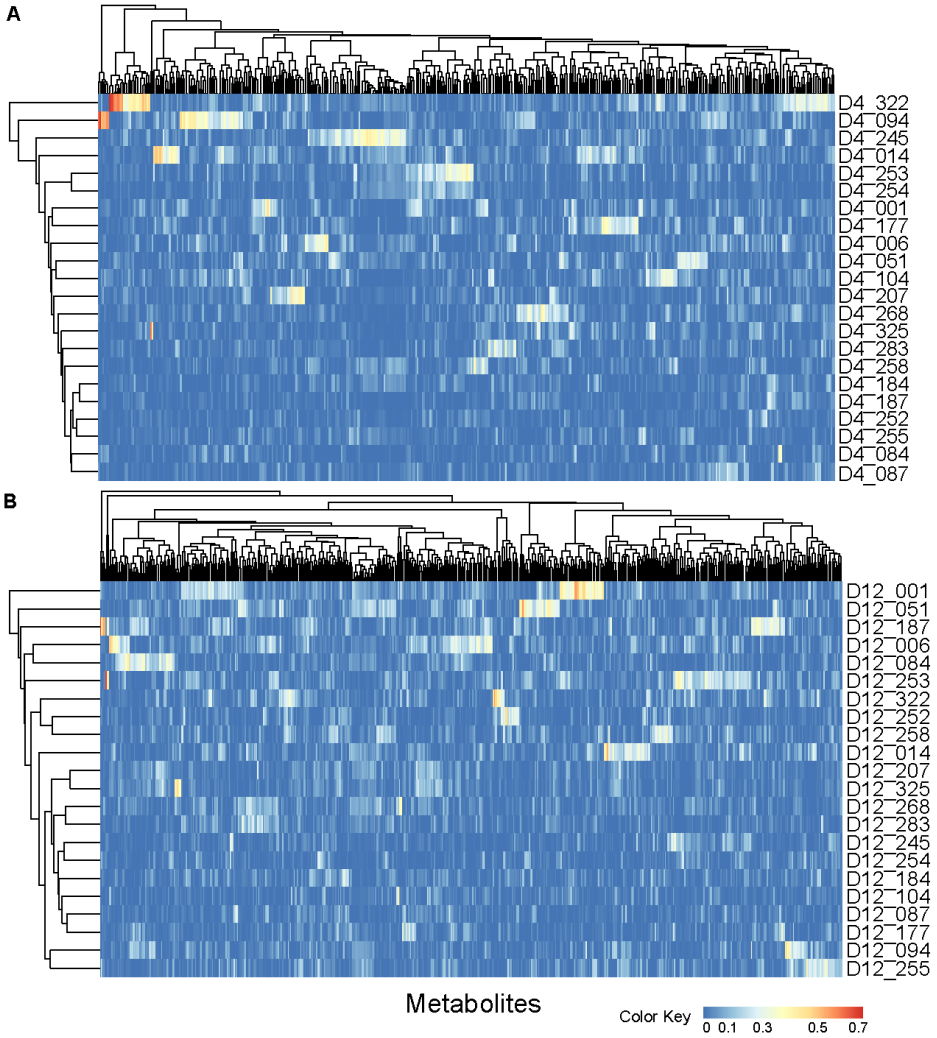


Figure 4.13. Heatmap showing the normalized values based on Haygood measure for (A) Exponential phase, (B) Stationary phase. The columns represent individual metabolites and the rows represent strains. Heatmap is plotted with Euclidean distance with average linkage clustering using the *pheatmap* function in 'pheatmap' package in R. High metabolite association values are indicated in red shades in the plot.

In (Figure 4.13), strains which previously showed high metabolite divergence or were categorized as the best or worst performers based on biochemical straits had distinct metabolite association patterns. For example D4_322 at exponential phase (Figure 4.13A) and D12_001 or D12_187 at stationary phase (Figure 4.13B) has a high

number of metabolites showing significant associations (increasing metabolite association is shaded from blue to red).

Expectedly the metabolic pathways which were linked to these strain-specific metabolite associations were mainly related to energy, lipid or fatty acid biosynthesis (Table 4.3). Taken together, our results suggest that strains UMACC 187, UMACC 254 and UMACC 322 to be the most efficient strains which can be targeted for further analysis for determining efficiency of large-scale biofuel production. Furthermore, UMACC 001 has a radically different metabolic and biochemical profile, and requires further analysis to determine the reprogramming strategies that allow it to have one of the highest total protein contents.

4.4. Conclusions

This survey of naturally varying strains measuring genetic diversity, biochemical characteristics and metabolic profiles allowed us to perform robust multivariate statistics to analyse the metabolic potential 22 *Chlorella* strains. We compared the metabolic divergence, genotype-metabotype associations and identified metabolic correlates of biotechnology related traits. This integrative analysis resulted in the identification of strain-specific metabolic biomarkers that can be developed for chemotaxonomic classification of oleaginous algae.

Our results suggest that the metabolic differences among the strains might be due to differences in gene expression levels or variation in genome sequences not covered by the 18S rRNA sampling. In this study, the strains were isolated and had been grown in lab conditions. However, they still exhibit significant metabolic and biochemical differences. Furthermore, the metabolic and biochemical profiles were largely invariant based on ribosomal sequence-based genetic distances, thus, suggesting that the strain might have undergone habitat-influenced genetic adaptions. Such adaptions result in changes to gene function and expression levels which can affect the gene product and therefore its phenotypic evolution (Whittkopp, 2013).

Furthermore, the regulatory mechanisms involved in modulating metabolic functions remain incompletely defined. Thus, while biochemical methods provide a glimpse into an organism's metabolic potential, the underlying mechanisms linking the genotype to its metabolic phenotype remain largely understudied. These require targeted transcriptomics and genomic analysis to narrow down the traits to specific genes. Such complementary information on gene expression levels can help build metabolic models and aid metabolic engineering techniques to probe symbiotic associations between genotype-metabotype (discussed using a model system in Chapter 5), their biochemical traits and influence of environmental factors.

Overall, the results from this study provide insights on the effect of genotypic differences and habitat-specific factors that produce large metabolic diversity between phylogentically similar strains. It is interesting to note that the variation in metabolic phenotypes is due to specific metabolites, thus indicating a highly selective metabolic strategy. These results also highlight the importance of untargeted analysis to identify natural variants. Such untargeted metabolomics approach in a diversity-oriented study provides a conceptual framework for effectively screening and classifying algal species without relying on genome information. The identification of naturally efficient strains provides the opportunity to selectively mine and isolate species of interest containing the desired biochemical and metabolic traits from known locations. As these strains have a natural tendency for producing desired compounds, they provide a favourable starting position for metabolic engineering strategies. Taken together, the approaches described in this study can help understand the effects of genetic or environmental perturbations on the metabolic diversity of biological systems.

Additionally, next generation sequencing techniques can be used to identify genomic potential and expression level variation for elucidating 'hotspots' based on gene-metabolic QTLs (Wen et al., 2014). These hotspots can be used for estimating the evolutionary divergence and can provide insights into to phenotypic buffering (Fu et

al., 2009) among the 22 strains. Such analysis will be useful for transforming systems level information info a functional large-scale production setup.

5. Data-dependent multi-omics approach to uncover effects of genetic perturbation on metabolic network

"Nothing unconnected ever occurs,
and anything unconnected
would instantly perish"

… Emanuel Swedenborg, 18th century Swedish scientist and philosopher

5.1. Introduction

Organisms being complex systems, regulate their physiological responses through inter-connected, non-linear, and dynamic interactions between multiple biological layers such as DNA, RNA, proteins and metabolites. Elucidation of such molecular mechanisms orchestrated via regulatory changes to gene expression and gene products can provide important clues in assessing cellular responses. With the metabolome representing the closest biochemical phenotype of an organism, analysing gene-metabolite relationships using an integrative omics approach can provide a systems level understanding of these cellular processes (Joyce and Palsson, 2006).

From the previous study, we determined that there is enormous metabolic diversity in naturally varying organisms. For example, in the plant kingdom, estimates indicate the presence of nearly 200,000 different types of metabolites with diverse physical and chemical properties (Fiehn, 2002). This diversity is generated via biochemical processes such as conjugations, hydroxylations, methylations, decarboxylations, oxidation/reduction and acyltransfer reactions. Among various conjugation processes, glycosylation which results in addition of sugar moieties to aglycone (acceptor) and deglycosylation are the most prominent.

Part of the results presented in this Chapter have been submitted for publication, *Amit Rai*, Shivshankar Umashankar*, Megha, Lim Boon Kiat, and Sanjay Swarup*, "TT8 affects innate immunity in Arabidopsis by reprogramming hormone biosynthesis and glycosylation of metabolites". Results from this Chapter are included as a part of a patent.

In plants, glycosylation or deglycosylation processes provide crucial modifications to the physicochemical properties of metabolites required for growth, development and stress response (Vaistij et al., 2009). This regio- and stereo- selective modification of aglycone molecules produce glycosides with diverse chemical structures and properties. Flavonoids consisting of almost 5,000 different structures, have important ecological, economic and pharmaceutical properties (Wink, 2010) and are one of the major classes of secondary metabolites to be influenced by glycosylation process. For example, glycosylation of a single flavonol metabolite, quercetin results in over 300 glycoside forms (Harborne and Baxter, 1999).

Glycosylation and deglycosylation processes are regulated by a specialized set of enzymes known as carbohydrate active enzymes (CAZy) (Lombard et al., 2014). CAZy comprises of five enzyme classes. For deglycosylation processes, they are glycoside hydrolases (GHs), polysaccharide lyases (PLs) and carbohydrate esterases (CEs) and while glycosylation is mediated via glycosyltransferases (GTs) and auxiliary activity enzymes (AAs). Though, glycosylated forms of metabolites and their aglycones along with the enzymes mediating such processes have been studied, there are still gaps in understanding how (i) the coordination between different molecular entities such as genes and metabolites occur during sugar conjugation (glycosylated) processes, and (ii) processes such as plant innate immunity, stress response and growth are related to glycosylation of metabolites (Wink, 2010).

In order to gain a holistic understanding of the molecular mechanisms co-ordinately regulated during glycosylation, we systematically perturbed a basic helix-loop-helix transcription factor, TRANSPARENT TESTA 8 (TT8) (Nesi et al., 2000) , in the model plant *Arabidopsis thaliana*. Using a model plant in a laboratory setup provided us with a robust experimental strategy devoid of unwanted variation due to environmental factors.

TT8 forms a ternary complex with two other proteins, TT2 and TTG1. This complex in turn regulates the expression of BAN and DFR genes from the ‘lower’

phenylpropanoid metabolic network (Baudry et al., 2004), thereby, co-ordinately controlling flavonoid biosynthesis. Recent reports (Xu et al., 2013) suggested that TT8 might be involved in regulating the expression of genes in flavonoid biosynthesis. Interestingly, TT8, was also shown to alter the chromatographic profiles of both the aglycone (Pelletier et al., 1999) and glycosylated (Narasimhan et al., 2003) forms of kaempferol and quercetin. However, it is not yet known whether TT8 directly regulates glycosylation of flavonoids and/or other metabolites (glycosylated and/or aglycone forms). These results indicate that TT8 might be a putative flavonoid glycosylation regulator and would be a suitable model to investigate: (i) glycosylation of flavonoids, (ii) the coordinated regulation of sugar conjugation, and, (iii) other processes that might be co-regulated with sugar conjugation. Furthermore, with glycosylation of flavonoids and other secondary metabolites regulated by CAZy genes, we hypothesize that TT8 affects glycosylation of flavonoids by regulating CAZy.

To discover novel glycosylation targets of TT8 and its regulatory network, we measured the effect of TT8 loss on the *Arabidopsis* metabolic network by performing non-targeted metabolomics profiling. Furthermore, to identify coordinated changes in the gene expression levels, we measured RNA levels using microarray-based expression profiling. To utilize the systems-level information from these high throughput datasets, I developed a novel integrative omics strategy that combines the genomic relationships with gene expression and metabolite abundances. In this study, I used multivariate statistical methods to analyse promoter sequences and performed enrichment analysis by integrating both genomics and metabolomics datasets.

To the best of our knowledge, this study represents the first approach to utilize orthogonal and complementary levels of biological information provided via genomics (analysis of shared promoter motifs), transcriptomics and non-targeted metabolomics in a systematic manner to understand genetic basis for glycosylation of metabolites.

5.2. Materials and methods

5.2.1. Plant materials and growth conditions

Arabidopsis thaliana plants of ecotype Wassilewskija (Ws-2) and *tt8-3* (deb122) were obtained from Versailles (INRA, France), while *tt8-2* and Ler-0 were obtained from the Arabidopsis seed stock center (ABRC, USA). Dexamethasone (dex) inducible *TT8:GR* overexpression lines were created by transforming 2x35S:*TT8:GR* promoter in Wt-Col (CS60000) background.

To generate 2x35S:TT8-GR construct, we amplified TT8 cDNA from PGWB20-TT8 clone using primers TT8-cla1(5'-AACTCGAGATGGATGAATCAAGTATTATTC-3') and TT8-xho1(5'-ATTATCGATTAGATTAGTATCATGTATTATGAC-3'), digested by Cla1, Xho1 and were introduced in HY109 vector between 2x35S promoter and glucocorticoid coding gene. Plant transformation was performed as described earlier (Bechtold and Pelletier, 1998). Twelve independent transformed lines were tested for expression levels of BAN and DFR, known direct targets of TT8 with and without dex treatment. Three lines that showed maximum up-regulation of known target genes were selected for analysing the phenotypes resulting from induced overexpression of TT8. For further studies, we selected inducible *TT8:GR* overexpression line that showed maximum up-regulation of BAN and DFR on dex treatment.

Seeds were surface sterilized using 30% Clorox with 7 minute incubation, followed by 6 times washing with autoclaved water and transferred to MS (Duchefa Biochemie) agar plate [1x MS media, 20 g/liter sucrose, and 6 g/liter phytoagar (pH 5.7)] for 4 days before placing them in growth chamber at 22°C for 16 h light/8 h dark cycle. For inducible lines, transgenic seeds were sown directly on MS agar plates with 30µM dex (stock solution prepared in ethanol) or equivalent amount of ethanol for mock treatment. Photon flux density was set at 50 µmol 2 min 1 s. Seedlings were harvested after 6 days for metabolites or RNA extraction.

5.2.2. Metabolome profiling

5.2.2.1. Metabolite extraction

Metabolite extraction for non-targeted metabolic profiling was performed as described in the protocol (Rai et al., 2013). Briefly, whole plants were snap frozen by liquid N2 and homogenized using mortar and pestle. 100mg of homogenized samples were transferred to a 1.5ml eppendrof tube and 0.5ml of ice cold 80% methanol was added immediately, centrifuged twice at 4°C, 13,500 rpm for 20 min and the supernatant collected were used to analyse through directly to mass spectrometry. Metabolite extracts for four biological replicates of each line, with each replicate injected thrice as technical replicates were used in Q-TOF, while Orbitrap had three biological replicates of each line, with three replicates. The sample run were not randomized, however, the entire extraction and analysis was performed in one batch.

5.2.2.2. LCMS analysis

We used two MS platforms to broaden coverage of ionized metabolites from samples. In the first platform, Agilent 1290 UHPLC system was used in-line with Agilent QTOF 6540. Six microliters of each metabolite extract sample was chromatographed on a Zorbax Eclipse Plus-C18 column (10cm length, 2.1 cm diameter, 1.8 µm particle size) with column temperature fixed at 50°C. Flow rate was maintained at 0.3 ml/min in an 18 min run with a gradient mobile phase: A) 0.1% FA in water; B) 0.1% FA in ACN (t = 0–0.5 min, B = 10%; t = 0.5–12 min, B = 100%; t = 12–15 min, B = 100%; t = 15–15.1 min, B = 10%; t = 15.1–18 min, B = 10%). The eluent was introduced directly into the mass spectrometer by electrospray, and during the whole period of injection samples were maintained at 4°C. Untargeted mass profiling was performed with Agilent Q-TOF 6540 using ESI probe in positive mode of ionization. Parameter for MS were: drying gas temperature- 350ºC with 10L/min (nitrogen) flow rate, sheath gas temperature -400ºC with 12L/min (nitrogen) flow rate, capillary voltage- 4000V, data acquisition in centroid mode, resolution- 30,000,

acquisition rate- 4 spectra/s, mass range- 50-1200 m/z. The system was controlled by Mass-Hunter Data Acquisition Software (Agilent Technologies, Santa Clara, CA, USA).

In the second platform, we used Acquity UHPLC system (Waters) in-line with LTQ-Orbitrap Velos (Thermo Fisher Scientific). Acquity UPLC BEH C18 column (10cm length, 2.1cm diameter, 1.7 µm particle size) was used with column temperature fixed at 50°C, injection volume was10 µl with sample vials maintained at 7°C. Solvents used were the same as in first approach with solvent flow rate of 0.4ml/min and a linear gradient of 13 min. The following gradient was used: 5% B for 0.5 min, 5–100% B in 9 min, holding at 100% B for 2 min and re-equilibration at 5% B for 2 min. MS profiling was performed in positive mode using in-line LTQ-Orbitrap Velos equipped with a heated electrospray probe (H-ESI II). The system was controlled by Xcalibur 2.2 (Thermo Fisher Scientific). ESI and MS parameters used for Orbitrap were: spray voltage 5.0 kV, sheath gas and auxiliary nitrogen pressures 50 and 10 arbitrary units, respectively, capillary and heater temperatures 300 and 350℃, respectively, tube lens voltages 110 V. The resolution was set at 60,000 for full MS scan (50-1200 m/z) with acquisition rate of 3 scan/s. Data were acquired in profile mode with external calibration.

Data-dependent MS/MS was performed using Agilent Q-TOF 6540 with ESI probe in positive mode of ionization. UHPLC system with column was setup in-line with mass spectrometer, with an 18 min long separation method same as described above for untargeted metabolic profiling. Parameters used were: drying gas temperature at 350ºC with 10L/min (nitrogen) flow rate, sheath gas temperature at 400ºC with 12L/min (nitrogen) flow rate, capillary voltage at 4000V, nozzle voltage 1500V, skimmer voltage 65.0V, fragmentor voltage 150V and octopole RFPeak voltage 750V. Parameters for precursor selections were - fixed collision energy at 50eV, max precursors per cycle at 10, threshold (absolute) at 100cps, active exclusion enabled with exclusion after 2 occurrences and release of active exclusion after 30 s.

Data acquisition was performed in centroid mode at the resolution of 30,000 with MS scan rate set at 8 spectra/s and MS/MS scan rate set at 2 spectra/s.

5.2.3. Metabolomics data analysis

Raw data files from both Q-TOF (.d files) and Orbitrap (.RAW files) were converted into mzXML format using msconvert of the ProteoWizard suite (Kessner et al., 2008). Using parameters defined for Q-TOF (method = 'centWave', ppm = 30, peakwidth = c(10,60), prefilter = c(0,0); peak grouping: bw = 5, minfrac = 0.3, mzwid = 0.025; retention time correction algorithm: 'obiwarp') and Orbitrap (method = 'centWave', ppm = 2.5, peakwidth = c(10,60), prefilter = c(3,5000); peak grouping: bw = 5, minfrac = 0.3, mzwid = 0.015; retention time correction algorithm: 'obiwarp') a total of 8,734 (QTOF), 5,969 (Orbitap) features were extracted using XCMS package(Patti et al., 2012; Smith et al., 2006) (version 1.38) in R software (R Core Team, 2014). Exploratory data analysis was performed using R and datPAV. Metabolite profiles were analysed using XCMS and m/z values of putative metabolites were checked again using Agilent's Mass Hunter Software. TICs for metabolite profiles obtained through Q-TOF showed a distinct divergence in the metabolite profiles for the biological replicate 1 for both *tt8* and wild-type (Figure 5.1).

Raw data was then log-transformed and quantile normalized. Similar to the TICs, for samples analysed in Q-TOF, we observed that biological replicate 1 from each genotype, *tt8* and wild-type formed a separate cluster (Figure 5.1). We then analysed the remaining three biological replicates of Q-TOF to ensure that they were statistically robust. All three biological replicates analysed using Orbitrap showed good reproducibility. This was visualized by performing Principal Coordinates Analysis (PCoA with Bray-Curtis dissimilarity) in PRIMER6 (Figure 5.2) on the replicates of Q-TOF and Orbitrap. Figure 5.2A shows the outlier biological replicate 1 of both *tt8* and wild-type in Q-TOF. After removing the outlier from Q-TOF samples, the analysis revealed that TT8 loss resulted in a difference in the metabolite profiles

between wild-type and *tt8* to be over 47% in Q-TOF (Figure 5.2B) and 23% in Orbitrap (Figure 5.2C) . We concur that any non-biological source of variation could have influenced this separation, thus, for further analysis, biological replicate 1 from both sets in Q-TOF were removed. As the remaining 3 biological replicates for each condition in Q-TOF showed strong analytical similarity, multivariate analysis was then performed on a total of 36 samples (3 biological replicates, 3 technical replicates for each genotype) from each MS. It is important to note that the analysis in the MS platforms were experimental replicates, meaning that the replicates used in each platform were from different batches of plants. Thus, biological replicate 1 of Q-TOF was not related to biological replicate 1 of Orbitrap. Furthermore, samples from Q-TOF could be used even after removing the outlier, only because these samples still had remaining three biological replicates in genotype, thus, they provided the required statistical robustness. Therefore, when designing such experiments, it is always useful to have a minimum of four biological replicates as even if one replicate is an outlier, the analysis can still be performed using the remaining three. Batch effect correction procedure described in Chapter 3, was not required in this scenario as all replicates were run in a single batch.

Mann-Whitney test with *p*-values adjusted (p-value < 0.05) using Benjamini-Hochberg false discovery rates was performed on the two groups yielding 1,259 and 611 differential features with p-value < 0.05 and fold change >2 on the Q-TOF and Orbitrap, respectively.

5.2.4. Metabolite identification

A total of 101 differential features were then recursively mapped (with mass tolerance < 5 and 10 ppm) onto KEGG Arabidopsis database (2014 version) (Kanehisa and Goto, 2000; Kanehisa et al., 2014) using PCDL manager (Agilent). Putative metabolites were predicted for each m/z feature following the same strategy described in Chapter 4. The metabolites whose mass difference (less than 10 ppm) between the

given m/z was minimum was selected to represent that feature. A list of all m/z features with their retention time, predicted metabolites and the ppm error is provided in Dataset 3. Of these, 46 metabolites were validated using data-dependent MS/MS-based fragmentation, with the fragmentation pattern matched against metabolites in the MassBank database (Horai et al., 2010).

5.2.5. Microarray-based expression profiling and analysis.

Total RNA was extracted using Omega plant RNA extraction kit (Omega BIO-TEK) following manufacturer's protocol. RNA was purified using an RNeasy purification kit (Qiagen). Total RNA was sent to Genomax technologies, Singapore for microarray analysis including RNA quality control using Bioanalyser (Agilent), reverse transcription and labelling, single color hybridization onto Agilent Arabidopsis 4×44k Array and preliminary data analysis. The experiments were performed for two biological replicates each for Ws and *tt8* lines, with two technical replicates (of each biological replicate) of each line being hybridized on single slide. Analysis was performed using Genespring 12.0 (Agilent) software.

All probes were normalized using percentile shift to 75% and were baseline corrected to the median. Data were filtered by taking 20^{th} to 100^{th} percentile of signal value followed by filtering data based on coefficient of variance with < 20% as cut off. Similar to the differences in metabolic profiles of Ws and *tt8* lines, their respective gene expression profiles also exhibited a clear difference (Figure 5.3). Filtered data were used for statistical analysis to identify differentially expressed genes using unpaired t-test with asymptotic *p*-value computation with Benjamini-Hochberg-based FDR (False discovery Rate) correction. 1,284 differentially expressed genes were selected based on p-value < 0.05 and absolute fold changes > 2 with respect to control.

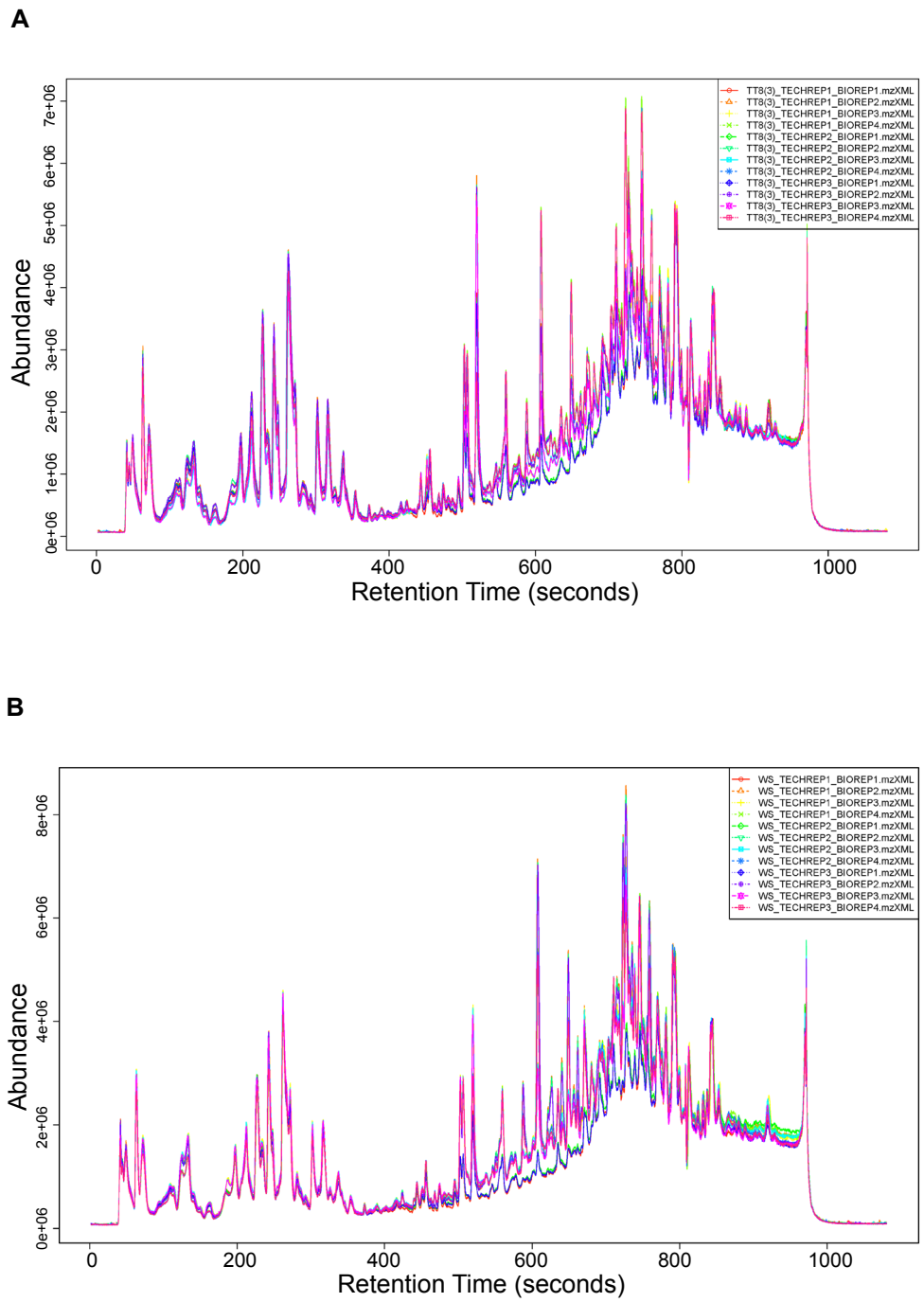


Figure 5.1. Raw Total Ion Chromatograms showing four biological replicates with each having 3 analytical replicates of (A) *tt8* (B) wild-type. The TICs of the biological replicate 1 of both *tt8* and wild-type show distinct profiles compared to the other replicates.

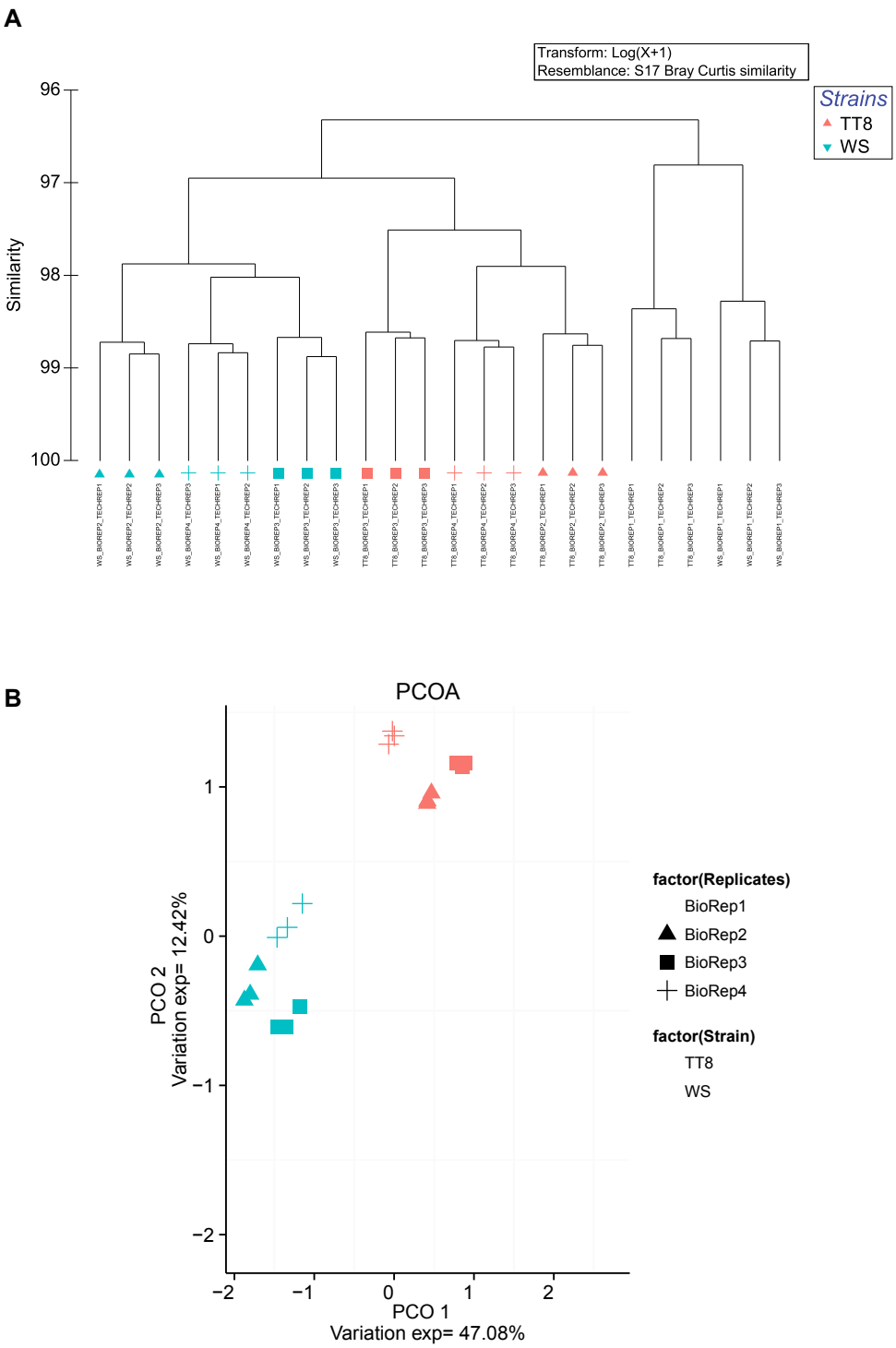


Figure 5.2A. Exploratory data analysis depicting similarity in metabolic profiles within biological replicates and differences between genotypes in Q-TOF. The first row shows the hierarchical clustering between samples and the second row ordination plots. Orange and blue shaded points represent TT8 and wild-type, respectively.

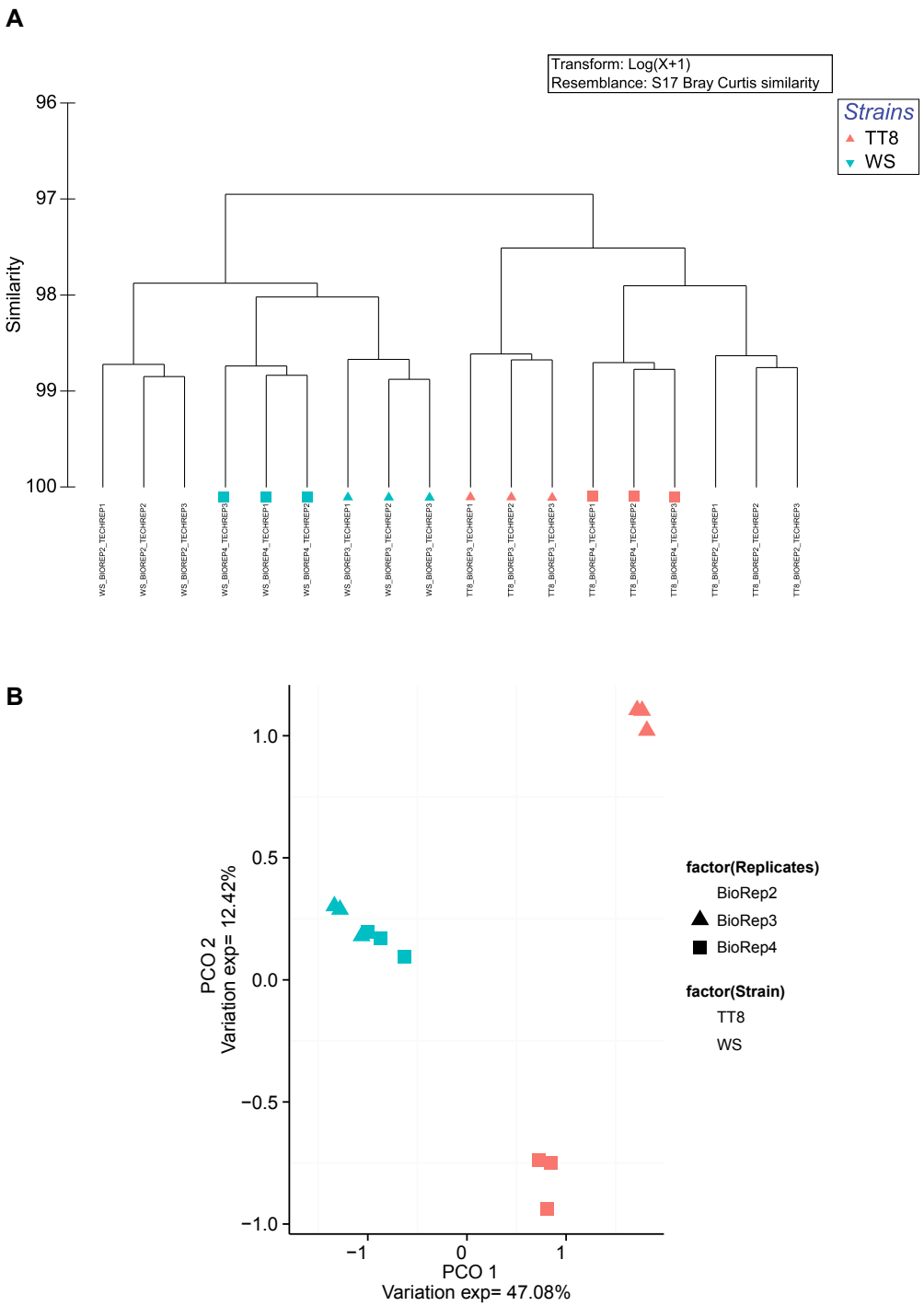


Figure 5.2B. Exploratory data analysis depicting similarity in metabolic profiles within biological replicates and differences between genotypes in Q-TOF after removing one outlier biological replicate from each genotype.

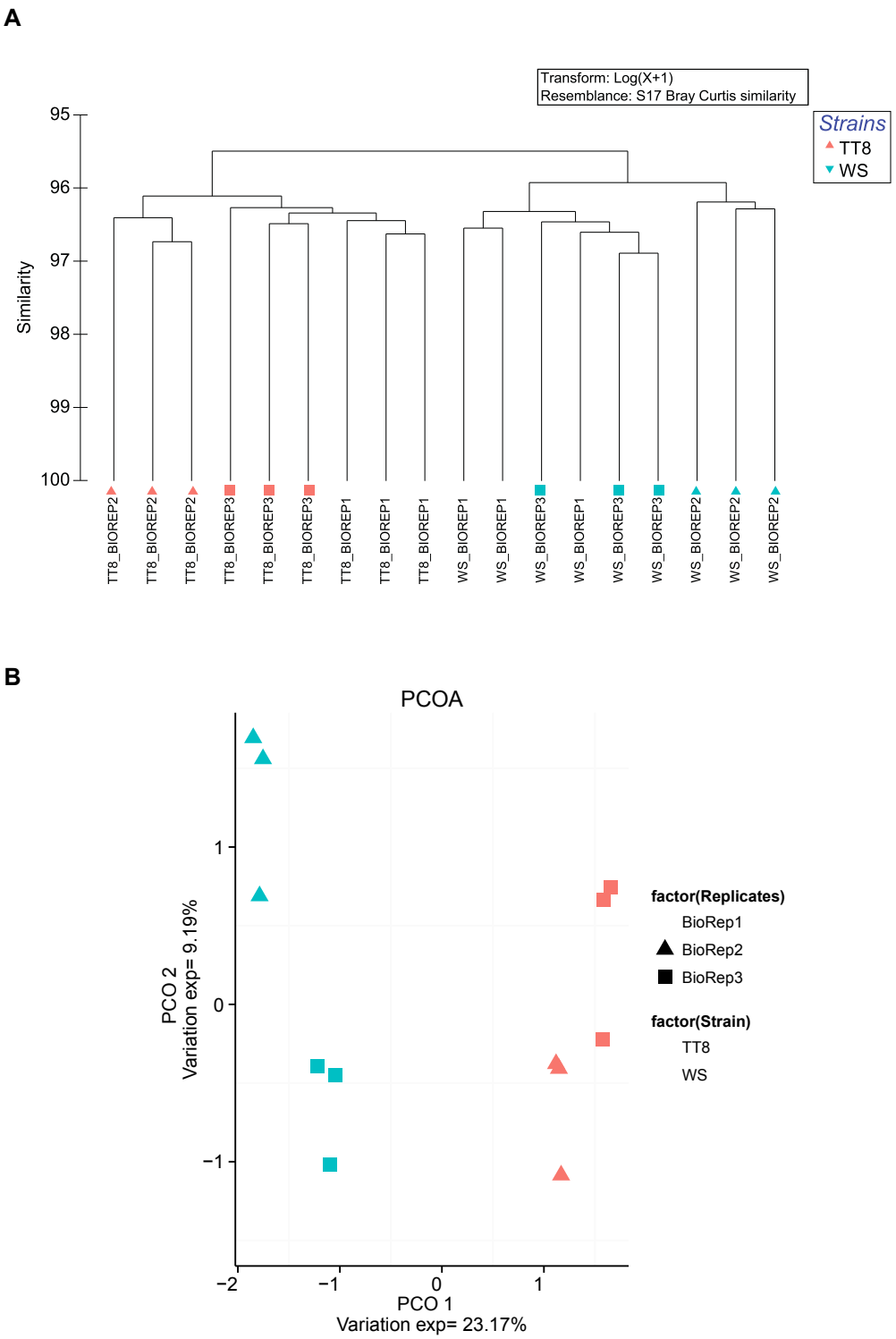


Figure 5.2C. Exploratory data analysis depicting similarity in metabolic profiles within biological replicates and differences between genotypes in Orbitrap.

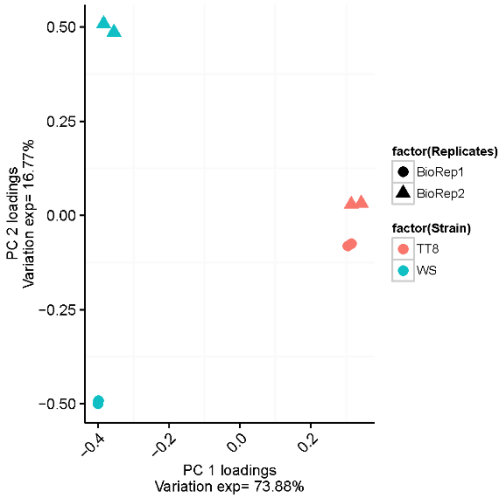


Figure 5.3. PCA using *princomp* function in R, shows similar trends between technical and biological replicates of wild-type and *tt8*.

Gene Set Enrichment Analysis (GSEA) enrichment analysis was performed using the web-based tool PlantGSEA (Yi et al., 2013). Subpathway enrichment was performed on differential genes and metabolites in R using iSubpathwayMiner package (Li et al., 2013). Microarray data has been uploaded onto the GEO database [under embargo until publication]. All *p*-values mentioned in study are FDR corrected.

5.2.6. Promoter regulatory network

Differential genes were analysed using network-guided guilt-by-association approach in AraNet (Vandepoele et al., 2009) to identify target gene group (TGG). Among the 1,284 differential genes, those that have minimal set of information in The Arabidopsis Information Resource (TAIR, https://www.arabidopsis.org/) were further selected. This removed genes (306 genes) which did not contain any annotation in *Arabidopsis* or even functional/sequence similarity to other organisms.

Amadeus (Linhart et al., 2008) platform was used for de novo motif discovery, we scanned for enriched motifs in the 968 differential genes with six different combinations, defined as: Promoter length: 1500, 1000 and 500 bps upstream of transcription start site, and, Motif length: 8-mer's and 10-mer's motifs. An adjacency matrix, which showed the number of shared motifs between any two genes, revealed

the maximum motif similarity between the 23 CAZY genes and other differential genes to be 19 motifs. Using an approach similar to (Vandepoele et al., 2009), we tested different constraints, such as, genes sharing at least 95%, 90%, 75% and 50% of the maximum motif similarity, these gene clusters contained 15 (sharing 18 motifs), 27 (sharing 17 motifs), 170 (sharing 14 motifs) and 596 (sharing 10 motifs) genes, respectively. To limit false positives but still have the potential to uncover new interactions, genes were selected based on a highly stringent condition i.e., those sharing a minimum of 14 motifs (at least 75% motif similarity). The promoter regulatory network was constructed using this list in Cytoscape (Shannon et al., 2003). PScan (Zambelli et al., 2009) was then used to identify enriched plant promoter motifs in 170 genes from the glycosylation regulatory network.

5.2.7. Growth assays for stress tolerance

For stress tolerance assays, *Arabidopsis* seeds were sown onto MS agar plate [1x MS media, 20 g/litre sucrose, and 6 g/litre phytoagar (pH 5.7)] with or without different stress conditions [Salt stress (100mM, 150mM, 200mM), Methyl jasmonate (MeJA) (50µM, 100µM, 150µM), Mannitol (250mM, 500mM), Deoxynivalenol (DON) (5ppm, 10ppm, 20ppm) and Abscisic (ABA) (1µM, 5µM, 10µM)], stratified for 4 days, and then incubated at 22℃ for 1 week. Ws and tt8 lines were sown directly on MS agar plates with or without salt, mannitol, ABA, MeJA and DON. *TT8:GR* seeds were sown on MS agar plates with all selected stress conditions in addition with 30 µM dex or equivalent volume of ethanol for TT8 induction or mock treatment, respectively. MeJA is a part of innate immune response for biotic stress(Navarro et al., 2008). DON is a fungal toxin that is inactivated by glycosylation to impart resistance against *Fusarium graminearum* (Poppenberger et al., 2003), and was used as a representative of biotic stress. For germination test under different stress conditions, 28 seeds for each transgenic line were sown on three petri plates and percentage germination was

measured. The germination of seeds was scored as positive when the tip of the radical had fully penetrated the seed coat.

5.2.8. ChIP assay and Quantitative real time PCR

ChIP assay using plant materials (0.8g) from six-day-old seedlings of *TT8:GR* lines, with or without dex treatment were performed as described previously (Kaufmann et al., 2009) with minor modifications. Promoters of target genes were scanned for putative motifs and primers were designed to cover entire promoter. Quantitative real-time PCR was performed in triplicates using Bio-Rad CFX384 real time PCR system (Bio-Rad) using Maxima SYBR Green qPCR mix (Thermo Fisher Scientific). The comparative Ct method (ΔΔCt) for relative quantification of gene expression was used for calculating the fold change using TUB2 as endogenous control.

Above experiments were performed by colleagues from Metabolite Biology Lab, NUS - Dr. Amit Rai, Megha and Lim Boon Kiat. I performed systems level analysis of genomics, transcriptomics and metabolomics datasets and developed and implemented the integrative omics approach mainly using scripts in R and computational tools described in Section 5.2 (Materials and methods). I generated Figures 5.1 to 5.12, except for parts of Figure 5.11. For Figure 5.11, images of *Arabidopsis* plants were produced by Dr. Amit Rai.

5.3. Results and discussion

5.3.1. Integrative omics approach to identify direct targets and the regulatory network mediated by a putative glycosylation regulator

This integrative omics strategy (Figure 5.4) uses three complementary biological data measurements, namely, metabolomics, transcriptomics and transcriptome dependent in-silico genomics. Using non-targeted mass spectrometry-based metabolomics, we aimed to capture a comprehensive portion of the *Arabidopsis* metabolome affected by TT8 loss. We then developed a targeted data-dependent

MS/MS approach to determine glycosylated and non-glycosylated metabolites (shaded grey on the left in Figure 5.4). For determining gene expression levels, we performed microarray to identify differential genes in *tt8* w.r.t its wild-type control. PlantGSEA tool was used to identify enriched functional categories. A metabolite set enrichment analysis using iSubpathwayMiner was performed.

Overrepresentation approaches (ORA) such as Fisher's exact test or a hypergeometric test, are usually performed to understand the significance of entities, such as metabolites and genes that have been mapped onto pathways. These tests compare the number of differential entities that have been mapped onto a pathway, against the probability of such mappings occurring purely by chance. A pathway is then deemed to be enriched (usually in the form of *p*-values, such as those seen in Section 4.3.2.1), if the number of entities that have been mapped onto it is significantly different from those expected by chance alone.

A number of such methods, including ORA and GSEA are widely used for identifying pathways whose entities have significant changes in their levels. However, the predictive power generated by such methods are limited, as they focus only on genes or metabolites and do not factor in simultaneous changes to levels of both of genes and metabolites. Furthermore, standard correlation-based methods are unable to distinguish between direct and indirect associations when comparing gene and metabolite abundances (due to confounding variables such as batch effects). Thus, an integrative pathway analysis strategy that can utilize the combined power of both genes and metabolites provides an excellent solution to interpret the underlying biological phenomena.

IMPaLA (Kamburov et al., 2011) and iSubpathwayMiner (Li et al., 2013) are two such tools that provide the option to use both gene and metabolite information for identifying significant pathways. The enrichment statistics for metabolic pathways in IMPaLA is generated by multiplying the *p*-values for genes and metabolites (obtained separately for each entity). This approach is certainly advantageous than analysing

genes and metabolites separately. However, IMPaLA does not consider the topology or the presence of hub-nodes in its analysis. These are taken into account in iSubpathwayMiner, which analyses the structure of the metabolic pathway.

If a biological entity has significant changes to its levels, then it might also induce corresponding changes to the levels of neighbouring entities in the metabolic pathway. This scenario is factored into the enrichment analysis in iSubpathwayMiner as it integrates information from genes, metabolites, and their relative positions in the sub pathway along with their metabolic cascade regions (Li et al., 2013). Furthermore, in a large metabolic pathway, only core part of the pathway, depending on the biological phenomena, might have an effect. Thus, iSubpathwayMiner performs integrative sub pathway enrichment analysis that can result in a higher predictive power for understanding metabolic responses. Specifically, iSubpathwayMiner analyses lenient distance similarities of key nodes within the metabolic pathway structure to identify important metabolic cascade sub-pathway regions. The enrichment scores for such analyses are then generated using a hypergeometric test. Furthermore, iSubpathwayMiner also enables direct import of the *Arabidopsis* metabolic network from KEGG into R. Therefore, in order to fully utilize the (i) strength of the combined gene and metabolite analysis, (ii) statistical power generated by analysing enriched sub-pathways, (iii) most updated Arabidopsis metabolic network, we selected iSubpathwayMiner for performing integrative analysis.

To determine the effect of TT8 loss on the metabolic network architecture we analysed genes and metabolites together in a sub-pathway enrichment analysis using $SubpathwayGM$ function in 'iSubpathwayMiner' package (Li et al., 2013) in R. The output from this analysis was then mapped onto AraCyc version 8.0 (Mueller et al., 2003) pathways.

Gene expression is affected by the combinatorial arrangement and sharing of motifs in the promoter region. These motifs act as transcription factor binding sites.

Hence, the presence or absence of certain motifs plays a significant role in recruiting transcription factors and initiation of transcription, thereby affecting gene expression levels. The information encoded in these promoter sequences such as type and number of motifs can help understand genotype-phenotype relationships by providing a mechanistic interpretation of the genetic basis that govern metabolite regulation (Levo and Segal, 2014). Thus, we analysed the promoter sequences of differentially expressed genes. We selected only those genes that had evidence (experimental or computational)-based on annotations in AraNet (Lee et al., 2010) to understand promoter relationships (shaded grey on the right in Figure 5.4). Furthermore, to identify conserved regions between differential genes, we performed de novo motif identification implemented using Amadeus (Linhart et al., 2008). This approach allowed us to identify motifs which had not yet been encoded to a transcription factor. Then, a constraint-based promoter regulatory network was developed based on shared enriched motifs between these genes. This subset of genes affected by TT8 loss formed the glycosylation regulatory network. The pathways associated with these genes were cross-referenced with the AraCyc metabolic pathways where the metabolites had earlier been mapped. This metabolic pathway level analysis helped us identify common targets affected by TT8. Taken together, by combining high-throughput metabolomics, genomics and transcriptomics data, we generated testable hypothesis on the role of TT8 in *Arabidopsis*. The hypothesis was then validated using TT8 loss-of-function and inducible overexpression lines.

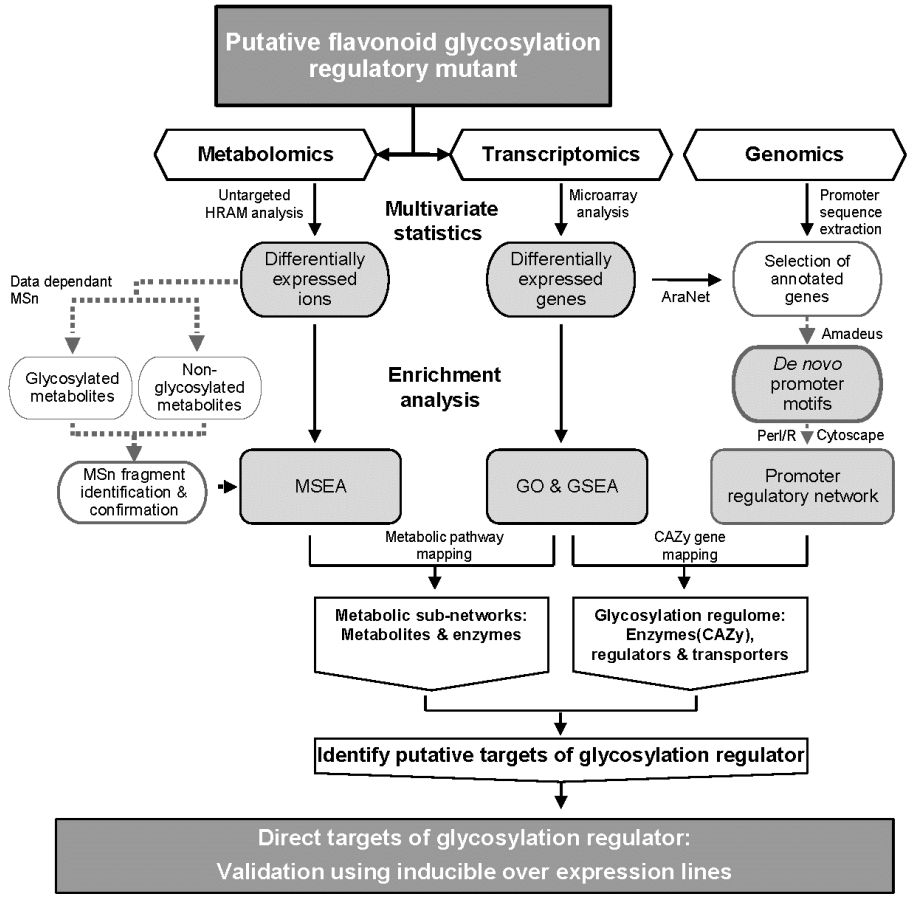


Figure 5.4. Integrative omics approach to identify direct targets of a flavonoid glycosylation regulator. Differential metabolites and genes were identified by comparing the metabolic and transcriptome profiles of *tt8* and wild-type. Glycosylated metabolites were identified using a targeted MS/MS approach. Promoter networks analysis was performed using differential genes. Enriched metabolic sub-networks in KEGG *Arabidopsis* metabolic network were identified by the combination of differential metabolite and genes

5.3.2. TT8 loss affects glycosylation of flavonoids and nucleotides

To determine the effects of TT8 loss on the metabolome, we performed untargeted metabolic profiling. As glycosylation is dynamic during the early growth stages, we chose to perform the experiments using 6-day old seedling metabolite extracts from *tt8* and wild-type (Ws background). Two high resolution mass spectrometry platforms were used to improve the coverage of differential metabolites. Figure 5.5 illustrates the metabolites detected using the two mass spectrometers on a mass-by-charge versus retention time plot.

The x-axis represent the retention time, while the y-axis represents the mass-by-charge ratio of the putative metabolites detected using database searches. From this figure, we observed that analytical methods were complementary rather than redundant as it covers unique regions of the metabolome. For example, the Q-TOF-based mass spectrometer (shaded green in Figure 5.5) covers the entire elution time, whereas Orbitrap which had a shorter run time (shaded red in Figure 5.5) was enriched towards the initial parts of the chromatography run. The differences in the types of metabolites being measured in the two MS arise mainly due to the differences in the chromatographic run (see Section 5.2.3). While Orbitrap had a LC run time for 15 mins, Q-TOF was run for 18 mins. Furthermore, there was also differences in the column make. These differences, plus the fact that Orbitrap slightly favours metabolites at lower mass range, eluting first, compared to Q-TOF, resulted in both the MS capturing different m/z ranges. Importantly, from Figure 5.2 we also know that both the mass spectrometers captured significant biological differences between wild-type and mutant. Thus, by using two mass spectrometers, we were able to obtain comprehensive coverage of the changes in the *Arabidopsis* metabolome as a result of TT8 loss.

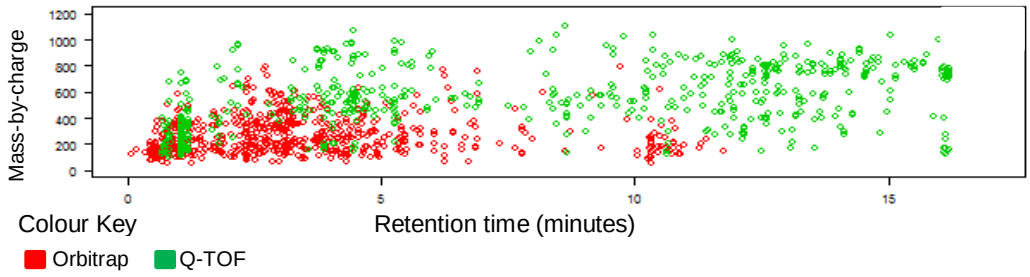


Figure 5.5. Comprehensive coverage of the perturbed metabolome. Two mass spectrometers were used for profiling complementary regions of the Arabidopsis metabolome. The points indicate the detected metabolite's location based on mass (y axis) and retention time (x axis).

From the differential metabolite analysis, we observed that flavonoids including flavonols, flavanones and anthocyanins were the largest class of metabolites affected by the loss of TT8 (Table 5.1). Similar to previous reports, quercetin and kaempferol aglycones were up-regulated in TT8 loss-of-function line (Nesi et al., 2000;

Pelletier et al., 1999). Our comprehensive metabolomics approach enabled the detection of thirteen additional flavonoid aglycones, whose glycosylated forms have been not yet been shown to be affected in *tt8* (Table 5.1). Interestingly, TT8 loss had differential effects on the glycosylation of aglycones mainly based upon the metabolite class. For example in Table 5.1A, the first group of metabolites had both aglycones and their corresponding glycosylated forms being affected in *tt8*, whereas, only the glycosylated forms were affected in the second group.

Another interesting observation, is that the trends in the glycosylated metabolites were mainly dependent on the nature of the attached sugar moiety. For example, most of the glucoside conjugated quercetin and kaempferol were down-regulated, while the rhamnoside conjugated forms were mostly up-regulated in *tt8*.

We also observed a significant number of nucleotides and their glycosylated forms being affected in *tt8* (Table 5.1B). The sugar conjugated nucleotides numbering around 13 out of 20 were mostly down-regulated whereas only 3 out of 10 aglycone nucleotides were down-regulated. Taken together, we observed that sugar conjugated forms of both primary and secondary metabolites were majorly affected by TT8 loss.

Table 5.1. Metabolites affected in *tt8*. Relative log2-fold levels of differential putative glycosylated metabolites and their aglycones belonging to (A) flavonoids, and (B) nucleotides in both TT8 loss and induced overexpression lines are shown. * indicates metabolite was confirmed using MS/MS. (shown in the next page)

A

Aglycone	*Glycone*	*tt8*	*TT8:GR*
*cyanidin**		2.4	-2.2
	cyanidin 3,5-diglucoside*	-2.4	2.9
	cyanidin 3-O-beta-D-glucoside*	-3	2
	cyanidin 3-O-sophoroside*	-2.4	1.3
*kaempferol**		9.6	-1.4
	kaempferol 3-sophorotrioside*	-5.2	3.1
	kaempferol 3-O-glucoside*	-4.6	2.1
	kaempferol 7-O-glucoside*	-4.2	1.8
	kaempferol 3,7-O-diglucoside*	-8.8	2
	kaempferol 3-O-glucosyl-(1-2)-glucoside*	1.6	-1.1
	kaempferol 3-O-glucosylgalactoside*	7	-2.3
	kaempferol 3-O-gentiobioside-7-O-rhamnoside*	2.7	-2.3
	kaempferol 3-rhamnoside*	2.8	-1.5
	kaempferol 3-rhamnoside-7-rhamnoside*	6.9	1.4
*quercetin**		9.5	-6.3
	quercetin 3,7-O-diglucoside*	-3.8	2.6
	quercetin 3-O-[xylosyl-(1-2)-glucoside]*	-7.8	3.1
	quercetin 3-O-glucoside*	-9.7	8.8
	quercetin 3-O-glucosyl-(1-2)-glucoside*	0.8	-2.1
	quercetin 3-O-rhamnoside*	9	-2
	quercetin 3-O-rhamnoside-7-O-glucoside*	-8.8	2.3
	quercetin 3-rhamnoside-7-rhamnoside*	9.9	-3.2
*luteolin**		12.9	-1.3
	luteolin 7-O-glucoside*	-10	0.8
*vitexin**		15.2	N.D
	vitexin 2-O-D-glucoside*	-8.6	N.D
*isovitexin**		-1.9	2.1
	isovitexin 2''-O-beta-D-glucoside*	-0.8	-1.9
salicylate		-4.6	1.8
	salicylate -D-glucose ester	7.5	-2.2
	salicylate 2-O-D-glucoside	-0.5	-1.1
apigenin		3.4	-1.7
	apigenin 7-O-beta-glucoside*	-9.5	N.D
	apigenin 7-O-neohesperidoside*	-1.7	N.D
	genistein 7-O-glucoside*	-0.8	2.1
	hesperetin 7-O-glucoside*	-7.6	3.3
	pelargonidin-3,5-diglucoside*	5	-3.8
	naringenin 7-O-glucoside*	8.6	-2.5
	bracteatin 6-O-glucoside*	0.7	-1.3
	coniferaldehyde glucoside	-1.3	1.4
	dihydroconiferyl alcohol glucoside	7.14	14.1
eriodictyol		8.3	-0.3
2',4,4',6'-tetrahydroxychalcone		8.3	-0.3
4-coumarate		-2.3	8.8
caffeoylshikimate		8.1	-3
camalexin		-6.4	13
chorismate		-6.5	10.4
dihydroconiferyl alcohol		1.1	-2.7
dihydromyricetin		27.4	-0.3

B

Aglycone	*Glycone*	*tt8*	*TT8:GR*
*UDP**		6.8	N.D
	UDP-2,3-bis(3-hydroxytetradecanoyl)glucosamine*	8.6	-2.3
	UDP-4-dehydro-6-deoxy-D-glucose*	2.8	-1.9
	UDP-D-apiose	-7.7	1.2
	UDP-D-glucose	-6.6	1.8
	UDP-D-glucuronate	-6.9	3.1
	UDP-D-xylo-4-keto-hexuronate	-11.8	3.3
	UDP-D-xylose	-7.7	3.9
	UDP-galactose	-6.6	2.3
	UDP-L-arabinofuranose*	8.5	-3.1
	UDP-L-arabinose	-7.7	3.5
	UDP-N-acetyl-D-glucosamine	-12	N.D
UMP		1.7	N.D
UTP		-10.8	3.4
dUMP		6.6	-1.7
dUTP		8.5	-2.1
dTDP		6.4	-3.3
dTTP		-3	3.1
	dTDP-alpha-L-rhamnose	2.8	N.D
	dTDP-D-glucose	-15.9	2.3
	TDP-rhamnose	-6.4	4
	dTDP-4-acetamido-4,6-dideoxy-D-galactose	2.7	-2.3
	dTDP-4-dehydro-6-deoxy-D-glucose	-9.2	7.5
*dGTP**		-3.9	1.1
	GDP-D-glucose*	-1.7	1.3
	GDP-L-fucose*	-15.9	2.3
	CDP-4-dehydro-3,6-dideoxy-D-glucose	9	-1.7
	CDP-4-dehydro-6-deoxy-D-glucose	7.5	-2.3
dCTP		N.D	-2.7
dGMP		N.D	6.3

5.3.3. TT8 loss affects genes associated with sugar metabolism and glycosylation

Microarray-based gene expression profiling was performed using *tt8* and wild-type seedlings to determine the effects of TT8 loss at transcript levels. As expected, the levels of TT8 were down-regulated in the mutant compared to the wild-type. Consistent with previous reports, targets of TT8 such as *BAN* and *DFR* were down-regulated by 1.3- and 1.67-fold, respectively, while homologs of TT8 namely, *EGL3* and *GL3*, were u*p*-regulated by 1.45- and 2.3-fold, respectively. The u*p*-regulation of homologs possibly explain the marginal but significant down-regulation of *BAN* and *DFR*.

Several biosynthetic genes from phenylpropanoid and flavonoid pathway showed down-regulation in their gene expression trends compared to wild-type. Similar results were also seen at their corresponding metabolite levels. As we observed a number of metabolites being glycosylated, we focused on the enzymes mediating glycosylation processes, specifically CAZy. Expectedly, 34 genes from the CAZy category were differentially expressed by more than 2-fold (Figure 5.6). We also observed around 10% of the total CAZy genes in *Arabidopsis* being significantly affected by more than 1.5-fold. The expression levels for CAZy genes associated with glycosylation of specific metabolites showed similar trends. For example, *UGT78D1*, regulates the glycosylation of aglycone forms of kaempferol by adding rhamnosides (Jones et al., 2003) was up-regulated by over 6.5-fold, showing the same trend as that of rhamnoside conjugated forms of kaempferol in *tt8* (Table 5.1A). Similarly, *UGT88A1*, which glucosylates quercetin was down-regulated by 2.5-fold (Figure 5.6), and so were the levels of three forms of quercetin glucosides in *tt8* (Table 5.1A). These two genes are also associated with several sugar conjugation processes.

To identify coordinated response of genes and metabolites in enriched sub pathways in the KEGG Arabidopsis metabolite network, we mapped differential genes and metabolites onto their respective KEGG identifiers and used iSubpathwayMiner in R. The most enriched process was inositol phosphate metabolism which is related to signalling in defence responses in *Arabidopsis*. However, the above analysis is limited by the fact that KEGG maps does not provide exhaustive coverage of the glycosylated forms. Using our targeted MS/MS analysis we observed that a number of metabolites from the flavonoid pathway being glycosylated. Upon incorporating results from MS/MS analysis, flavonoid or nucleotide glycosylation networks had the maximum number of affected entities based mainly upon the number of differential metabolites (glycosylated and non-glycosylated) in those pathways (Table 5.1A). Thus, sub-networks associated with flavonoid metabolism were the most enriched sub-network affected by *tt8*. The next most enriched pathways were related to nucleotides and sugar

metabolism (Figure 5.8 and Supplementary Dataset 3). Sugar metabolism networks of fructose, mannose and pyruvate pathways were enriched together with glycolysis and TCA cycle (Supplementary Dataset 3). Branches of TCA cycle that lead to amino acid pathways such as arginine, proline, and cysteine and methionine metabolism were also enriched.

Sugar associated pathways, such as pentose phosphate pathways, and amino acid pathways, such as alanine and proline metabolism contribute precursors to the components of nucleotide network, namely purine and pyrimidines. Sub-networks with these pathway inputs were significantly enriched, suggesting indirect influence of enriched sugar metabolism network on nucleotide metabolism. These results suggest that TT8 loss affects nucleotide and flavonoid sugar conjugation process at both gene and metabolite levels

In addition to the sugar and nucleotide metabolic sub-network, a number of sugar transporters and sucrose synthases were also affected in *tt8*. Sugar transporters SUC6 and SUC7, sucrose synthases SUS2, SUS3 and SUS5 along with sugar binding proteins STP6 were down–regulated by more than 1.5-fold. Taken together, gene expression analysis shows disruption of sugar conjugation machinery in *tt8* with members of sugar metabolism and glycosylation machinery such as sucrose synthases, sugar binding proteins, sugar transporters, glycosyltransferases and hydrolases being affected, thus explaining the perturbation in the levels of metabolite glycosylation.

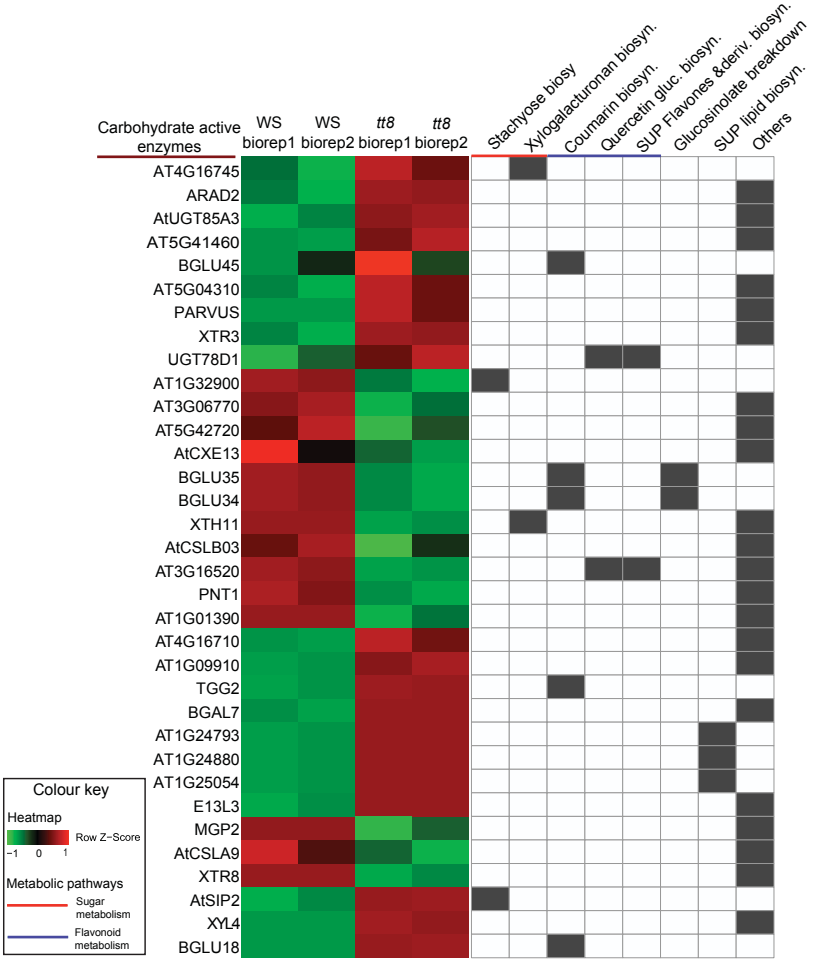


Figure 5.6. Differentially expressed CAZy genes. Heatmap on the left showing relative CAZy gene expression levels computed as z-scores using heatmap2 function in R. In the above plot, the each column represents two biological replicates each of tt8 and its wild-type. Each column is an average of two technical replicates. Each row represents the relative abundance of genes. In the matrix on the right, each column represents a pathway from AraCyc. The presence of a gene in that pathway is indicated in grey shade. Genes not classified under these pathways are annotated as others.

5.3.4. Abiotic-biotic stress response together with jasmonate and brassinosteroid biosynthesis network is enriched in *tt8*

GSEA revealed 14 biological processes encompassing around 30% of the differential genes to be significantly enriched. Interestingly, abiotic and biotic stress response, hormone response, metabolic pathways and developmental functions were the major categories that were enriched (Figure 5.7). From Figure 5.7, we observed that genes associated with biotic stress response showed overall up-regulation (relatively higher proportion of red in the figure), while abiotic stress categories showed a mixed trend, with hormone response and biosynthesis processes being effectively up-regulated in TT8 loss.

Furthermore, from our sub-network enrichment analysis, we observed jasmonate sub-networks to be significantly enriched (p-value < 0.01) at both gene and metabolite levels. Interestingly, the sugar or amino conjugated forms of jasmonates, were not affected in *tt8*. Expression levels of genes from jasmonic acid biosynthesis pathways, such as, *AOC1*, *AOC2*, *AOC3*, *LOX2*, *LOX3*, *HPL1*, *OPCL1*, *OPR3* and *ST2A* were all up-regulated by more than 2-fold in *tt8,* this was the same trend witnessed in their metabolic intermediates (Table 5.2). Genes associated with jasmonic acid responses, such as, key regulators - *JAZ1* to *JAZ12* (Thines et al., 2007), JMT, *and CEJ1* and expect *JAZ3* and *JAR1*, were all u*p*-regulated in *tt8*. Additionally, JAR1 regulates jasmonic acid-dependent processes, whereas *CEJ1* expression is co-regulated by jasmonic acid.

Previous reports (Chico et al., 2008), state that increase in jasmonic acid biosynthesis leads to u*p*-regulation of *MYC2*, a bHLH transcription factor, which has an important role in jasmonate signalling pathways. This was also observed in our study with *MYC2* being u*p*-regulated by 2.3-fold in response to increased jasmonic acid biosynthesis in *tt8*. TT8 loss not only affected jasmonic acid biosynthesis, but also extended upstream to its fatty acid precursor, such as α-linolenate. Our integrative

omics analysis also revealed enrichment of fatty acid biosynthesis and lipid pathways (Supplementary Dataset 3), which is further supported by recent biochemical evidence that shows that TT8 exudes inhibitory effects fatty acid biosynthesis (Chen et al., 2014).

Another major phytohormone pathway affected by *tt8* is brassinosteroid biosynthesis. Brassinosteroid sub-networks showed significant enrichment at the metabolite levels. Furthermore, the key genes of this pathway, namely *BR6OX1* and *AT4G27440* were down-regulated by more than 2-fold. Biosynthesis of brassinolide from campesterol in *Arabidopsis* occurs through two routes, one through (6α)-hydroxycampestanol and second through 6-deoxycathasterone (Figure 5.8). Although, both branches for brassinolide biosynthesis are active in wild-type (Noguchi et al., 2000), TT8 loss resulted in a switch, with (6α)-hydroxycampestanol branch being up-regulated while 6-deoxycathasterone being down-regulated, thus suggesting a preferred route of brassinolide biosynthesis in *tt8* (Figure 5.8).

SERK4 (AT2G13790), SERK5 (AT2G13800) which are part of the brassinosteroid signalling cascade were up-regulated by over 2-fold, while FLS2 was down-regulated by 1.5-fold change. These results suggest that the signalling mechanisms in brassinosteroid pathway were not majorly affected. However, there were significant changes observed in the genes and metabolites belonging to brassinosteroid biosynthesis pathway.

These results clearly indicate that TT8 loss affects two major phytohormone biosynthesis pathways in *Arabidopsis*. Furthermore, these pathways are also known to play an important role in stress response.

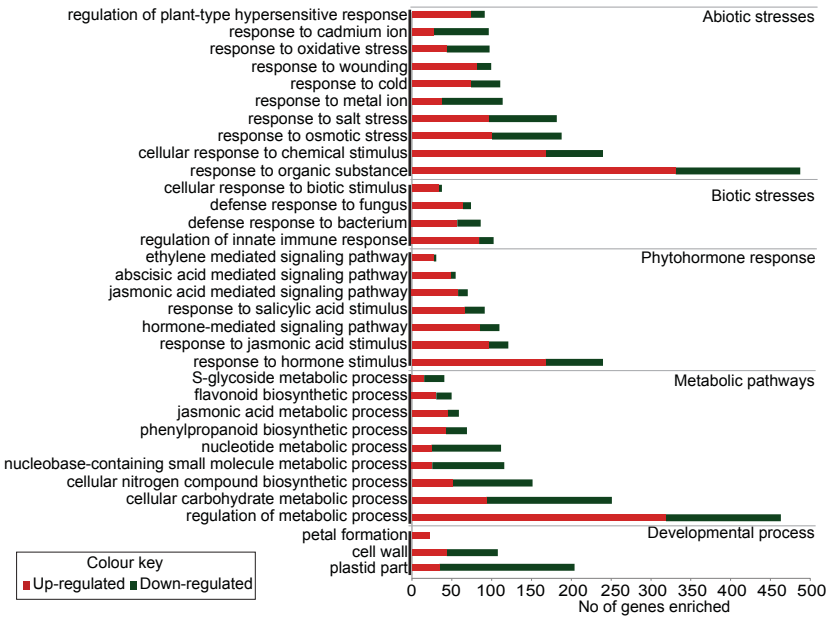


Figure 5.7. Gene Set Enrichment Analysis. Top 33 enriched GO categories (p-value < 0.05) with the number of differential genes affected in each category is shown here. Within each of the enriched GO category, the number of overexpressed and down-regulated genes are indicated in red and green colour, respectively.

Table 5.2. Metabolites belonging to phytohormone pathways affected in *tt8*. * indicates metabolite was confirmed using MS/MS.

Aglycone	***Glycone***	***tt8***	***TT8:GR***
	Jasmonic acid biosynthesis		
*volicitin **		-11.7	N.D
12-OPDA		-6.1	3.4
jasmonic acid		9.1	-1.8
7-Isomethyljasmonate		8.5	-3.2
17-hydroxylinolenic acid		8.9	-3.3
2(R)-HOT		11.3	-2.5
2(R)-HPOT		11.5	-2.6
methyl jasmonate		5.6	-3.7
OPC4-trans-2-enoyl-CoA		2.1	-1.1
OPC6-trans-2-enoyl-CoA		-6.5	-2
alpha-linolenate		2	1.5
(9Z,11E,15Z)-(13S)-hydroperoxyoctadeca-9,11,15-trienoate		7.1	1.8
	Brassinosteroid biosynthesis		
*campest-4-en-3-one **		-5.9	6.1
22 alpha-hydroxy-campest-4-en-3-one		2.2	-0.3
26-hydroxycastasterone		8.2	-5.4
*brassinolide **		1.3	-1.9
*castasterone **		8.6	-1.1
6 alpha-hydroxy-castasterone		-11.5	2.9
6-deoxocastasterone		-5.3	9.4
6-deoxotyphasterol		-6.5	6.9
26-hydroxybrassinolide		13.3	-3.4
	brassinolide-23-O-glucoside	-8.3	2.5
	castasterone-23-O-glucoside	-5.1	0.1

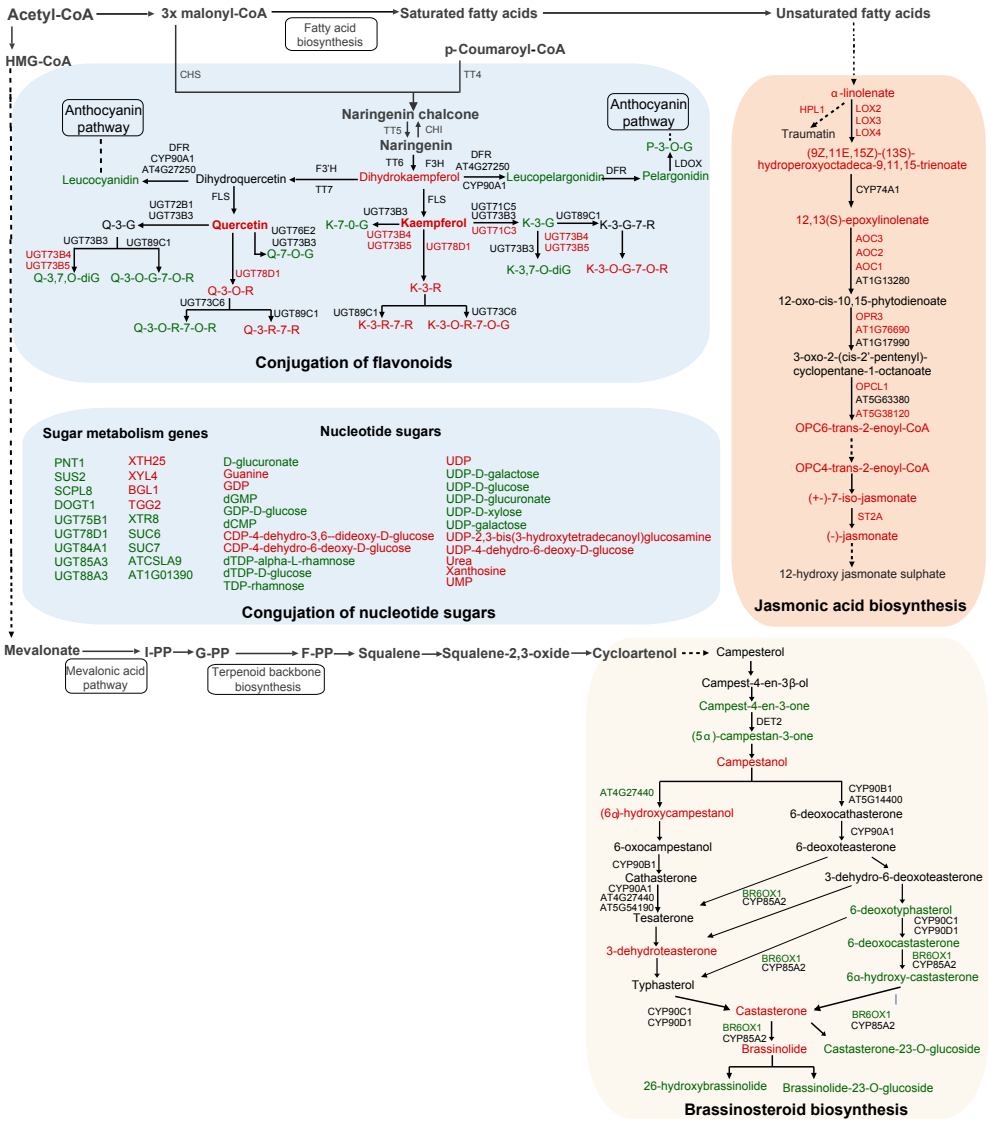


Figure 5.8. The top 3 enriched pathways (redrawn based on AraCyc and KEGG), namely biosynthesis of flavonoids and their conjugation, jasmonic acid and brassinosteroids and their associated differential metabolites and genes are shown. TT8 leads to increase in both metabolite and gene levels in jasmonic acid pathways, while brassinosteroid pathway shows a mixed trend in metabolite levels.

5.3.5. TT8 regulatory network links genes associated with carbohydrate active enzymes to innate immunity

We analysed the promoter sequences of differentially expressed genes to determine whether these genes shared common enriched motifs. Transcription Factor Binding Sites (TFBS) by controlling the timing and location of transcriptional activity, play an important role in regulating gene expression levels (Leister et al., 2011; Meier et al., 2008; Vandepoele et al., 2009; Vidal et al., 2013). Thus, genes sharing motifs might be under common regulatory mechanisms. We used a network-guided guilt-by-association approach to uncover relationships at a regulome level and analysed enriched 8-mer and 10-mer de novo promoter motifs (Linhart et al., 2008) to obtain shared motifs between differentially expressed genes. As described in the previous sections, glycosylation of flavonoids and nucleotides and CAZy associated genes were significantly affected in *tt8*. Furthermore, we also witnessed TT8 having significant effects on processes associated with stress response and hormone biosynthesis (Figure 5.7). Therefore, in order to identify whether these genes share common regulatory mechanisms, we constrained the promoter network to show only relationships between genes that were connected with CAZy genes.

We then developed an undirected network with nodes representing genes and edges representing the shared motifs (Figure 5.9). The width of the edges is proportional to the number of shared motifs between any two genes (nodes), with high similarity indicated by thicker edges. This network represents the genes in the TT8-glycosylation regulome. This glycosylation regulome consists of 18 CAZy genes that connect mainly with genes associated with stress response (13 genes) and phytohormone biosynthesis (13 genes). CAZy genes also share motifs with genes involved in sugar metabolism such as transporters (5 genes) and transferases, hydrolases, oxidases (9 genes) (Figure 5.9). Genes involved in initial brassinosteroid

biosynthesis and response, such as the cytochrome P450 genes (Figure 5.9) also shared motifs with jasmonate and CAZy genes.

Genes from the TT8-glycosylation regulome analysed using PScan (Zambelli et al., 2009) tool revealed bZI*P*-related binding sites, EmB*P*-1, MYB, and Squamosa, TFBS over-represented (Table 5.3). Interestingly, recent reports have suggested these TFBS to be enriched in the presence of regulators such as JAZ and a number of stress associated CAZy genes and sugar-related pathways (Kang et al., 2010; Kazan and Manners, 2012; Qi et al., 2011). These relationships suggest the possibility of a common regulatory mechanisms among CAZy genes and between CAZy and jasmonic acid-associated genes. The results from transcriptome and metabolite profiling also support the novel relationships identified in the network, such as those between CAZy and jasmonic acid biosynthesis associated genes. Taken together, our results suggest that glycosylation of metabolites and processes associated with stress response might be co-regulated in the TT8-glycosylation regulome.

Table 5.3. Enriched plant transcription factors of the TT8-glycosylation regulome. The co-regulated gene sets were assessed for the occurrence of 21 known plant transcription factor binding sites.

Transcription factor binding site	***500bp***	***1000bp***
Jaspar		
EmBP-1	1.00E-04	2.00E-05
HAT5	0.001	0.0001
squamosa	0.002	0.015
ATHB-5	0.016	0.001
PEND	0.018	0.003
bZIP910	0.021	0.164
AGL3	0.023	0.001
Jaspar-Fam		
MADS	4.00E-12	7.00E-07
Homeobox	1.00E-11	2.00E-11
Forkhead	4.00E-08	3.00E-05
bZIP(bZIP cEBP-like subclass)	2.00E-06	8.00E-04
bHLH(zip)	1.00E-05	2.00E-09
bZIP(bZIP CREB/G-box-like subclass)	0.028	8.00E-04
Transfac		
P$PIF3_02	8.00E-06	2.00E-05
P$PIF3_01	1.00E-04	3.00E-06
P$AGL3_01	3.00E-03	0.007
P$AGL3_02	0.01	0.007
P$SBF1_01	0.029	7.00E-05
P$ATHB5_01	0.04	0.001
P$DOF1_01	0.769	0.007
P$ATHB1_01	0.157	3.00E-04
P$PBF_01	0.173	0.009

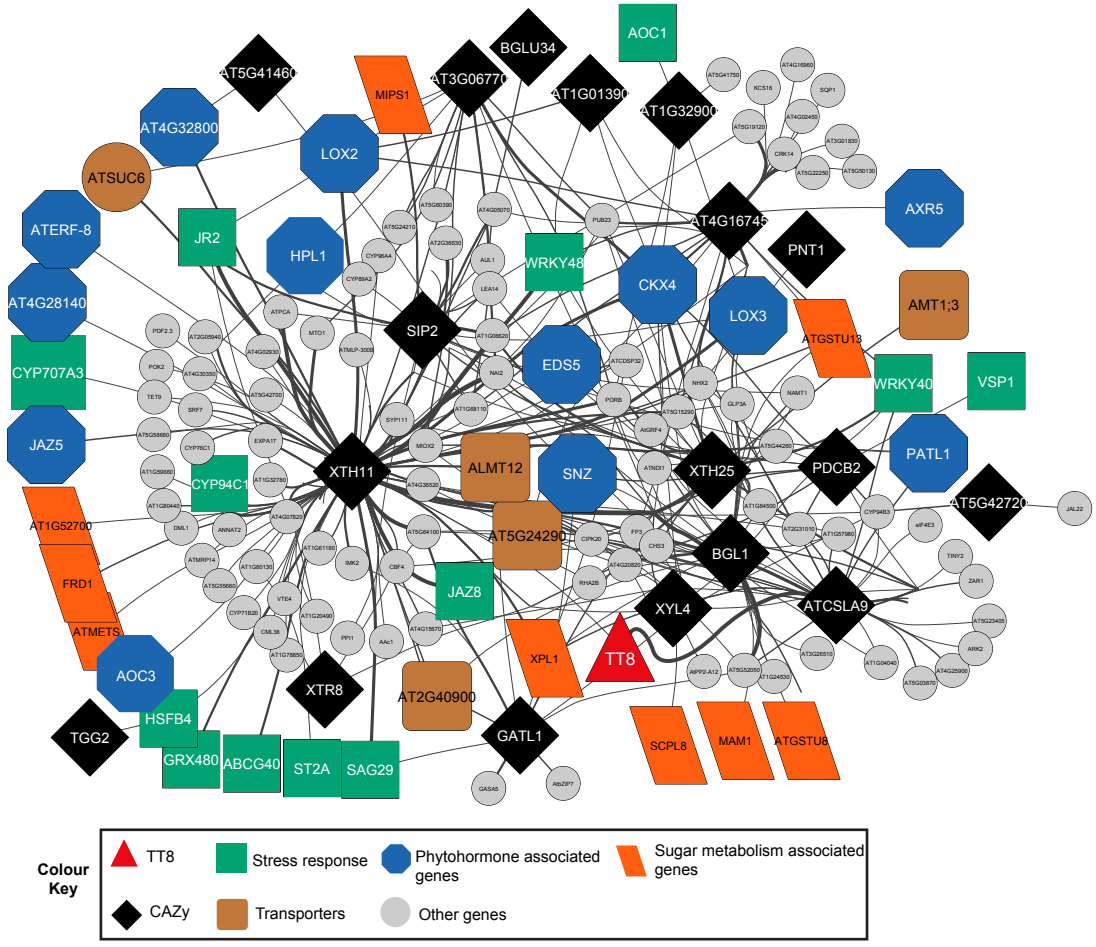


Figure 5.9. Promoter network showing CAZy genes that share motif similarity with stress response and phytohormone-associated genes. Differential genes formed the nodes, while the edge width indicates the number of shared motifs between any two genes.

5.3.6. TT8 reprograms hormone biosynthesis and sugar conjugations by physically binding to their promoters

Direct associations of TT8 with genes associated with jasmonic acid and brassinosteroid biosynthesis, glycosylation of metabolites and stress response pathways were tested using ChIP-PCR and RT-PCR in inducible overexpression lines. Representative genes from each process in the promoter network were selected and their expression levels were analysed in overexpression and mutant lines using RT-PCR-based relative quantification.

Interestingly a number of CAZy genes showed reciprocal expression trends in TT8 loss and induced overexpression lines, suggesting direct association of the expression of these genes with TT8 activity (Figure 5.10A). Hydrolases which were down-regulated in dex-induced overexpression lines were u*p*-regulated in *tt8*. Conversely, transferases and sugar transporters, down-regulated in TT8 loss-of-function lines were up-regulated in dex-induced overexpression lines. TT8 also showed direct binding to the promoters of the CAZy genes, UGT84A1, UGT85A3 and UGT88A1 in dex-induced lines (Figure 5.10A). In TT8 loss-of-function lines, twelve jasmonic acid-associated genes were up-regulated, while the same set of genes were down-regulated in the dex-inducible system (Figure 5.10B). TT8 also showed direct binding to the promoters of six jasmonic acid biosynthesis genes, namely *AOC2*, *AOC3*, *LOX2*, *OPCL1*, *ST2A* and *HPL1* in dex-induced overexpression lines.

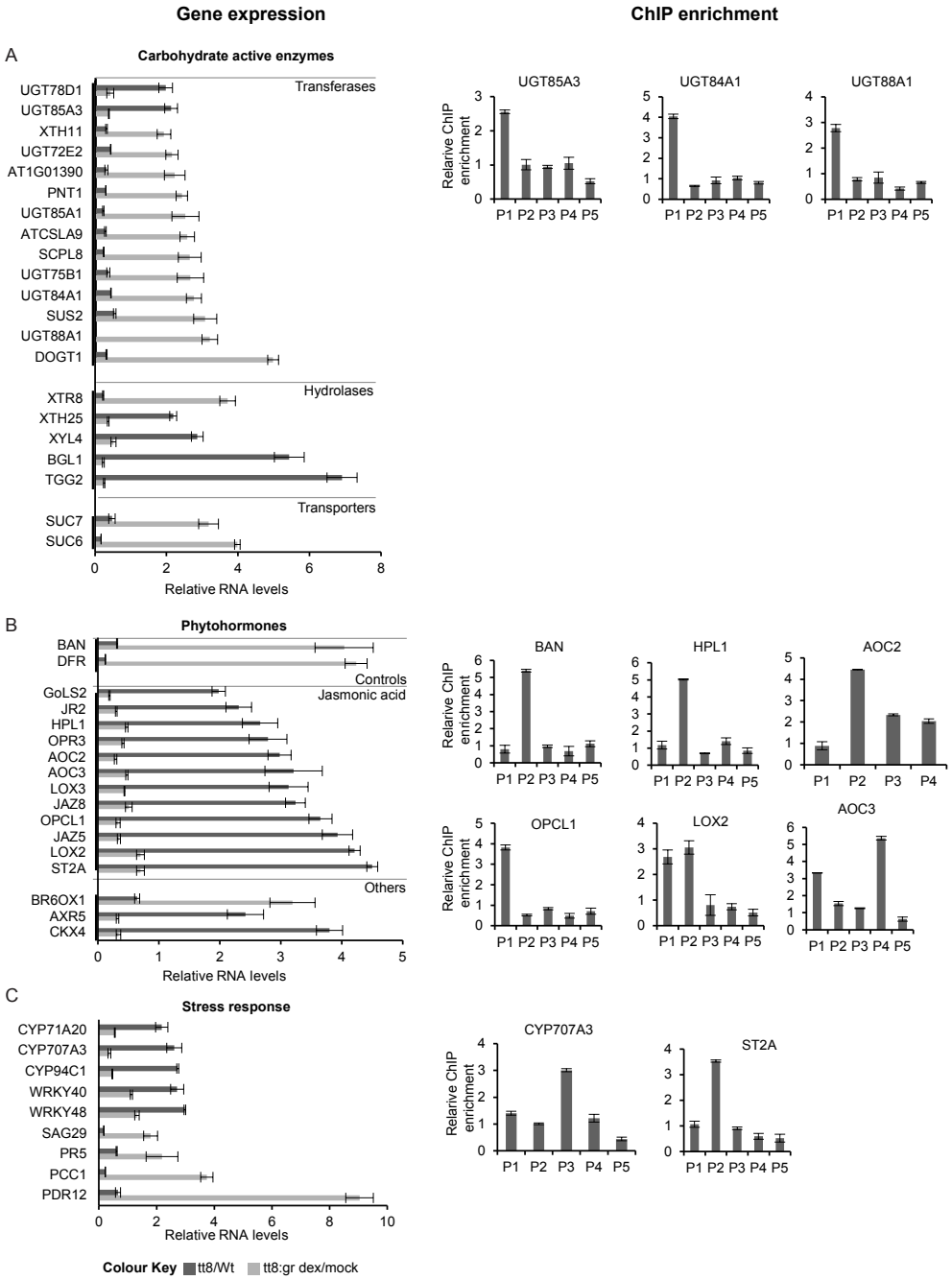


Figure 5.10. Expression trends of genes which shared promoter motifs in tt8 and induced overexpression lines. Dark grey bar represents relative gene expression levels in tt8/Ws, whereas light indicates TT8:GR dex/TT8:GR mock. (A) Genes associated with sugar metabolism, (B) phytohormones, and (C) stress response-associated genes.

The key gene from brassinosteroid biosynthesis pathway, BR6ox1 was up-regulated in *tt8* and down-regulated in dex-induced overexpression lines. Genes associated with stress response also showed reversal in gene expression trends in mutant and overexpression lines (Figure 5.10C). Interestingly, the overall gene and metabolite expression trends in dex-induced overexpression lines were similar to *tt8* but showed exact inverse trends. These results taken together show that TT8 is a potential master regulator that directly binds to and reprograms genes associated with stress-related hormone biosynthesis, sugar conjugation processes and indirectly regulates stress-associated genes in *Arabidopsis thaliana*.

5.3.7. TT8 overexpression enhances stress tolerance

The results from gene expression and metabolomics analyses showing that TT8 controls stress response and hormone biosynthesis overwhelmingly suggest a regulatory role for TT8 in mediating plant innate immunity. Thus, we tested *tt8* and its dex-induced overexpression lines with salt, mannitol and ABA treatments for abiotic stress conditions and for biotic stress, we used MeJA and DON. Response to these stress conditions was analysed by counting the number of germinated seeds and recording images of 6-day-old seedlings. The germination percentage of seeds was calculated to provide a quantitative measure of the stress response.

We observed a large reduction in the germination rates for all treatments in *tt8* when compared to its wild-type (Figure 5.11A). Remarkably, dex-induced *TT8:GR* showed nearly 20-30% improvement in germination rates compared to its wild-type (Figure 5.11B). The germination rates of dex-induced *TT8:GR* compared with its wild-type in stress-free conditions were similar. Therefore, the increased germination rates witnessed in *TT8:GR* lines were not induced by dex, but rather are a result of increased TT8. The effect of stress treatments to *tt8* were most pronounced in salt, mannitol and DON (we observed effects on germination between days 1 to 3), with *tt8* showing nearly 50% lower germination at the end of week one.

Effects of MeJa and ABA were observed between days two or three and culminated in nearly 25% lower germination rates compared to wild-type after one week. Under severe salt stress (200 mM NaCl), *tt8* germinated as late as five days after sowing. These results establish that TT8 plays a direct role in tolerance towards multiple abiotic and biotic stresses tested here.

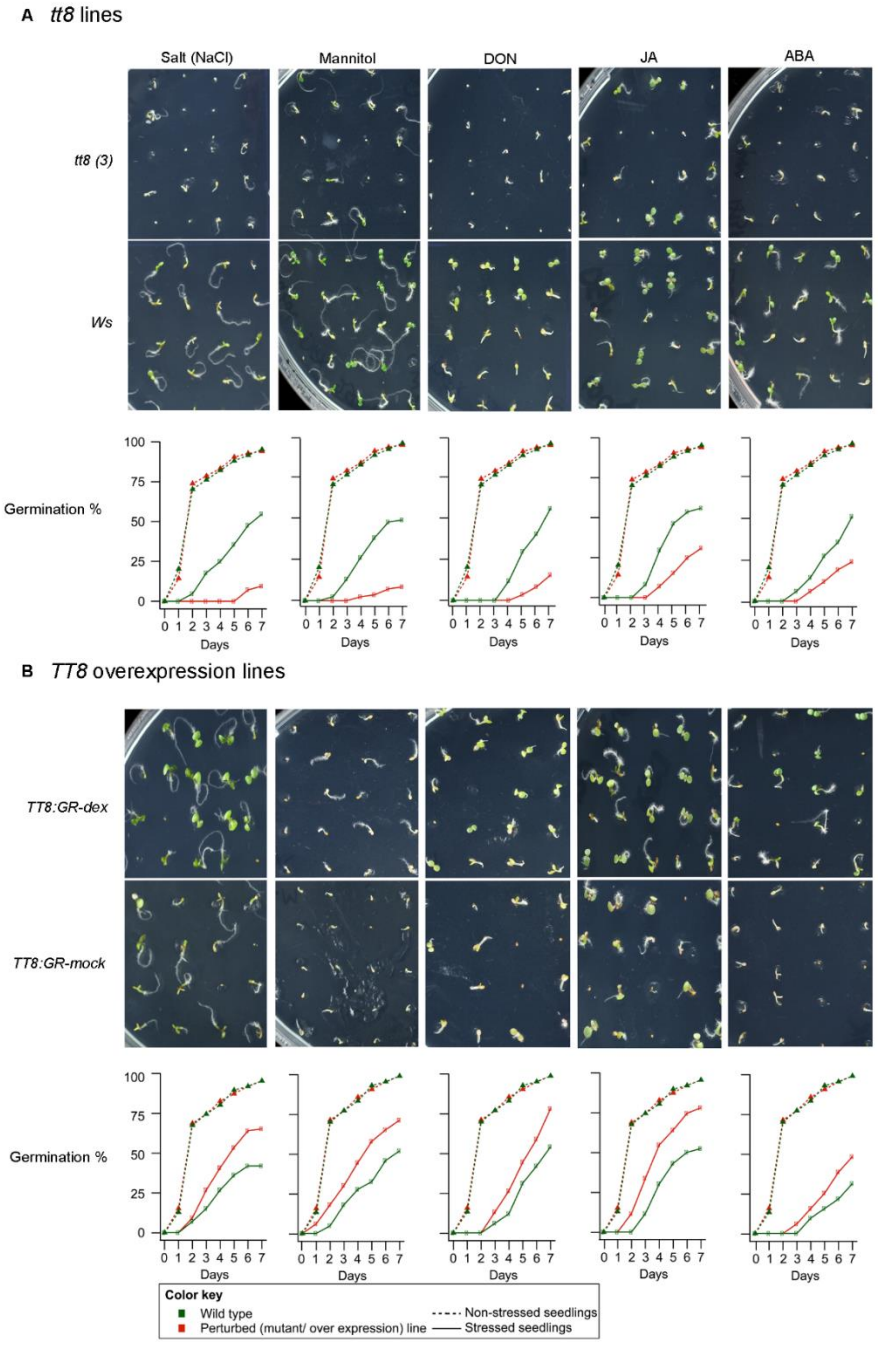


Figure 5.11. Effect of selected stress conditions on (A) TT8 loss of function, and (B) mock/dex treated induced TT8 overexpression lines. TT8 overexpression lines show increased germination rates under all stress conditions compared to its wild-type.

5.4. Conclusion

In this study, we provide computational and biochemical evidence to establish the role of TT8 as a key regulator of glycosylation of metabolites. We have also shown for the first time, the role of TT8 in mediating,

(i) Coordinated regulation of glycosylation of both primary and secondary metabolites through a common mechanism. The utilization of a common mechanism to regulate both primary and secondary metabolites suggests that glycosylation processes that generate diverse metabolites require wide range of cellular resources to be regulated in a coordinated manner.

(ii) Direct relationship between metabolite conjugation and stress hormone biosynthesis through coordinated regulation via direct binding to the promoters or indirectly affecting expression levels.

(iii) Plant stress response against multiple biotic and abiotic stress factors by directly or indirectly regulating the gene expression levels of eight stress-associated genes. These genes have important roles in enabling the plants to survive salt and drought stress (Liu et al., 2013; Mir et al., 2013; Seo et al., 2008).

Interestingly, TT8 overexpression improves stress tolerance, while its loss renders the plant to be more sensitive to abiotic and biotic stresses. Recent reports had suggested that TT8 might play a role in stress response based on increase TT8 levels witnessed in roots in response to salt and osmotic stress (Jiang et al., 2009). Based on TT8's roles in coordination of glycosylation of metabolites and stress response processes, we have now established TT8 as a key regulator of plant stress response. Finally, we propose a model highlighting the role of TT8 in co-ordinately regulating sugar conjugation processes and innate immunity (Fig. 5.12)

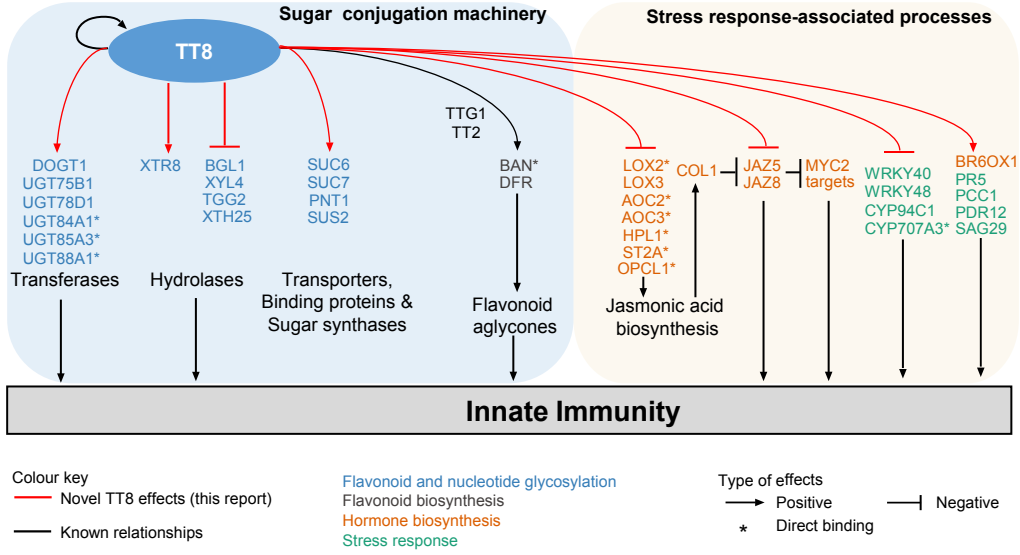


Figure 5.12. Model depicting the role of TT8 in regulating glycosylation of metabolites and mediating plant innate immunity. Glycosylation of nucleotides and flavonoids are positively regulated by TT8, while jasmonic acid biosynthesis is negatively regulated. Expression level of several genes associated with stress response are also regulated by TT8. Thus, TT8 acts as an integrator of secondary metabolism and innate immunity.

The model shows that TT8 directly binds to the promoters of genes associated with glycosylation of metabolites (such as UGT84A1, UGT85A3 and UGT88A1) and hormone biosynthesis (LOX2, AOC2, AOC3, HPL1, ST2 and OPCL1).

Furthermore, the expression levels of members of sugar conjugation machinery (highlighted in blue in Figure 5.12) are also affected by TT8. Thus the regulatory control exhibited through TT8-mediated processes results in increased metabolite diversity. Finally, several genes associated with stress response are either up- or down-regulated as a result of TT8 activity.

Taken together, these results highlight the importance of processes that generate metabolite diversity and reveal the underlying complex mechanistic interactions involved in plant defence strategies.

6. Overall conclusions and future perspectives

"Fill the brain with high thoughts, highest ideals,
place them day and night before you, and
out of that will come great work."
...Swami Vivekananda, Indian philosopher

Deriving meaningful biological information from complex systems requires robust statistical methods. It is clear from these studies that metabolomics approaches can provide valuable insights into cellular responses by linking genotype to phenotypes. Furthermore, we also show that integrating datasets with complementary information about the biological system can generate strong testable biological hypothesis, which can further be validated.

The major conclusions from the three studies are:

(i) Large-scale experiments are prone to unwanted non-biological sources of variation that confound the outcome. If these variations are not corrected, they lead to erroneous biological interpretations. Furthermore, off-the-shelf solutions do not work for complex experimental designs, thus, requiring tailor-made statistical solutions to investigate possible batch effects using exploratory data analysis. This study also highlights the importance of storing all possible meta-information that can be used for investigating unwanted variations. Finally, the statistical methods in this study can also be used for others types of omics datasets.

(ii) While direct induced perturbations can be used in model organisms to understand the metabolic processes, natural varying species that contain genetically intractable systems can be studied using a non-targeted metabolomics approach. As the first study of natural variation in the microalga-*Chlorella* using metabolomics approach, we observed distinct relationships

between habitat and metabolic diversity. This diversity is greater compared to the metabolic divergence at species level between *Chlorella* and *Parachlorella* strains. Physiological and environmental factors therefore outweigh genetic influence on metabolic phenotypes. Furthermore, we showed that by associating growth and physiochemical parameters with metabolic profiles, we can derive biomarkers and associated metabolic pathways. Such associations can be used for predicting and optimizing the behaviour of non-model systems and help in bioprospecting of natural products in naturally varying systems.

(iii) We show that organisms respond to perturbations by modulating their gene expression levels, thus, exuding regulatory controls involving coordinated gene-metabolite changes that reprogram the *Arabidopsis* metabolite network. By utilizing the combined power derived by integrating genomic relationships and gene expression outcomes with metabolite profiling, we were able to uncover TT8's role in increasing metabolite diversity and in regulating stress response and phytohormone biosynthesis. This systems level understanding of the regulatory control in reprogramming of perturbed metabolic networks could only be detected using both transcriptome and metabolome measurements, thus highlighting the utility of a multi-omics approach. Furthermore, we show that plant stress responses are highly dependent on biochemical processes that generate metabolic diversity.

The approaches developed in this study integrate experimental design, environmental factors and direct measurements of metabolic network components (i.e., enzyme and metabolite levels) to assess multiple sources of influence on metabolic phenotypes. These are then analysed using statistical methods to provide valuable biological interpretations. Taken together, the current research work contributes important results that have the potential to be developed into useful applications for human health (Appendix 2), environment (Chapters 4 and 5) and energy (Chapter 4).

Future perspectives from these results include:

- We will be including recommendations for the minimal meta-data structure based on the metabolomics data standards initiative (Fernie et al., 2011; Steinbeck et al., 2012) for reporting and analysing batch-specific variations. To facilitate open access research, we are also planning to provide open access to R scripts upon publication of the paper.
- Secondly, we have identified efficient strains for biofuel production based on the results of metabolomics data that was generated from a survey of naturally varying oleaginous microalgae in Malaysia. Currently we are working on a small-scale batch culture of the same strains in collaboration with Prof. Phang Siew Moi (University of Malaya). Furthermore, as a part of metabolic engineering strategies, multi-omics approaches used in *Arabidopsis* can be extended to to*p*-performing *Chlorella* for enhancing the naturally existing biological design principles leading to efficient biofuel production.
- We identified important components and mechanisms that affect bioactive properties of flavonoids, hormones, stress response, and generate increased metabolite diversity in *Arabidopsis*. Additionally, by establishing TT8, which has homologs in crop plants, as a key integrator of biochemical processes that lead to increased diversity of secondary metabolites and increase plant stress response, we have uncovered untapped potential for the application of TT8 in agriculture and pharmaceutical industries. We are now at an advanced stage in filing for a patent detailing these mechanisms and their applications.

7. Bibliography

Aboud, O.A., and Weiss, R.H. (2013). New Opportunities from the Cancer Metabolome. Clin Chem *59*, 138-146.

Alisdair, R.F., Richard, N.T., Arno, J.K., and Lothar, W. (2004). Innovation: Metabolite Profiling: From Diagnostics to Systems Biology. Nat Rev Mol Cell Bio *5*.

Allen, E., Moing, A., Ebbels, T.M., Maucourt, M., Tomos, A.D., Rolin, D., and Hooks, M.A. (2010). Correlation Network Analysis Reveals a Sequential Reorganization of Metabolic and Transcriptional States During Germination and Gene-Metabolite Relationships in Developing Seedlings of Arabidopsis. BMC Syst Biol *4*, 62.

Allwood, J., and Goodacre, R. (2010). An Introduction to Liquid Chromatography-Mass Spectrometry Instrumentation Applied in Plant Metabolomic Analyses. Phytochem Anal *21*, 33-47.

Alter, O., Brown, P.O., and Botstein, D. (2000). Singular Value Decomposition for Genome-Wide Expression Data Processing and Modeling. Proc Natl Acad Sci U S A *97*, 10101-10106.

Álvarez-Sánchez, B., Priego-Capote, F., and Castro, M.D.L.d. (2010). Metabolomics Analysis Ii. Preparation of Biological Samples Prior to Detection. rac-Trend Anal Chem *29*, 120-127.

Arbib, Z., Ruiz, J., Alvarez-Diaz, P., Garrido-Perez, C., and Perales, J.A. (2014). Capability of Different Microalgae Species for Phytoremediation Processes: Wastewater Tertiary Treatment, Co2 Bio-Fixation and Low Cost Biofuels Production. Water Res *49*, 465-474.

Baker, M. (2011). Metabolomics: From Small Molecules to Big Ideas. Nat Meth *8*, 117-121.

Baudry, A., Heim, M.A., Dubreucq, B., Caboche, M., Weisshaar, B., and Lepiniec, L. (2004). Tt2, Tt8, and Ttg1 Synergistically Specify the Expression of Banyuls and Proanthocyanidin Biosynthesis in Arabidopsis Thaliana. Plant J *39*, 366-380.

Bechtold, N., and Pelletier, G. (1998). In Planta Agrobacterium-Mediated Transformation of Adult Arabidopsis Thaliana Plants by Vacuum Infiltration. Methods Mol Biol *82*, 259-266.

Bedair, M., and Sumner, L.W. (2008). Current and Emerging Mass-Spectrometry Technologies for Metabolomics. Trac-Trend Anal Chem *27*, 238-250.

Bhalla, R., Narasimhan, K., and Swarup, S. (2005). Metabolomics and Its Role in Understanding Cellular Responses in Plants. Plant Cell Rep *24*, 562-571.

Bin, Z., Jun Feng, X., Leepika, T., and Habtom, W.R. (2012). Lc-Ms-Based Metabolomics. Mol Biosyst *8*.

Bino, R.J., Hall, R.D., Fiehn, O., Kopka, J., Saito, K., Draper, J., Nikolau, B.J., Mendes, P., Roessner-Tunali, U., Beale, M.H., *et al.* (2004). Potential of Metabolomics as a Functional Genomics Tool. Trends Plant Sci
9, 418-425.

Biswas, A., Mynampati, K.C., Umashankar, S., Reuben, S., Parab, G., Rao, R., Kannan, V.S., and Swarup, S. (2010). Metdat: A Modular and Workflow-Based Free Online Pipeline for Mass Spectrometry Data Processing, Analysis and Interpretation. Bioinformatics *26*, 2639-2640.

Biswas, A., Rao, R., Umashankar, S., Mynampati, K.C., Reuben, S., Parab, G., and Swarup, S. (2011). Datpav--an Online Processing, Analysis and Visualization Tool for Exploratory Investigation of Experimental Data. Bioinformatics *27*, 1585-1586.

Blanc, G., Duncan, G., Agarkova, I., Borodovsky, M., Gurnon, J., Kuo, A., Lindquist, E., Lucas, S., Pangilinan, J., Polle, J., *et al.* (2010). The Chlorella Variabilis Nc64a Genome Reveals Adaptation to Photosymbiosis, Coevolution with Viruses, and Cryptic Sex. Plant Cell *22*, 2943-2955.

Boccard, J., and Rudaz, S. (2014). Harnessing the Complexity of Metabolomic Data with Chemometrics. J Chemometrics *28*, 1-9.

Boccard, J., Veuthey, J.L., and Rudaz, S. (2010). Knowledge Discovery in Metabolomics: An Overview of Ms Data Handling. J Sep Sci *33*, 290-304.

Bouslimani, A., Sanchez, L.M., Garg, N., and Dorrestein, P.C. (2014). Mass Spectrometry of Natural Products: Current, Emerging and Future Technologies. Nat Prod Rep *31*, 718-729.

Brennan, L., and Owende, P. (2010). Biofuels from Microalgae—a Review of Technologies for Production, Processing, and Extractions of Biofuels and Co-Products. Renew Sust Energ Rev *14*, 557-577.

Breunig, J.S., Hackett, S.R., Rabinowitz, J.D., and Kruglyak, L. (2014). Genetic Basis of Metabolome Variation in Yeast. PLoS Genet *10*, e1004142.

Broadhurst, D., and Kell, D. (2006). Statistical Strategies for Avoiding False Discoveries in Metabolomics and Related Experiments. Metabolomics *2*, 171-196.

Brohee, S., Faust, K., Lima-Mendez, G., Sand, O., Janky, R., Vanderstocken, G., Deville, Y., and van Helden, J. (2008). Neat: A Toolbox for the Analysis of Biological Networks, Clusters, Classes and Pathways. Nucleic Acids Res *36*, W444-451.

Brown, P.D., Tokuhisa, J.G., Reichelt, M., and Gershenzon, J. (2003). Variation of Glucosinolate Accumulation among Different Organs and Developmental Stages of Arabidopsis Thaliana. Phytochemistry *62*, 471-481.

Bundy, J., Davey, M., and Viant, M. (2009). Environmental Metabolomics: A Critical Review and Future Perspectives. Metabolomics *5*, 3-21.

Cai, Y.D. (2012). Editorial. The Application of Systems Biology and Bioinformatics Methods in Proteomics, Transcriptomics and Metabolomics. Protein Pept Lett *19*, 2-3.

Cevallos-Cevallos, J.M., Reyes-De-Corcuera, J.I., Etxeberria, E., Danyluk, M.D., and Rodrick, G.E. (2009). Metabolomic Analysis in Food Science: A Review. Trends Food Sci Tech *20*, 557-566.

Chambers, M.C., Maclean, B., Burke, R., Amodei, D., Ruderman, D.L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., *et al.* (2012). A Cross-Platform Toolkit for Mass Spectrometry and Proteomics. Nat Biotechnol *30*, 918-920.

Chan, E.K.F., Rowe, H.C., Hansen, B.G., and Kliebenstein, D.J. (2010). The Complex Genetic Architecture of the Metabolome. PLoS Genet *6*, e1001198.

Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., and Liu, C. (2011). Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. PLoS One *6*, e17238.

Chen, M.D., Xuan, L., Wang, Z., Zhou, L., Li, Z., Du, X., Ali, E., Zhang, G., and Jiang, L. (2014). Transparent Testa 8 Inhibits Seed Fatty Acid Accumulation by Targeting Several Seed Development Regulators in Arabidopsis. Plant Physiol.

Chico, J.M., Chini, A., Fonseca, S., and Solano, R. (2008). Jaz Repressors Set the Rhythm in Jasmonate Signaling. Curr Opin Plant Biol *11*, 486-494.

Chisti, Y. (2007). Biodiesel from Microalgae. Biotechnol Adv *25*, 294-306.

Choi, H., and Pavelka, N. (2012). When One and One Gives More Than Two: Challenges and Opportunities of Integrative Omics. FGENE *2*.

Chung, N.C., and Storey, J.D. (2014). Statistical Significance of Variables Driving Systematic Variation in High-Dimensional Data. Bioinformatics.

Clardy, J., and Walsh, C. (2004). Lessons from Natural Molecules. Nature *432*, 829-837.

Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., *et al.* (2007). Integration

of Biological Networks and Gene Expression Data Using Cytoscape. Nat Protoc *2*, 2366-2382.

De Livera, A.M., Dias, D.A., De Souza, D., Rupasinghe, T., Pyke, J., Tull, D., Roessner, U., McConville, M., and Speed, T.P. (2012). Normalizing and Integrating Metabolomics Data. Anal Chem *84*, 10768-10776.

de Oliveira Dal'Molin, C.G., Quek, L.E., Palfreyman, R.W., Brumbley, S.M., and Nielsen, L.K. (2010). Aragem, a Genome-Scale Reconstruction of the Primary Metabolic Network in Arabidopsis. Plant Physiol *152*, 579-589.

De Vos, R., Moco, S., Lommen, A., Keurentjes, J., Bino, R., and Hall, R. (2007). Untargeted Large-Scale Plant Metabolomics Using Liquid Chromatography Coupled to Mass Spectrometry. Nat Protoc *2*, 778-791.

Dervilly-Pinel, G., Courant, F., Chereau, S., Royer, A.L., Boyard-Kieken, F., Antignac, J.P., Monteau, F., and Le Bizec, B. (2012). Metabolomics in Food Analysis: Application to the Control of Forbidden Substances. Drug Test Anal *4 Suppl 1*, 59-69.

Dettmer, K., Aronov, P., and Hammock, B. (2007). Mass Spectrometry-Based Metabolomics. Mass Spectrom Rev *26*, 51-78.

Dixon, R.A., Gang, D.R., Charlton, A.J., Fiehn, O., Kuiper, H.A., Reynolds, T.L., Tjeerdema, R.S., Jeffery, E.H., German, J.B., Ridley, W.P., *et al.* (2006). Applications of Metabolomics in Agriculture. J Agric Food Chem *54*, 8984-8994.

Doan, T.T.Y., Sivaloganathan, B., and Obbard, J.P. (2011). Screening of Marine Microalgae for Biodiesel Feedstock. Biomass Bioenergy *35*, 2534-2544.

Dunn, W.B., Wilson, I.D., Nicholls, A.W., and Broadhurst, D. (2012). The Importance of Experimental Design and Qc Samples in Large-Scale and Ms-Driven Untargeted Metabolomic Studies of Humans. Bioanalysis *4*, 2249-2264.

Eckardt, N.A. (2010). The Chlorella Genome: Big Surprises from a Small Package. Plant Cell *22*, 2924.

Eliasson, M., Rannar, S., and Trygg, J. (2011). From Data Processing to Multivariate Validation--Essential Steps in Extracting Interpretable Information from Metabolomics Data. Curr Pharm Biotechnol *12*, 996-1004.

Ernest, B., Gooding, J.R., Campagna, S.R., Saxton, A.M., and Voy, B.H. (2012). Metabr: An R Script for Linear Model Analysis of Quantitative Metabolomic Data. BMC Res Notes *5*, 596.

Ernst, M., Silva, D.B., Silva, R.R., Vencio, R.Z., and Lopes, N.P. (2014). Mass Spectrometry in Plant Metabolomics Strategies: From Analytical Platforms to Data Acquisition and Processing. Nat Prod Rep *31*, 784-806.

Fehrmann, R.S.N., Jansen, R.C., Veldink, J.H., Westra, H.-J., Arends, D., Bonder, M.J., Fu, J., Deelen, P., Groen, H.J.M., Smolonska, A., *et al.* (2011). *Trans*-Eqtls Reveal That Independent Genetic Variants Associated with a Complex Phenotype Converge on Intermediate Genes, with a Major Role for the Hla. PLoS Genet *7*, e1002197.

Fernie, A.R. (2007). The Future of Metabolic Phytochemistry: Larger Numbers of Metabolites, Higher Resolution, Greater Understanding. Phytochemistry *68*, 2861-2880.

Fernie, A.R., Aharoni, A., Willmitzer, L., Stitt, M., Tohge, T., Kopka, J., Carroll, A.J., Saito, K., Fraser, P.D., and DeLuca, V. (2011). Recommendations for Reporting Metabolite Data. Plant Cell *23*, 2477-2482.

Fiehn, O. (2002). Metabolomics--the Link between Genotypes and Phenotypes. Plant Mol Biol *48*, 155-171.

Fiehn, O., Wohlgemuth, G., Scholz, M., Kind, T., Lee do, Y., Lu, Y., Moon, S., and Nikolau, B. (2008). Quality Control for Plant Metabolomics: Reporting Msi-Compliant Studies. Plant J *53*, 691-704.

Fott, B., and Nováková, M. (1969). A Monograph of the Genus Chlorella. The Freshwater Species. Studies in Phycology, 10-74.

Fu, J., Keurentjes, J.J., Bouwmeester, H., America, T., Verstappen, F.W., Ward, J.L., Beale, M.H., de Vos, R.C., Dijkstra, M., Scheltema, R.A., *et al.* (2009). System-Wide Molecular Evidence for Phenotypic Buffering in Arabidopsis. Nat Genet *41*, 166-167.

Gagnon-Bartsch, J.A., and Speed, T.P. (2012). Using Control Genes to Correct for Unwanted Variation in Microarray Data. Biostatistics *13*, 539-552.

Gary, J.P., Oscar, Y., and Gary, S. (2012). Innovation: Metabolomics: The Apogee of the Omics Trilogy. Nat Rev Mol Cell Bio *13*.

Georgianna, D.R., and Mayfield, S.P. (2012). Exploiting Diversity and Synthetic Biology for the Production of Algal Biofuels. Nature *488*, 329-335.

Gibon, Y., and Rolin, D. (2012). Aspects of Experimental Design for Plant Metabolomics Experiments and Guidelines for Growth of Plant Material. In Plant Metabolomics, N.W. Hardy, and R.D. Hall, eds. (Humana Press), pp. 13-30.

Godzien, J., Ciborowski, M., Angulo, S., and Barbas, C. (2013). From Numbers to a Biological Sense: How the Strategy Chosen for Metabolomics Data Treatment May Affect Final Results. A Practical Example Based on Urine Fingerprints Obtained by Lc-Ms. Electrophoresis *34*, 2812-2826.

Goldinger, A., Henders, A.K., McRae, A.F., Martin, N.G., Gibson, G., Montgomery, G.W., Visscher, P.M., and Powell, J.E. (2013). Genetic and Nongenetic Variation Revealed for the Principal Components of Human Gene Expression. Genetics *195*, 1117-1128.

Goodacre, R., Broadhurst, D., Smilde, A., Kristal, B., Baker, J.D., Beger, R., Bessant, C., Connor, S., Capuani, G., Craig, A., *et al.* (2007). Proposed Minimum Reporting Standards for Data Analysis in Metabolomics. Metabolomics *3*, 231-241.

Goodacre, R., Vaidyanathan, S., Dunn, W.B., Harrigan, G.G., and Kell, D.B. (2004). Metabolomics by Numbers: Acquiring and Understanding Global Metabolite Data. Trends Biotechnol *22*, 245-252.

Gromski, P.S., Xu, Y., Kotze, H.L., Correa, E., Ellis, D.I., Armitage, E.G., Turner, M.L., and Goodacre, R. (2014). Influence of Missing Values Substitutes on Multivariate Analysis of Metabolomics Data. Metabolites *4*, 433-452.

Halabalaki, M., Bertrand, S., Stefanou, A., Gindro, K., Kostidis, S., Mikros, E., Skaltsounis, L.A., and Wolfender, J.L. (2014). Sample Preparation Issues in Nmr-Based Plant Metabolomics: Optimisation for Vitis Wood Samples. Phytochem Anal *25*, 350-356.

Hall, R. (2006). Plant Metabolomics: From Holistic Hope, to Hype, to Hot Topic. New Phytol *169*, 453-468.

Hamilton, J.J., and Reed, J.L. (2014). Software Platforms to Facilitate Reconstructing Genome-Scale Metabolic Networks. Environ Microbiol *16*, 49-59.

Harborne, J.B., and Baxter, H. (1999). The Handbook of Natural Flavonoids Vol 1 (John Wiley & Sons).

Hartmann, T. (1996). Diversity and Variability of Plant Secondary Metabolism: A Mechanistic View. Entomol Exp Appl *80*, 177-188.

Hartmann, T. (2007). From Waste Products to Ecochemicals: Fifty Years Research of Plant Secondary Metabolism. Phytochemistry *68*, 2831-2846.

Haug, K., Salek, R.M., Conesa, P., Hastings, J., de Matos, P., Rijnbeek, M., Mahendraker, T., Williams, M., Neumann, S., Rocca-Serra, P., *et al.* (2013). Metabolights--an Open-Access General-Purpose Repository for Metabolomics Studies and Associated Meta-Data. Nucleic Acids Res *41*, D781-786.

Haygood, R., Fedrigo, O., Hanson, B., Yokoyama, K.D., and Wray, G.A. (2007). Promoter Regions of Many Neural- and Nutrition-Related Genes Have Experienced Positive Selection During Human Evolution. Nat Genet *39*, 1140-1144.

Hegeman, A. (2010). Plant Metabolomics--Meeting the Analytical Challenges of Comprehensive Metabolite Analysis. Brief Funct Genomics *9*, 139-148.

Hendriks, M.M.W.B., van Eeuwijk, F.A., Jellema, R.H., Westerhuis, J.A., Reijmers, T.H., Hoefsloot, H.C.J., and Smilde, A.K. (2011). Data-Processing Strategies for Metabolomics Studies. Trac-Trend Anal Chem *30*, 1685-1698.

Hirai, M.Y., Klein, M., Fujikawa, Y., Yano, M., Goodenowe, D.B., Yamazaki, Y., Kanaya, S., Nakamura, Y., Kitayama, M., Suzuki, H., *et al.* (2005). Elucidation of Gene-to-Gene and Metabolite-to-Gene Networks in Arabidopsis by Integration of Metabolomics and Transcriptomics. J Biol Chem *280*, 25590-25595.

Hirai, M.Y., Sugiyama, K., Sawada, Y., Tohge, T., Obayashi, T., Suzuki, A., Araki, R., Sakurai, N., Suzuki, H., Aoki, K., *et al.* (2007). Omics-Based Identification of Arabidopsis Myb Transcription Factors Regulating Aliphatic Glucosinolate Biosynthesis. Proc Natl Acad Sci U S A *104*, 6478-6483.

Holme, P. (2011). Metabolic Robustness and Network Modularity: A Model Study. PLoS One *6*, e16605.

Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., *et al.* (2010). Massbank: A Public Repository for Sharing Mass Spectral Data for Life Sciences. J Mass Spectrom *45*, 703-714.

Houshyani, B., Kabouw, P., Muth, D., de Vos, R.C., Bino, R.J., and Bouwmeester, H.J. (2012). Characterization of the Natural Variation in Arabidopsis Thaliana Metabolome by the Analysis of Metabolic Distance. Metabolomics *8*, 131-145.

Huerlimann, R., de Nys, R., and Heimann, K. (2010). Growth, Lipid Content, Productivity, and Fatty Acid Composition of Tropical Microalgae for Scale-up Production. Biotechnol Bioeng *107*, 245-257.

Hummel, J., Selbig, J., Walther, D., and Kopka, J. (2007). The Golm Metabolome Database: A database for Gc-Ms Based Metabolite Profiling. In Metabolomics, J. Nielsen, and M. Jewett, eds. (Springer Berlin Heidelberg), pp. 75-95.

Huss, M., and Holme, P. (2007). Currency and Commodity Metabolites: Their Identification and Relation to the Modularity of Metabolic Networks. IET Syst Biol *1*, 280-285.

Ideker, T., Galitski, T., and Hood, L. (2001). A New Approach to Decoding Life: Systems Biology. Annu Rev Genomics Hum Genet *2*, 343-372.

Ivosev, G., Burton, L., and Bonner, R. (2008). Dimensionality Reduction and Visualization in Principal Component Analysis. Anal Chem *80*, 4933-4944.

Jaeger, R.G., and Halliday, T.R. (1998). On Confirmatory Versus Exploratory Research. Herpetologica *54*, S64-S66.

Jansen, R.C. (2003). Studying Complex Biological Systems Using Multifactorial Perturbation. Nat Rev Genet *4*, 145-151.

Jari Oksanen, F.G.B., Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Helene Wagner (2013). Vegan: Community Ecology Package.

Jiang, Y., Yang, B., and Deyholos, M.K. (2009). Functional Characterization of the Arabidopsis Bhlh92 Transcription Factor in Abiotic Stress. Mol Genet Genomics *282*, 503-516.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods. Biostatistics *8*, 118-127.

Jones, D.P., Park, Y., and Ziegler, T.R. (2012). Nutritional Metabolomics: Progress in Addressing Complexity in Diet and Health. Annu Rev Nutr *32*, 183-202.

Jones, P., Messner, B., Nakajima, J., Schaffner, A.R., and Saito, K. (2003). Ugt73c6 and Ugt78d1, Glycosyltransferases Involved in Flavonol Glycoside Biosynthesis in Arabidopsis Thaliana. J Biol Chem *278*, 43910-43918.

Joyce, A.R., and Palsson, B.O. (2006). The Model Organism as a System: Integrating 'Omics' Data Sets. Nat Rev Mol Cell Biol *7*, 198-210.

Junker, B.H., Klukas, C., and Schreiber, F. (2006). Vanted: A System for Advanced Data Analysis and Visualization in the Context of Biological Networks. BMC Bioinformatics *7*, 109.

Kaddurah-Daouk, R., Kristal, B.S., and Weinshilboum, R.M. (2008). Metabolomics: A Global Biochemical Approach to Drug Response and Disease. Annu Rev Pharmacol Toxicol *48*, 653-683.

Kamburov, A., Cavill, R., Ebbels, T.M., Herwig, R., and Keun, H.C. (2011). Integrated Pathway-Level Analysis of Transcriptomics and Metabolomics Data with Impala. Bioinformatics *27*, 2917-2918.

Kanehisa, M., and Goto, S. (2000). Kegg: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res *28*, 27-30.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, Information, Knowledge and Principle: Back to Metabolism in Kegg. Nucleic Acids Res *42*, D199-205.

Kang, S.G., Price, J., Lin, P.C., Hong, J.C., and Jang, J.C. (2010). The Arabidopsis Bzip1 Transcription Factor Is Involved in Sugar Signaling, Protein Networking, and DNA Binding. Mol Plant *3*, 361-373.

Karpievitch, Y.V., Dabney, A.R., and Smith, R.D. (2012). Normalization and Missing Value Imputation for Label-Free Lc-Ms Analysis. BMC Bioinformatics *13 Suppl 16*, S5.

Katajamaa, M., and Oresic, M. (2007). Data Processing for Mass Spectrometry-Based Metabolomics. J Chromatogr A *1158*, 318-328.

Kazan, K., and Manners, J.M. (2012). Jaz Repressors and the Orchestration of Phytohormone Crosstalk. Trends Plant Sci *17*, 22-31.

Kell, D., and Mendes, P. (2000). Snapshots of Systems. In Technological and Medical Implications of Metabolic Control Analysis, A. Cornish-Bowden, and M. Cárdenas, eds. (Springer Netherlands), pp. 3-25.

Kell, D.B. (2004). Metabolomics and Systems Biology: Making Sense of the Soup. Curr Opin Microbiol *7*, 296-307.

Kessner, D., Chambers, M., Burke, R., Agus, D., and Mallick, P. (2008). Proteowizard: Open Source Software for Rapid Proteomics Tools Development. Bioinformatics *24*, 2534-2536.

Kim, H.K., Choi, Y.H., and Verpoorte, R. (2011). Nmr-Based Plant Metabolomics: Where Do We Stand, Where Do We Go? Trends Biotechnol *29*, 267-275.

Kim, H.K., and Verpoorte, R. (2010). Sample Preparation for Plant Metabolomics. Phytochem Anal *21*, 4-13.

Kind, T., Scholz, M., and Fiehn, O. (2009). How Large Is the Metabolome? A Critical Analysis of Data Exchange Practices in Chemistry. PLoS One *4*, e5440.

Kresnowati, M.T., van Winden, W.A., Almering, M.J., ten Pierick, A., Ras, C., Knijnenburg, T.A., Daran-Lapujade, P., Pronk, J.T., Heijnen, J.J., and Daran, J.M. (2006). When Transcriptome Meets Metabolome: Fast Cellular Responses of Yeast to Sudden Relief of Glucose Limitation. Mol Syst Biol *2*, 49.

Krug, D., and Muller, R. (2014). Secondary Metabolomics: The Impact of Mass Spectrometry-Based Approaches on the Discovery and Characterization of Microbial Natural Products. Nat Prod Rep *31*, 768-783.

Kuhl, C., Tautenhahn, R., Bottcher, C., Larson, T.R., and Neumann, S. (2012). Camera: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. Anal Chem *84*, 283-289.

Lazar, C., Meganck, S., Taminau, J., Steenhoff, D., Coletta, A., Molter, C., Weiss-Solis, D.Y., Duque, R., Bersini, H., and Nowe, A. (2013). Batch Effect Removal Methods for Microarray Gene Expression Data Integration: A Survey. Brief Bioinform *14*, 469-490.

LeBlanc, G.A. (2007). Phase Ii—Conjugation of Toxicants. In Molecular and Biochemical Toxicology (John Wiley & Sons, Inc.), pp. 219-237.

Lee, I., Ambaru, B., Thakkar, P., Marcotte, E.M., and Rhee, S.Y. (2010). Rational Association of Genes with Traits Using a Genome-Scale Gene Network for Arabidopsis Thaliana. Nat Biotechnol *28*, 149-156.

Lee, Y.J., Narasimhan, K., and Swarup, S. (2013). Enhancement of Plant–Microbe Interactions Using Rhizosphere Metabolomics-Driven Approach and Its Application in the Removal of Polychlorinated Biphenyls. In Molecular Microbial Ecology of the Rhizosphere (John Wiley & Sons, Inc.), pp. 1191-1198.

Leek, J.T. (2014). Svaseq: Removing Batch Effects and Other Unwanted Noise from Sequencing Data. Nucleic Acids Res *42*.

Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., and Storey, J.D. (2012). The Sva Package for Removing Batch Effects and Other Unwanted Variation in High-Throughput Experiments. Bioinformatics *28*, 882-883.

Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., and Irizarry, R.A. (2010). Tackling the Widespread and Critical Impact of Batch Effects in High-Throughput Data. Nat Rev Genet *11*, 733-739.

Leek, J.T., and Storey, J.D. (2007). Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. PLoS Genet *3*, 1724-1735.

Lei, Z., Huhman, D., and Sumner, L. (2011). Mass Spectrometry Strategies in Metabolomics. J Biol Chem *286*, 25435-25442.

Leister, D., Wang, X., Haberer, G., Mayer, K.F., and Kleine, T. (2011). Intracompartmental and Intercompartmental Transcriptional Networks Coordinate the Expression of Genes for Organellar Functions. Plant Physiol *157*, 386-404.

Levo, M., and Segal, E. (2014). In Pursuit of Design Principles of Regulatory Sequences. Nat Rev Genet *15*, 453-468.

Lewis, N.E., Nagarajan, H., and Palsson, B.O. (2012). Constraining the Metabolic Genotype-Phenotype Relationship Using a Phylogeny of in Silico Methods. Nat Rev Microbiol *10*, 291-305.

Li, C., Han, J., Yao, Q., Zou, C., Xu, Y., Zhang, C., Shang, D., Zhou, L., Zou, C., Sun, Z., *et al.* (2013). Subpathway-Gm: Identification of Metabolic Subpathways Via Joint Power of Interesting Genes and Metabolites and Their Topologies within Pathways. Nucleic Acids Res *41*, e101.

Liland, K.H. (2011). Multivariate Methods in Metabolomics – from Pre-Processing to Dimension Reduction and Statistical Analysis. Trac-Trend Anal Chem *30*, 827-841.

Linhart, C., Halperin, Y., and Shamir, R. (2008). Transcription Factor and Microrna Motif Discovery: The Amadeus Platform and a Compendium of Metazoan Target Sets. Genome Res *18*, 1180-1189.

Liu, W.X., Zhang, F.C., Zhang, W.Z., Song, L.F., Wu, W.H., and Chen, Y.F. (2013). Arabidopsis Di19 Functions as a Transcription Factor and Modulates Pr1, Pr2, and Pr5 Expression in Response to Drought Stress. Mol Plant *6*, 1487-1502.

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M., and Henrissat, B. (2014). The Carbohydrate-Active Enzymes Database (Cazy) in 2013. Nucleic Acids Res *42*, D490-495.

Marx, V. (2013). Biology: The Big Challenges of Big Data. Nature *498*, 255-260.

Medema, M.H., Breitling, R., Bovenberg, R., and Takano, E. (2011). Exploiting Plug-and-Play Synthetic Biology for Drug Discovery and Production in Microorganisms. Nat Rev Microbiol *9*, 131-137.

Meier, S., Gehring, C., MacPherson, C., Kaur, M., Maqungo, M., Reuben, S., Muyanga, S., Shih, M.-D., Wei, F.-J., Wanchana, S., *et al.* (2008). The Promoter Signatures in Rice Lea Genes Can Be Used to Build a Co-Expressing Lea Gene Network. Rice *1*, 177-187.

Members, M.S.I.B., Sansone, S.A., Fan, T., Goodacre, R., Griffin, J.L., Hardy, N.W., Kaddurah-Daouk, R., Kristal, B.S., Lindon, J., Mendes, P., *et al.* (2007). The Metabolomics Standards Initiative. Nat Biotechnol *25*, 846-848.

Mir, R., Hernandez, M.L., Abou-Mansour, E., Martinez-Rivas, J.M., Mauch, F., Metraux, J.P., and Leon, J. (2013). Pathogen and Circadian Controlled 1 (Pcc1)

Regulates Polar Lipid Content, Aba-Related Responses, and Pathogen Defence in Arabidopsis Thaliana. J Exp Bot *64*, 3385-3395.
Mithani, A., Preston, G.M., and Hein, J. (2010). A Bayesian Approach to the Evolution of Metabolic Networks on a Phylogeny. PLoS Comput Biol *6*.
Moco, S., Vervoort, J., Bino, R.J., De Vos, R.C., and Bino, R. (2007). Metabolomics Technologies and Metabolite Identification. Trac-Trend Anal Chem *26*, 855-866.
Mueller, L.A., Zhang, P., and Rhee, S.Y. (2003). Aracyc: A Biochemical Pathway Database for Arabidopsis. Plant Physiol *132*, 453-460.
Nah, J.H., Kim, H.J., Lee, H.N., Lee, M.J., Choi, S.S., and Kim, E.S. (2013). Identification and Biotechnological Application of Novel Regulatory Genes Involved in Streptomyces Polyketide Overproduction through Reverse Engineering Strategy. Biomed Res Int *2013*, 549737.
Narasimhan, K., Basheer, C., Bajic, V.B., and Swarup, S. (2003). Enhancement of Plant-Microbe Interactions Using a Rhizosphere Metabolomics-Driven Approach and Its Application in the Removal of Polychlorinated Biphenyls. Plant Physiol *132*, 146-153.
Navarro, L., Bari, R., Achard, P., Lison, P., Nemri, A., Harberd, N.P., and Jones, J.D. (2008). Dellas Control Plant Immune Responses by Modulating the Balance of Jasmonic Acid and Salicylic Acid Signaling. Curr Biol *18*, 650-655.
Naz, S., Vallejo, M., Garcia, A., and Barbas, C. (2014). Method Validation Strategies Involved in Non-Targeted Metabolomics. J Chromatogr A *1353*, 99-105.
Nesi, N., Debeaujon, I., Jond, C., Pelletier, G., Caboche, M., and Lepiniec, L. (2000). The Tt8 Gene Encodes a Basic Helix-Loop-Helix Domain Protein Required for Expression of Dfr and Ban Genes in Arabidopsis Siliques. Plant Cell *12*, 1863-1878.
Nguyen, Q.T., Merlo, M.E., Medema, M.H., Jankevics, A., Breitling, R., and Takano, E. (2012). Metabolomics Methods for the Synthetic Biology of Secondary Metabolism. FEBS Lett *586*, 2177-2183.
Nielsen, J.H., and Jewett, M.C. (2007). Metabolomics : A Powerful Tool in Systems Biology (Berlin: Springer).
Noguchi, T., Fujioka, S., Choe, S., Takatsuto, S., Tax, F.E., Yoshida, S., and Feldmann, K.A. (2000). Biosynthetic Pathways of Brassinolide in Arabidopsis. Plant Physiol *124*, 201-209.
Obata, T., and Fernie, A.R. (2012). The Use of Metabolomics to Dissect Plant Responses to Abiotic Stresses. Cell Mol Life Sci *69*, 3225-3243.
Okazaki, Y., and Saito, K. (2012). Recent Advances of Metabolomics in Plant Biotechnology. Plant Biotechnol Rep *6*, 1-15.
Okuda, S., Yamada, T., Hamajima, M., Itoh, M., Katayama, T., Bork, P., Goto, S., and Kanehisa, M. (2008). Kegg Atlas Mapping for Global Analysis of Metabolic Pathways. Nucleic Acids Res *36*, W423-426.
Oliver, S.G., Winson, M.K., Kell, D.B., and Baganz, F. (1998). Systematic Functional Analysis of the Yeast Genome. Trends Biotechnol *16*, 373-378.
Onofrejova, L., Vasickova, J., Klejdus, B., Stratil, P., Misurcova, L., Kracmar, S., Kopecky, J., and Vacek, J. (2010). Bioactive Phenols in Algae: The Application of Pressurized-Liquid and Solid-Phase Extraction Techniques. J Pharm Biomed Anal *51*, 464-470.
Pace, N., Stahl, D., Lane, D., and Olsen, G. (1986). The Analysis of Natural Microbial Populations by Ribosomal Rna Sequences. In Advances in Microbial Ecology, K.C. Marshall, ed. (Springer US), pp. 1-55.
Pages, H., Aboyoun, P., Gentleman, R., and DebRoy, S. Biostrings: String Objects Representing Biological Sequences, and Matching Algorithms.
Palsson, B. (2006). Systems Biology: Properties of Reconstructed Networks (Cambridge University Press).

Parab, G.S., Rao, R., Lakshminarayanan, S., Bing, Y.V., Moochhala, S.M., and Swarup, S. (2009). Data-Driven Optimization of Metabolomics Methods Using Rat Liver Samples. Anal Chem *81*, 1315-1323.

Patti, G. (2011). Separation Strategies for Untargeted Metabolomics. J sep sci *34*, 3460-3469.

Patti, G., Tautenhahn, R., and Siuzdak, G. (2012). Meta-Analysis of Untargeted Metabolomic Data from Multiple Profiling Experiments. Nat Protoc *7*, 508-516.

Pelletier, M.K., Burbulis, I.E., and Winkel-Shirley, B. (1999). Disruption of Specific Flavonoid Genes Enhances the Accumulation of Flavonoid Enzymes and End-Products in Arabidopsis Seedlings. Plant Mol Biol *40*, 45-54.

Phang, S.-M. (1990). Algal Production from Agro-Industrial and Agricultural Wastes in Malaysia. Ambio, 415-418.

Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2010). Understanding Mechanisms Underlying Human Gene Expression Variation with Rna Sequencing. Nature *464*, 768-772.

Pir, P., Kirdar, B., Hayes, A., Onsan, Z.Y., Ulgen, K.O., and Oliver, S.G. (2006). Integrative Investigation of Metabolic and Transcriptomic Data. BMC Bioinformatics *7*, 203.

Pluskal, T., Castillo, S., Villar-Briones, A., and Oresic, M. (2010). Mzmine 2: Modular Framework for Processing, Visualizing, and Analyzing Mass Spectrometry-Based Molecular Profile Data. BMC Bioinformatics *11*, 395.

Pollard, K.S., Dudoit, S., and van der Laan, M.J. (2005). Multiple Testing Procedures: The Multtest Package and Applications to Genomics. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor, R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit, eds. (Springer New York), pp. 249-271.

Poppenberger, B., Berthiller, F., Lucyshyn, D., Sieberer, T., Schuhmacher, R., Krska, R., Kuchler, K., Glossl, J., Luschnig, C., and Adam, G. (2003). Detoxification of the Fusarium Mycotoxin Deoxynivalenol by a Udp-Glucosyltransferase from Arabidopsis Thaliana. J Biol Chem *278*, 47905-47914.

Pribyl, P., Cepak, V., and Zachleder, V. (2012). Production of Lipids in 10 Strains of Chlorella and Parachlorella, and Enhanced Lipid Productivity in Chlorella Vulgaris. Appl Microbiol Biotechnol *94*, 549-561.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies. Nat Genet *38*, 904-909.

Price, N.D., Papin, J.A., Schilling, C.H., and Palsson, B.O. (2003). Genome-Scale Microbial in Silico Models: The Constraints-Based Approach. Trends Biotechnol *21*, 162-169.

Prosser, G.A., Larrouy-Maumus, G., and de Carvalho, L.P. (2014). Metabolomic Strategies for the Identification of New Enzyme Functions and Metabolic Pathways. EMBO Rep.

Ptashne, M. (1988). How Eukaryotic Transcriptional Activators Work. Nature *335*, 683-689.

Qi, T., Song, S., Ren, Q., Wu, D., Huang, H., Chen, Y., Fan, M., Peng, W., Ren, C., and Xie, D. (2011). The Jasmonate-Zim-Domain Proteins Interact with the Wd-Repeat/Bhlh/Myb Complexes to Regulate Jasmonate-Mediated Anthocyanin Accumulation and Trichome Initiation in Arabidopsis Thaliana. Plant Cell *23*, 1795-1814.

R Core Team (2014). R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).

Raamsdonk, L.M., Teusink, B., Broadhurst, D., Zhang, N., Hayes, A., Walsh, M.C., Berden, J.A., Brindle, K.M., Kell, D.B., Rowland, J.J., *et al.* (2001). A Functional

Genomics Strategy That Uses Metabolome Data to Reveal the Phenotype of Silent Mutations. Nat Biotech *19*, 45-50.
Radakovits, R., Jinkerson, R.E., Darzins, A., and Posewitz, M.C. (2010). Genetic Engineering of Algae for Enhanced Biofuel Production. Eukaryot Cell *9*, 486-501.
Rai, A., Umashankar, S., and Swarup, S. (2013). Plant Metabolomics: From Experimental Design to Knowledge Extraction. In Legume Genomics, R.J. Rose, ed. (Humana Press), pp. 279-312.
Rasmussen, S., Parsons, A.J., and Jones, C.S. (2012). Metabolomics of Forage Plants: A Review. Ann Bot *110*, 1281-1290.
Redestig, H., Fukushima, A., Stenlund, H., Moritz, T., Arita, M., Saito, K., and Kusano, M. (2009). Compensation for Systematic Cross-Contribution Improves Normalization of Mass Spectrometry Based Metabolomics Data. Anal Chem *81*, 7974-7980.
Rhee, K.Y., de Carvalho, L.P., Bryk, R., Ehrt, S., Marrero, J., Park, S.W., Schnappinger, D., Venugopal, A., and Nathan, C. (2011). Central Carbon Metabolism in Mycobacterium Tuberculosis: An Unexpected Frontier. Trends Microbiol *19*, 307-314.
Rhodes, M. (1994). Physiological Roles for Secondary Metabolites in Plants: Some Progress, Many Outstanding Problems. Plant Mol Biol *24*, 1-20.
Riedelsheimer, C., Lisec, J., Czedik-Eysenberg, A., Sulpice, R., Flis, A., Grieder, C., Altmann, T., Stitt, M., Willmitzer, L., and Melchinger, A.E. (2012). Genome-Wide Association Mapping of Leaf Metabolic Profiles for Dissecting Complex Traits in Maize. Proc Natl Acad Sci U S A *109*, 8872-8877.
Rochfort, S. (2005). Metabolomics Reviewed: A New "Omics" Platform Technology for Systems Biology and Implications for Natural Products Research. J Nat Prod *68*, 1813-1820.
Rodgers, J.L., and Nicewander, W.A. (1988). Thirteen Ways to Look at the Correlation Coefficient. Am Stat *42*, 59-66.
Rodolfi, L., Chini Zittelli, G., Bassi, N., Padovani, G., Biondi, N., Bonini, G., and Tredici, M.R. (2009). Microalgae for Oil: Strain Selection, Induction of Lipid Synthesis and Outdoor Mass Cultivation in a Low-Cost Photobioreactor. Biotechnol Bioeng *102*, 100-112.
Ruppin, E., Papin, J.A., de Figueiredo, L.F., and Schuster, S. (2010). Metabolic Reconstruction, Constraint-Based Analysis and Game Theory to Probe Genome-Scale Metabolic Networks. Curr Opin Biotechnol *21*, 502-510.
Saccenti, E., Hoefsloot, H.J., Smilde, A., Westerhuis, J., and Hendriks, M.W.B. (2014). Reflections on Univariate and Multivariate Analysis of Metabolomics Data. Metabolomics *10*, 361-374.
Saito, K., and Matsuda, F. (2010). Metabolomics for Functional Genomics, Systems Biology, and Biotechnology. Annu Rev Plant Biol *61*, 463-489.
Saito, N., Ohashi, Y., Soga, T., and Tomita, M. (2010). Unveiling Cellular Biochemical Reactions Via Metabolomics-Driven Approaches. Curr Opin Microbiol *13*, 358-362.
Sakakibara, H. (2006). Cytokinins: Activity, Biosynthesis, and Translocation. Annu Rev Plant Biol *57*, 431-449.
Salek, R.M., Haug, K., Conesa, P., Hastings, J., Williams, M., Mahendraker, T., Maguire, E., Gonzalez-Beltran, A.N., Rocca-Serra, P., Sansone, S.A., *et al.* (2013a). The Metabolights Repository: Curation Challenges in Metabolomics. Database (Oxford) *2013*, bat029.
Salek, R.M., Haug, K., and Steinbeck, C. (2013b). Dissemination of Metabolomics Results: Role of Metabolights and Cosmos. Gigascience *2*, 8.
Sana, T.R., Roark, J.C., Li, X., Waddell, K., and Fischer, S.M. (2008). Molecular Formula and Metlin Personal Metabolite Database Matching Applied to the Identification of Compounds Generated by Lc/Tof-Ms. J Biomol Tech *19*, 258-266.

Schumacher, M., Kelkel, M., Dicato, M., and Diederich, M. (2011). A Survey of Marine Natural Compounds and Their Derivatives with Anti-Cancer Activity Reported in 2010. Molecules *16*, 5629-5646.

Schweiger, R., Baier, M.C., Persicke, M., and Muller, C. (2014). High Specificity in Plant Leaf Metabolic Responses to Arbuscular Mycorrhiza. Nat Commun *5*, 3886.

Segata, N., Boernigen, D., Tickle, T.L., Morgan, X.C., Garrett, W.S., and Huttenhower, C. (2013). Computational Meta'omics for Microbial Community Studies. Mol Syst Biol *9*, 666.

Seo, P.J., Lee, A.K., Xiang, F., and Park, C.M. (2008). Molecular and Functional Profiling of Arabidopsis Pathogenesis-Related Genes: Insights into Their Roles in Salt Response of Seed Germination. Plant Cell Physiol *49*, 334-344.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res *13*, 2498-2504.

Shlens, J. (2014). A Tutorial on Principal Component Analysis. arXiv preprint arXiv:14041100.

Shulaev, V. (2006). Metabolomics Technology and Bioinformatics. Brief Bioinform *7*, 128-139.

Shuman, J.L., Cortes, D.F., Armenta, J.M., Pokrzywa, R.M., Mendes, P., and Shulaev, V. (2011). Plant Metabolomics by Gc-Ms and Differential Analysis. Methods Mol Biol *678*, 229-246.

Smart, A.G., Amaral, L.A., and Ottino, J.M. (2008). Cascading Failure and Robustness in Metabolic Networks. Proc Natl Acad Sci U S A *105*, 13223-13228.

Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R., and Siuzdak, G. (2006). Xcms: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. Anal Chem *78*, 779-787.

Spectrometry, A.S.f.M. (2009). Metabolomics Asms Workshop Survey 2009.

Spratlin, J.L., Serkova, N.J., and Eckhardt, S.G. (2009). Clinical Applications of Metabolomics in Oncology: A Review. Clin Cancer Res *15*, 431-440.

Sreekumar, A., Poisson, L.M., Rajendiran, T.M., Khan, A.P., Cao, Q., Yu, J., Laxman, B., Mehra, R., Lonigro, R.J., Li, Y., *et al.* (2009). Metabolomic Profiles Delineate Potential Role for Sarcosine in Prostate Cancer Progression. Nature *457*, 910-914.

Steen, E.J., Kang, Y., Bokinsky, G., Hu, Z., Schirmer, A., McClure, A., Del Cardayre, S.B., and Keasling, J.D. (2010). Microbial Production of Fatty-Acid-Derived Fuels and Chemicals from Plant Biomass. Nature *463*, 559-562.

Steinbeck, C., Conesa, P., Haug, K., Mahendraker, T., Williams, M., Maguire, E., Rocca-Serra, P., Sansone, S.A., Salek, R.M., and Griffin, J.L. (2012). Metabolights: Towards a New Cosmos of Metabolomics Data Management. Metabolomics *8*, 757-760.

Stengel, D.B., Connan, S., and Popper, Z.A. (2011). Algal Chemodiversity and Bioactivity: Sources of Natural Variability and Implications for Commercial Application. Biotechnol Adv *29*, 483-501.

Stranger, B.E., Montgomery, S.B., Dimas, A.S., Parts, L., Stegle, O., Ingle, C.E., Sekowska, M., Smith, G.D., Evans, D., Gutierrez-Arcelus, M., *et al.* (2012). Patterns of Cis Regulatory Variation in Diverse Human Populations. PLoS Genet *8*, e1002639.

Sugimoto, M., Kawakami, M., Robert, M., Soga, T., and Tomita, M. (2012). Bioinformatics Tools for Mass Spectroscopy-Based Metabolomic Data Processing and Analysis. Curr Bioinform *7*, 96-108.

Suhre, K., Wallaschofski, H., Raffler, J., Friedrich, N., Haring, R., Michael, K., Wasner, C., Krebs, A., Kronenberg, F., Chang, D., *et al.* (2011). A Genome-Wide Association Study of Metabolic Traits in Human Urine. Nat Genet *43*, 565-569.

Takahashi, H., Morioka, R., Ito, R., Oshima, T., Altaf-Ul-Amin, M., Ogasawara, N., and Kanaya, S. (2011). Dynamics of Time-Lagged Gene-to-Metabolite Networks of Escherichia Coli Elucidated by Integrative Omics Approach. OMICS *15*, 15-23.

Tautenhahn, R., Cho, K., Uritboonthai, W., Zhu, Z., Patti, G.J., and Siuzdak, G. (2012). An Accelerated Workflow for Untargeted Metabolomics Using the Metlin Database. Nat Biotech *30*, 826-828.

Teusink, B., Baganz, F., Westerhoff, H.V., and Oliver, S.G. (1998). 17 Metabolic Control Analysis as a Tool in the Elucidation of the Function of Novel Genes. In Methods in Microbiology, J.P.B. Alistair, and T. Mick, eds. (Academic Press), pp. 297-336.

Thines, B., Katsir, L., Melotto, M., Niu, Y., Mandaokar, A., Liu, G., Nomura, K., He, S.Y., Howe, G.A., and Browse, J. (2007). Jaz Repressor Proteins Are Targets of the Scf(Coi1) Complex During Jasmonate Signalling. Nature *448*, 661-665.

Tohge, T., Nishiyama, Y., Hirai, M.Y., Yano, M., Nakajima, J., Awazuhara, M., Inoue, E., Takahashi, H., Goodenowe, D.B., Kitayama, M., *et al.* (2005). Functional Genomics by Integrated Analysis of Metabolome and Transcriptome of Arabidopsis Plants over-Expressing an Myb Transcription Factor. Plant J *42*, 218-235.

Vaistij, F., Lim, E.-K., Edwards, R., and Bowles, D. (2009). Glycosylation of Secondary Metabolites and Xenobiotics. In Plant-Derived Natural Products, A.E. Osbourn, and V. Lanzotti, eds. (Springer US), pp. 209-228.

van den Berg, R.A., Hoefsloot, H.C., Westerhuis, J.A., Smilde, A.K., and van der Werf, M.J. (2006). Centering, Scaling, and Transformations: Improving the Biological Information Content of Metabolomics Data. BMC Genomics *7*, 142.

Van Der Kloet, F.M., Bobeldijk, I., Verheij, E.R., and Jellema, R.H. (2009). Analytical Error Reduction Using Single Point Calibration for Accurate and Precise Metabolomic Phenotyping. J Proteome Res *8*, 5132-5141.

Vandepoele, K., Quimbaya, M., Casneuf, T., De Veylder, L., and Van de Peer, Y. (2009). Unraveling Transcriptional Control in Arabidopsis Using Cis-Regulatory Elements and Coexpression Networks. Plant Physiol *150*, 535-546.

Vello, V., Phang, S.-M., Chu, W.-L., Abdul Majid, N., Lim, P.-E., and Loh, S.-K. (2014). Lipid Productivity and Fatty Acid Composition-Guided Selection of Chlorella Strains Isolated from Malaysia for Biodiesel Production. J Appl Phycol *26*, 1399-1413.

Viant, M., and Sommer, U. (2013). Mass Spectrometry Based Environmental Metabolomics: A Primer and Review. Metabolomics *9*, 144-158.

Vidal, E.A., Moyano, T.C., Riveras, E., Contreras-Lopez, O., and Gutierrez, R.A. (2013). Systems Approaches Map Regulatory Networks Downstream of the Auxin Receptor Afb3 in the Nitrate Response of Arabidopsis Thaliana Roots. Proc Natl Acad Sci U S A *110*, 12840-12845.

Vranova, E., Coman, D., and Gruissem, W. (2012). Structure and Dynamics of the Isoprenoid Pathway Network. Mol Plant *5*, 318-333.

Vuckovic, D. (2012). Current Trends and Challenges in Sample Preparation for Global Metabolomics Using Liquid Chromatography-Mass Spectrometry. Anal Bioanal Chem *403*, 1523-1548.

Wang, J., Zhang, Y., Marian, C., and Ressom, H.W. (2012). Identification of Aberrant Pathways and Network Activities from High-Throughput Data. Brief Bioinform *13*, 406-419.

Wang, S.Y., Kuo, C.H., and Tseng, Y.J. (2013a). Batch Normalizer: A Fast Total Abundance Regression Calibration Method to Simultaneously Adjust Batch and Injection Order Effects in Liquid Chromatography/Time-of-Flight Mass Spectrometry-Based Metabolomics Data and Comparison with Current Calibration Methods. Anal Chem *85*, 1037-1046.

Wang, Y., Zhang, X.S., and Chen, L. (2013b). Computational Systems Biology in the Big Data Era. BMC Syst Biol *7 Suppl 2*, S1.

Want, E.J., Masson, P., Michopoulos, F., Wilson, I.D., Theodoridis, G., Plumb, R.S., Shockcor, J., Loftus, N., Holmes, E., and Nicholson, J.K. (2013). Global Metabolic Profiling of Animal and Human Tissues Via Uplc-Ms. Nat Protoc *8*, 17-32.

Weckwerth, W., Loureiro, M.E., Wenzel, K., and Fiehn, O. (2004). Differential Metabolic Networks Unravel the Effects of Silent Plant Phenotypes. Proc Natl Acad Sci U S A *101*, 7809-7814.

Wen, W., Li, D., Li, X., Gao, Y., Li, W., Li, H., Liu, J., Liu, H., Chen, W., Luo, J., *et al.* (2014). Metabolome-Based Genome-Wide Association Study of Maize Kernel Leads to Novel Biochemical Insights. Nat Commun *5*, 3438.

Werner, E., Croixmarie, V., Umbdenstock, T., Ezan, E., Chaminade, P., Tabet, J.C., and Junot, C. (2008a). Mass Spectrometry-Based Metabolomics: Accelerating the Characterization of Discriminating Signals by Combining Statistical Correlations and Ultrahigh Resolution. Anal Chem *80*, 4918-4932.

Werner, E., Heilier, J.F., Ducruix, C., Ezan, E., Junot, C., and Tabet, J.C. (2008b). Mass Spectrometry for the Identification of the Discriminating Signals from Metabolomics: Current Status and Future Trends. J Chromatogr B Analyt Technol Biomed Life Sci *871*, 143-163.

Whittkopp, P.J. (2013). Evolution of Gene Expression. In The Princeton Guide to Evolution, Jonathan B. Losos, David A. Baum, Douglas J. Futuyma, Hopi E. Hoekstra, Richard E. Lenski, Allen J. Moore, Catherine L. Peichel, Dolph Schluter, and M.J. Whitlock, eds. (Princeton University Press), p. 848.

Wijffels, R.H., and Barbosa, M.J. (2010). An Outlook on Microalgal Biofuels. Science *329*, 796-799.

Wikoff, W.R., Anfora, A.T., Liu, J., Schultz, P.G., Lesley, S.A., Peters, E.C., and Siuzdak, G. (2009). Metabolomics Analysis Reveals Large Effects of Gut Microflora on Mammalian Blood Metabolites. Proc Natl Acad Sci U S A *106*, 3698-3703.

Wink, M. (2010). Introduction: Biochemistry, Physiology and Ecological Functions of Secondary Metabolites. In Annual Plant Reviews Volume 40: Biochemistry of Plant Secondary Metabolism (Wiley-Blackwell), pp. 1-19.

Winkel-Shirley, B. (2001). Flavonoid Biosynthesis. A Colorful Model for Genetics, Biochemistry, Cell Biology, and Biotechnology. Plant Physiol *126*, 485-493.

Wishart, D. (2009). Bioinformatics for Metabolomics. In Bioinformatics for Systems Biology, S. Krawetz, ed. (Humana Press), pp. 581-599.

Wishart, D.S. (2008a). Metabolomics: Applications to Food Science and Nutrition Research. Trends Food Sci Tech *19*, 482-493.

Wishart, D.S. (2008b). Quantitative Metabolomics Using Nmr. Trac-Trend Anal Chem *27*, 228-237.

Wishart, D.S., Jewison, T., Guo, A.C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., *et al.* (2013). Hmdb 3.0--the Human Metabolome Database in 2013. Nucleic Acids Res *41*, D801-807.

Xia, J., Mandal, R., Sinelnikov, I.V., Broadhurst, D., and Wishart, D.S. (2012). Metaboanalyst 2.0--a Comprehensive Server for Metabolomic Data Analysis. Nucleic Acids Res *40*, W127-133.

Xin, L., Hu, H.Y., Ke, G., and Sun, Y.X. (2010). Effects of Different Nitrogen and Phosphorus Concentrations on the Growth, Nutrient Uptake, and Lipid Accumulation of a Freshwater Microalga Scenedesmus Sp. Bioresour Technol *101*, 5494-5500.

Xu, W., Grain, D., Le Gourrierec, J., Harscoet, E., Berger, A., Jauvion, V., Scagnelli, A., Berger, N., Bidzinski, P., Kelemen, Z., *et al.* (2013). Regulation of Flavonoid Biosynthesis Involves an Unexpected Complex Transcriptional Regulation of Tt8 Expression, in Arabidopsis. New Phytol *198*, 59-70.

Yeang, C.H. (2009). Integration of Metabolic Reactions and Gene Regulation. Methods Mol Biol *553*, 265-285.

Yen, H.W., Hu, I.C., Chen, C.Y., Ho, S.H., Lee, D.J., and Chang, J.S. (2013). Microalgae-Based Biorefinery--from Biofuels to Natural Products. Bioresour Technol *135*, 166-174.

Yi, X., Du, Z., and Su, Z. (2013). Plantgsea: A Gene Set Enrichment Analysis Toolkit for Plant Community. Nucleic Acids Res *41*, W98-103.

Zambelli, F., Pesole, G., and Pavesi, G. (2009). Pscan: Finding over-Represented Transcription Factor Binding Site Motifs in Sequences from Co-Regulated or Co-Expressed Genes. Nucleic Acids Res *37*, W247-252.

Zelezniak, A., Sheridan, S., and Patil, K.R. (2014). Contribution of Network Connectivity in Determining the Relationship between Gene Expression and Metabolite Concentration Changes. PLoS Comput Biol *10*, e1003572.

Zhang, P., Foerster, H., Tissier, C.P., Mueller, L., Paley, S., Karp, P.D., and Rhee, S.Y. (2005). Metacyc and Aracyc. Metabolic Pathway Databases for Plant Research. Plant Physiol *138*, 27-37.

Zhang, W., Li, F., and Nie, L. (2010). Integrating Multiple 'Omics' Analysis for Microbial Biology: Application and Methodologies. Microbiology *156*, 287-301.

Zhang, W.C., Shyh-Chang, N., Yang, H., Rai, A., Umashankar, S., Ma, S., Soh, B.S., Sun, L.L., Tai, B.C., and Nga, M.E. (2012). Glycine Decarboxylase Activity Drives Non-Small Cell Lung Cancer Tumor-Initiating Cells and Tumorigenesis. Cell *148*, 259-272.

Appendix 1: datPAV- A web-based exploratory data analysis tool

This tool was developed to provide user-friendly informatics programs for analysis of large omics data sets. The input data is a standard two dimensional matrix with rows and columns. This tool explores organization of the data, detect errors and supports basic statistical analyses whose results can be visualized using a suite of programs. The functions are placed as individual modules as well as in a customizable workflow.

Statistical techniques in datPAV help to determine the distribution of data, establish correlations to explore experimental consistency or instrument reliability, perform fold-change analysis and fit a relation between variables using linear regression techniques. My contribution to the development of datPAV is the implementation of statistical data analysis (Table A1) modules written in PERL and integration with the R statistical environment.

Table A1. Statistical tools in datPAV

Process/Analyse data	***With further options to select columns and set parameters for selected processing option***
Mean centering (auto scale)	Subtract column mean and divide by column σ
Pareto scaling	Divide by column standard deviation (σ)
Column normalization	Divides each value by its column maximum
Global normalization	Divides each value by the dataset maximum
Distribution	Computes % distribution for 10 intervals for each column
Filter(moving average)	Useful for time series data
Variable correlation	Depicts dependency between every pair of attributes
Noise correction	Helps in filtering the predefined noise in the data
Fold change (between columns)	Log2 transformation of ratio between two columns
t-test	*P*-value calculation between columns
Auto correlation	Depicts similarity between observations as a function of time
Cross correlation	Depicts dependency between every pair of attributes as a measure of time

Biswas, A., Rao, R., Umashankar, S., Mynampati, K.C., Reuben, S., Parab, G., and Swarup, S. (2011). datPAV--an online processing, analysis and visualization tool for exploratory investigation of experimental data. Bioinformatics *27*, 1585-1586.

Appendix 2: Metabolic reprogramming in Cancer

Lung cancer is the leading cause of cancer-related mortality in both men and women worldwide with over 1 million deaths each year, and a 5-year survival below 15%. Non-small cell lung cancer accounts for approximately 85% of all lung cancers. Our collaborator Dr. Bing Lim (Genome Institute of Sciences) identified the metabolic enzyme glycine decarboxylase (GLDC) as critical for tumour initiating cells (TIC). However the effect of GLDC in initiating metabolic reprogramming of cancer cells was not known.

We performed non-targeted metabolomics on perturbed (overexpression or knock down) cells from human lung cancer cell line (A549), mouse embryonic fibroblasts (3T3), and normal human adult lung fibroblasts (HLF) using LC-MS to identify the global perturbations in the metabolome. Using the statistical techniques discussed in this thesis, I performed metabolomics data analysis which lead to the identification of glycine, serine and threonine metabolism, glycolysis and pyrimidine pathways as being significantly perturbed. These were then verified using targeted metabolomics approach (Tandem MS/MS and Multiple Reaction Monitoring analysis).

Taken together, our metabolomics approach helped us determine that GLDC over expression induces dramatic changes in glycolysis and glycine/serine metabolism which then lead to changes in pyrimidine metabolism to regulate cancer cell proliferation (Zhang et al., 2012). In human patients, GLDC overexpression is significantly associated with higher mortality from lung cancer, and aberrant GLDC expression is observed in multiple cancer types. Our findings helped establish a novel link between glycine metabolism and tumorigenesis, and may provide novel targets for advancing anti-cancer therapy.

Zhang, W.C., Shyh-Chang, N., Yang, H., Rai, A., Umashankar, S., Ma, S., Soh, B.S., Sun, L.L., Tai, B.C., Nga, M.E., et al. (2012). Glycine decarboxylase activity drives non-small cell lung cancer tumor-initiating cells and tumorigenesis. Cell *148*, 259-272.