

Integrated omics analyses of the
interaction between
Brachypodium distachyon and
Magnaporthe oryzae

Jasen Peter Finch BSc.

A thesis submitted at the Institute of Biological,
Environmental and Rural Sciences,
Aberystwyth University,
for the degree of Doctor of Philosophy.

September 2016

Declaration

Word count of thesis: 41,531

This work has not previously been accepted in substance for any degree and is not concurrently submitted in candidature for any degree.

Signed.....(Candidate)

Date.....

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated. Where *correction services have been used, the extent and nature of the correction is clearly marked in a footnote(s).

Other sources are acknowledged by footnotes giving explicit references A bibliography is appended.

Signed.....(Candidate)

Date.....

[*this refers to the extent to which the text has been corrected by others]

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed.....(Candidate)

Date.....

NB: Candidates on whose behalf a bar on access (hard copy) has been approved by the University should use the following version of Statement 2:

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loans after expiry of a bar on access approved by Aberystwyth University.

Signed.....(Candidate)

Date.....

Abstract

Fungal biotrophic phytopathogens such as *Magnaporthe oryzae*, the causal agent of rice blast disease, are becoming increasingly important in crop losses worldwide with the onset of anthropogenic climate change and monoculture farming practices. The use of model organisms such as the grass species *Brachypodium distachyon* provides the opportunity undertake large-scale integrated omics analyses that would otherwise be infeasible. These can provide insights into the system-wide responses of plants to pathogen infection as well as identify areas of the host plant system that could be targets for pathogen manipulation.

A spectral binning method for high resolution metabolome fingerprinting was developed along with the R package binneR as a software implementation for routine application of the method. This was applied to investigate experimental control and robustness in plant-pathogen interactions, with the development of a new inoculation strategy for appropriately controlling inoculum derived responses unrelated to *M. oryzae* pathogenesis. Large-scale, high resolution metabolomic fingerprinting and profiling, along with RNA-Seq transcriptomic analyses, were conducted for the interaction between *B. distachyon* and *M. oryzae* and identified dynamic changes during the pre-symptomatic biotrophic phases. Both data and knowledge driven omics integration strategies identified associations between the transcriptomic and metabolomic changes during the interaction, with chloroplasts and nitrogen metabolism found as key response areas. A disease resistance locus Rbr1 was identified using QTL analyses on chromosome 4 of *B. distachyon* for resistance to *M. oryzae* utilising computer vision based phenotyping. Eight candidate NB-LRR resistance genes were found within this locus.

Acknowledgements

Firstly I would like to thank Professor John Draper for the plethora of opportunity he has provided throughout my studies. Thank you to Dr Manfred Beckmann for his treasure trove of experience and quirky ideas. A special thanks to Dr Thomas Wilson his invaluable assistance in debugging code and for ironing out the creases of understanding. Thank you to Kathleen Taillart for always ensuring my samples could be run when needed. Also thanks to Rob Darby for always sourcing necessary kit when needed.

Thanks to Dr Matthew Moscou and Jan Bettgenhaeuser of the Sainsbury Laboratory and Professor John Doonan for providing seed for mapping populations and for their advice and guidance in all things genetic. Thanks to Tom Thomas for the cups of tea and his ever dependable *Brachypodium* cultivation. Thanks to Dr Martin Vickers for his support in the use of the IBERS HPC and image analysis advice.

Thank you to the Earlham Institute for the exceptional speed and quality of sequencing services provided.

Finally I would like to thank all my family and friends for their encouragement and belief and for putting up with my nonsensical utterances about grass and fungi over the past few years.

Contents

1	General introduction	1
1.1	The host: <i>Brachypodium distachyon</i>	2
1.2	The pathogen: <i>Magnaporthe oryzae</i>	3
1.3	The infection of cycle of <i>M. oryzae</i>	3
1.4	Pathogen infection strategies and plant defences at the molecular level	5
1.4.1	The ‘zigzag’ model of the plant pathogen interactions . . .	6
1.4.2	Pathogen subversion by effectors	7
1.4.3	Plant defences against pathogens	8
1.5	Metabolite and gene expression changes associated with biotrophic fungal infection of cereal hosts	10
1.6	The modern plant pathologist’s omics toolbox	12
1.7	Systems biology and biological networks	14
1.8	Aims	15
2	General materials and methods	18
2.1	<i>B. distachyon</i> growth conditions	18
2.2	<i>M. oryzae</i> maintenance and growth conditions	19
2.3	<i>M. oryzae</i> inoculum preparation and <i>B. distachyon</i> inoculation . .	19
2.4	Harvesting of plant material	20
2.5	Large-scale inoculations to investigate the pre-symptomatic phases of the <i>B. distachyon</i> and <i>M. oryzae</i>	20
2.6	Metabolite Extraction	21
2.7	FIE-HRMS analysis and data processing	21

2.8	LC-HRMS analysis and data processing	22
2.9	FIE-HRMS and LC-HRMS data mining	24
2.10	Putative metabolite annotation	24
2.10.1	FIE-HRMS	24
2.10.2	LC-HRMS	25
2.11	RNA extraction, library preparation and sequencing of RNA-Seq samples	25
2.12	Transcriptomic data mining	26
3	Spectral binning for untargeted FIE-HRMS metabolome fin- gerprinting	27
3.1	Introduction	27
3.1.1	High resolution mass spectrometry	28
3.1.2	Spectral binning for signal processing	30
3.1.3	Untargeted FIE-MS metabolome fingerprinting	30
3.2	Aims	32
3.3	Materials and Methods	33
3.3.1	Preparation and mass spectrometry (MS) analysis of exam- ple <i>B. distachyon</i> sample	33
3.3.2	Performance testing for the R package binneR	34
3.3.3	Investigation of missing value imputation	34
3.4	Results and Discussion	34
3.4.1	Development of the R package binneR	34
3.4.2	Optimal bin size for FIE-HRMS metabolome fingerprinting	37
3.4.3	FIE-HRMS data pre-treatment	39
3.4.4	Metabolite annotation using FIE-HRMS metabolome fin- gerprinting data	48
3.4.5	A general workflow for spectral binning based FIE-HRMS metabolomic fingerprinting analyses	51
3.5	Concluding remarks	54

4	Experimental control and robustness in omics analyses of plant-pathogen interactions	57
4.1	Introduction	57
4.1.1	Experimental considerations for omics experiments involving plant-pathogen interactions	58
4.1.2	Random Forest for metabolomic data mining	59
4.1.3	Robustness and validation in omics analyses	61
4.2	Aims	63
4.3	Materials and Methods	64
4.3.1	Inoculations to investigate the inoculum related metabolomic changes in <i>B. distachyon</i> as a result of <i>M. oryzae</i> inoculation	64
4.3.2	Inoculum preparation and LC-MS analyses to investigate the effect of centrifugation on inoculum constituents	65
4.3.3	Independent inoculations of <i>B. distachyon</i> with <i>M.oryzae</i> and Random Forest classification using an external validation re-sampling strategy	65
4.4	Results and Discussion	66
4.4.1	Experimental control of the <i>B. distachyon</i> and <i>M. oryzae</i> interaction	66
4.4.2	Assessing the robustness of patho-system metabolomic changes	75
4.4.3	Potential sources of experimental variability and feature instability	84
4.5	Concluding remarks	86
5	Metabolomic and transcriptomic analyses of the pre-symptomatic phases of the <i>B. distachyon</i> and <i>M. oryzae</i> interaction	89
5.1	Introduction	89
5.1.1	Untargeted LC-MS profiling for metabolomic investigations	90
5.1.2	Metabolomic analyses of plant pathogen interactions . . .	92
5.1.3	Whole transcriptome sequencing analyses using RNAseq transcriptomics	93

5.1.4	Transcriptomic analyses of plant pathogen interactions . .	96
5.2	Aims	96
5.3	Materials and Methods	97
5.3.1	Inoculation and harvesting of plant tissue	97
5.3.2	Metabolite and RNA extraction	98
5.3.3	Mass spectral metabolomic analyses	98
5.3.4	Metabolomic data mining	98
5.3.5	RNA-Seq and transcriptomic data mining	99
5.4	Results and discussion	99
5.4.1	Omics-level differences between the <i>B. distachyon</i> ecotypes ABR6 and Bd21	99
5.4.2	Metabolomic changes during early phases of the <i>B. dis-</i> <i>tachyon</i> and <i>M. oryzae</i> interaction	100
5.4.3	Transcriptomic changes during early phases of the <i>B. dis-</i> <i>tachyon</i> and <i>M. oryzae</i> interaction	119
5.5	Concluding remarks	130
6	Omics integration to elucidate key pathways in the <i>B. dis-</i> <i>tachyon</i> and <i>M. oryzae</i> interaction	131
6.1	Introduction	131
6.1.1	Strategies for integrating omics data	132
6.1.2	Omics integration for plant stress responses	134
6.2	Aims	136
6.3	Materials and Methods	137
6.3.1	Metabolomic and transcriptomic data preparation for inte- grative analyses	137
6.3.2	Integrative correlation network analysis and pathway map- ping of metabolomic and transcriptomic data	137
6.4	Results and Discussion	138
6.4.1	Integration of metabolomic analyses	138
6.4.2	Integration of metabolomic and transcriptomic analyses . .	144

6.4.3	Key pathways in the early phases of the interaction between <i>M. oryzae</i> and <i>B. distachyon</i>	154
6.5	Concluding Remarks	162
7	Identification of the Rbr1 disease resistance locus using com- puter vision based phenotyping	164
7.1	Introduction	164
7.1.1	Quantitative plant phenotyping using computer vision . .	165
7.1.2	The genetic basis of plant resistance to disease	167
7.1.3	Linking phenotype to genotype: QTL mapping to identify disease resistance loci	169
7.2	Aims	172
7.3	Materials and Methods	174
7.3.1	Inoculation of F _{4:5} ABR6 x Bd21 population with rice blast	174
7.3.2	Image acquisition and processing	174
7.3.3	Manual scoring and QTL analyses	174
7.4	Results and Discussion	177
7.4.1	Computer vision based analysis for quantitatively assessing <i>B. distachyon</i> and <i>M. oryzae</i> interaction phenotypes	177
7.4.2	The identification of linked loci for rice blast disease resis- tance <i>B. distachyon</i>	186
7.5	Concluding remarks	191
8	General conclusions and future work	192
8.1	Summary of general conclusions	192
8.2	Future work	195
	References	199

List of Figures

1.1	The host and pathogen	2
1.2	The initial host colonisation phases of <i>Magnaporthe oryzae</i>	5
3.1	An example <i>B. distachyon</i> FIE-MS ion chromatogram	32
3.2	Mass deviation of the base peak malic acid $[M-H]^{1-}$ between scans and injections	40
3.3	Example of peak splitting at bin width 0.001	41
3.4	The example <i>B. distachyon</i> sample fingerprint m/z intensities and variability	44
3.5	The effect of TIC normalisation and \log_{10} transformation on fin- gerprint m/z	45
3.6	Distributions of bin occupancy and signal intensity	49
3.7	A general workflow for FIE-HRMS analyses	55
4.1	Example m/z trends of explanatory features between inoculation treatment comparisons	71
4.2	Negative mode base peak ion chromatograms of inoculum treat- ments analysed by LC-HRMS	76
4.3	Venn diagrams comparing explanatory features from individual in- oculation experiments	77
4.4	Feature similarity between experiment comparisons	82
4.5	Example box plots of feature experimental variability at 60 hpi . .	83
4.6	Heat map of negative mode ABR6 m/z trends at 60 hpi	87
5.1	Metabolome and transcriptome differences between the <i>B. dis-</i> <i>tachyon</i> ecotypes ABR6 and Bd21.	101

5.2	Random Forest margins for control and infected tissue comparisons between 0 and 60 hpi for both ABR6 and Bd21	104
5.3	Intersection plot of FIE-HRMS explanatory m/z identified by Ran- dom Forest in comparisons between control and infected tissue at each time point	113
5.4	Intersection plot of LC-HRMS explanatory m/z identified by Ran- dom Forest in comparisons between control and infected tissue at each time point	114
5.5	K -means clustering of ABR6 explanatory m/z identified using FIE-HRMS	115
5.6	K -means clustering of Bd21 explanatory m/z identified using FIE- HRMS	116
5.7	K -means clustering of ABR6 explanatory m/z identified using LC- HRMS	117
5.8	K -means clustering of Bd21 explanatory m/z identified using LC- HRMS	118
5.9	Intersection plot of RNAseq DEGs for in comparisons between control and infected tissue at each time point	121
5.10	K -means clustering to identify co-expression clusters of ABR6 DEGs	126
5.11	K -means clustering to identify co-expression clusters of Bd21 DEGs	127
5.12	Functional enrichment analysis of ABR6 co-expression clusters . .	128
5.13	Functional enrichment analysis of Bd21 co-expression clusters . .	129
6.1	Correlation network of ABR6 metabolomic analyses	142
6.2	Correlation network of Bd21 metabolomic analyses	143
6.3	Correlation network of ABR6 metabolomic and transcriptomic anal- yses	147
6.4	Correlation network of Bd21 metabolomic and transcriptomic anal- yses	148
6.5	Mapman visualisations of metabolism for 12 hpi metabolite and gene expression changes	151

6.6	Mapman visualisations of metabolism for 24 hpi metabolite and gene expression changes	152
6.7	Mapman visualisations of metabolism for 48 hpi metabolite and gene expression changes	153
6.8	Profiles of explanatory photosynthesis related metabolites and transcripts	157
6.9	Profiles of explanatory lipoxygenases, NADPH oxidases and unsaturated fatty acids	158
6.10	Profiles of explanatory nitrogen metabolism transcripts and metabolites	161
7.1	A general workflow in computer vision analyses	166
7.2	Rice blast disease response phenotypes in the <i>B. distachyon</i> ecotypes Bd21 and ABR6	172
7.3	The ABR6 x Bd21 F ₄ genetic map	173
7.4	RGB image channels of and example ABR6 leaf showing incompatible disease symptoms	180
7.5	Lesion segmentation using image subtraction and thresholding . .	181
7.6	Segmentation results of Rice Blast disease response phenotypes. .	182
7.7	Ecotype comparisons of density, size and shape features extracted from leaf images	184
7.8	Ecotype comparisons of colour based features extracted from leaf images	184
7.9	Ecotype comparisons of texture based features extracted from leaf images	185
7.10	Validation of lesion area measurements using response scores in the ABR6 x Bd21 mapping population	187
7.11	QTL analysis of lesion area data using the ABR6 x Bd21 F ₄ genetic map	189
7.12	LOD support intervals of significant markers found in the Rbr1 locus	190

List of Tables

3.1	Computational requirements for spectral binning of FIE-HRMS metabolomic fingerprinting data using the R package binneR . . .	36
3.2	Spectral binning width: variable numbers and missing values . . .	40
3.3	The effect of class occupancy filtering and kNN imputation on binary classification	49
3.4	Common m/z relationships found within FIE-MS metabolomic fin- gerprints	52
3.5	Bin n132 correlations	53
3.6	Bin n132.03 correlations	53
4.1	Random forest classification results for comparisons of inoculation treatment responses	69
4.2	Putative annotations of inoculum associated metabolites	73
4.3	ABR6 random forest classification performance using an external validation resampling approach	79
4.4	Bd21 random forest classification performance using an external validation resampling approach	79
4.5	Random Forest classification results of comparisons of 0 hpi control treatments between experiments	86
5.1	Putative annotations of FIE-HRMS explanatory m/z	108
5.2	Putative annotations LC-HRMS explanatory m/z	110
6.1	Correlations of explanatory m/z found in both FIE-HRMS and LC-HRMS analyses	141

7.1	Manual scoring of rice blast disease responses in the ABR6 x Bd21	
	RIL population	176
7.2	NB-LRRs identified within the 2-LOD support interval of the Rbr1	
	locus	190

List of Equations

3.1 Mass Resolving Power	28
4.1 Probability of variable selection at a given node	61
4.2 Binomial probability of selection of a given feature across all nodes in a forest	61
4.3 Jaccard's Index	63
4.4 Canberra Distance	63

Abbreviations

m/z mass to charge ratio.

amu atomic mass units.

ANOVA analysis of variance.

AUC area under the receiver operator characteristic (ROC) curve.

CIM composite interval mapping.

cM centiMorgans.

DEG differentially expressed gene.

DI direct injection.

FIE flow infusion electrospray.

FPR false positive rate.

FT Fourier transform.

GC gas chromatography.

GO gene ontology.

hpi hours post inoculation.

HR high resolution.

ICR ion-cyclotron resonance.

IR infrared.

kNN k-nearest neighbour.

LC liquid-chromatography.

LOD logarithmic of odds.

MS mass spectrometry.

NB-LRR nucleotide-binding leucine-rich repeat.

NBS nucleotide-binding site.

NGS next generation sequencing.

NMR Nuclear Magnetic Resonance.

PAMP pathogen-associated molecular pattern.

PC-LDA principle component linear discriminant analysis.

PCA principle component analysis.

PCR polymerase chain reaction.

PDA potato dextrose agar.

ppm parts per million.

QC quality control.

QTL quantitative trait loci.

Rbr1 Rice Blast Resistance 1.

RIL recombinant inbred lines.

RNA-Seq RNA sequencing.

ROC receiver operator characteristic.

ROI regions of interest.

ROS reactive oxygen species.

RPKM reads per kilobase per million mapped reads.

RuBisCo ribulose-bisphosphate carboxylase.

SIM simple interval mapping.

SNP single nucleotide polymorphism.

TIC total ion count.

TOF time of flight.

Preface

This thesis presents research with the intention of integrating omics analyses for the investigation of the interaction between *Brachypodium distachyon* and *Magnaporthe oryzae*. **Chapter 1** is a general introduction to cover key themes in plant pathogen interactions and introduce omics technologies. **Chapter 2** contains the general methodologies used throughout the thesis. **Chapter 3** presents the development of a spectral binning method for high resolution metabolome fingerprinting. **Chapter 4** investigates elements of experimental control and robustness in omics analyses of plant-pathogen interactions. **Chapter 5** presents results of metabolomic and transcriptomic analyses of the pre-symptomatic phases of the interaction between *B. distachyon* and *M. oryzae*. **Chapter 6** aims to build on the results of **Chapter 5**, by directly integrating data from these omics analyses to identify key pathways involved in the interaction. **Chapter 7** presents the identification of a disease resistance locus to *M. oryzae* in *B. distachyon*, utilising computer vision based phenotyping. Finally **Chapter 8** summarises the general conclusions and presents opportunities for future research.

In the interest of open and reproducible research, only open-source bioinformatics software has been used. This thesis has been written using a combination of L^AT_EX and the R package knitr, to allow the direct integration of both experimental data and analysis code into its compilation, ensuring both reproducibility and transparency. All L^AT_EX and R source code and data underlying the figures and tables presented here can be found in Appendix A along with compilation instructions. All other analysis code and workflows used can be accessed at <https://github.com/jasenfinch>.

Chapter 1

General introduction

With the onset of anthropogenic climate change, widespread monoculture farming practices and increasing human populations, plant pathogens present an increasing threat to global food security. Rice blast, the most destructive disease of rice, destroys between 10 and 30% annually of the global rice crop. This is enough rice to feed the equivalent of the UK population each year (Dean et al. 2005).

Not only has there been greater incidence of emergent pathogens such as *Phytophthora ramorum*, but there is also an increase in virulence of already established pathogens (Fisher et al. 2012). It is vital to understand the molecular basis of how pathogens colonise plant tissues and in turn how plants defend themselves so that appropriate methods of control and resistant varieties can be developed.

Model organisms have had a central role across the biological sciences in investigating fundamental aspects of biological systems. From the use of *Saccharomyces cerevisiae* for investigating genetics and cell biology to the wide application of *Arabidopsis thaliana* for plant development and light sensing. The field of plant pathology is no different. Model organisms are essential for investigating plant-pathogen interactions. Their use can simplify practical constraints such as specific growth needs, allowing the investigation of fundamental questions that can be inferable upon other plant-pathogen interactions.

The interaction between the model grass species *Brachypodium distachyon* and the model plant pathogen *Magnaporthe oryzae* has been developed as a model

(a) *Brachypodium distachyon*



(b) *Magnaporthe oryzae*



Figure 1.1: **The host and pathogen.** a) 21 day old *B. distachyon*. b) 16 day old *M. oryzae* cultures on PDA media.

interaction to study dynamic host-pathogen interactions (Parker et al. 2008).

1.1 The host: *Brachypodium distachyon*

B. distachyon is temperate grass species whose natural range spans from the Mediterranean to the Middle East. It has status as a model grass species due to its small size, fast growth rate, short life cycle minimal growth requirements and is readily genetically transformable (Draper et al. 2001; Vogel et al. 2006). *B. distachyon* is diploid with 5 chromosomes and genome sequencing of the ecotype Bd21 has revealed a small genome of only 272 Mb (Brachypodium Initiative 2010). Phylogenetically it has a close proximity to important crop species such as wheat, barley and rice.

Well over 40 ecotypes have been collected that show variation in flowering time, vernalisation requirements and domestication traits such spikelet and grain morphology. It is self-pollinating making it suitable for developing mapping populations for the investigation of segregating traits. This has led to its application as a model for bioenergy crops whose genomes are normally large and complex (Opanowicz et al. 2008)

B. distachyon also shows diversity in responses of ecotypes to a number of important phytopathogens that cause significant crop losses. These include the rust disease causing *Puccinia* and *Magnaporthe oryzae* (Draper et al. 2001).

1.2 The pathogen: *Magnaporthe oryzae*

M. oryzae, the causal agent of rice blast disease, is a haploid ascomycete fungus. It has widespread occurrence and is present in all rice growing regions throughout the world.

M. oryzae is heterothallic with two mating types present, MAT1-1 and MAT1-2, the pairing of isolates carrying opposite mating types will form sexual fruiting bodies. It can be transformed using a number of selective markers including complementary auxotrophic markers. This makes it readily amenable to genetic analyses.

It has model phytopathogen status and its genome has been sequenced yielding a 40.3Mb genome (Dean et al. 2005). *M. oryzae* is able to infect all aerial parts of rice and its spore germination and infection mechanisms are well characterized. It is also able to infect roots and systematically spread through the tissue of susceptible hosts (Sesma and Osbourn 2004).

Formally known as *M. grisea*, *M. oryzae* has been identified as a distinct species through multilocus gene genealogy and host preference. *M. grisea* is associated with specificity to the grass genus *Digitaria* (Couch and Kohn 2002). *Magnaporthe* has a wide host range and has been reported to occur on more than 50 grass species including other important cereal crops such as wheat and barley. Its interaction with that of *B. distachyon* has been found to closely resemble those with rice (Routledge et al. 2004). The *B. distachyon* ecotype Bd21 shows a compatible response (susceptible) and the ecotype ABR6 shows an incompatible response (resistance) with *M. oryzae*.

1.3 The infection of cycle of *M. oryzae*

The infection cycle of *M. oryzae* begins with aerially dispersed three-celled conidia landing upon the leaf cuticle. Spore germination is triggered by high humidity or the presence of dew. A strong adhesive is secreted from the spore tip that is used to stick to the leaf surface. This contains α -linked-mannosyl and glucosyl residues

as well as protein and lipid components. These are released upon hydration of the spore.

The emergence of a germ tube quickly ensues within 30 minutes of attachment (Hamer et al. 1988). There is a short period of host recognition where the presence of cutin and lipid monomers is sensed prior to committing to appressorium development. The spore will subsequently develop into a dome shaped appressorium; the structure with which the fungal pathogen breaches the cuticle layer and enters the host tissue.

Turgor pressure is generated within the appressorium by the accumulation of glycerol, drawing in water by osmosis. This can generate up to 8 MPa of pressure and forces a penetration peg through the cell wall (de Jong et al. 1997). The conidium then undergoes programmed cell death, which is essential to the penetration process (Veneault-Fourrey et al. 2006). Penetration of the host cell wall allows the proliferation of a bulbous fungal invasion hyphae and invagination of the host cell membrane. This differentiates into a specialised feeding structure known as a haustoria; the nutritional interface through which host manipulation can occur (O’Connell and Panstruga 2006). It begins the biotrophic phase of the colonisation process and occurs within 30 hours of initial spore germination. These initial phases of the interaction are shown in Figure 1.2.

From the initial colonisation of the primary host cell, secondary spread into adjacent cells occurs through plasmodesmata and *M. oryzae* proliferates through the host tissue. Then, around 4 days after spore germination, *M. oryzae* shifts into a necrotrophic phase and will actively kill host cells producing the characteristic blast lesions. This shift to necrotrophy is thought to release a burst of nutrients that enable it to sporulate (Talbot 1995). Sporulation occurs under high humidity and spores are disseminated by wind, completing its infection cycle.

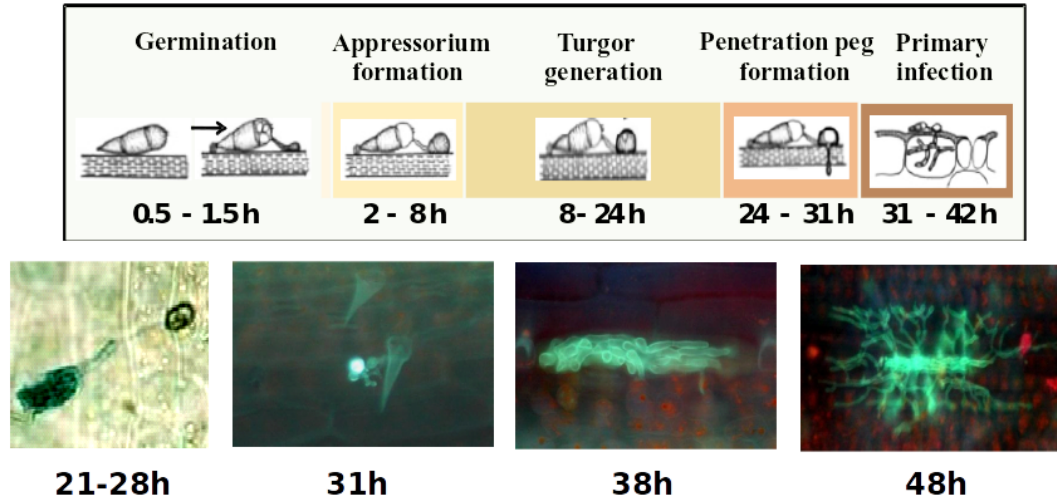


Figure 1.2: The initial host colonisation phases of *Magnaporthe oryzae*.

1.4 Pathogen infection strategies and plant defences at the molecular level

Plant pathogens have a diverse range of life strategies across all plant tissue types. Biotrophic pathogens will maintain host viability during infection, subverting it in order to gain nutrition. Necrotrophs actively kill host tissue, deriving nutrition from breakdown products released during this process. Hemi-biotrophs such as *M. oryzae* have both biotrophic and necrotrophic phases to their life cycle with biotrophy during initial infection phases, switching to necrotrophy prior to sporulation. Bacteria will inhabit apoplastic spaces, entering through stomatal pores or wounding. Alternatively many fungal pathogens will directly penetrate epidermal cells, forming intimate interfaces with the plant cells.

Unlike animals, plants lack an adaptive mobile immune system of specialized cells, the only purpose of which is to neutralise invading pathogens. Instead they rely upon the innate immunity of each individual cell to be able to appropriately recognise and defend against these diverse pathogen strategies (Dodds and Rathjen 2010). These defences consist of both constitutive and inducible responses.

Successful pathogens have to successfully suppress plant defences and subvert the host system in order to gain the nutrients required to complete their life cycle. They do this through the use of specialised effector proteins that are secreted into

the host tissue during infection.

1.4.1 The ‘zigzag’ model of the plant pathogen interactions

Plant-pathogen interactions can be represented by a four phased ‘zigzag’ model (Jones and Dangl 2006). This describes the ‘zigzag’ between successful pathogen colonisation and successful host resistance as infection proceeds.

The initial phase is the recognition of pathogen-associated molecular pattern (PAMP)s by the plant. PAMPs are highly conserved molecules that are widely distributed between microbial species. They will often carry out essential functions within the pathogen species, but will be absent from the host species. Identified PAMPs include flagellin in bacteria and chitin in fungi. The ability of plants to recognise these PAMPs will vary between species, depending on the co-evolution of species with particular pathogens (Chisholm et al. 2006).

The recognition of PAMPs is mediated by transmembrane pattern recognition receptors. These trigger signalling cascades that initiate host defences in response to the pathogen presence. The characterisation of pattern recognition receptors has proved difficult, however they have been hypothesised to be analogous to Toll-like receptors in mammals. These contain extracellular leucine rich repeat and intracellular TIR domains that mediate signal transduction upon pattern recognition (Ingle, Carstens, and Denby 2006). FLS2, a receptor like kinase containing an extracellular leucine-rich repeat domain in *Arabidopsis* has been shown to bind the flagellin constituent flg22 (Chinchilla et al. 2006).

The second phase is essential for successful pathogens and that is the suppression of the PAMP triggered immune responses by the pathogen. Effector proteins are deployed into the host tissue that allow the pathogen to overcome the plant defences and continue with infection.

The third phase is the specific recognition of the pathogen effectors intracellularly by host nucleotide-binding leucine-rich repeat (NB-LRR) proteins and results in effector triggered immunity. NB-LRRs and the genetic basis of plant

resistance to disease is discussed further in Section 7.1.2. Any subsequent phases will be a repetition of pathogen suppression of defences and host recognition as a result of the plant and pathogen arms race during co-evolution.

1.4.2 Pathogen subversion by effectors

Effectors are molecules secreted by phytopathogens that modulate the interaction between a pathogen and its host. Their role is to suppress host defences and subvert the host system for nutritional gain. They can function both inter and intra cellularly.

Fungal plant pathogens have extensive effector repertoires, with the functional constituents dependent on the lifestyle of the pathogen. Biotrophic fungi require effectors that will suppress host defences; whereas necrotrophs will require effectors that will directly kill plant cells. The proportions of cell wall degrading enzymes in necrotrophs or hemi-biotrophs are higher than in biotrophs but are comparable to those of saprotrophs. *M. oryzae* has a repertoire of over 1500 putatively secreted proteins (Lo Presti et al. 2015).

Knowledge is still limited on how fungal plant pathogens translocate effector proteins into host cells. *M. oryzae* has two distinct secretion systems for translocating effectors within host tissues. Effectors that are secreted extracellularly follow conventional secretory pathways of filamentous fungi. However, those that are translocated into the host tissue utilise a biotrophic interfacial complex. This is a plant-derived interfacial structure, lying outside of the fungal cell membrane and wall. It is associated with a secretion system that involves exocyst components and the Sso1 t-SNARE to deliver effectors into the host cell (Giraldo et al. 2013).

Effectors have diverse functions for both suppressing host defences as well as subverting host cells. Those such as the *M. oryzae* effector Slp1 can bind to chitin and is able to suppress chitin elicited PAMP triggered immunity (Mentlak et al. 2012). The Avr-Piz-t effector suppresses the generation of reactive oxygen species (ROS) during PAMP triggered responses by inhibiting the rice RING E3

ubiquitin ligase APIP6 (Park et al. 2012).

Ustilago maydis secretes high amounts of chorismate mutase into the cytoplasm of the host plant cells during colonisation. This reduces the levels of chorismate within the cells, which serves as a precursor to the salicylic acid synthesis. Therefore, the cell's ability to utilize salicylic acid for defence responses is diminished and promotes the virulence of *U. maydis* (Djamei et al. 2011).

In *M. oryzae*, an avirulence gene ACE1 (a polyketide synthase) is up-regulated during penetration phases; however, it is not predicted to be secreted and localises in appressorial cytoplasm. This suggests that an unknown secondary metabolite that is synthesised by Ace1 could be acting as an effector (Bohnert 2004). Secondary metabolite effectors are also speculated to exist in *Colletotrichum higginsianum*. Twelve secondary metabolism clusters have been found to be induced before penetration and during biotrophy (O'Connell et al. 2012).

Other fungal pathogen effector targets include protease inhibition and the disruption of plant immune receptors. No effectors have yet been identified that redirect plant metabolism that would allow fungal pathogens to meet their nutritional needs (Lo Presti et al. 2015).

1.4.3 Plant defences against pathogens

Plants have a number of defences, both constitutive and induced, in order defend themselves against pathogen invasion. The plant cuticle and cell wall presents the first obstacle that a prospective fungal phytopathogen must overcome in order to begin colonising the host tissue. Thick cuticle and wax layers can provide a defence against fungal pathogens that directly penetrate the host tissue (Hématy, Cherk, and Somerville 2009).

Plants cells are also able to induce the formation of papillae at sites of active pathogen penetration. Although the exact form and constituents of papillae vary between plant species, these are areas of cell wall thickening that form plugs or collars inhibit the penetration of the pathogen or restrict the activity of haustorial feeding structures (Meyer et al. 2009). Typical constituents in the formation

of papillae include callose, phenolics, phenolic polyamines as well as pectin and xyloglucans. ROS are also required in this process with hydrogen peroxide essential to allowing phenolic crosslinking for cell wall strengthening (Mellersh et al. 2002).

Arguably the most important of the induced host defences is that of the hypersensitive response. Host plant cells being colonised by a pathogen will elicit a form of programmed cell death in order to contain the pathogen and prevent its further spread into surrounding tissues. There are forms of both micro and macro responses where many or few cells will be involved. The hypersensitive response varies among plant species and pathogen interaction but can be associated with granulation of the cell cytoplasm, an oxidative burst through the generation of ROS and the appearance of localised, non-spreading necrotic lesions (Mur et al. 2008). The response is induced through the recognition of pathogen effectors during initial pathogen infection.

ROS form an essential part of plant defence responses to pathogens. There is rapid accumulation after pathogen recognition that is known as an oxidative burst. This substantially alters the cellular environment, altering pH, ion fluxes and protein phosphorylation. It can create an antimicrobial environment, acting directly against pathogen colonisation. However, ROS also acts in cellular signalling and can induce the hypersensitive response and the expression of defence related genes (Desikan, Neill, and Hancock 2000).

ROS can be produced by a range of cellular processes and includes superoxide radicals, hydrogen peroxide, hydroxyl radical and nitric oxide. Superoxide radicals can be produced by NADPH oxidase, leakage from electron transport chains in both mitochondria and chloroplasts and xanthine oxidase. Peroxidases can form hydrogen peroxide from superoxide dismutase or dismutation can occur spontaneously (O'Brien et al. 2012).

Phytoalexins are low molecular weight secondary metabolites with antimicrobial activity, that are synthesised upon the recognition of pathogen attack. They are a heterogeneous and diverse group of compounds, with repertoires varying

greatly between plant species and their induction is also pathogen specific.

1.5 Metabolite and gene expression changes associated with biotrophic fungal infection of cereal hosts

There is evidence to suggest commonality in the reprogramming of cereal host metabolism during biotrophic fungal infection. Similar metabolite changes between different cereal hosts interacting with the same fungal biotroph as well as between different fungal biotrophs and the same host. Parker et al. (2009) found identical patterns of metabolomic change during compatible interactions between *M. oryzae* and rice, barley and *B. distachyon*. Accumulations of malate and polyamines suggested a disruption to the generation of defensive ROS. Also accumulations of quinate and non-polymerised lignin precursors suggested diversion of the shikimic acid pathway and modulation of the phenylpropanoid pathway.

Voll (2011) also found commonality between metabolite changes in compatible interactions, but used a diverse set of fungal biotroph pathosystems; *U. maydis* and *C. graminicola* on maize and *Blumeria graminis* f.sp. hordei on barley. Stages of early and late biotrophic phases could be identified associated between the pathosystems using natural clustering based on 42 water soluble metabolites. Increases in glutamine, asparagine and glucose and reductions in phosphoenol pyruvate and 3-phosphoglycerate were found to be common post-penetration host changes between the pathosystems.

Interestingly, transcriptome data for these pathosystems was unable to discriminate the biotrophic phases in a way similar to the metabolites. However, alterations to genes involved in the TCA cycle, nucleotide energy metabolism and amino acid metabolism were found to be consistent between the interactions. This highlighted the importance of metabolic energy and alterations of amino

acid pools during early biotrophy.

Alongside these commonalities between pathosystems, there have been a diverse range of gene expression and metabolite pathways associated with the fungal infection of cereal hosts. These include areas of primary metabolism, secondary metabolism and hormone signalling. Increases in both sucrose and hexoses have been found around the infection sites of compatible interactions between *B. graminis* and barley (Swarbrick, Schulze-Lefert, and Scholes 2006). This was associated with the preferential uptake of hexoses by the fungal pathogen and the subsequent transition of infected host cells from carbon sources to carbon sinks. Pools of amino acids such as alanine were found to be less affected in incompatible interactions of *M. oryzae* and rice compared to compatible interactions (Jones et al. 2011).

Fungal biotroph infection induces many secondary metabolite pathways that include genes and metabolites involved in the production of phytoalexins, antioxidants and lignin precursors. Biosynthesis of phytoalexins involves the induction of a number of primary metabolic pathways such as the shikimate pathway and acetate-malonate pathways. They are commonly derived from the phenylpropanoids, flavonoids and isoflavonoids, sesquiterpenes and polyketides (Ahuja, Kissen, and Bones 2012). Momilactones are diterpene compounds that have been found to be synthesised in rice, in response to *M. oryzae* infection and severely restricted the pathogen growth *in vitro* (Hasegawa et al. 2010).

Antioxidants such as glutathione, ascorbate and polyphenols such as flavanoids have been shown to be important regulators of oxidative stress in many plant pathogen interactions. Doehlemann et al. (2008) found an induction of seven glutathione S-transferase genes after just 12 hours of infection of maize by *U. maydis*. Elevated levels of glutathione were found 24 hours post inoculation (hpi) and a high reduction state of the glutathione pool was also found to be maintained throughout the rest of the interaction. Levels of ascorbate and tocopherol were unaffected.

As mentioned previously, Parker et al. (2009) found evidence of modulation in

the production of lignin precursors in compatible interactions of *M. oryzae* with cereal hosts. Lignin precursors such as sinapoyl alcohol, caffeoylquinic acid, ferulate were found to accumulate. Along with the lack of primary cell wall thickening that was observed during the interaction, it was hypothesised that this accumulation was as a result of insufficient ROS production for mono-lignan polymerisation and potentially a diversion of the phenylpropanoid pathway by pathogen effectors.

Hormone signalling plays an important part in cereal host responses to fungal biotrophic infection. A number of hormones have been implicated which include salicylic and jasmonic acid, auxins and gibberellins (Pieterse et al. 2012). They mediate the induction of signalling pathways that alters the expression many genes and metabolite pathways. Salicylic acid levels in resistant interactions of *Fusarium graminearum* and wheat were found to be elevated 3 hpi (Ding et al. 2011). Concordant with this elevation was increases in the expression of the phenylalanine ammonia lyase gene involved in salicylic acid biosynthesis. β -(1,3;1,4)-glucanase-2 expression, which responds exclusively to salicylic acid signalling, also showed the same increased profile at 3 hpi.

1.6 The modern plant pathologist’s omics toolbox

Omics technologies encompass the holistic analysis of cellular environments. There are now numerous omics layers that describe individual aspects of the cellular hierarchy. This includes the transcriptome, which describes the total mRNA molecules expressed from an organisms genes and the metabolome that describes the entire metabolite complement present within an organism.

Recent technological advances such as next generation sequencing (NGS) have revolutionised omics research and has provided unprecedented volumes of genomic and transcriptomic information. The development of high resolution (HR) MS has also revolutionised the fields of metabolomics and proteomics allowing increasing profiling capabilities (Mochida and Shinozaki 2011). Metabolomic and

transcriptomic techniques are further discussed in Sections 3.1, 5.1.1 and 5.1.3.

Technological advances in omics introduces new challenges with respect to data processing, analysis and storage. They are often concomitant with magnitude increases in sample throughput, volumes of acquired data and the numbers of measured variables. This requires the development of software tools that are not only able to process the data using methods appropriate to the new techniques, but are also able to execute it efficiently with respect to processing time and required computational resources as well as being user-friendly. These tools should also be open-source freeware, allowing easy access and adoption by the scientific community. An example is the development of Bowtie 2, a read alignment tool, developed as a result of the increasing throughput of NGS and the need for fast and efficient alignment of sequencing reads to reference genomes (Langmead and Salzberg 2012).

High performance computing is now an essential part of omics analyses. The routine use clusters of compute nodes provides the necessary memory and processing resources that not only makes the analysis of large omics data sets possible but routine (O’Driscoll, Daugelaite, and Sleator 2013). Software tools need to be able to take advantage of parallel processing; where computation is distributed across many processors, substantially reducing the necessary computational time. Both long and short term storage of the copious volumes of data produce by omics technologies also requires appropriate infrastructure.

The holistic nature of omics techniques means that many variables are simultaneously measured. This provides high dimensional data sets that require powerful data mining techniques to extract information relevant to the biological question. Machine learning algorithms that construct representations by learning from the data, are commonly applied in metabolomics analyses with the aim of deriving relationships between groups of biological observations (classes) and measured variables (features) (Enot et al. 2008). These algorithms include Random Forest and Support Vector Machines and Artificial Neural Networks, all of which have previously been applied to metabolomic analyses (Ward et al. 2010; Mahadevan

et al. 2008; Brougham et al. 2011). Random Forest is discussed further in Section 4.1.2.

With the high dimensional nature of omics data sets, consideration needs to be given to the concept of false discovery and the strategies required to avoid it. When many variables are simultaneously measured, the chance that any one of these is coincidentally related to the biological question also increases. This can be further enhanced by bias, inadequate sample size and the inappropriate choice of analytical technique (Broadhurst and Kell 2006). These factors need to be considered when designing omics experiments to ensure that the results obtained are valid and that resources are not squandered. Closely linked to this is the need for adequate validation of the results obtained to ensure that they are both relevant and reproducible. Omics validation is discussed further in Section 4.1.3.

The integration of omics data is becoming increasingly important to broaden the holistic views beyond that of the individual omics layers. Omics integration is further discussed in Section 6.1.1.

1.7 Systems biology and biological networks

There is still contention as to the exact definition of the field of systems biology. One definition is the study of interactions among biological components using models or networks to integrate genes, metabolites, proteins, regulatory elements and other cellular components (Yuan et al. 2008). In investigating biological systems, models of system characteristics can be used to predict the outcomes of system alterations to produce a phenotype of interest. Currently omics analyses provide best technological answer to holistically investigating biological systems. However, even though these analyses provide copious amounts of data, they are often incomplete and only focus on a single aspect of the biological system.

Biological systems can be represented as networks containing a series of nodes connected by edges. Nodes can represent genes, proteins or metabolites. The edges are the connections between the nodes that could represent co-expression, protein-protein interactions or metabolite correlations. They can also be either

directed or undirected depending on the type of association being represented. For instance, an irreversible enzyme catalyzed metabolite reaction would represent a directed edge where as a correlation between two gene expression profiles would represent an undirected edge (Gehlenborg et al. 2010).

Measures of network topology can be used to compare the structure and connectivity of a network. The degree distribution is the probability that a selected node has k links. This is useful for determining a networks type, whether a network is randomly structured or scale-free. Clustering coefficients can be used to characterize the tendency of nodes to form clusters within networks. Biological networks often have a scale-free topology, where the connectivity within the network is characterised by a power-law degree distribution. A result of this is that they will contain highly connected nodes known as hubs as well as distinct groups of connected nodes known as modules (Barabási and Oltvai 2004). The identification of hubs and modules within biological networks is an important part of inferring biological function. Hubs can represent key regulatory bottlenecks, such as transcription factors that can be responsible for the functional regulation of large modules of other co-regulated genes.

In plant-pathogen interactions, hubs and modules are not only key in the initiation of host defences but are also important in the pathogen subversion of the host network. Proteins involved in pathogen recognition signalling pathways are also likely hubs that will then initiate modules of genes involved in defence initiation (Pritchard and Birch 2011). However, because of the highly connected nature of these hubs, they can often form efficient pathogen effector targets. Perturbation of a hub, over other targets, will have an effect on the broadest regions of function; that system robustness is unlikely to overcome.

1.8 Aims

The central aim of this thesis is to apply integrative omics analyses to investigate the interaction between *B. distachyon* and *M. oryzae*; with emphasis on its biotrophic, pre-symptomatic phases. This will utilize both HR metabolomic

techniques as well as RNA sequencing (RNA-Seq) transcriptomic analyses. Alterations in metabolite and gene expression levels as a result of *M. oryzae* colonisation will be identified during these phases in both compatible and incompatible interactions.

The identification of metabolic and transcriptional alterations occurring during the interaction can provide insight into the infection and defence strategies employed by *M. oryzae* and *B. distachyon*. *M. oryzae* has to successfully suppress host defences and subvert host metabolism in order to acquire the nutrition needed to complete its life cycle. Conversely *B. distachyon* has to successfully recognise the presence of the pathogen and initiate appropriate defence responses to halt the infection. Understanding the system changes that occur during these process can provide useful targets for further research and potentially the development of durable resistance in these interactions.

In an addition to the integrative omics analyses for studying this interaction, there will also be application of computer vision phenotyping and quantitative trait loci (QTL) analyses to investigate the genetic basis of resistance to *M. oryzae* in *B. distachyon*. This forms an important aspect of understanding the context of the alterations in metabolite and gene expression changes during the interaction.

The overall aims and objectives can be summarised as follow:

- Develop methods for the processing of flow infusion electrospray (FIE)-HR MS data, suitable for metabolomic fingerprinting.
- Determine appropriate control measures for experiments involving plant-pathogen interactions and investigate the robustness of system changes between independent inoculations during asymptomatic phases
- Use metabolomic and transcriptomic analyses to investigate system perturbation during pre-symptomatic phases of the *B. distachyon* and *M. oryzae* interaction.
- Integrate metabolomic and transcriptomic data to elucidate key pathways involved during the biotrophic phases of the interaction.

-
- Use computer vision based phenotyping to identify gene loci in *B. distachyon* linked to *M. oryzae* resistance.

Chapter 2

General materials and methods

2.1 *B. distachyon* growth conditions

Seed of the *B. distachyon* ecotypes ABR6 and Bd21 were sown in sterilized Levington's Universal Compost (Levington Horticulture, Suffolk, UK) mixed with gravel (50:50 v/v) into modular plastic trays with 6 seeds per module. Plants were grown to 21 days old in environmentally controlled growth rooms (Polysec, R. J. Hicks Refrigeration, Aberystwyth, UK) at 23 °C under a 16 hour light period. Plants were watered daily, never being allowed to stand in water. Plants were illuminated using 55 W high-frequency lighting tubes and supplemented with 30 W clear tube cooled lighting with plants placed 60 cm from the light bank.

Seed stocks were maintained by allowing self pollination. The ecotype ABR6 required vernalisation to induce flowering. Vernalisation was induced by placing plants in a 4 °C cold room for six weeks, prior to being returned to the normal growth conditions; flowering occurred 3-4 weeks post vernalisation. Seed was collected and placed in dry storage at room temperature until required.

2.2 *M. oryzae* maintenance and growth conditions

The *M. oryzae* strain Guy11 (mating type MAT1-2) was maintained on potato dextrose agar (PDA) (Oxoid, Hampshire, UK) prepared as 39 g L⁻¹ and autoclaved at 121 °C for 15 minutes. Plates were cultured in a temperature controlled incubator (Gallenkamp Illuminated Cooled Incubator 9, Loughborough, UK) at 23 °C for 14 days under a 16 hour light period. Pathogenicity was maintained by re-isolation of *M. oryzae* spores from infected *B. distachyon* leaf tissue as described by Parker et al. (2008).

2.3 *M. oryzae* inoculum preparation and *B. distachyon* inoculation

Conidial suspensions of *M. oryzae* were prepared as described in Parker et al. (2008) by scraping mycelia from the surface of culture plates in 0.2% (w/v) gelatine solution. Suspensions were then centrifuged (RT7, Sorvall) at 2500 rpm for 5 minutes at room temperature. The supernatant was then poured off, re-suspended in gelatine solution and re-centrifuged. The final suspension was re-suspended in gelatine solution using a volume of 1.4 mL/plate giving a conidial density of 10¹⁰ conidia/mL. Conidial concentrations were adjusted accordingly after estimating density using a microscope and haemocytometer.

A non-pathogenic control was subsequently prepared by subjecting a portion of the conidial suspension to two rounds of snap freezing in liquid nitrogen, thawing in a 35 °C water bath and sonication for 5 minutes to completely neutralise the fungal spores.

B. distachyon plants were inoculated at 21 days old. Plants were spray inoculated with approximately 2.5 mL of conidial suspension per plant using an artists airbrush (Model 250-2, Badger, USA). Plants were then placed into plastic propagator trays to maintain high humidity and removed no sooner than 2 days post

inoculation. Control and infected plants to be harvested at the same time point were placed into the same plastic propagator ensuring no contact between plants of differing treatment.

2.4 Harvesting of plant material

B. distachyon leaf tissue was harvested by detaching the 2nd fully developed leaf from the base of each plant. The middle 4 cm segment was excised and immediately placed in a 2 mL Eppendorf tube containing a 4 mm steel ball bearing and snap frozen in liquid nitrogen. Samples were stored at -80°C until extraction. Where more than one leaf was sampled, details are given where appropriate. Unless otherwise stated, plants were removed from high humidity conditions immediately before harvesting. All harvesting was conducted in the growth room to avoid environmental fluctuations.

2.5 Large-scale inoculations to investigate the pre-symptomatic phases of the *B. distachyon* and *M. oryzae*

Three independent, large scale inoculations of the *B. distachyon* ecotypes ABR6 (incompatible) and Bd21 (compatible) were undertaken simultaneously with control and infected tissue harvested at 12 hour intervals from 0-60 hpi. The independent inoculations were conducted at weekly intervals. 12 metabolomics replicates were harvested as described in Section 2.4 and combining two leaf sections per replicate for each treatment class at each time point in each experiment giving 36 total replicates for each class. Concurrently, one transcriptomics replicate was harvested for each treatment class by combining 10 *B. distachyon* leaf sections per replicate, at each time point in each experiment, giving 3 total replicates for each class.

2.6 Metabolite Extraction

For global metabolite extraction, frozen samples were milled using a Retsch mm 301 Mixer Mill at 30 Hz for 30 seconds then placed on crushed ice. Pre-chilled extraction solvent (CHCl_3 :MeOH:H₂O; 1:2.5:1; v:v:v) was immediately added using 700 μL /leaf and suspended by vortexing (Scientific Industries Vortex Genie-2). Samples were then placed on a orbital shaker (FATSN002, Favorgen Biotech Corp) for 20 minutes at 1,400 rpm and a temperature of 4 °C. After shaking, samples were centrifuged (EBA 12R, Hettich) at 13,000 rpm for 6 minutes at 0 °C. The supernatant was then transferred to a clean 2 mL Eppendorf tube and the pellet discarded. Samples were stored at −80 °C until MS analysis.

Samples for liquid-chromatography (LC)-HRMS analysis were prepared firstly by complete drying of 400 μL of sample in a centrifugal vacuum evaporator (Univapo 150H + Unijet II refrigerated aspirator, Uniequip) for approximately 2 hours. The samples were then reconstituted in 40 μL of pre-chilled ultra-pure water (18 Ω), vortexed and then sonicated (Ultra wave) at 60 Hz for 15 minutes. Samples were shaken in an orbital shaker for 20 minutes at 1,400 rpm at 4 °C and centrifuged at 14,000 rpm for 8 minutes at 0 °C. The supernatant was then carefully transferred into a 200 μL vial. The samples were made up no more than 4 hours prior to analysis and were stored a 4 °C.

Samples for quality control (QC) were prepared prior to the drying down of samples by combining an aliquot of each of samples that were to be analysed. These samples were then prepared in the same way as all of the other samples.

2.7 FIE-HRMS analysis and data processing

FIE-HRMS analysis was performed in an Exactive Orbitrap mass spectrometer (ThermoFinnigan, San Jose, CA) coupled to an Accela (ThermoFinnigan, San Jose, CA) ultra-performance liquid chromatography front-end. 20 μL of sample was delivered to the electrospray ionisation source in a flow solvent of pre-mixed HPLC grade methanol and ultra-pure water (7:3). Data was acquired for a total

of 3 minutes. The flow rate was maintained at $200\text{ }\mu\text{L min}^{-1}$ for the first 1.5 minutes then raised to $600\text{ }\mu\text{L min}^{-1}$ up to 3 minutes.

Both positive and negative ionisation modes were acquired simultaneously using polarity switching. Scans for each ionisation mode consisted of a single scan event, ranging from 55-1000 mass to charge ratio (m/z) in positive mode and 63-1000 m/z in negative mode at a scan rate of 1 Hz. An automatic gain control target of 5×10^5 was used and the resolution set at 100,000 and a maximum injection time of 250 ms.

Raw data was acquired as profile data in the proprietary ThermoFinnigan file format (.RAW). These were converted to the mzXML format and centroided using the msconvert tool (TransProteomicPipeline, <http://proteowizard.sourceforge.net/tools.shtml>).

Scans representing the infusion peak of each sample were selected and data was spectrally binned using the R package binneR (<https://github.com/jasenfinch/binneR>) using a bin width of 0.01 atomic mass units (amu). The data was then total ion count (TIC) normalised and m/z filtered based on a 66% maximum class occupancy threshold across all classes.

2.8 LC-HRMS analysis and data processing

LC-HRMS analyses were performed on an Orbitrap Fusion Tribrid mass spectrometer (Thermo Scientific) that was coupled to a Unimate 3000 liquid chromatography tower (Dionex, Thermo Scientific). A Hypersil Gold reverse phase C_{18} column (2.1 mm x 150 mm; particle size $1.9\text{ }\mu\text{m}$) was used for chromatography, maintained at a temperature of $60\text{ }^{\circ}\text{C}$. The mobile phases for gradient elution consisted of ultra-pure water - formic acid (100:0.1) (**A**), LC-MS grade methanol -formic acid (100:0.1) (**B**) and HPLC grade isopropyl alcohol - LCMS grade methanol (1:1) (**C**). The initial condition was *A:B* (99:1.0) for 0.5 minutes and the percentage of **B** increased linearly over 6 minutes, to 60.0%. The percentage of **B** was increased further for another 4 minutes to 100% and held for 2 minutes. **C** was introduced over 0.1 minutes to 100% and held for 1.4 minutes.

Prior to re-equilibration, **C** 100% was switched to **B** 100% in 0.1 minutes, and then the starting conditions **A:B** (99:1) held for 3.4 minutes. This gave a total method time of 17 minutes. The flow rate was kept at 400 $\mu\text{L min}^{-1}$ except for between 12.1 and 13.6 minutes where it was reduced to 300 $\mu\text{L min}^{-1}$.

Similar to the FIE-HRMS analyses raw data was acquired as profile data in the proprietary ThermoFinnigan file format (.RAW). These were converted to the mzXML format, centroided and the ionisation modes separated using the msconvert tool (TransProteomicPipeline, <http://proteowizard.sourceforge.net/tools.shtml>).

Signal processing of converted LC-HRMS data was performed using XCMS (Smith et al. 2006). The presence of technical outliers was assessed prior to peak detection by ensuring the stability of sample TIC. Any sample with a TIC ± 3 standard deviations of the batch median were removed from subsequent analyses. The continuous wavelet transform algorithm was used for peak picking (Tautenhahn, Böttcher, and Neumann 2008). The tolerated deviation between consecutive scans was set at 1.5 parts per million (ppm) with a minimum and maximum peak width of 2 and 40 seconds respectively. The ordered bijective interpolated warping algorithm was used for retention time correction (Prince and Marcotte 2006) following peak detection.

Peak grouping was performed using a density matching method across all the samples with a tolerated m/z size of ± 0.015 amu and a chromatographic bandwidth of 5 seconds. A minimum threshold for group occupancy was set at 60% for each class. Each peak group represented a unique m/z and retention time value, each to 2 decimal places. For peak infilling, the XCMS *fillPeaks* method was used to integrate intensity regions missed during the initial peak detection.

QC sample variable filtering applied by calculating the coefficient of variance for each peak across the QC samples. Peaks with a coefficient of variance below 25% were removed to ensure analytical reproducibility of identified features. The data was subsequently log10 transformed.

2.9 FIE-HRMS and LC-HRMS data mining

The data mining of both FIE-HRMS and LC-HRMS was performed using custom workflows found at https://github.com/jasenfinch/Orbi_FIE and <https://github.com/jasenfinch/LC-HRMS> respectively. Data quality and structure was initially assessed using principle component analysis (PCA). Outlier samples were removed if their Mahalanobis distance of the first two principle components was outside the 95% confidence interval. An initial assessment of class discrimination was conducted using principle component linear discriminant analysis (PC-LDA) prior to supervised machine learning based classification.

Unless otherwise stated, binary classification was performed using the *AC-CEST* function of the FIEmspro R package (<http://users.aber.ac.uk/jhd>) using Random Forest. The classification accuracy, margin of classification and area under the ROC curve (AUC) were calculated based on 100 bootstrapped re-sampling iterations, the data partitioned at 63.2% for the creation of test and training sets. Variable importance was based on variable selection frequency using 1% threshold for approximate false positive rate based on the binomial distribution (Konukoglu and Ganz 2014)).

2.10 Putative metabolite annotation

2.10.1 FIE-HRMS

Accurate masses of 0.01 amu spectrally binned explanatory features were extracted by creating electronic class master mixes after spectrally binning the unprocessed data to 0.00001 amu. For each 0.01 amu explanatory bin, the most intense accurate mass signal across all samples was extracted for that bin. Semi-automated annotation was performed using R routines found at <https://github.com/jasenfinch/mzAnnotation>. This included putative ionisation product database searches using the annotation tool MZedDB (Draper et al. 2009) and molecular formula generation based on the ‘Seven Golden Rules’ using

a 5 ppm window (Kind and Fiehn 2007). Correlation analyses were also applied between all identified spectral 0.01 amu bins. Mass difference matching was applied to significantly correlated bins ($p < 0.05$) to identify isotopic, adduct and metabolic associations based on relationships found in Table 3.4.

2.10.2 LC-HRMS

For LC-HRMS data the R package CAMERA was used for initial identification of adduct and isotopic relationships (Kuhl et al. 2012). Pseudo-spectra were formed at 0.7% of full width half maximum and within group correlations between extracted ion chromatograms filtered based on a threshold of 0.8. The set of adduct and isotope rules used for association can be found in Appendix B. Putative ionisation product database searches and molecular formula generation for accurate masses was conducted as for the FIE-HRMS features.

2.11 RNA extraction, library preparation and sequencing of RNA-Seq samples

RNA-Seq samples were extracted using the RNEasy Plant Mini Kit (Qiagen) for total RNA extracts eluted in RNase free water. Total RNA extract concentrations were quantified using a Nanodrop 1000 spectrophotometer (Thermo Scientific). RNA integrity assessment, library preparation and sequencing were performed by the Earlham Institute (Norwich, UK). cDNA libraries were prepared and sequenced using the Illumina Truseq HT protocol on a HiSeq 2000 Sequencing System (Illumina) with a 126 bp paired end metric. 36 samples were sequenced in total from three time points (12, 24 and 48 hpi) for both ecotypes and multiplexed across 3 lanes.

2.12 Transcriptomic data mining

Fastqc was used to assess sequencing read quality (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). The Trimmomatic tool was subsequently used to remove the first 10 bp from each read due to over-representation in the per base sequence content (Bolger, Lohse, and Usadel 2014). Tophat was used to align reads using assembly v3.0 of *B. distachyon* genome and v3.1 annotation. A maximum and minimum intron length of 50000 and 70 were used based on Mandadi and Scholthof (2015). Cufflinks was used for transcriptome assembly using the default parameters. Cuffdiff was used for differential expression analysis, again using the default parameters. The Mapman Mercator tool was used for ontological assignment of identified genes for use in enrichment analyses. The Fischer exact test was used in enrichment analyses to test for over-representation of functional ontologies.

Chapter 3

Spectral binning for untargeted FIE-HRMS metabolome fingerprinting

3.1 Introduction

With no single analytical technique able to survey the entire chemical composition of complex biological matrices, investigators have to utilize multiple techniques to study an organisms metabolome. This includes both profiling and fingerprinting techniques.

Metabolome fingerprinting is the global and high-throughput analysis of crude sample extracts for classification or screening (Dunn, Bailey, and Johnson 2005). It provides a key branch point in the metabolomic pipeline as a first pass analytical tool that is able to inform further more targeted analyses, based on the identification of compounds of interest. It is primarily used for classification or regression with sample metadata (Draper et al. 2013).

There are a number of analytical techniques that can be used to provide a fingerprinting platform; including Fourier transform (FT)-infrared (IR) spectroscopy, Nuclear Magnetic Resonance (NMR) spectroscopy, and direct injection (DI)-MS or FIE-MS. Mass spectrometric techniques have the advantage over

other analytical methods in that they allow a greater potential for metabolite identification (Beckmann et al. 2008). Added to this is the recent addition of high resolution mass spectrometry instrumentation can provide fingerprints with orders of magnitude finer detail and thus are more representative of an organisms metabolome.

3.1.1 High resolution mass spectrometry

MS is the measurement of the m/z of ionized chemical species. As MS has developed so too has the resolution capabilities of the instrumentation. The last decade has seen the introduction of high resolution mass analyzers that have revolutionised the capabilities and applications of MS instrumentation in the laboratory.

For a single m/z spectral peak (m), resolving power (RP) can be defined as:

$$RP = \frac{m}{\Delta m_{50\%}} \quad (3.1)$$

Using full width at half maximum as the specified fraction. Mass analyzers with a resolving power $m/\Delta m_{50\%} > 10,000$ are considered to be high resolution (Xian, Hendrickson, and Marshall 2012). This includes time of flight (TOF) and ultra-high resolution, FT-MS analyzers such as ion-cyclotron resonance (ICR) and Orbitrap analyzers. Orbitrap based analyzers are becoming the standard instrumentation for HRMS and will be the primary focus here due to their ultra-high resolution capabilities, low cost, low maintenance and bench top size compared to ICR mass analyzers.

FT-MS is based on the measurement of frequencies emitted by ion motional amplitude rather than ion deflection and stability that TOF and quadrupole analyzers use respectively. Frequencies are first transformed from the time domain to the frequency domain, then to the mass domain to give the final mass spectrum. Signals from a wide m/z range can be detected simultaneously allowing the entire spectrum to be yielded in a single scan (Marshall and Hendrickson 2008).

In an Orbitrap, radio frequencies are generated by axial oscillation of ions in an electrostatic quadrupole potential well. Ions rotate around a central electrode creating oscillations between 50 and 150 kHz for m/z 200-2000. Differential image-current detection is achieved by an outer electrode that is split into two halves (Hu et al. 2005). Orbitraps require the uniform injection of ions into the analyzer, which is enabled by a C-trap. A C-trap alongside the Orbitrap not only allows storage of ions prior to analysis but ensures the required coherent motion to be achieved during injection. The small size of the Orbitrap allows close proximity of the C-trap and so reduces the potential for TOF discrimination (Hardman and Makarov 2003). Dynamic range is affected by the amount of ions injected from the C-trap due to space charge repulsions. Mass resolving power is proportional to $1/(m/z)^{1/2}$ and unlike ICR, decreases more slowly with m/z . Orbitrap scan time is typically 1 second when both ionisation modes are acquired simultaneously.

The accuracy of Orbitrap mass analyzers is typically < 5 ppm but < 2 ppm is achievable depending on whether internal or external calibration is used. Variability in accuracy when external calibration is used is mainly temperature dependent due to instability of the inner electrode voltage. Internal calibration is limited by space charge effects that change between scans (Makarov et al. 2006).

There are a number of benefits to using high resolution as opposed to low resolution instruments. Higher resolution allows better deconvolution of peaks and therefore signal intensities that are reflective of a single metabolite, as opposed to the combination of multiple metabolites with similar masses. The use of ultra high mass accuracy allows the calculation of elemental composition and the generation of empirical molecular formulas. However, as the m/z increases, the number of possible molecular formulas also exponentially increases. The application of the seven golden rules for molecular formula generation provides a method for substantial filtering of molecular formulas, although the use of isotopic ratio patterns is still essential for narrowing down candidate molecular formulas (Kind and Fiehn 2007). A resolution of greater than 100,000 allows the separation of

isobaric peaks which allows their use for molecular formula elucidation. It has been shown previously that isotopic abundance measurements in both FT-ICR-MS and Orbitrap-MS have the ability to increase the number of single empirical molecular formula assignments (Weber et al. 2011).

3.1.2 Spectral binning for signal processing

The spectral binning of data, where spectral data are grouped based on a common interval, is a common practice among many signal processing disciplines. It is a form of quantization that allows a reduction of data complexity which can then facilitate further analyses. It has common application in spectral radiometry where spectral regions can be binned to reduce data complexity (Dell’Endice et al. 2009).

Within metabolomics, spectral binning has previously been applied to the processing NMR spectroscopy and nominal mass FIE-MS fingerprinting (Wishart 2008; Beckmann et al. 2008). In both cases, spectral binning was achieved using the rounding of measurements and subsequent averaging or summing. It allows small deviations in measurements between samples to be overcome. This allows comparability between observations and therefore the use of statistical analyses.

3.1.3 Untargeted FIE-MS metabolome fingerprinting

FIE-MS fingerprinting is an analytical technique that provides a global overview of total sample metabolite composition, that does not incorporate chromatographic separation (Goodacre et al. 2004). Electrospray ionisation is the most common of ionisation techniques and is a ‘soft’ technique that induces the loss or gain of a proton or adduct, with ions tending to only have a single or double charge (Draper et al. 2013). Minimal sample preparation is needed and crude, global extracts containing both polar and non-polar metabolites are suitable.

With electrospray ionisation, a number of techniques can be used to introduce the sample to the mass spectrometer. For direct injection the sample is manually injected directly into the ion source using a syringe pump. Flow infusion incor-

porates the use of an auto-sampler which allows automated introduction of the sample to the electrospray ion source by infusion into a mobile phase. A ‘plug’ flow is created, across which signal intensities can be averaged to give a samples chemical fingerprint (Figure 3.1; Beckmann et al. 2008). This provides relatively simple data pre-processing compared to chromatographic techniques.

Further to the use of auto-samplers for flow infusion sample introduction is the use of chip-based nano infusion devices such as the NanoMateTM. These do not require the addition of a mobile phase during sample introduction (Southam et al. 2007).

The lack of chromatographic separation substantially reduces analysis time (< 5 mins per sample) and cost compared to LC-MS or gas chromatography (GC)-MS. Analyses involving chromatography will also be subject to ‘drift’ over large sample batches, further complicating data processing and subsequent analysis. Without this constraint in FIE-MS fingerprinting, large batch sizes can be consistently run and so routine analyses of 1000’s of samples becomes feasible.

Hierarchical based FIE-MS fingerprinting techniques incorporate the use of both high and low resolution instrumentation. Low resolution instrumentation is first used for initial FIE-MS fingerprinting. This allows classification and feature selection analyses to firstly be conducted to identify explanatory features relevant to the biological question. Then HRMS instrumentation is used to analyse class master mixes to provide high mass accuracy m/z information for explanatory feature structural identification.

The primary goal of metabolome fingerprinting is the classification or discrimination of biological samples of different origin (Enot et al. 2008). Owing to its relative simplicity and high-throughput nature, FIE-MS fingerprinting has been applied to a wide range biological problems and sample matrices (Draper et al. 2013). These include a variety of plant based matrices discriminating potato cultivars (Beckmann et al. 2007), polyphenol content in berry fruits (McDougall, Martinussen, and Stewart 2008) and wound responses in *Arabidopsis thaliana* (Grata et al. 2007).

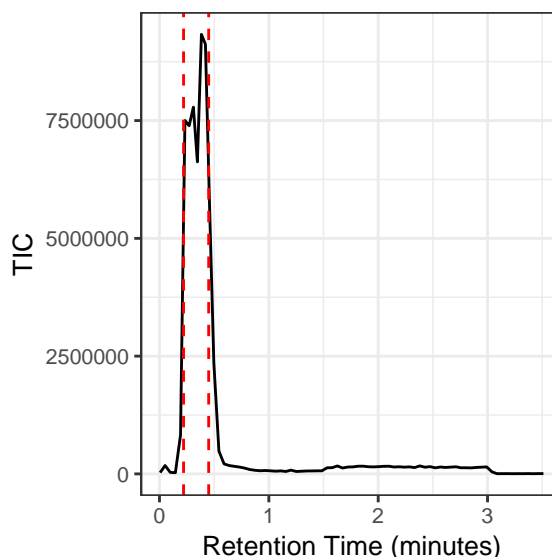


Figure 3.1: **An example *B. distachyon* FIE-MS ion chromatogram.** Dashed red lines show retention times between which the sample ‘plug’ flow is analysed.

3.2 Aims

During the early phases of this PhD project, there was a transition from using low-resolution, nominal mass based MS instrumentation towards the use of HR Orbitrap-MS instruments. This required the development of new data processing methods and bespoke software tools able to deal with the magnitude increases in data volume and complexity to allow routine FIE-HRMS metabolome fingerprinting analyses. The aim of this chapter therefore is to present the development of spectral binning as a pragmatic solution to processing FIE-HRMS fingerprinting data. This allows increased resolution and data density over previous nominal mass based methods but can still retain low computational requirements and therefore processing time that are characteristic of fingerprinting techniques. Further to this, the necessary data pre-treatments will be considered with respect to data transformation and normalisation as well as missing values and class occupancy. The increased resolution also improves the potential for metabolite annotation compared to nominal mass techniques. This potential will be investigated in terms how increased resolution and the correlations between adducts, isotopes and metabolically related metabolites allows more confidence in annota-

tion assignment. This provides the following questions to address:

- Present the R package binneR developed for the spectral binning of FIE-HRMS fingerprinting data.
- Determine optimal bin size for FIE-HRMS fingerprinting data.
- Assess the data pre-treatments that are required prior to further analyses.
- Investigate the considerations that are needed for missing values and class occupancy.
- Determine the impact of increased bin resolution improve the potential for metabolite assignment.

3.3 Materials and Methods

3.3.1 Preparation and MS analysis of example *B. distachyon* sample

In order to investigate aspects of spectral binning, an example *B. distachyon* sample was prepared using the *B. distachyon* ecotype ABR1. Plants were grown as described in Section 2.1 to 21 days old. A total of 7 *B. distachyon* plants were harvested and metabolites extracted using a global extraction method as described in Sections 2.4 and 2.6 giving a total extract volume of 4.9 mL. 55 consecutive technical injections of the example *B. distachyon* sample were analysed by FIE-HRMS, centroided and converted to mzML using msconvert as described in Section 2.7. The files were spectrally binned using the binneR package (available at <https://github.com/jasenfinch/binneR>), developed as part of this project that will be further discussed in Section 3.4.1. Differing numbers of these injections were used to investigate the different aspects of spectral binning for FIE-HRMS. The numbers of injections used is stated where appropriate.

3.3.2 Performance testing for the R package binneR

To test the computational requirements for use of the binneR package the R packages `profvis` (<https://cran.r-project.org/web/packages/profvis/index.html>) and `rbenchmark` (<https://cran.r-project.org/web/packages/rbenchmark/index.html>) were used for peak memory usage and performance benchmarking respectively of the binneR package. The data used was that of FIE-HRMS fingerprinting of the pre-symptomatic phases of the interaction between *B. distachyon* and *M. oryzae* described in more detail in Section 5.3.

3.3.3 Investigation of missing value imputation

To test the effects of missing value imputation upon spectrally binned FIE-HRMS fingerprints, treatment and control data from the 36 hpi time point of the resistant ABR6 interaction from the large scale FIE-HRMS fingerprinting of the pre-symptomatic phases of the interaction between *B. distachyon* and *M. oryzae* described in Sections 2.5 and 2.7. All the samples were TIC normalised and differing class occupancy filtering treatments were applied (none, minimum and maximum) using a two-thirds threshold. kNN imputation was then used to impute all missing values still present within the matrices using the FIEmspro R package. The margin value between the treatment and control classes of the imputed and un-imputed data were computed using Random Forest classification, as described in Section 2.9, to assess the impact of imputation upon the differing pre-treatment conditions.

3.4 Results and Discussion

3.4.1 Development of the R package binneR

This section discusses the the development of a software package called binneR (available at <https://github.com/jasenfinch/binneR>) during this project for spectral binning of FIE-HRMS data as an implementation in the statistical pro-

programming environment R. R is open source freeware that is widely used for data analysis and visualization in the biological sciences with a dedicated package repository Bioconductor (Crawley 2013; Huber et al. 2015).

The package provides utility for building a 2-dimensional intensity matrix for one or multiple FIE-HRMS data files with columns as m/z and rows as samples, using the functions *sampProcess* and *readFiles*. This intensity matrix can then be used for downstream analyses such as classification and feature selection. The Bioconductor R package *mzR* is used for file parsing so a range of file formats are supported such as the open source formats *mzXML* or *mzML* (Pedrioli et al. 2004; Martens et al. 2011). The use of centroided data is preferable over profile data as this substantially reduces data volume and therefore computational time. This also reduces the effect ‘bin splitting’, an artifact of spectral binning that is discussed in Section 3.4.2. Data files containing single or multiple scan modes are supported as well as multiple scan ranges. Other parameters include the bin width by specifying the number of decimal places for rounding and the range of scans over which the data should be averaged. Parallel computing is also supported allowing the parallelized parsing of files and processing of individual acquisition modes.

The spectral binning of an individual sample firstly requires m/z intensities within each scan to be rounded and sum aggregated to the required bin width. Next, any missing bins within a given scan need to be added and filled with zero intensity values. Bin intensities can then be averaged across the scans to give the intensity matrix for the given sample. For processing multiple samples, further addition of missing bins is required before the intensity matrix can be constructed.

An important requirement of the use of spectral binning for FIE-HRMS metabolome fingerprinting is that it requires little computational time and resources. As shown in Table 3.1 the processing of 50 data files at bin widths of 0.01 and 0.00001 amu can be done in less than 1 minute and requires little memory. Even parallel processing 1000 files at a bin width 0.01 can be achieved in

Table 3.1: **Computational requirements for spectral binning of FIE-HRMS metabolomic fingerprinting data using the R package binneR.** Data files from the large-scale *B.distachyon/M. oryzae* described in Section 5.3. The processing time is based on 10 replications.

No. Samples	Bin Width (amu)	Processing Time (minutes)		Peak Memory Usage (MB)
		Single Core	8 Cores	
50	0.01	0.39	0.29	59.85
50	0.00001	0.96	0.79	190.84
1000	0.01	7.65	4.08	1059.22
1000	0.00001	30.97	62.49	167379.70

under 5 minutes with memory requirements well within that of a modern standard desktop PC. This makes the routine processing of 1000's of samples easily feasible if binning at a width of 0.01 amu.

Computing requirements become much more substantial when trying to process 1000 data files at a bin width of 0.00001 amu. Here memory usage increases due to the exponential increase in the number of bins as the bin width decreases (Table 3.2). Also the use of parallel processing becomes infeasible with a increase in processing time. This is due to the increased overheads caused by increased memory requirements.

The package also contains a Shiny application for viewing raw mzXML or mzML files named *viewSpectrum*. Shiny is an R package that allows the development of interactive, web browser based applications (Chang et al. 2016). The application allows the visualization of ion chromatograms and mass spectrums with selectable retention time and mass ranges. Bin width can be selected by choosing the number of decimal places by which to round. This application is particularly useful for identifying the correct 'plug flow' scans for averaging across when spectral binning (Figure 3.1). A screenshot of the application can be found in Appendix C.

3.4.2 Optimal bin size for FIE-HRMS metabolome fingerprinting

The selection of the optimal bin width for spectral binning of FIE-HRMS metabolome fingerprinting data is a compromise between retaining as high resolution as possible without reducing the quality of the data. A reduction of data quality would include an increase in the proportion of missing data as well as the introduction of processing artifacts.

As the bin width decreases the number of variables exponentially increases in both modes (Table 3.2). Firstly, this imposes a computational constraint which has been discussed in Section 3.4.1 and is highly important when considering optimal bin width. Table 3.2 also shows that the proportion of missing values greatly increases. This is due to the deviation of peaks both between scans of a sample and between samples. These deviations are as a result of changes in parameters such as temperature and space charge compensation during Fourier transformation which is calculated on a scan by scan basis (Hu et al. 2005). These parameter changes are difficult to accurately account for and so makes correcting for and aligning deviating peaks difficult. An example of deviation is shown in Figure 3.2 where there is a deviation of up to 0.00032 amu between samples. This means that peaks are able to freely shift between bins when a bin width of 0.0001 or less is used and so introduce a high degree of artificial missing data. This rules out bin widths below 0.0001 amu as appropriate widths to use for FIE-HRMS data.

Another consideration for bin width choice is the introduction of artifacts such as peak splitting. This would be caused by the deviation of a bin such as in Figure 3.2 occurring near to a bin boundary. The result of this is that two adjacent bins would be obtained for what in reality is a single peak. An example of this is shown in Figure 3.3a which shows the density of peaks within a single bin at a width of 0.01 amu. It is likely that this bin contains two real peaks; however, also projected upon this is the bin boundaries if a width of 0.001 amu was used. It can clearly be seen that the larger peak would be ‘split’ between

two bins, thus adding artificial data.

The relationship between two adjacent ‘split’ bins would be negative as shown in Figure 3.3b. This is likely to be as a result of the centroided data and the averaging across scans so that a peak will only fall into one or other of the bins. Due to the random nature of the deviation, when the scans are averaged one of the bins will have a higher intensity than the other, proportional to the actual peak intensity. It is especially likely if an odd number of scans is used. The likelihood of peak splitting is also increased as the m/z increases due to the reduction in resolution and the increase in peak width.

A common issue in FT-MS spectra is the presence of Gibbs oscillations or ‘ripples’ that can occur either side of a high intensity peak. These are echos of the main peak and are a result of the FT (Marshall and Hendrickson 2008). The presence of these would further complicate the application of peak picking routines upon these high resolution data. Detection of these peaks would result the addition of artifacts into the data, although compensating for them would not be trivial.

The introduction of these artifacts into the data could have an effect of over-inflating discrimination between classes if a split bin happens to be explanatory in the context of the biological question. This would be likely to happen at bin widths of 0.01 and 0.001, however with the order of magnitude increase in resolution at 0.001 amu the likelihood of this occurring is much greater.

Increasing the bin width also increases the likelihood that one bin will contain peaks from multiple metabolites. This has implication for downstream data analysis as feature trends become convoluted and difficult to interpret in the context of the biological question. It also makes putative annotation of features more difficult as the correlations between bins become less reflective of the underlying metabolite, isotope and adduct relationships (see Section 3.4.4).

A bin width of 0.01 amu provides the best compromise between retaining resolution and not introducing substantial amounts of missing data or artifacts. Therefore this can be considered the optimal bin size for FIE-HRMS metabolome

fingerprinting data.

3.4.3 FIE-HRMS data pre-treatment

Post raw data processing and prior to downstream statistical analyses, pre-treatment of metabolome fingerprinting data is an essential step to ensure data quality and integrity (Enot et al. 2008). It is important that appropriate pre-treatment strategies are used not only for the analytical technique being applied but are also suitable for the statistical or machine learning analyses that are to be used on the data. For instance, it has been previously identified that the scaling of metabolomic data can greatly influence the accuracy of model classification depending on the classifier being used (Gromski et al. 2015b). Here, data normalisation, transformation, variable occupancy filtering and missing value imputation will be discussed in the context of spectrally binned FIE-HRMS metabolomic fingerprinting.

3.4.3.1 Normalisation and transformation

Not only are FIE-HRMS fingerprints highly dimensional but are also heteroscedastic in nature and contain a range of variable intensity magnitudes (van den Berg et al. 2006). The majority of variables within a typical FIE-HRMS fingerprint are of low abundance with 93% of the fingerprint m/z constituting only 1% of the TIC (Figure 3.4a). This means that fingerprints are dominated by a relatively small number of highly intense m/z . Also the lower the intensity of a given m/z , the higher its relative variance (Figure 3.4b). This presents a major challenge for data pre-treatment without distorting or introducing artifacts into the data structure that would then affect the outcome of downstream analyses.

Normalisation can be used to account for technical variance introduced during an experiment, sample preparation or mass spectrometric analysis. The type of normalisation used is often experiment dependent. The sample context will dictate where artificial variance is likely to be introduced and therefore needs accounting for. There is a need in mass spectrometry based metabolomics to

Table 3.2: **Spectral binning width: variable numbers and missing values.** Variable numbers and missing value percentages are given for 10 technical injections of the example *B. distachyon* sample scanning between 70 and 1000 m/z .

Mode	Bin Width (amu)	No. Variables	% Missing Values
Negative	1	806	42
	0.1	2631	69.8
	0.01	4315	73.2
	0.001	6885	74.8
	0.0001	14900	81.6
	0.00001	23753	86.7
Positive	1	868	31.3
	0.1	3354	56.7
	0.01	7387	67.9
	0.001	14277	74.7
	0.0001	31529	83
	0.00001	48331	87.1

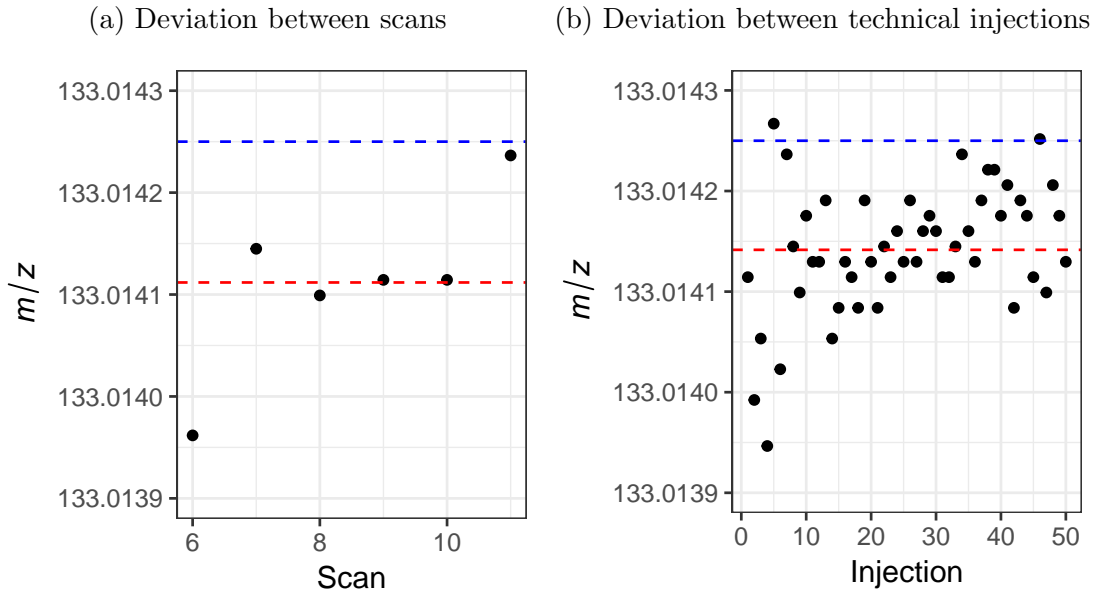
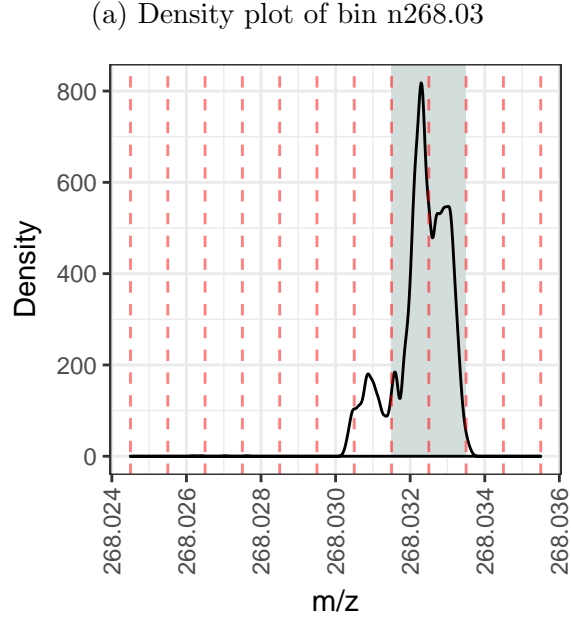


Figure 3.2: **Mass deviation of the base peak malic acid $[M-H]^{1-}$ between scans and injections.** Based on technical injections of the example *B. distachyon* sample. The dashed red line is the average measured m/z . The dashed blue line is the theoretical m/z of malic acid $[M-H]^{1-}$.



(b) Relationship between adjacent bins n268.033 and n268.032

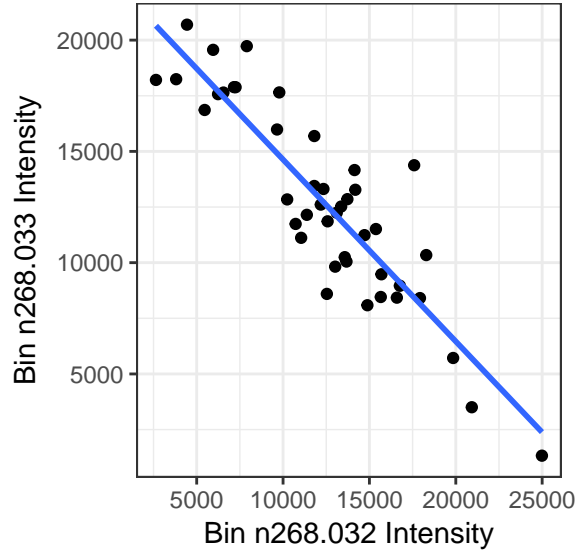


Figure 3.3: **Example of peak splitting at bin width 0.001.** (a) Density of peaks within bin n268.03 taken as an average across 43 technical injections of the example *B. distachyon* sample. Red lines denote 0.001 bin width boundaries. Shaded areas denote the bins that are plotted in (b). (b) Scatter plot of bin intensities for 43 technical injections of the example *B. distachyon* sample. The line of best fit denotes the negative linear relationship ($R^2 = 0.798$, $p < 0.001$)

account for instrument variability during sample run time. In metabolome profiling, where chromatographic separation is used prior to MS analysis, internal standards are often introduced to samples, the variability of which can be used to account for instrument variability over the course of a sample batch (Fiehn et al. 2000). However, for FIE-MS metabolome fingerprinting this is not an option due to the effects of ion suppression.

TIC normalisation is often used in metabolome fingerprinting to account for variability in instrument sensitivity and sample concentration (Enot et al. 2008). The effect of TIC normalisation on m/z can be seen in Figure 3.5a. The presence of a few dominant m/z that account for a high proportion of the TIC within the fingerprint (Figure 3.4a) means that only these m/z will be affected as they are most likely to be diagnostic of instrument sensitivity and sample concentration. TIC normalisation on an individual sample basis becomes an issue when a high intensity signal is also explanatory for the biological question the experiment is trying to answer. Not only could this accentuate how explanatory the m/z is but is also likely to introduce spurious knowledge into the fingerprint data. Therefore care needs to be taken when applying this normalisation.

There are two potential strategies for applying TIC normalisation when sample TIC becomes class dependent. The first from Enot et al. (2008) allows a sample's TIC to be corrected to directly remove class dependency by removing the difference between a class's TIC and the average TIC prior to correction. The second suggested by Draper et al. (2013) allows TIC normalisation to be applied to only correct for instrument variability within or between batches by averaging the TIC across all the samples in a randomised block or batch prior to normalisation. This relies on samples being analysed in a randomised block manner. The most suitable strategy for applying TIC will be dependent on experiment context and the source of the variability.

Data transformation and scaling are statistical practices used to allow data to meet the assumptions of a particular statistical inference such as those of parametric tests. They are commonly used in metabolomics prior to downstream analyses

such as PCA. Common transformation practices in metabolomics include auto-scaling, Pareto scaling and \log_{10} transformation. These techniques aim to reduce the magnitude differences in scale between m/z and reduce heteroscedasticity. However, they are often poor at handling variables with high relative standard deviation and can inflate measurement errors such as instrument variability (van den Berg et al. 2006). Figure 3.5b shows the effect of \log_{10} transformation on the example *B.distachyon* sample metabolome fingerprint. Here it is the noisy, low intensity signals (Figure 3.4b) whose trends are affected.

The importance of the data pre-treatment steps in the metabolomics workflow should not be overlooked as it can have a significant effect on the outcome of both multivariate and univariate analysis techniques (Gromski et al. 2015b). Commonly used techniques such as PCA and partial least squares discriminant analysis (PLS-DA) are sensitive to magnitude differences variance between variables. Random Forest is insensitive to variable magnitude and does not have the assumption requirements of parametric tests (Breiman 2001). It has been found to outperform many other classifiers when dealing with metabolomics data sets (Scott et al. 2013). Using this technique would not require significant data pre-treatment which is an important consideration when deciding upon appropriate downstream analyses.

Lastly, another consideration is the order in which these data pre-treatments are performed. Enot et al. (2008) suggests a \log_{10} transformation prior to a TIC normalisation. The effect of this can be seen in Figure 3.5c. This has had an effect of substantially distorting the data and likely contribute spurious information. The pseudo scaling effect of the \log_{10} transformation would have substantially increased the contribution of the noisy, low intensity signals to the TIC, post transformation. This in turn would reduce the ability of the TIC to correct for sample and machine variability.

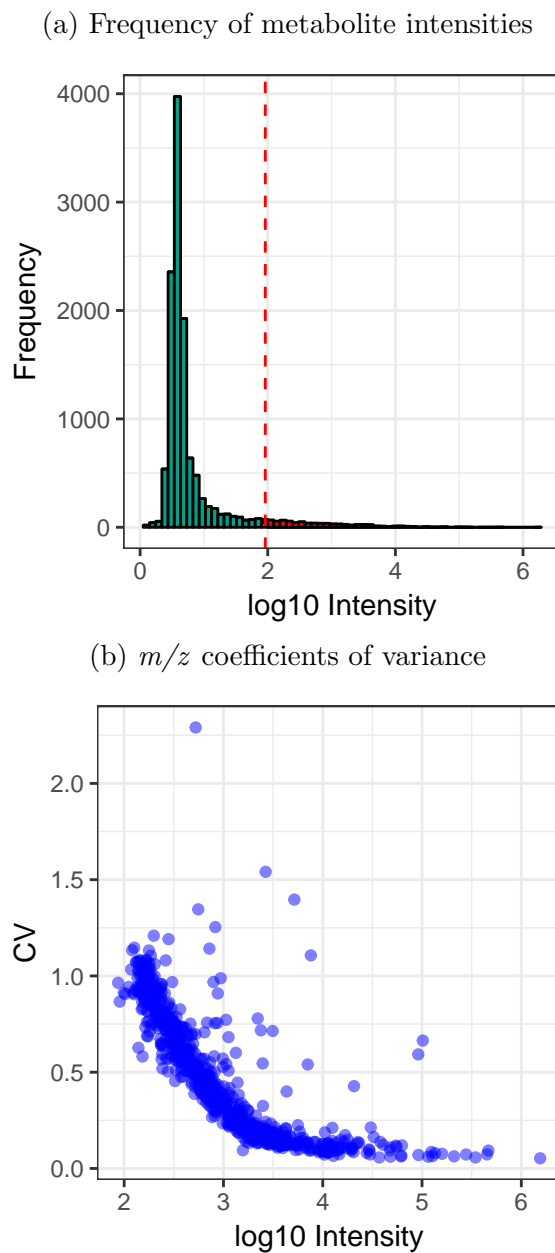


Figure 3.4: T

he example *B. distachyon* sample fingerprint m/z intensities and variability. Based on 43 injections of the example *B. distachyon* sample binned to 0.01 amu giving a total of 12180 bins. Intensities were averaged across all injections for (a). The dashed line in (a) indicates the point above which intensities above (bars coloured red) constitute 99% of the fingerprint TIC.

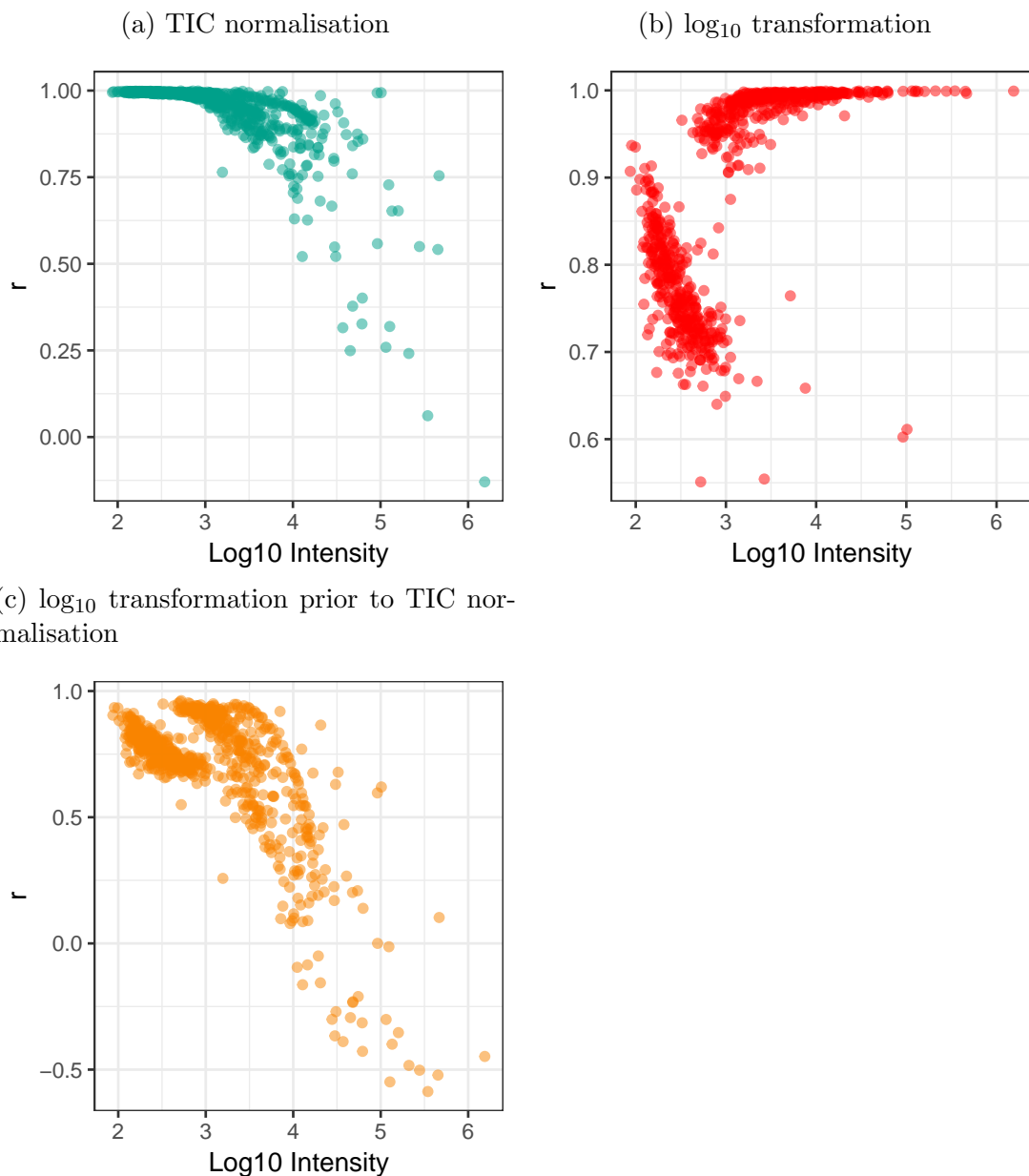


Figure 3.5: **The effect of TIC normalisation and \log_{10} transformation on fingerprint m/z .** Based on 43 injections of the example *B. distachyon* sample binned to 0.01 amu. r calculated by pearsons correlation between 765 raw and pre-treated negative ion mode bins.

3.4.3.2 Bin occupancy and missing value imputation

As can be seen in Table 3.2, zero or missing values can make up a large proportion (approx. 70% at a bin width of 0.01 amu) of spectral binned metabolome fingerprints. There are likely to be a number of sources of these including biological variation, variability in detection and the instrument limit of detection. Instrumental sources present a problem for analysing metabolome fingerprint data. A zero value in a fingerprint does not necessarily equate to an absence of a metabolite in the sample. Suitable treatment of missing values is crucial for removing noise and improving the interpretability of the metabolome fingerprinting data as many analysis techniques are heavily influenced by zero values (Hrydziusko and Viant 2012).

Figure 3.6 shows the distribution of missing values by bin occupancy for 10 technical injections of the example *B. distachyon* sample. Even without the addition of biological variation, 83% of bins have an occupancy of 50% or less. This represents a substantial proportion of the fingerprint that is likely to be background noise and variability in detection due to the co-occurrence of low occupancy and low signal intensity (Figure 3.6b).

A simple solution to dealing with missing values would be to exclude bins below a threshold occupancy from further analyses. However, this can be complicated by the presence of class structure within a data set as this could exclude bins showing presence and absence between classes. A more suitable solution in this case would be to only exclude bins whose maximum occupancy across all classes is below a certain threshold. This way, presence and absence trends will be retained; although a higher proportion of noise will also be retained. The former strategy is likely to be better suited to regression based analyses where there is no discrete class structure within the data.

The effect of these occupancy filtering strategies on a simple binary classification problem is shown in Table 3.3. The use of all bins retains a high proportion of missing values and so substantially reduces the Random Forest margin. Margins are a measure of classification performance and are discussed in Section

4.1.3. Using full occupancy filtering substantially reduces the number of variables, completely removes all missing values and improves the discrimination between classes. There is little difference between the Random Forest margins when class occupancy thresholding is used. Using a maximum occupancy threshold increases the number of variables by 139 compared to using a minimum threshold. However this also almost doubles the proportion of missing values present.

These results highlight the need for suitable variable filtering prior to classification as it can substantially improve the results by the removal of missing values and highly variable data from the analyses. However, the way in which this filtering is performed can also have an effect on the results of analyses and is also likely to be dependent on the experimental context. The presence of explanatory features within a data set that are analytically reproducible and have low numbers of missing values across all classes will be less affected by variable filtering than a data set where the presence of missing values has a biological origin. An example would include presence and absence metabolites, potentially present in plant pathogen interactions.

Similar issues with missing values have long been recognised in microarray based transcriptomics and a plethora of imputation algorithms exist that use a wide variety of strategies on which to base imputed values. These include global, local and knowledge based strategies (Moorthy, Mohamad, and Deris 2014). Missing values in metabolomics data sets have had comparatively little attention; however, it has been previously identified that k-nearest neighbour (kNN) imputation has provided an optimal method in direct injection Fourier transform ion cyclotron resonance mass spectrometry based fingerprinting (Hrydziuszko and Viant 2012).

Table 3.3 shows the effect of kNN imputation using different occupancy filtering strategies on Random Forest classification. Imputation using all bins, completely removed all discriminatory power between the classes compared to the un-imputed data. This likely reflects a reduction in the accuracy of the imputation algorithm due to the high proportion of noise and missing values (Jörnsten

et al. 2005). Imputation on minimum class occupancy thresholded data yielded a very similar Random Forest margins to the un-imputed. Interestingly, the imputation on maximum class occupancy thresholded data showed a drop in Random Forest margin that produced almost identical results to the minimum class occupancy thresholded data.

These results show that suitable variable filtering is needed prior to imputation. However, the similarity between imputed maximum and minimum class occupancy thresholded data reflects the presence of highly explanatory features that also contain very few missing values. Therefore, similar to the need to occupancy filter FIE-HRMS metabolome fingerprinting data, the suitability of missing value imputation will also be dependent on the experimental context.

3.4.4 Metabolite annotation using FIE-HRMS metabolome fingerprinting data

One of the key uses of metabolome fingerprinting is as a branch point in the metabolomics pipeline to inform decisions for further more targeted metabolite analyses. The putative annotation of explanatory features identified within this data thus provides the potential to further improve the utilisation of this technique within this context.

FIE-HRMS data has previously been used for putative metabolite identification in hierarchical approaches where a low resolution instrument is used for initial sample analysis and classification. Explanatory features are then targeted using a high resolution instrument on which class master mixes are analysed to obtain accurate mass information and allow putative metabolite identification (Draper et al. 2013).

A major advantage to using the spectral binning approach with high resolution data over these hierarchical approaches is that the accurate mass data is already available without the need for further instrument analyses. Sample electronic master mixes can be created by spectral binning the FIE-HRMS data using a bin width of 0.00001 amu. Accurate masses for explanatory features identified from

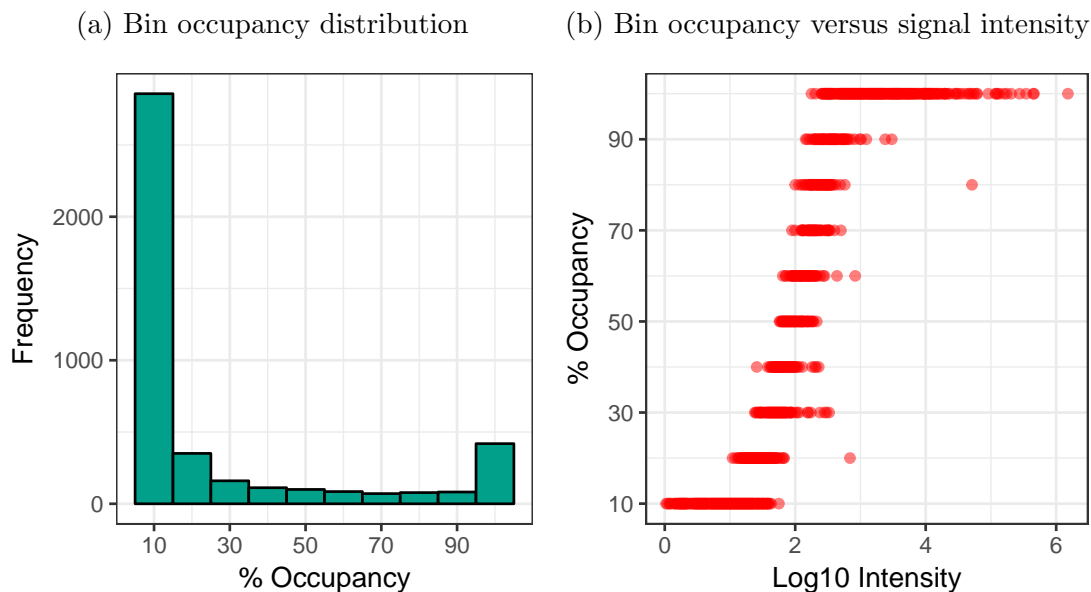


Figure 3.6: **Distributions of bin occupancy and signal intensity.** Percentage occupancy was calculated for negative mode across 10 technical injections of the example *B. distachyon* sample that included 4315 bins. a) shows the distribution of occupancy. b) shows occupancy plotted against intensity.

Table 3.3: **The effect of class occupancy filtering and kNN imputation on binary classification.** Negative ion mode FIE-HRMS metabolome fingerprint data was spectrally binned using a width of 0.01. Random Forest classification margins are shown using data from infected and control leaf tissue at 36 hours post inoculation from a resistant interaction between the *B.distachyon* ABR6 and *M. oryzae*. Further experimental details can be found in Section 5.3.

Data Matrix	No. Variables	% Missing Values	Random Forest Margin
All Bins	15632	89.3	0.20
Full Occupancy	374	0.0	0.35
Min. 2/3 Occupancy	660	6.6	0.30
Max. 2/3 Occupancy	799	12.3	0.33
All Bins Imputed	15632	0.0	0.07
Min. 2/3 Occupancy Imputed	660	0.0	0.29
Max. 2/3 Occupancy Imputed	799	0.0	0.28

analyses using 0.01 amu spectrally binned data can then be extracted and plotted similarly to what is shown in Figure 3.3a. These can then be assessed for the presence of multiple peaks or potentially split bins. Further binning could then be used to extract peak trends if multiple peaks are present to further identify which peak is explanatory. Alternatively algorithms such as the continuous wavelet transform could be used to peak pick within the given region and extract accurate masses for putative metabolite identification (Tautenhahn, Böttcher, and Neumann 2008).

The use of correlations to find associations between m/z within metabolome fingerprints is the first step in annotating a feature. Associations can include isotopic and adduct relationships as well as metabolically related metabolites such as those that are found within the same metabolic pathway (biotransformations) (Overy et al. 2008). With the increased resolution that spectral binning using a width of 0.01 amu provides, compared to that nominal mass binning, allows the potential for improved metabolite relationships to be identified. The more deconvolved nature of the bins means that differences between adduct, isotope and biotransformations will be better resolved.

Table 3.4 shows the mass changes for some common isotope, adduct and biotransformations found within FIE-MS fingerprints. For the most part, many of these relationships can be resolved at nominal mass, especially the adducts. However, there are a few important relationships that can only be resolved at a bin width of 0.01 amu and provide valuable information when annotating m/z features. One example is the ability to resolve the difference between a ^{13}C isotopic relationship and two related metabolites with a mass difference equivalent to the loss of an ammonia group and gain of an amine group. At nominal mass, both associations would have a difference of 1. However, using a bin width 0.01 would give the former a difference again of 1 but the later would give a difference of 0.98. This allows this relationship to be properly assigned without having to assess signal intensity ratios. This is shown in Tables 3.5 and 3.6 where the bin n133 is highly correlated, with a difference of 1, with n132 at nominal mass. At a bin width of 0.01 however, this difference can be resolved to 0.98 which can rule

out the ^{13}C isotopic relationship and give a likely molecular formula difference of an alcohol loss and amine group gain between the two metabolites in question. This difference has been deduced without the need to compare signal intensity ratios with relative isotopic abundance ratios. What can also be seen in Tables 3.5 and 3.6 is that these lists of correlations are very similar. The added 100 fold increase in resolution using a bin width of 0.01 allows substantially more confidence in assigning likely molecular relationships between the bins.

With the added *a priori* knowledge gained from correlation analyses using high resolution data, tools such as molecular formula generators and ionisation databases such as MZedDB can be used to assign putative molecular formula or metabolite identities with substantially greater confidence (Draper et al. 2009). This in turn greatly increases the quality of hypotheses that can be generated from this kind of metabolome fingerprinting data.

The confirmation beyond that of putative metabolite identities in metabolome fingerprints requires the comparison of MS/MSⁿ spectra derived from standards with those acquired from the sample in question (Overy et al. 2008). Due to the increased resolution obtained by FIE-HRMS metabolome fingerprinting, explanatory features become more deconvolved and metabolite identity can be more readily be assigned. However, the use of FIE-MS/MSⁿ for metabolite confirmation is limited as mass analyzers such as quadrupoles are only able to isolate a minimum m/z window of 0.1 amu. This means that fragmentation of m/z in the context of FIE-MS is infeasible and would require chromatographic separation to ensure the reliability of MS/MSⁿ spectra.

3.4.5 A general workflow for spectral binning based FIE-HRMS metabolomic fingerprinting analyses

The aspects of processing and analysing FIE-HRMS using a spectral binning approach that have been discussed previously, allows the development of a generalisable workflow that should be applicable to most experimental questions (Figure 3.7). Initial sample collection and preparation will likely be highly dependent on

Table 3.4: **Common m/z relationships found within FIE-MS metabolomic fingerprints.** Adduct mass changes are relative to a protonated or deprotonated parent ion depending on the acquisition mode.

Name	Type	Mass Change (amu)	MF Change	Mode
Dephosphorylation	Biotransformation	-79.96633	-[PO ₃ H ₂]+[H]	+/-
Decarboxylation	Biotransformation	-43.98983	-[CHO ₂]+[H]	+/-
H ₂ O loss	Adduct	-18.01056	[M+H-H ₂ O]1+	+
Dehydration	Biotransformation	-18.01056	-[H ₂ O]	+/-
Dehydrogenation	Biotransformation	-2.01565	-[H ₂]	+/-
Transamination 1	Biotransformation	-0.02381	-[O]+[NH ₂]	+/-
Ammonia ligation	Biotransformation	0.98402	-[OH]+[NH ₂]	+/-
¹³ C	Isotope	1.00335	iC13	+/-
³⁴ S	Isotope	1.99579	iS34	+/-
³⁷ Cl	Isotope	1.99704	iCl37	-
⁴¹ K	Isotope	1.99812	iK41	+/-
¹⁸ O	Isotope	2.00425	iO18	+/-
Hydrogenation	Biotransformation	2.01565	+ [H ₂]	+/-
Alcohol to carboxylic group	Biotransformation	13.97926	-[H ₂]+[O]	+/-
Methylation	Biotransformation	14.01565	-[H]+[CH ₃]	+/-
Transamination 2	Biotransformation	15.0109	-[H]+[NH ₂]	+/-
K and Na adduct difference	Adduct difference	15.97394	K-Na	+
Hydroxylation	Biotransformation	15.99491	-[H]+[OH]	+/-
NH ₄ adduct	Adduct	17.026	[M+NH ₄]1+	+
Hydration	Biotransformation	18.01056	+ [H ₂ O]	+/-
Na adduct	Adduct	21.9814	[M+Na]1+	+
Methyl to carboxylic acid	Biotransformation	29.97418	-[H ₃]+[HO ₂]	+/-
Cl adduct	Adduct	35.97723	[M+Cl]1-	-
K adduct	Adduct	37.95533	[M+K]1+	+
Carboxylation	Biotransformation	43.98983	-[H]+[CHO ₂]	+/-
Sulphation	Biotransformation	79.95681	-[H]+[SO ₃]	+/-
Phosphorylation	Biotransformation	79.96633	-[H]+[PO ₃ H ₂]	+/-

the experimental design and requirements of the biological tissue to be analysed. For instance, plant leaf tissue will require milling prior to the addition of extraction solvent whereas aqueous matrices such as urine will not. In most cases, a global extraction solvent of chloroform, methanol and water (1:2.5:1) will be suitable (Beckmann et al. 2008). Orbitrap FIE-HRMS fingerprinting methods are likely to differ little between sample types. Cleaning of the electrospray ionisation source and instrument calibration would be required between batches of samples.

Data processing, analysis and annotation can be achieved using a pseudo-hierarchical approach where different spectral bin sizes are used for statistical analyses and extraction of accurate mass peaks for annotation of explanatory features. This can be seen as a pseudo-hierarchical approach as the accurate mass information has already been acquired during the Orbitrap FIE-HRMS analyses and so the samples do not require further analysis on a high resolution instrument.

Explanatory bins relevant to the biological question should be identified by applying classification and feature selection techniques such as Random Forest on spectrally binned data using a bin width of 0.01 amu. Accurate masses for

Table 3.5: **Bin n132 correlations.** Top 15 significant correlations shown ($p < 0.05$) based on 43 injections of the example *B. distachyon* sample spectrally binned to nominal mass.

Bin	r	m/z Difference
n146	0.94	14.00
n97	0.93	-35.00
n145	0.91	13.00
n79	0.91	-53.00
n135	0.86	3.00
n115	0.83	-17.00
n102	0.83	-30.00
n341	0.82	209.00
n91	0.82	-41.00
n99	0.81	-33.00
n342	0.80	210.00
n133	0.80	1.00
n71	0.80	-61.00
n128	0.80	-4.00
n306	0.79	174.00

Table 3.6: **Bin n132.03 correlations.** Top 15 significant correlations shown ($p < 0.05$) based on 43 injections of the example *B. distachyon* sample spectrally binned to using a bin width of 0.01.

Bin	r	m/z Difference
n146.05	0.94	14.02
n96.97	0.93	-35.06
n96.96	0.91	-35.07
n145.06	0.91	13.03
n78.96	0.91	-53.07
n135.03	0.86	3.00
n379.16	0.85	247.13
n115	0.83	-17.03
n102.06	0.83	-29.97
n90.99	0.82	-41.04
n133.01	0.80	0.98
n71.01	0.80	-61.02
n128.04	0.79	-3.99
n306.08	0.79	174.05
n191.06	0.78	59.03

these explanatory bins can then be extracted from 0.00001 amu spectrally binned data as described in Section 3.4.4. This can then be used alongside the 0.01 amu binned data for putative annotation.

The putatively annotated, spectrally binned data would then be suitable for further analyses such as their integration with other omics data for hypothesis generation and biological interpretation. Alternatively, this data can be used to inform the use of further metabolomics analyses such as gas or liquid chromatography based MS or quantitative MRM analyses. FIE-HRMS analyses could also be used to ensure the integrity of samples or the identification of particular chemical classes within explanatory feature lists can be used to inform these further analyses.

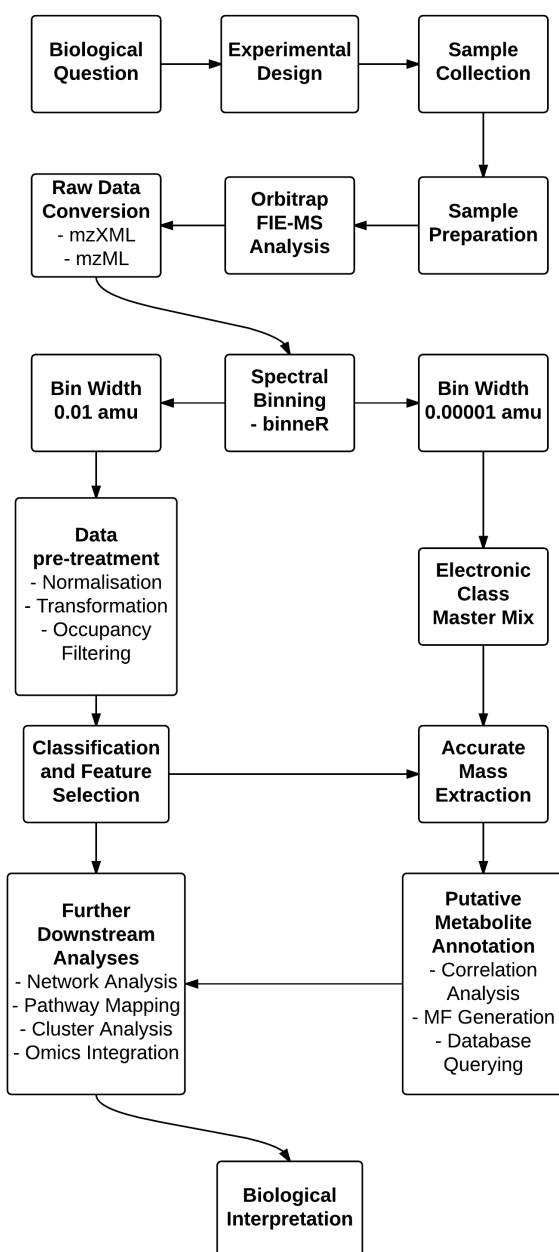
3.5 Concluding remarks

The central theme of this chapter was to present use of spectral binning as a viable and pragmatic technique for FIE-HRMS metabolomic fingerprinting. The development of the R package `binneR` allows implementation of this the technique. A bin width of 0.01 amu was established as the optimal bin width for spectral binning. Low computation requirements means that routine analysis in excess 1000 samples becomes feasible on a standard desktop PC.

With the analytical and data processing advantages that spectrally binned FIE-HRMS metabolome fingerprinting provides, it has been identified here that much of the raw data that is obtained is noisy and can be highly variable. Therefore rigorous data pre-treatment strategies need to be applied in order account for sample variability and missing data, without introducing bias or spurious knowledge into the data that would affect downstream analyses. The application of data pre-treatments is likely to be dependent on both experiment and sample context. However, it has also been shown that the order in which data pre-treatments are applied can have a significant effect on the data.

The increased resolution associated with FIE-HRMS fingerprinting also provides the potential for increased confidence in putative metabolite identification.

Figure 3.7: A general workflow for FIE-HRMS analyses.



The obvious result of this is the improved quality in hypothesis generation that can inform decisions regarding further chemical analyses and experimentation thus further cementing the role of FIE-HRMS fingerprinting as an important branch point in the metabolomics pipeline.

To conclude, this chapter has achieved its five aims as well as a proposed workflow in which the technique can be applied. The next chapter provides an example of the types of real world questions for which FIE-HRMS fingerprinting is suitable. Specifically, this involves the application of the technique in answering questions related to experimental control and robustness in the interaction between *B. distachyon* and *M. oryzae*.

Chapter 4

Experimental control and robustness in omics analyses of plant-pathogen interactions

4.1 Introduction

When designing experiments involving plant-pathogen interactions, it is important that all parameters are appropriately controlled to ensure that results are both reproducible and the hypotheses generated are relevant to the biological question. This becomes more crucial with large scale integrated omics experiments, where there will be a high investment of both time and resources at all stages; from the experimental set-up to sample analysis. As the number of measured variables increases, so too does the likelihood that an unknown confounding factor will contribute to those variables and give rise to false discovery (Smith, Ventura, and Prince 2013).

The utility of applying omics technologies to problems such as plant-pathogen interactions is to allow inductive hypothesis generation that can direct further scientific investigation through more traditional deductive methods. Without careful appreciation of all the factors that underpin these complex interactions, the quality of hypothesis generation can be misinformed.

4.1.1 Experimental considerations for omics experiments involving plant-pathogen interactions

Due to the dynamic nature of plant-pathogen interactions a number of different parameters need to be accounted for in order to effectively control for interacting plant processes and reduce variability between samples. Different omics analyses can also have different sampling and preparation requirements which have to be accounted for when designing integrated analyses.

Sources of variability within experiments of plant pathogen interactions can occur at all stages of the experimental process, from initial pathogen inoculation, to sample collection, sample preparation and sample analysis (Beckmann et al. 2008). These sources will also be different depending on the patho-system in question as different inoculation strategies are required for different plant pathogens.

For the interaction between *B. distachyon* and *M. oryzae* there are a number of patho-system requirements that are integral to ensuring successful and uniform pathogenesis. For instance, the fungal spores need to be applied to the leaf surface, aurally and in suspension, to initiate spore germination (Parker et al. 2008). This requires a surfactant to allow adherence to the highly hydrophobic *B. distachyon* leaf surface. The fungal spores also require high humidity conditions (> 80%) in the initial phases of infection up to initial host cell penetration (Li, Uddin, and Kaminski 2014). Parker et al. (2008) used the addition of gelatine to the inoculum as a surfactant and placed inoculated plants into clear plastic bags as solutions maintain high humidity. These are artificial experimental factors added to this interaction that would not be found in the natural environment; they are however, requirements for successful pathogenesis under laboratory conditions. Therefore, they need to be appropriately controlled for by ensuring that there are gelatine solution inoculated control plants that are placed under the same high humidity conditions as the pathogen inoculated plants.

Plants are also under the influence of circadian rhythms and cellular development that are likely to influence host responses to pathogen infection (Roden and Ingle 2009). It is therefore important that plant tissues of the same devel-

opmental stage are sampled and that infected and control inoculated plants are sampled at the same point during the diurnal cycle to control for these changes.

Plant tissue harvesting and sample preparation can also influence the quality and variability of samples for omics analyses. Care needs to be taken not to introduce artificial variance that could alter or mask true changes in metabolites, proteins or transcripts depending on the omics level being studied. Immediate snap freezing of samples in liquid nitrogen is essential for halting all cellular processes to provide a snap shot of the cellular state at the time of sampling (De Vos et al. 2007). Also, for non quantitative metabolomics techniques such as FIE-MS, uniform sample dry weight is essential to ensure that relative metabolite concentrations are comparable between samples (Beckmann et al. 2008). This is less important for omics such as RNA-seq based transcriptomics as RNA extract concentration can be adjusted prior to cDNA library preparation and differences in sequencing depth accounted for by normalisation (Dillies et al. 2013).

4.1.2 Random Forest for metabolomic data mining

Essential to extracting valuable information from metabolomic data sets is the use of data mining techniques that are able handle the their complex and high dimensional nature. Random Forest is ensemble machine learning method based on the use of forests of decision trees, grown to make predictions, yielding both classification accuracy and variable importance (Breiman 2001).

A single tree is grown using a bootstrapped subset of samples. At each node, the variable that best splits the node is selected from a random subset of all variables, based on the classification and regression trees split criterion. Tree leaves (terminal nodes) contain a fixed pre-specified number of observations. Bagging allows aggregation of predictions across the forest, yielding classification accuracy and variable importance values (Breiman 1996).

Random Forest is unaffected the scaling of data and are able to handle missing values. It is also able to handle both high dimensional and correlated data without the need for prior dimension reduction, both of which are common in

metabolomics data sets (Enot et al. 2008). It can yield both classification and variable importance results that other machine learning techniques such as support vector machines and kNN are unable to do (Sandri and Zuccolotto 2006). Random Forest has been shown to out perform other classifiers such as PC-LDA and support vector machines when applied to chemometric data sets (Gromski et al. 2015b; Scott et al. 2013)

There are three ways that variable importance can be measured with Random Forest. These include the mean decrease in accuracy, the Gini impurity index and selection frequencies. The mean decrease in accuracy or permuted importance, uses the out of bag error estimate. When a given variable is left out, if classification error increases then it can be assumed that the variable contributes to the true classification of the observations. It can therefore be considered important.

The Gini impurity index can be determined by how well variables can split the observations at individual nodes. It measures the performance of a variable in separating the observations at the parent node into the left and right daughter nodes. An increase in the Gini impurity index between two nodes indicates a strong association with the true classification. A decrease in Gini impurity index for a variable between nodes indicates an increase in the extent of splitting and therefore a high variable rank.

An issue with the use of the mean decrease in accuracy and Gini impurity index for variable importance is that they are relative to each individual forest. They require heuristic thresholds to be set in order to compare feature subsets (Konukoglu and Ganz 2014). Selection frequencies can be used as an alternative method for variable importance. At each node within each individual tree of a forest, the variable that is best able to split the data is selected. The selection frequency is the number of times a given feature is selected across all nodes in the entire forest. If there is no significant relationships between observation labels and variables present in the forest, the probability of any given variable being selected is:

$$P(f_n^* = f) = \frac{1}{F} \quad (4.1)$$

Where P is the probability and F is the total number of variables. The independence of trees within the forest means that variable selection for a given node does not influence subsequent nodes. The probability of a variable being selected then extends to any node within the forest (Konukoglu and Ganz 2014). An approximate false positive rate for a given feature can be estimated using binomial distribution using the equation:

$$P(C_{k,T}^f) = \binom{TK}{k} \left(\frac{1}{F}\right)^k \left(1 - \frac{1}{F}\right)^{TK-k} \quad (4.2)$$

Where T is the number of trees in the forest, K is the average number of nodes per tree across the forest, k is the variable selection frequency and F is the total number of variables.

4.1.3 Robustness and validation in omics analyses

Omics analyses by their very nature generate complex, high dimensional data sets. Adding to this, most omics experiments have a very low sample to feature ratio, usually due to high cost limiting sample analysis (Ioannidis and Khoury 2011). This severely limits the statistical power available to investigators, therefore robust data mining techniques such as Random Forest and validation strategies become critical for identifying relevant explanatory features, allowing confidence in biological interpretation and an assessment of experimental robustness.

A common practice for supervised classification in bioinformatics and chemometrics is to use re-sampling to account for sample paucity; where sample sets are split into separate training and test sets. The training set can be used to construct the model with *a priori* knowledge of sample class. The model can then be used to predict the class labels of the test set and allow evaluation of

the model performance. This process can be repeated to improve the precision of performance estimates. A number of re-sampling strategies exist; cross validation is popular in chemometrics but others such as bootstrapping are widely used in statistics (Scott et al. 2013).

It has been identified previously that internal validation via re-sampling can potentially overestimate classifier performance and lead to bias when validating molecular classifiers (Castaldi, Dahabreh, and Ioannidis 2011). External validation using independently collected samples is the only sure method for avoiding bias and over-fitting; however, its use in the validation of omics data sets is still very limited with the biomedical sciences being one of the few areas where external validation is essential (Collins et al. 2014).

In the case of plant-pathogen interaction experiments, different levels of sample independence exist. This could be separately performed inoculations and sample analysis by the same investigators or inoculations and sample analyses performed by independent investigators.

A binary classification model's performance can be assessed by its ability to correctly or incorrectly predict sample class labels. These measures include the accuracy, AUC, Cohen's Kappa coefficient and the margin. Accuracy, which is the proportion of correctly predicted samples, assesses a model's overall effectiveness. AUC accounts for the relationship between the model's sensitivity and specificity and assesses a model's ability to avoid false classification (Obuchowski, Lieber, and Wians 2004). Cohen's Kappa coefficient assesses a model's accuracy in relation to the expected accuracy and accounts for random chance (Ben-David 2008). The margin is the difference between the true positives and maximum number of votes for another class. This gives a finer estimate for model performance (Enot and Draper 2007).

Not only is it important to assess classifier performance when comparing models but it is also important to assess the feature stability underlying these models. Two models could have similar performance measures, however it may be different features responsible for this. This is important in omics analyses as it is the

explanatory features that are of interest in the biological interpretation of these data (He and Yu 2010). There are numerous methods available for assessing feature stability. Methods such as the Jaccard's Index can be used to compare feature subsets (Saeys, Abeel, and Peer 2008), with the equation shown below:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (4.3)$$

Here A and B are feature subsets of arbitrary cardinality. This is useful for comparing lists of explanatory features above a certain threshold. Other methods, such as the Canberra distance, are useful for comparing the rankings of entire feature lists (Jurman et al. 2008). The equation is shown below:

$$d(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^n \frac{|p_i - q_i|}{|p_i| + |q_i|} \quad (4.4)$$

Where \mathbf{p} and \mathbf{q} are vectors of rankings with identical cardinality. Relative Canberra distance can be calculated by dividing the Canberra distance by the maximum possible Canberra distance for the given vector lengths. This measure has the advantage over others such as Spearman's rank coefficient as it gives weight to the top ranked features that are of more interest (He and Yu 2010).

4.2 Aims

The aim of this chapter is to objectively assess key elements of the *B. distachyon* and *M. oryzae* patho-system related to appropriate experimental control and the robustness of the metabolomic changes that are observed during the interaction. Specifically, this relates to the potential of the fungal inoculum to elicit plant responses other than those specifically related to infection. This could have major implications for the quality and relevance of hypotheses generated from omics data from this interaction. Also important to the relevance of generated hypotheses is the reproducibility of the data underpinning these. If the experimental

conditions cannot be sufficiently well replicated between independent inoculation events then hypotheses generated cannot have practical relevance in elucidating the molecular mechanisms underlying the interaction. This provides us with the following aims:

- Assess the extent of impact the inoculum has on the interaction metabolome.
- Develop methods for controlling potentially inoculum related responses.
- Assess the robustness of interaction metabolomic responses across multiple inoculations.

4.3 Materials and Methods

4.3.1 Inoculations to investigate the inoculum related metabolomic changes in *B. distachyon* as a result of *M. oryzae* inoculation

The *B. distachyon* ecotype Bd21 and *M. oryzae* were grown as described in Sections 2.1 and 2.2. Three inoculum treatments were used; gelatine only solution as a surfactant control, a non-pathogenic inoculum containing inviable *M. oryzae* spores to control for the inoculum constituents and a pathogenic inoculum containing viable *M. oryzae* spores. The preparation of the pathogenic and non-pathogenic inoculum are described in Section 2.3. The gelatine solution was prepared as 0.2% (w/v). Plants were inoculated and placed under high humidity conditions as described in Section 2.3. 10 replicate plants were harvested as described in Section 2.4 at 0, 24 and 48 hpi.

Samples were globally extracted as described in Section 2.6. FIE-HRMS metabolome fingerprinting analysis and data pre-processing was performed as described in Section 2.7. PCA and PC-LDA were used to initially assess FIE-HRMS data quality. Random Forest classification and feature selection were used to assess class discrimination and explanatory features as described in Sec-

tion 2.9. A 1% selection frequency false positive rate (FPR) was used as a cut off for explanatory features.

4.3.2 Inoculum preparation and LC-MS analyses to investigate the effect of centrifugation on inoculum constituents

To investigate the effect of centrifugation on inoculum constituents, a pathogenic inoculum was prepared which was then centrifuged and washed with the supernatant retained. This gave four inoculum components for investigation. The gelatine solution and pathogenic inoculum were prepared as described in Sections 4.3.1 and 2.3 respectively. The inoculum supernatant and centrifuged inoculum were prepared by taking a portion of the pathogenic inoculum and centrifuging at 2500rpm for 5 minutes with the supernatant poured off and kept. The remaining pellet was then re-suspended in gelatine solution, vortexed and re-centrifuged as before. The supernatant was poured off and discarded and the pellet re-suspended in gelatine solution.

Samples were extracted by adding 500 µl of methanol to 500 µl of inoculum component. The samples were then sonicated for 5 minutes, shaken for 20 minutes at 1400 rpm and centrifuged at 14000 rpm for 4 minutes. The supernatant was then pipetted into a fresh 2ml eppendorf. Samples were prepared for LC-HRMS analysis as described in Section 2.6. LC-HRMS analysis was performed as described in Section 2.8.

4.3.3 Independent inoculations of *B. distachyon* with *M. oryzae* and Random Forest classification using an external validation re-sampling strategy

For investigation of the robustness of metabolome responses during the pre-symptomatic phases of the susceptible and resistant *B. distachyon* and *M. oryzae* interactions, metabolome fingerprinting data was used from the three indepen-

dent inoculations described in Section 2.5. All six time points were used (0 - 60 hpi) for both ecotypes (ABR6 and Bd21) with 12 replicates for each treatment and control class at each time point. Sample extraction, FIE-HRMS fingerprinting, spectral binning and pre-treatment were performed as described in Sections 2.6 and 2.7 respectively. Samples from all three inoculations were randomised together into equally represented blocks for sample extraction and again for FIE-HRMS fingerprinting.

To compare metabolome responses in each of the independent inoculations, Random Forest classification was performed using an external validation re-sampling strategy using R (version 3.2.3) and the randomForest package with a custom written script (Appendix D). 10 re-sampling iterations of Random Forest using 1000 trees were performed for each of the 3 possible experiment training and test combinations (1+2~3, 1+3~2, 2+3~1). Each combination consisted of binary comparisons between infected and control treatments for each of the 6 time points for each of the ecotypes. Training and test partitions were sampled without replacement using an external validation strategy where 8 replicates were taken from each of 2 experiments for each treatment for training and 8 replicates were sampled without replacement from the remaining experiment for each treatment for testing. Random Forest model performance measures (accuracy, Cohen's Kappa coefficients, AUC and margin) were calculated at each re-sampling iteration and mean aggregated across iterations. False positive rates for feature selection frequencies were also calculated at each re-sampling iteration (as described in Section 4.1.2) and mean aggregated across iterations.

4.4 Results and Discussion

4.4.1 Experimental control of the *B. distachyon* and *M. oryzae* interaction

Previous metabolomic studies using this interaction by Allwood et al. (2006), Parker (2006) and Zubair (2014) have used control plants that have been inoc-

ulated with only gelatine solution, to control for its use as a surfactant in the preparation of the inoculum. Inoculations conducted at the commencement of this PhD project indicated that the inoculum was having an additive effect, other than those of the intended pathogen infection. It was observed that significant discrimination between inoculated and control plants could be obtained at 0 hpi (see Section 4.4.1.2). This would not be expected as infection would not have yet taken place at this time. Compounding this effect, is that for studying the pre-symptomatic phases of this interaction, a higher density of spores is needed to elicit a detectable metabolomic response due to the relatively low number of cells involved during initial host colonisation (approximately 1-3 host cells per fungal spore) (O’Connell and Panstruga 2006). It was inferred that this discrimination was as a result of initial differences between the inoculum and the gelatine solution control due to the inoculum preparation. Therefore, an experiment was undertaken in order to test this hypothesis and to allow the effect of the inoculum upon the host to be assessed. Three treatment types were used: the standard gelatine solution, the standard pathogenic inoculum at two times spore density and a non-pathogenic inoculum containing non-viable *M. oryzae* spores (see Section 4.3.1).

4.4.1.1 Control treatments for inoculating *B. distachyon* with *M. oryzae*

Common sterilisation methods include heat, radiation and chemical treatments. The use of heat to sterilise the inoculum was considered unsuitable as the temperatures that would be needed in order to effectively kill the spores would also be likely to cause a breakdown of molecules within the inoculum. This would be detectable in the chemical fingerprints and add to discrimination between the inoculum treatments at 0 hpi. Radiation treatments were also considered inappropriate for similar reasons. Ultra violet light is commonly used for sterilisation. This would cause breakdown of sensitive molecules within the inoculum and thus chemically alter it (Braga et al. 2015).

There are numerous chemical treatments for sterilising fungal spores. Many of these such as formaldehyde are carcinogenic and likely to not only be toxic to the spores but also to the plant when the inoculum is applied so is likely to elicit a host response. Also the addition of a chemical to the inoculum will alter its composition.

Another option would to be to use genetically modified spores that arrest in development prior to primary host cell penetration (Xu and Hamer 1996). This would not require pre-treatment of the inoculum and provide a control for fungal constituents other than the fungal spores. However, it would be difficult to control the spore concentration as well as inoculum concentration so that it would be exactly comparable to that of the pathogenic inoculum, as two separate fungal cultures would be necessary to produce the pathogenic and non-pathogenic inoculum.

The treatment that was chosen in order to effectively sterilize the inoculum, with the least impact upon it, was to use two rounds of snap freezing in liquid nitrogen, thawing using a water bath and sonication for 5 minutes. A similar method has previously been used for quenching yeast cultures prior to extraction for metabolomic analysis (Murray, Beckmann, and Kitano 2007). The effect of this treatment is to rupture the spores; however, this was not considered to be detrimental. Any enzymes released that have the potential to chemically alter the inoculum would likely to have already been released during the scraping of the fungal hyphae from the surface of the PDA media, so will have also affected the untreated virulent inoculum. Bd21 plants did not show any visible signs of infection 5 days after inoculation, neither was *M. oryzae* growth obtained when PDA media was inoculated with this inoculum confirming its complete neutralisation.

4.4.1.2 Host responses to the pathogen inoculum

Table 4.1 shows the Random Forest classification results of pairwise comparisons between each treatment type at each sampled time point. There were explana-

Table 4.1: **Random Forest classification results for comparisons of inoculation treatment responses.** GS = gelatine solution; PI = pathogenic inoculum; NPI = non-pathogenic inoculum.

hpi	Comparison	Accuracy	AUC	Margin
0	GS vs PI	0.96	1.00	0.33
	GS vs NPI	0.92	0.99	0.27
	PI vs NPI	0.32	0.29	-0.08
24	GS vs PI	0.99	1.00	0.46
	GS vs NPI	1.00	1.00	0.49
	PI vs NPI	0.57	0.65	0.03
48	GS vs PI	1.00	1.00	0.59
	GS vs NPI	1.00	1.00	0.56
	PI vs NPI	0.96	0.99	0.33

tory margin values (> 0.3) obtained for comparisons between the gelatine solution and pathogenic inoculum and the gelatine solution and the non-pathogenic inoculum at all time points, with the margins increasing as the hpi increases. Non-significant margin values were obtained in comparisons between the pathogenic inoculum and the non-pathogenic inoculum at both 0 and 24 hpi; however, significant margin values were obtained at 48 hpi.

These results indicate that not only is there a difference in chemical composition between both the non-pathogenic inoculum and the gelatine solution at 0 hpi, but the increased margin values at subsequent time points suggests that there is a differential response to this inoculum. These plant tissue responses cannot only be attributed to the presence of the spores as there are no viable spores within the non-pathogenic inoculum.

If metabolites present in the inoculum were acting passively upon the host leaf surface then only metabolite trends such as Figure 4.1a would be expected for explanatory m/z at 0 hpi. This m/z is showing a consistent increase between the inoculum and gelatine solution treatments across all time points. However, metabolite trends such as those in Figures 4.1 b, c & d indicate that constituents of the inoculum other than the spores are also contributing to the interaction metabolomic response. The metabolite trend shown in Figure 4.1b is similar to that of Figure 4.1a in that it is explanatory at all time points between the inoculum and the gelatine solution treatments; however, it also increases in level

as the experiment progresses. The metabolite trend shown in Figure 4.1c is not explanatory at 0 hpi but it is explanatory at both 24 and 48 hpi. This can likely be attributed to a response to the presence of the pathogenic inoculum as it is not explanatory in comparisons between the pathogenic inoculum and non-pathogenic inoculum at any time point. The metabolite trend shown in Figure 4.1d is most interesting at 48 hours where opposite trends are found in the infected tissue, depending on which control treatment is used. The response would be increased if taken relative to the gelatine solution but decreased if relative to the non-pathogenic inoculum.

Zubair (2014) advocated filtering out m/z that are explanatory at 0 hpi as a suitable strategy to account for the significant discrimination between inoculum and control at 0 hpi. This strategy would be suitable for effectively accounting for features showing trends such in Figure 4.1a removing some or most of the variance for which the inoculum is directly responsible. However, it would not account for the variance that is caused by the underlying host responses to inoculum constituents other than the pathogen spores. It would be an unsuitable strategy for accounting for metabolite trends shown in Figures 4.1 b, c & d.

4.4.1.3 Inoculum constituents

The putative annotations for inoculum associated metabolites that were found to be explanatory in both comparisons between the gelatine solution and inoculum or neutral inoculum are shown in Table 4.2. All these features are increased in the inoculated plants with positive fold changes. The main sources of these metabolites are likely to be through release from the break up of fungal mycelia and those dissolved from the PDA media, on which the *M. oryzae* is cultured during inoculum preparation.

The annotations include a number of fatty acids such as linoleic acid. Fatty acids are important signalling molecules within plants and have been associated with responses to both biotic and abiotic stresses (Hou, Ufer, and Bartels 2015). Lipxygenase oxidation products of fatty acids such as linoleic acid and have been

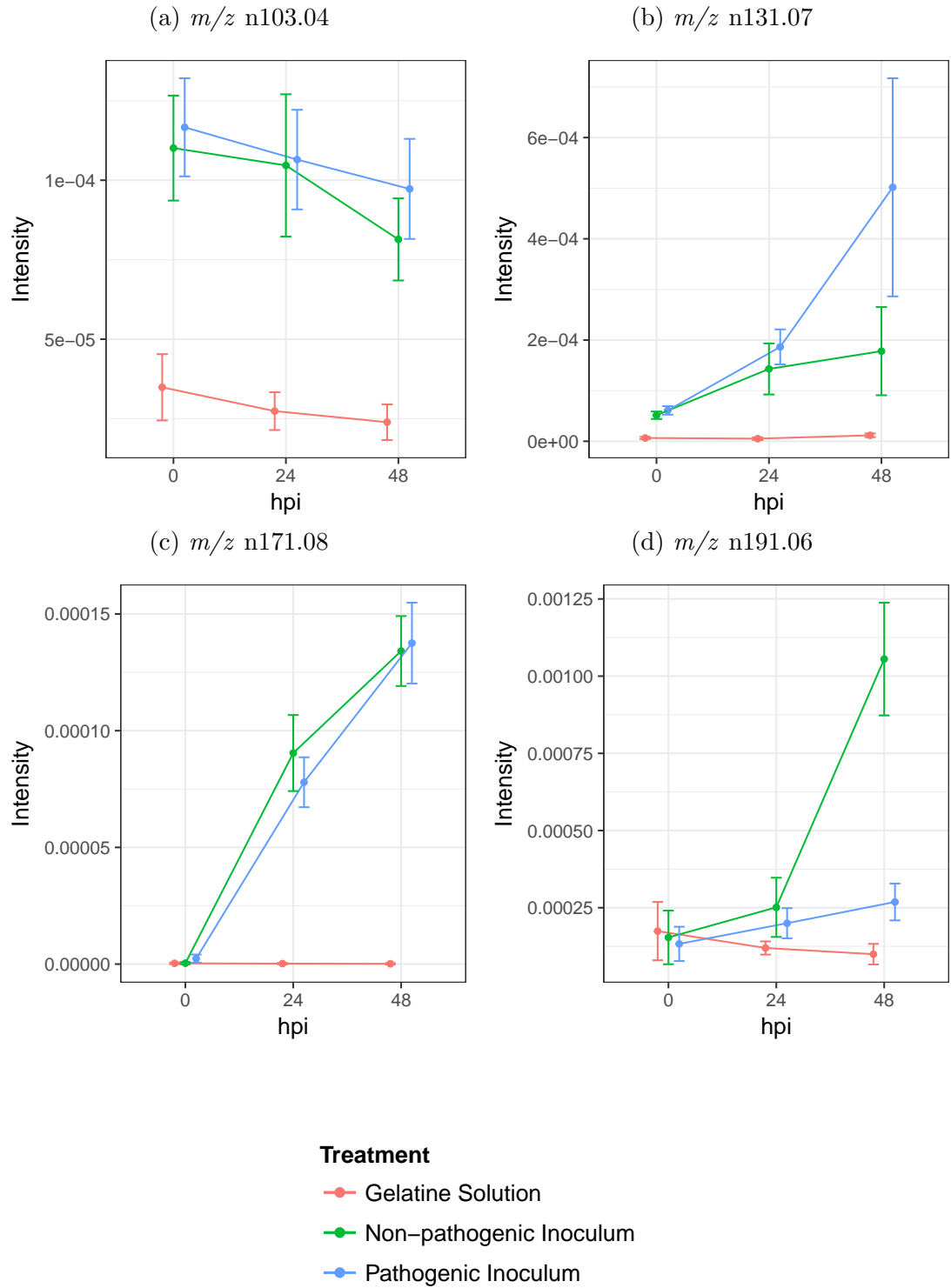


Figure 4.1: **Example m/z trends of explanatory features between inoculation treatment comparisons.** Points show class mean intensities ($N=10$). Error bars show 95% confidence intervals estimated using the t distribution.

found to accumulate in plant tissues during pathogen infection. There is evidence that they are linked to salicylic acid signalling, systemic acquired resistance and have been found to accumulate during the hypersensitive response in a number of plant pathogen interactions (Kachroo and Kachroo 2009). Also the accumulation hydroxy linolenic acid in rice has been found to enhance resistance to *M. oryzae* by inhibiting its growth (Yara et al. 2008).

Glucose was also identified as an inoculum associated metabolite (Table 4.2). Not only is it an important energy source within plant cells but is also recognised as an important signalling molecule (Rolland, Baena-Gonzalez, and Sheen 2006). Glucose levels are sensed by plant cells both intracellularly via hexokinase and extracellularly via G proteins. They allow the plant to monitor cellular homeostasis and energy status (Hanson and Smeekens 2009). Sugar sensing has been linked to both plant innate immunity and responses to stress (Baena-González and Sheen 2008; Moghaddam and Van Den Ende 2012). Alterations to sucrose:hexose ratios within plant tissues have been shown to induce the anthocyanin biosynthesis in *Arabidopsis* (Solfanelli et al. 2006). The application of foliar carbohydrates has also been suggested as a potential method for crop protection by priming the innate plant immune response (Trouvelot et al. 2014).

The presence of high levels of glucose on the leaf surface could stimulate the blooming of epiphytic microbes. Epiphytic bacteria have been shown to readily uptake simple sugars on the leaf surface (Mercier and Lindow 2000). If this blooming is sufficient, this is likely to add to the composition of the leaf metabolome as well as deplete inoculum glucose levels. This would add to treatment class discrimination and present glucose trends that would be unrepresentative of changes that could potentially be occurring as a result of *M. oryzae* infection.

Other inoculum constituents will be recognised by the plant cells as PAMPs that elicit innate immune responses (Liu et al. 2013). Although the neutral inoculum is unable to elicit a full PAMP triggered immune response, resulting in the hypersensitive response, just the presence of fungal cell wall fragments such as chitin are likely to be recognised and responded to by the plant cells (Liu et al.

Table 4.2: **Putative annotations of inoculum associated metabolites.** m/z bins explanatory at 0 hpi in both comparisons between gelatine solution and the inoculum or neutral inoculum comparisons. FC refers to \log_2 fold changes between the inoculum (I) or neutral inoculum (N) and gelatine solution (G). P/A is presence and absence with the bin absent in the gelatine solution. MF is the compound molecular formula.

Bin	FC (I/G)	FC (N/G)	m/z	Name	MF	Adduct	Theoretical m/z	PPM Error
n103.04	1.74	1.66	103.04008	Hydroxy-butyric acid	C4H8O3	[M-H] ⁻	103.04007	0.113
n122.02	3.91	3.49	122.02471	Nicotinic acid	C6H5NO2	[M-H] ⁻	122.02475	-0.347
n129.06	1.17	1.13	129.05569	Keto-hexanoic acid	C6H10O3	[M-H] ⁻	129.05572	-0.22
n131.07	3.24	2.99	131.07132	Leucic acid	C6H12O3	[M-H] ⁻	131.07137	-0.369
n147.07	5.19	5.04	147.06630	Mevalonic acid	C6H12O4	[M-H] ⁻	147.06628	0.113
n163.02	8.71	8.73	163.02473	Dehydro-xyloic acid	C5H8O6	[M-H] ⁻	163.02481	-0.511
n215.03	1.88	1.85	215.03300	Glucose	C6H12O6	[M+Cl] ⁻	215.03279	0.971
n217.03	2.73	2.78	217.03023	Glucose ³⁷ Cl	C6H12O6	[M+iCl] ⁻	215.03279	0.971
n218.1	2.87	2.6	218.10347	Pantothenic acid	C9H17NO5	[M-H] ⁻	218.1034	0.333
n227.1	P/A	P/A	227.10463	Unknown	-	-	-	-
n277.03	3.16	3.06	277.03262	Unknown	-	-	-	-
n279.23	4.69	4.79	279.23215	Linoleic acid	C18H32O2	[M-H] ⁻	279.233	-2.877
n280.24	4.87	5.41	280.23584	Linoleic acid ¹³ C	C17iCH32O2	[M-H] ⁻	280.23576	1.634
n283.08	4.24	4.2	283.07953	Unknown	-	-	-	-
n293.21	7.88	7.6	293.21140	Hydroxy-linolenic acid	C18H30O3	[M-H] ⁻	293.21222	-2.791
n295.23	12.29	11.67	295.22714	Keto-oleic acid	C18H32O3	[M-H] ⁻	295.22787	-2.467
n306.06	5.35	4.86	306.05908	Unknown	-	-	-	-
n309.21	7.88	7.65	309.20569	Dihydroxy-octadecatrienoic acid	C18H30O4	[M-H] ⁻	309.2071	-4.668
n311.22	11.72	11.26	311.22195	Hydroperoxy-linoleic acid	C18H32O4	[M-H] ⁻	311.22278	-2.678

2013).

The inoculum contains metabolites that can both be used as nutrients and as key parts of signalling cascades within plant cells. When the inoculum is sprayed upon the leaf surface, these metabolites still need to breach highly hydrophobic cuticular layer of the *B. distachyon* leaf. The permeability of plant surfaces to dissolved nutrients has been extensively studied. Liphophilic compounds are able to directly permeate the cuticular layer; however the mechanisms by which polar compounds can directly permeate are still poorly understood (Schönherr 2006). Trichomes and stomatal pores on the *B. distachyon* leaf surface are likely areas where uptake of dissolved inoculum constituents could occur (Fernández and Brown 2013).

The composition of apoplastic spaces are highly monitored by plant cell membranes, especially during plant pathogen invasion (Sattelmacher 2001; Pignocchi and Foyer 2003). Stomatal entry would allow these inoculum constituents to cause a compositional change in the apoplast and responses by cells in the mesophyll layers. Stomatal uptake is also likely to be exacerbated by stomatal opening caused by the high humidity conditions in which the plants are placed post inoculation to enhance *M. oryzae* spore germination.

4.4.1.4 Reducing the impact of inoculum associated metabolites

As has been shown in the preceding sections, the pathogen inoculum can have a substantial influence on host responses other than *M. oryzae* pathogenesis. Figure 4.2 shows C₁₈ LC-HRMS ion chromatograms of the separate inoculum components. There are substantial composition differences between the gelatine solution and the inoculum. Much of this composition can be removed simply by centrifuging the inoculum, removing the supernatant and re-suspending the debris in gelatine solution. This reduction in inoculum constituents can clearly be seen between the ion chromatogram of the inoculum and centrifuged inoculum.

Difficulty was found in trying remove all hyphal fragments from the inoculum. There was a three quarters reduction in inoculum spore density from

2×10^5 conidia/ml to 0.5×10^5 conidia/ml when filtering was used due to the matted nature of the hyphal fragments. This would require four times as many *M. oryzae* cultures to provide an inoculum of sufficient spore density. This was seen as logistically impractical, especially for large scale inoculations.

Combining this inoculum cleaning by centrifugation with neutralising the spores for control treatments provided an inoculation method that not only substantially reduces the host responses to it, but controls for any host responses elicited by PAMPs still present. PAMP triggered responses by the host would be seen as confounding to the biological question as they can be elicited even when plants are inoculated with non-viable spores.

4.4.2 Assessing the robustness of patho-system metabolomic changes

Individual analysis of independent inoculation experiments suggests that there is variability in the consistency of explanatory features found between each experiment using Random Forest feature selection in both compatible and incompatible interactions. In both interactions, only 19-20% of explanatory features were found consistently in all 3 experiments (Figure 4.3). The proportions of features shared by at least two experiments also varies. In the incompatible interaction of ABR6, only 19.2% and 20.5% of features were found to be shared between experiments 1 & 2 and 1 & 3 respectively. This is also true for the compatible interaction of Bd21 but the difference was less pronounced. The issue with assessing the interaction metabolome responses in this manner is that it does not account for the reproducibility of the underlying metabolite trends. Explanatory feature list occupancy doesn't describe whether a feature is either increased or decreased or to what extent. Also placing all the samples into one large re-sampling model to see if discrimination can still be achieved would also be ineffective. The training and test sets would not effectively account for the independence of the inoculation experiments and so potentially bias results. However, more replicates could be used in the training and test sets, although the re-sampling results would be less

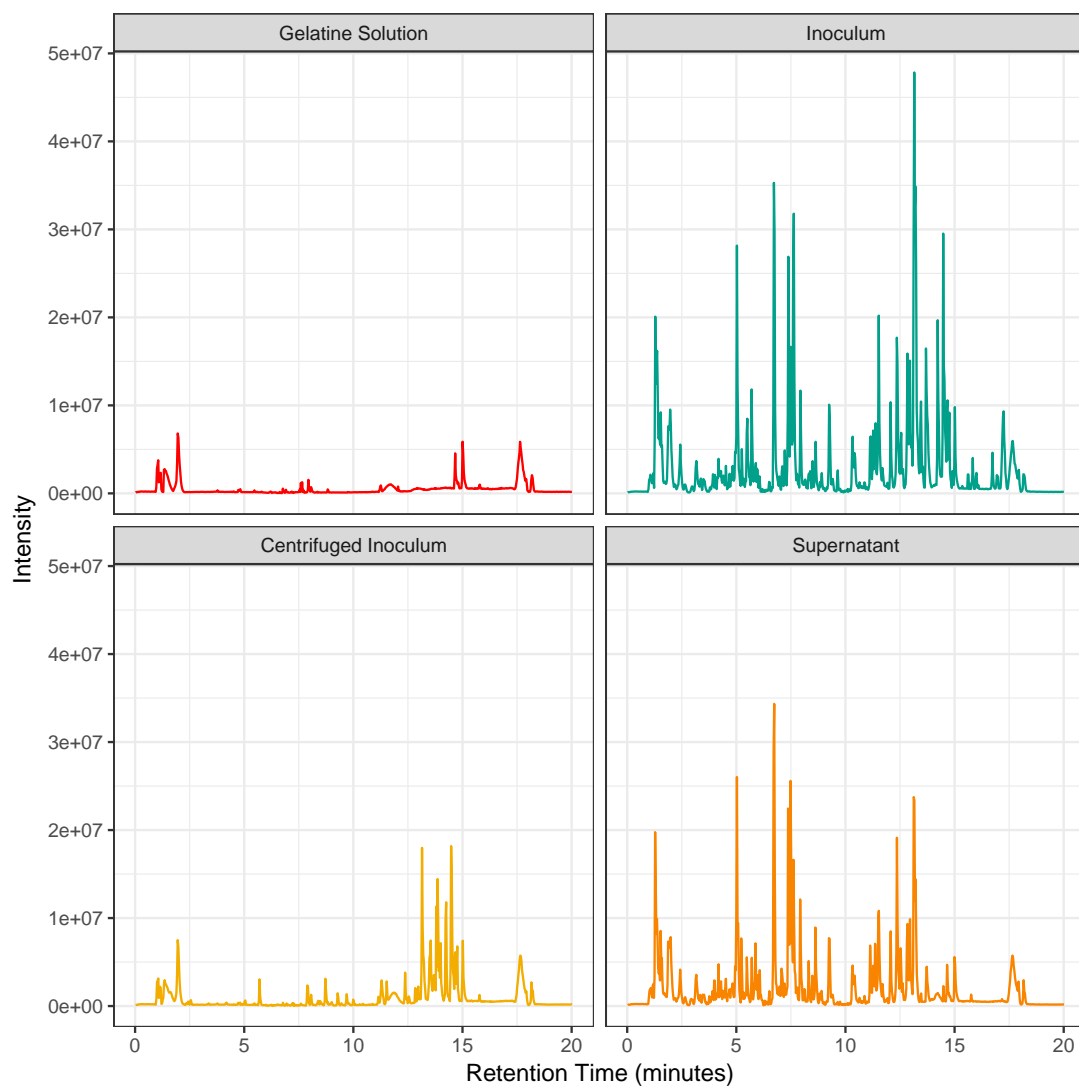


Figure 4.2: **Negative mode base peak ion chromatograms of inoculum components analysed by C₁₈ LC-HRMS.** See Section 4.3.2 for inoculum preparation details.

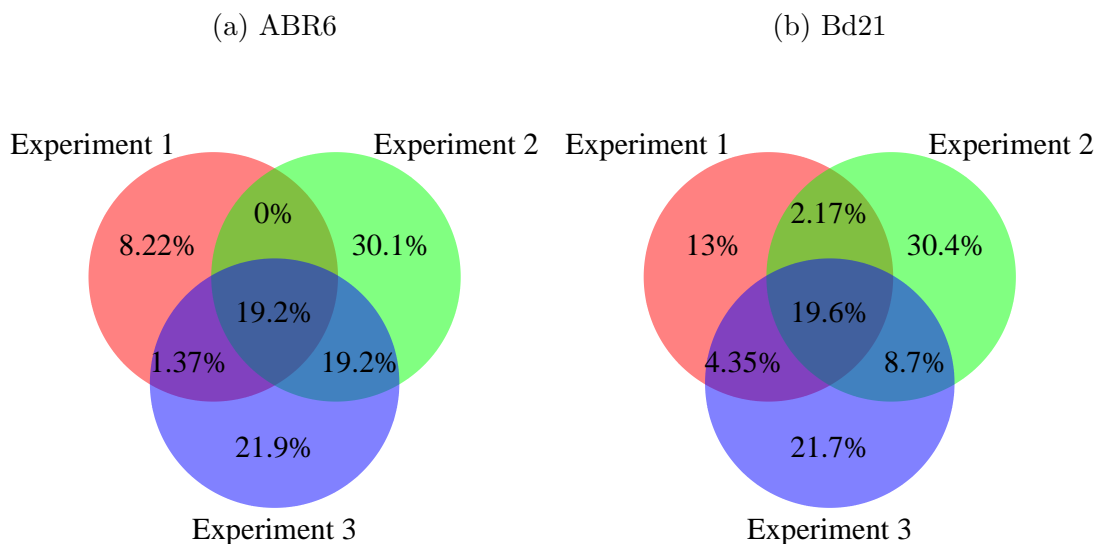


Figure 4.3: **Venn diagrams comparing explanatory features between individual inoculation experiments.** Explanatory features were taken from comparisons between infected and control treatments across time points 12, 24, 36, 48, 60 hpi that had a selection frequency FPR below 1%. Totals of 73 and 46 features were found for ABR6 and Bd21 respectively. Percentages are given to three significant figures.

interpretable. Using a re-sampling strategy that incorporates external validation accounts for this independence and allows more robust identification of consistently reproducible features. The following sections assess the results of applying this strategy to the *B. distachyon* and *M. oryzae* interaction.

4.4.2.1 Classification performance using external validation re-sampling

By comparing the Random Forest model performances for each combination of experiment comparison when using an external validation re-sampling strategy it can provide indications of the robustness of metabolome changes between treatment classes, between independent inoculation experiments. Metabolomic changes at a particular time point could be considered robust if consistent performance measures are obtained between all training and test combinations, irrespective of the extent of discrimination between treatments classes. However, time points with higher discrimination would be more likely to have greater variability in model performance as a greater proportion of robust explanatory features

would be required to maintain high model performance (i.e. poor treatment discrimination is easier to reproduce). Limits of acceptable deviation would depend on the extent of discrimination between the treatment classes at a particular time point and the consistency of performance measures for the other experiment comparisons. Poor robustness would be characterised by either an substantial increase or decrease in performance measures of a particular training and test combination (Sokolova and Lapalme 2009).

Tables 4.3 and 4.4 show the classification performance measures for each re-sampling comparison combination at each time point for ABR6 and Bd21 respectively. Interestingly, model performance of each comparison combination was not consistent between time points in both ecotypes. For instance, in the compatible interaction of Bd21 the comparison of experiments 1 and 2 versus experiment 3 performed worst at 12 hpi whereas at 24 hpi the comparison of experiments 2 and 3 versus experiment 1 performed worst (Table 4.4).

The 0 hpi time point shows consistently poor performance in both ecotypes for all training and test combinations. This is to be expected as no host or pathogen responses will have yet begun. The 60 hpi time point in ABR6 shows consistently high accuracy, Kappa and AUC measures with margins of 0.4 for comparisons 1+2~3 and 1+3~2. However, comparison 2+3~1 shows a substantial drop in performance with a margin of 0.2, half that of the other comparisons. This suggests the presence of a cohort of features that are explanatory between infected and control treatments in experiments 2 and 3 but are not explanatory in experiment 1. Therefore the training models constructed using samples from experiments 2 and 3 have over-fitted and poorly predict the samples from experiment 1. Although, the high discrimination found for the other comparisons at this time point suggest that there are still a proportion of explanatory features present in experiment 1 that are able to predict the features in experiments 2 and 3.

Table 4.3: **ABR6 random forest classification performance using an external validation resampling approach.** Comparisons are given as formulas for which experiments were used for model training (+) and which experiment was used for testing (\sim).

hpi	Comparison	Accuracy	Kappa	AUC	Margin
0	1+2 \sim 3	0.57	0.14	0.57	0.03
	1+3 \sim 2	0.61	0.21	0.62	0.03
	2+3 \sim 1	0.60	0.20	0.60	0.03
12	1+2 \sim 3	0.98	0.96	0.98	0.33
	1+3 \sim 2	1.00	1.00	1.00	0.37
	2+3 \sim 1	0.96	0.91	0.96	0.24
24	1+2 \sim 3	0.96	0.91	0.96	0.32
	1+3 \sim 2	0.99	0.99	0.99	0.36
	2+3 \sim 1	0.93	0.86	0.93	0.27
36	1+2 \sim 3	0.71	0.41	0.71	0.15
	1+3 \sim 2	0.79	0.59	0.79	0.14
	2+3 \sim 1	0.79	0.59	0.79	0.12
48	1+2 \sim 3	0.84	0.68	0.84	0.28
	1+3 \sim 2	1.00	1.00	1.00	0.37
	2+3 \sim 1	0.91	0.81	0.91	0.21
60	1+2 \sim 3	1.00	1.00	1.00	0.40
	1+3 \sim 2	1.00	1.00	1.00	0.42
	2+3 \sim 1	0.59	0.19	0.59	0.21

Table 4.4: **Bd21 random forest classification performance using an external validation resampling approach.** Comparisons are given as formulas for which experiments were used for model training (+) and which experiment was used for testing (\sim).

hpi	Comparison	Accuracy	Kappa	AUC	Margin
0	1+2 \sim 3	0.54	0.09	0.54	0.02
	1+3 \sim 2	0.61	0.21	0.61	0.03
	2+3 \sim 1	0.62	0.24	0.62	0.04
12	1+2 \sim 3	0.73	0.46	0.73	0.14
	1+3 \sim 2	0.85	0.70	0.85	0.23
	2+3 \sim 1	0.89	0.78	0.89	0.18
24	1+2 \sim 3	0.77	0.54	0.77	0.16
	1+3 \sim 2	0.89	0.78	0.89	0.18
	2+3 \sim 1	0.60	0.20	0.60	0.09
36	1+2 \sim 3	0.68	0.36	0.68	0.19
	1+3 \sim 2	0.94	0.89	0.94	0.23
	2+3 \sim 1	0.63	0.26	0.63	0.17
48	1+2 \sim 3	0.86	0.72	0.86	0.21
	1+3 \sim 2	0.96	0.91	0.96	0.23
	2+3 \sim 1	0.88	0.76	0.88	0.21
60	1+2 \sim 3	0.88	0.76	0.88	0.32
	1+3 \sim 2	0.87	0.74	0.87	0.28
	2+3 \sim 1	0.98	0.96	0.98	0.23

4.4.2.2 Feature stability between externally validated re-sample experiment combinations

Although the model performance measures discussed in the previous section give indications of the robustness of the metabolomic changes found between independent inoculations, assessing the stability of the underlying feature lists forms an important part of the assessment. The importance of feature selection stability has been identified in both biomarker discovery and differential gene expression analysis where it forms an essential part of assessing experimental reproducibility (Boulesteix and Slawski 2009; He and Yu 2010).

Feature selection stability in this context of plant/pathogen interactions is likely to reflect consistent changes in metabolites of both the plant and pathogen during pathogenesis. Stability will also be highly linked to the performance measures mentioned previously. Variability in model performance measures between training and test comparisons will lead to poor feature stability between comparisons (Davis et al. 2006).

Figure 4.4 shows the mean Jaccard's indexes and Canberra distances for the comparisons in Tables 4.3 and 4.4. Similar to the classification performance measures, feature stability between the comparisons varied between time points. Also the trends between the similarity measures differed for both ecotypes, with the incompatible interaction of ABR6 differing more than the compatible interaction of Bd21.

At 0 hpi, which had poor treatment discrimination in both ecotypes and therefore yielded few explanatory features, had the lowest Jaccard's index in both ecotypes (Figures 4.4 a & b). However, the relative Canberra distances were found to be highest. This suggests that in terms of the entire feature lists, 0 hpi is showing the greatest feature stability. A lack of underlying system perturbation by pathogenesis is the likely cause of this feature list stability but instability of relevant features (Kalousis, Prados, and Hilario 2007). Contrastingly, where significant class discrimination was obtained at 12 and 60 hpi in ABR6 and Bd21 respectively, a higher mean Jaccard's index and lower mean Canberra distance

was obtained.

Further understanding of the observed feature stability trends requires visualising the trends of the features between the individual experiments. Figure 4.5 shows box plotted examples of feature trends between experiments and their trends when all the data is combined for both ecotypes at 60 hpi.

If a feature was to be considered as robust and reproducible, it would need to be showing consistent intensity trends between treatments between each experiment. This can be seen in ABR6 for feature n132.03 (Figure 4.5b). This consistency can further be seen when all samples are combined where the feature appears to be highly explanatory. A similar trend in this feature is also seen in Bd21, however it is not as consistent as that seen in ABR6. Although experiment 3 is showing the same decrease in infected tissue as the other two experiments, the relative levels are slightly higher. In fact, the levels in the infected tissue overlap with the control levels of the other two experiments, reducing discrimination when the samples are combined.

The feature trend of n191.06 in ABR6 shown in Figure 4.5c shows an increase in infected tissue but only in experiments 2 and 3. Although experiment 1 shows a slight increase, it is not to the extent of experiments 2 and 3. This feature could still be considered robust but is likely to contribute to the drop in Random Forest model performance when experiment 1 is used as the test set (Table 4.3).

An explanatory feature with poor robustness would be expected to be explanatory in one of the experiments but show either no change in other inoculation or even an opposite trend. The feature n131.05 shown in Figure 4.5a shows poor reproducibility with it being explanatory in experiment 2 with an increase in infected tissue compared to control tissue. This trend is not seen in the other 2 experiments and overlapping interquartile ranges are seen when samples from all the experiments are combined.

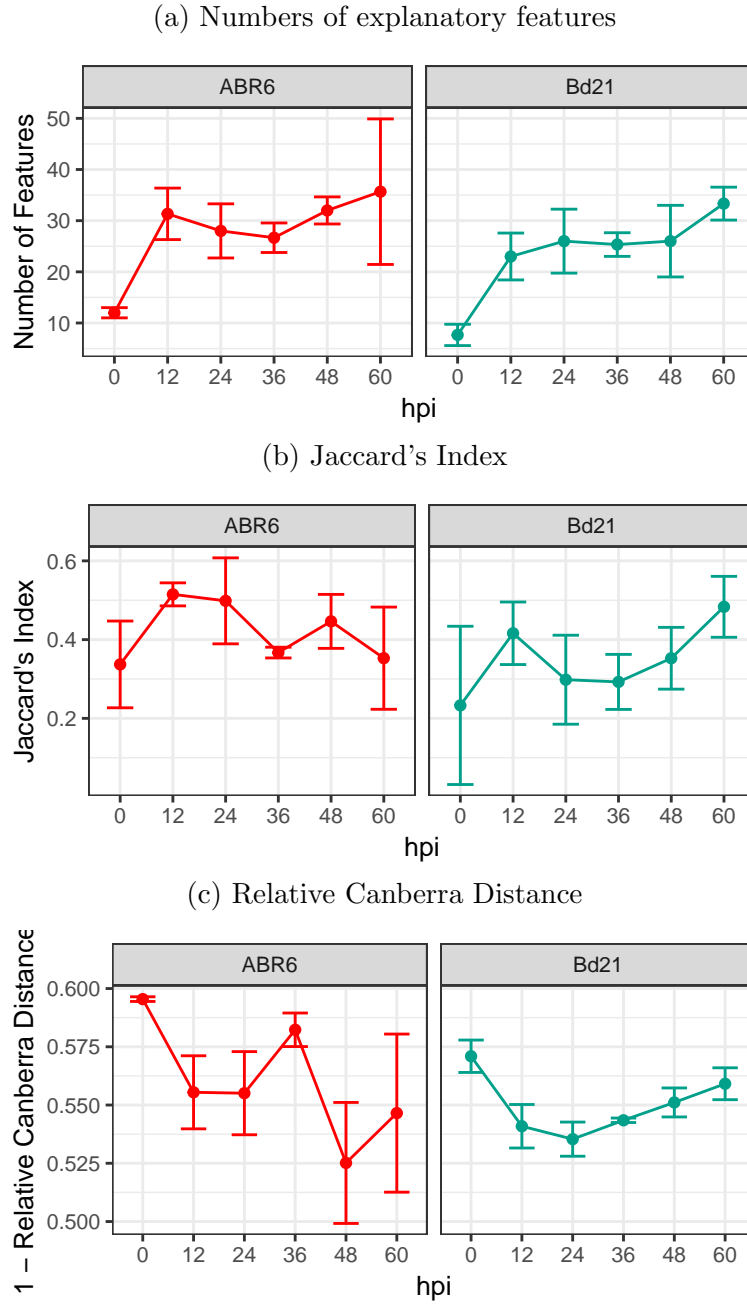


Figure 4.4: **Feature similarity between experiment comparisons.** Mean similarity measures were calculated using feature lists used from the comparisons shown in Tables 4.3 and 4.4. Error bars show ± 1 SD. Jaccard's index was calculated using feature with a selection frequency FPR below 1%. Relative Canberra distance was calculated using complete feature lists of 1560 features.

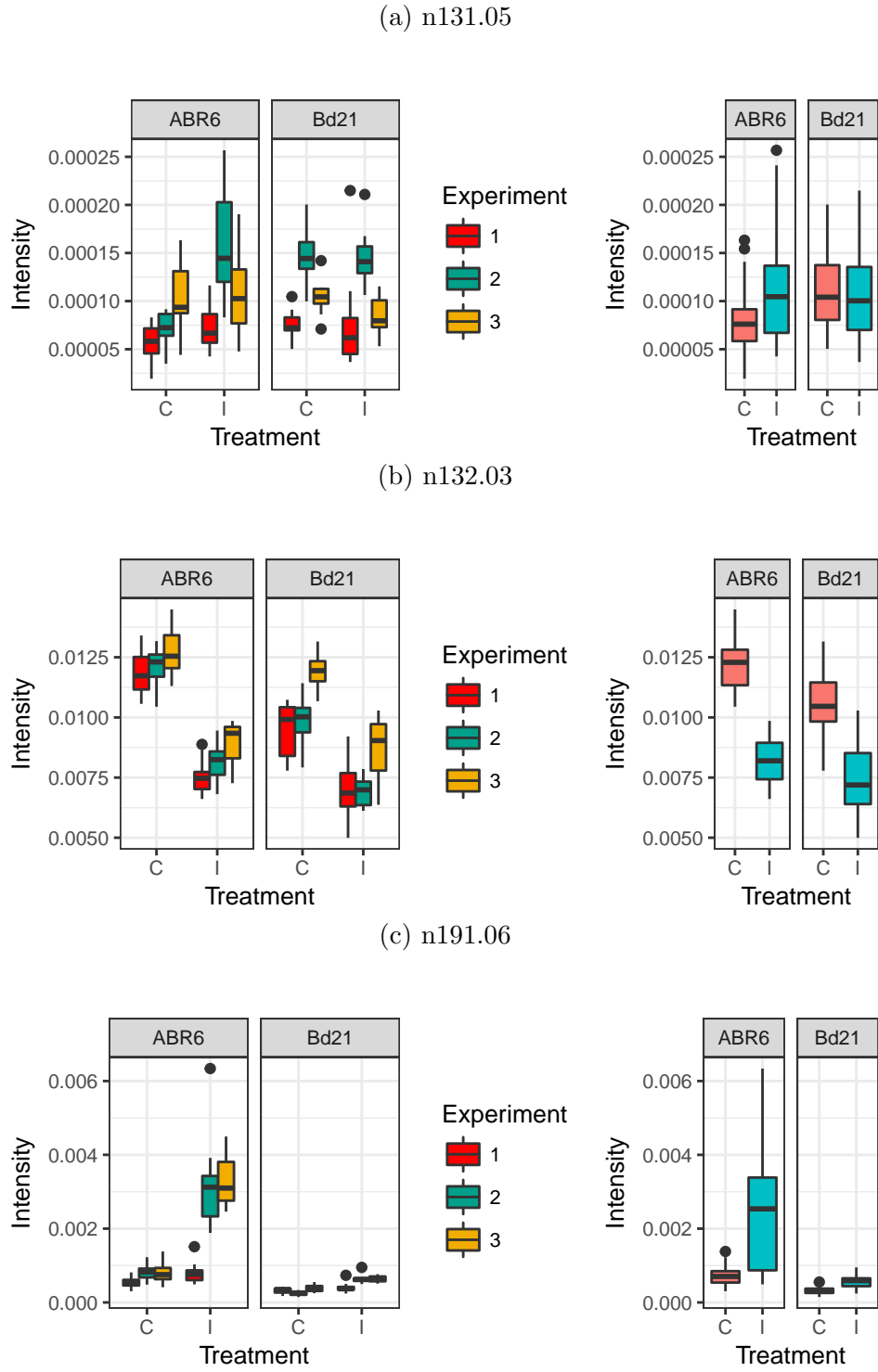


Figure 4.5: **Example box plots of feature experimental variability at 60 hpi.** Left hand plots show individual experiment trends. Right hand plots show trends when experiments are combined.

4.4.3 Potential sources of experimental variability and feature instability

While every effort can be made to limit sample variability through careful experimental design and adherence to protocol between inoculations, as shown in the previous sections, variability and feature instability can still occur. The sources of this variability are likely to have both technical and biological origins.

From a technical perspective the high dimensionality of metabolome fingerprint data is likely to introduce a degree of feature instability through the identification of false positives due to the relatively low feature to sample ratio (Broadhurst and Kell 2006). Also additive would be small yet unavoidable variability that can be introduced at almost all stages of the experiment from initial inoculation, sample collection, sample preparation and instrumental analysis.

With respect to variability of biological origin, this could have a number of sources. The plants used in these inoculations require 21 days of growth prior to inoculation. Although the plants are cultivated under temperature and light regulated conditions, small variations in seed germination time along with temperature and watering regimes over this period can cause variability in rates of both sucrose production and nitrogen uptake which would lead to variability in the extent of development between plants (Weitbrecht, Müller, and Leubner-Metzger 2011; Hikosaka et al. 2006; Gonzalez-dugo et al. 2010). Between inoculations, this would give differences in the cellular system starting points of the host leaves and likely affect the trajectories of host responses to pathogenesis (Pritchard and Birch 2011). This is shown by the high discrimination between 0 hpi control treatments in Table 4.5 and is unlikely due to control inoculum as no explanatory features were found in common when experiment comparisons are compared between ecotypes.

Linked to developmental variability is the spatial diversity of cell types on an individual host leaf. It has been shown that the transcriptome of maize leaves differs along the proximal and distal plane as well as between bundle sheath and mesophyll cells (Li et al. 2010b). The numerous cell types in the leaves of *B. dis-*

tachyon including parenchyma, mesophyll and phloem means that many cellular states are present. Also *M. oryzae* has tissue dependent infection strategies with a fully biotrophic strategy adopted when invading rice root tissue. Although the organ specific adaptation of *M. oryzae* has been shown, tissue specific adaptation of infection strategy within the same organ has not yet been investigated.

With the use of the spray application of the inoculum, there is the potential for heterogeneity in spore density across the leaf surface . This allows the potential for small areas to contain much higher densities of spores than other areas. It is likely that this could cause heterogeneity in the intensity of systemic signalling between cells and therefore the responses of these cells, depending on the sub-cellular origin of these signals, especially in the early stages of host colonisation (Mullineaux 2006).

Further to this is the temporal heterogeneity of pathogen colonisation (Parker et al. 2008). With the colonisation of host tissue by *M. oryzae* occurring not only horizontally, but also vertically through the leaf, there is the potential of different cell types to be coming into contact and responding to the pathogen colonisation at different times.

The sampling of leaf sections and subsequent homogenisation that has been used here will not take into account the variability in these spatial and temporal factors. This variability is most likely to affect the stability of marginally explanatory features rather than those that are highly explanatory.

Spatial and temporal heterogeneity of *B. distachyon* responses to *M. oryzae* has the potential to be responsible for the presence of cohorts of related explanatory features that, although showing the same trends in two of the three experiments, are not found in the other. Over-fitting was found in ABR6 at 60 hpi (Section 4.4.2.1) when experiment 1 was used for testing. Figure 4.6 shows the trends of the explanatory features that are responsible for this. These features form 2 distinct groups, those that are decreased and those that are increased in experiments 2 and 3.

In FIE-HRMS, a metabolite can be represented by more than one m/z due

Table 4.5: **Random Forest classification results of comparisons of 0 hpi control treatments between experiments.**

Ecotype	Comparison	Accuracy	AUC	Margin
ABR6	1 vs 2	0.88	0.98	0.27
	1 vs 3	0.98	1.00	0.42
	2 vs 3	1.00	1.00	0.43
Bd21	1 vs 2	0.94	0.99	0.32
	1 vs 3	0.89	0.98	0.36
	2 vs 3	0.79	0.91	0.19

to the presence of adducts and isotopes. This can be responsible for part of these groups. However, related metabolites could represent co-regulated modules present within these biological networks.

The regulation of biological networks is highly modular with regulatory hubs responsible of the co-ordination of cellular response to stress conditions (Barabási and Oltvai 2004). Hubs in plant defence include the MAPK induction of WRKY transcription factors (Meng and Zhang 2013). Instability in the activation of these key regulatory hubs between inoculations has important implications for the validity of omics results based on single inoculations. If the origin of this potential regulatory instability is biological rather than technical, it could provide evidence of plasticity in network responses during plant defense.

4.5 Concluding remarks

The overall aim of this chapter was to objectively assess three basic but key elements of successful omics investigation of plant pathogen interactions. The application of FIE-HRMS fingerprinting has allowed the identification that inoculum constituents other than that of pathogenic fungal spores can elicit substantial host responses. Not only can these confound host response caused *M. oryzae* pathogenesis, but also mask them. This simple yet profound insight not only has implications for the design of omics experiments involving the interaction between *B. distachyon* and *M. oryzae* but also for investigating the interactions of plant pathogen interactions in general, depending on the inoculum preparation and inoculation techniques used. These responses could be controlled for by applica-

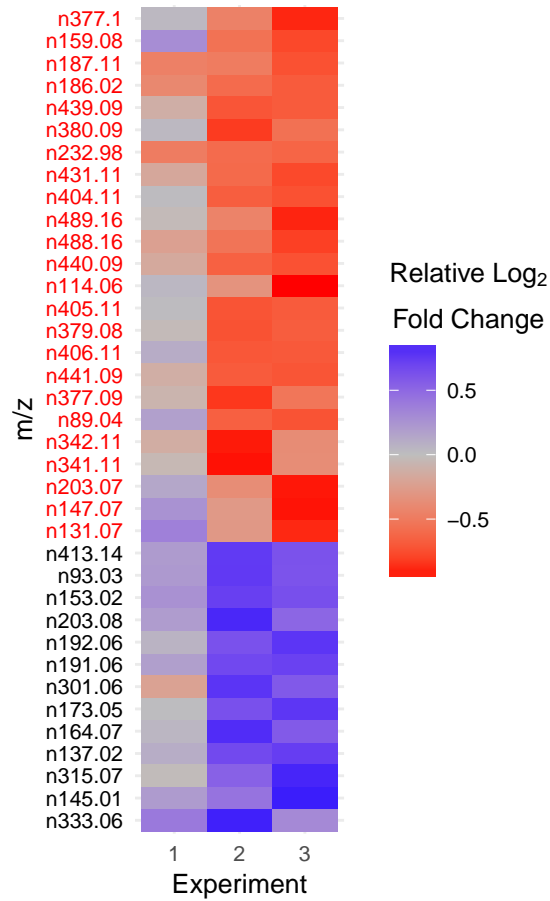


Figure 4.6: **Heat map of negative mode ABR6 m/z trends at 60 hpi.** Features shown are those that are explanatory in experiments 2 and 3 only. Relative log₂ fold change were calculated using the ratio of mean intensities between infected and control treatments. Ratios were then log₂ transformed and sum of squares scaled. m/z are ordered by hierarchical clustering and coloured by K-means cluster occupancy (K=2).

tion of a non-pathogenic inoculum that will elicit these responses but not cause infection of the host tissue. Careful consideration is needed for all aspects of the experimental process and thorough reporting of all aspects experimental set-up when using omics to investigate these interactions. This ensures that results are interpretable in the biological context for which they were intended.

An appropriate method of controlling for these inoculum induced responses was developed by the use of a non-pathogenic inoculum. Freeze treatment and agitation provided the most suitable method for neutralising *M. oryzae* spores, without altering the composition of the inoculum.

FIE-HRMS fingerprinting allowed an assessment of the robustness of metabolome changes observed between independent experiments to investigate the pre-symptomatic phases of the interaction between *B. distachyon* and *M. oryzae*. Random Forest classification and an external validation re-sampling strategy showed that metabolome changes can vary between inoculations. This variability can also be dynamic between time points; with some phases of pathogenesis more reproducible than others. It highlights the need for variability between inoculations to be taken into account when designing large scale omics experiments to investigate plant/pathogen interactions.

There is opportunity for the integration of external validation into omics investigations involving plant-pathogen interactions. The cost of conducting experiment repeats is substantially lower than for fields such as biomedicine where the use of external validation is common. The ability to identify robust and stable explanatory feature trends is key for ensuring the validity of further omics integration or more targeted pathway analyses involving quantitation. In the longer term this will provide more fruitful hypothesis generation and improve the value of the use of omics analyses.

Chapter 5

Metabolomic and transcriptomic analyses of the pre-symptomatic phases of the *B. distachyon* and *M. oryzae* interaction

5.1 Introduction

On the molecular level, the pre-symptomatic phases of biotrophic fungal plant-pathogen interactions represent a key point that can decide the outcome of the interaction. During these phases, pathogens are at their most vulnerable to host defences. Nutritional resources are scarce and their extent of colonisation is limited. Failure to colonise beyond that of the primary host cell will mean almost certain death. Biotrophic fungal pathogens have to successfully subvert the host metabolism and establish a nutritional interface in order for successful colonisation to occur through to sporulation (O’Connell and Panstruga 2006).

Investigating plant-pathogen interactions using omics level analyses provides a system wide overview of changes occurring during pathogen colonisation. As mentioned in Chapter 4, the power of omics analyses is the valuable information they provide for hypothesis generation. This can then allow further, more

targeted experiments to both validate and extend findings. Applying multiple levels of omics investigation further enhances interpretation of these data with respect to the underlying mechanisms involved in the plant-pathogen interaction. This chapter will apply both metabolomic and transcriptomic techniques to the pre-symptomatic phases of the *B. distachyon* and *M. oryzae* interaction. These techniques will include FIE-HRMS and LC-HRMS based metabolomics as well as RNA-Seq based transcriptomics.

5.1.1 Untargeted LC-MS profiling for metabolomic investigations

Unlike the FIE-HRMS fingerprinting techniques previously discussed in Chapter 3, LC-MS profiling allows the chromatographic separation of sample analyte prior to MS analysis. Electrospray ionisation is usually used as it is a soft ionisation technique that limits fragmentation of the parent ion. An advantage of profiling is that it allows the potential separation of compounds based on their chemical properties (depending on the LC column that is used) and therefore isomers or compounds of similar molecular mass are visible (Hagel and Facchini 2008). This greatly improves the potential for metabolite annotation over that of techniques such as FIE-MS. LC can be used in conjunction with a high resolution mass spectrometer such as Orbitrap analysers, that can further improve annotation potential.

Introduction of the time dimension requires prolonged acquisition time resulting in the potential for analytical variation. LC-MS profiling typically takes > 15 minutes per sample leading to a substantially lower throughput than that of fingerprinting techniques (Beckmann et al. 2008). Columns can also suffer from retention time drift over the course of an analytical run so making peak alignment difficult.

LC-MS isn't truly global a metabolomic technique as there will inevitably be some targeting of sample chemistry based on the chromatographic conditions that are used for analysis. These conditions will be dependent on the type of biological

matrix being analysed and the compound types that are of interest. C₁₈ reverse-phase columns, where more polar compounds elute first, are commonly used in the analysis of plant matrices due to their ability to separate phenolics and flavonoids that are highly abundant (De Vos et al. 2007). Also commonly applied is hydrophilic interaction liquid chromatography (HILIC) that better resolves compounds that are poorly retained using reverse phase columns. These are more suitable for separating amino acids and carbohydrate compounds within plant matrices (Tolstikov and Fiehn 2002).

Due to the potential for analytical conditions to change over the course of a sample batch, the use of QC measures are essential. These include the use of internal standards, test mixtures as well as the most commonly used QC samples. A QC sample consists of an aliquot of all biological samples that are to be analysed. It is therefore an aggregation of all the present biological variance. This QC sample is injected multiple times at the start of a run in order to condition the column, then injected at regular intervals throughout the analytical run. Samples are best randomised in blocks based on the class structure, ensuring that each class is equally represented in each block. QC injections can then flank each of these randomised blocks during the run (Dunn et al. 2012).

PCA can be used to validate the stability of the QC injections over the course of the analytical run. Such methods can also be used to ensure the analytical reproducibility of the individual acquired LC-MS features. This is based on the assumption that highly variable LC-MS features in the QC sample across an analytical run, are variable are not analytically reproducible and therefore should be excluded from the analysis. Thresholding of the coefficient of variation of the individual features across the QC injections allows these highly variable features to be removed prior subsequent statistical analyses. An upper limit of the coefficient of variation of 30% can be used as a suitable threshold, however lowering this will improve the reliability of the acquired features (Want et al. 2010).

Signal processing is required to align and extract m/z features from the LC-MS profiles so that statistical analyses can be applied to identify differences be-

tween the biological classes. There are numerous software packages that can be used for this including MZMine, MetAlign and XCMS. Most commonly used is XCMS that has both for which there is an R package and a cloud based interface (Smith et al. 2006; Tautenhahn et al. 2012). It includes numerous algorithms for peak detection, retention time correction, grouping of samples based on biological origin and peak infilling. Extracted feature are a combination of m/z regions of interest (ROI) and unique retention times.

As mentioned in Chapter 3 the data pre-treatment can affect the results of classification and feature selection techniques (Gromski et al. 2015b). Appropriate data pre-treatments will be dependent on both the statistical analyses to be applied and specific requirements of the experiment. Partial least squares discriminant analysis is the most commonly applied multivariate techniques for identifying biomarkers within metabolomics data sets (Gromski et al. 2015a).

Metabolite annotation strategies are similar to those as for FIE-MS and are improved by the use of high resolution mass analysers. The added retention time information also allows inference of the chemical properties of the compound in question, depending on the chromatographic conditions used. Additionally isotopic and adduct relationships will be restricted to the retention time of the parent ion. Metabolite databases such as MZedDB and Metlin can be used for putative ionisation product searches to identify candidate metabolite annotations and LC-MS/MSⁿ can be used to gain additional structural information.

5.1.2 Metabolomic analyses of plant pathogen interactions

There are numerous examples of metabolomic techniques being applied to investigate plant pathogen interactions. These include pathogens from across kingdoms with bacteria, fungi and viruses. These investigations have utilised a range of metabolomic techniques from LC and GC-MS profiling as well as MS and NMR fingerprinting (Allwood, Ellis, and Goodacre 2008).

A number of investigations have applied metabolomic techniques to investigate the interactions of *M. oryzae* using a number of hosts. Parker et al. (2009) found

similar metabolomic responses of rice, barley and *B. distachyon* between 1 and 5 days of infection. Allwood et al. (2006) applied FT-IR and identified phospholipids as key discriminatory non-polar metabolites. Jones et al. (2011) applied NMR fingerprinting, LC and GC-MS profiling to compare compatible and incompatible interactions of rice and *M. oryzae*, including a number of pre-symptomatic time points. No significant differences were found in the responses prior to 24 hpi. A number of amino acids including alanine were found to be diverging as the responses diverged.

5.1.3 Whole transcriptome sequencing analyses using RNAseq transcriptomics

Transcriptome profiling has been revolutionised in recent years by the application of NGS for whole transcriptome sequencing using RNA-Seq. Transcriptomics has previously been reliant on hybridisation and tag-based techniques, although each has its limitations.

Hybridisation based micro-arrays using fluorescently labelled cDNA are relatively inexpensive. However, they are reliant on prior sequence knowledge of the target organism and does not ensure total coverage of the transcriptome. Also this technique suffers from a low dynamic range due to a high background through cross-hybridisation and signal saturation. This can make it difficult to compare profiles between experiments and can require complex normalisation. Tag based techniques were developed to overcome these issues however they are expensive, many short reads cannot be mapped uniquely to the reference genomes and isoforms are indistinguishable (Morozova, Hirst, and Marra 2009).

RNA-Seq can use either total or fractionated populations of RNA that are fragmented and then converted to a library of cDNA. Adaptors are ligated to these fragments and they are then sequenced with or without amplification using NGS. Common sequencing technologies include Illumina HiSeq, Applied Biosystems SOLiD and Roche 454 Life Science. Sequence reads are typically 30-400 bp, depending on the sequencing technology that is used. Reads can be sequenced

from one end or both ends to give single or paired end reads respectively (Wang, Gerstein, and Snyder 2009).

Unlike DNA sequencing, transcriptome coverage from an RNA-Seq experiment is difficult to estimate due to differing levels of transcription within cells, depending on their state. Sample multiplexing, where different adapters are ligated to the cDNA fragments depending on the sample of origin, allows multiple samples to be run simultaneously that reduces cost. This however will also reduce the sequencing depth per sample (Trapnell et al. 2012). Adequate coverage is dependent on the intended use of the data and is a compromise with cost. Experiments intended to identify rare genes or isoforms will need considerably more depth than those aimed at expression profiling of common genes, where the sequencing of more replicates will be of value.

After sequencing, QC routines can be applied to remove poor quality reads or trim reads to remove sequenced adapters. The reads can then either be aligned to a reference genome, if one is available for the target organism, or *de novo* assembly can be used to assemble the transcriptome. Tophat and Cufflinks are commonly used open source software for reference genome alignment and transcriptome assembly (Kim et al. 2013; Trapnell et al. 2013). Trinity is commonly used for *de novo* assembly (Haas et al. 2013). The ability to assemble the transcriptome without a reference genome is one of the major attractions to using this technique. Considerable transcriptomic data can be obtained without first having to invest in producing the necessary genomic resources. With later development of these resources, previously *de novo* assembled RNA-Seq data sets can be revisited to potentially yield more information (Wang, Gerstein, and Snyder 2009).

Not only is RNA-Seq able to generate transcript expression levels but it also reveals complex information about the structure of the transcriptome. Exons can be mapped with single base precision, allowing the connectivity between exons to be investigated with regards to post transcriptional modifications such as splice isoforms. It is also able to reveal the presence of novel genes (Martin et al. 2013).

Due to its lack of background noise and it's sensitivity, RNA-Seq has a large

dynamic range. Transcriptional levels can be measured using the total number of reads that fall within a reading frame. As longer reading frames will produce a greater number fragments and therefore reads per mRNA, the total number of reads of a reading frame need to be normalised by it's length. Finally this is then normalised by the total number of mapped reads across the entire genome to account for variability in library size. The resulting measure of expression is the reads per kilobase per million mapped reads (RPKM) value (Trapnell et al. 2012). These values have been shown to correlate highly with expression values obtained from qPCR (Nagalakshmi et al. 2008).

Other methods have been developed for normalisation that include that of the DESeq Bioconductor package and the trimmed mean of M-values of the edgeR Bioconductor package. The method of the DESeq package assumes that most genes are not differentially expressed between samples. Similar genes will therefore have read ratios close to 1. The median of this ratio for the lane provides a correction factor that can be applied to all the read counts. The trimmed mean of M-values is based on a similar assumption, using a weighted mean of log ratios to estimate the correction factor (Dillies et al. 2013).

Differential expression analyses allow the identification of genes that are showing statistically different expression as a result of an experimental condition. There have been a number of computational methods for differential expression analyses developed for RNA-Seq. The Cuffdiff software is able to test for differential expression at individual transcript-level resolution and can control for variability within replicate libraries (Trapnell et al. 2013). As many reads within an RNA-Seq experiment can map to multiple regions of the genome, Cuffdiff is able to estimate this over-dispersion and account for greater variability across replicates than would be estimated by a simple Poisson distribution. Actual expression levels can then be estimated using a beta negative binomial distribution which can then be used to test for significance between experimental conditions (Rapaport et al. 2013).

5.1.4 Transcriptomic analyses of plant pathogen interactions

Similar to the application of metabolomics, there are numerous instances of the application of both microarray and RNA-Seq based transcriptomics to plant-pathogen interactions. These investigations have targeted transcriptomes of the pathogen or host plant. Soanes et al. (2012) used RNA-Seq to investigate the transcriptional changes occurring in *M. oryzae* during appressorium development. Genes relating to quinate uptake and utilization were found to be up-regulated as well as large scale gene expression changes relating to lipid metabolism, autophagy and melanin biosynthesis.

Bagnaresi et al. (2012) compared the transcriptional responses of resistant and susceptible rice genotypes using RNA-Seq at 24 hpi. Genes for diterpene phytoalexin synthesis, chitinases and flavin monooxygenases were found to be highly up-regulated in the resistant interaction. Similar gene ontology (GO) terms were found in common between the response types, however the gene sets contributing to these were found to be dissimilar.

The sensitivity and high dynamic range of RNA-Seq allows effective simultaneous profiling of both plant and fungal pathogen especially during early phases when host tissue considerably outweighs pathogen tissue. Kawahara et al. (2012) simultaneously analysed transcriptional changes in both rice and *M. oryzae* in both compatible and incompatible interactions at 24 hpi. A conidial suspension was used as a control with which to compare with infected leaf tissue. Large scale up-regulation of genes encoding secreted proteins were observed in *M. oryzae*. Greater transcriptional changes in rice were found in the incompatible interaction, which included phytoalexin biosynthetic genes.

5.2 Aims

The aim of this chapter is to assess the extent of metabolomic and transcriptomic changes that occur during the pre-symptomatic phases of the *B. distachyon* and

M. oryzae interaction; in both the compatible and incompatible situations of the ecotypes Bd21 and ABR6 respectively. FIE-HRMS fingerprinting and LC-HRMS profiling based metabolomic techniques along with RNA-Seq based transcriptomics will be utilised. It will mainly be concerned with the extent of discrimination between treatment and control groups across a range of time points and the relationship of these changes to the main microscopic events that occur during early colonisation. It is meant as a prelude to Chapter 6, which will be concerned with the integration of the multiple levels of omics data analysed here.

This provides us with the following aims with which to focus this chapter:

- Determine the extent of metabolome and transcriptome changes occurring during the pre-symptomatic phases of *M. oryzae* colonisation of *B. distachyon*.
- Identify if metabolomic and transcriptome changes are synchronous with key microscopic events during the *B. distachyon* and *M. oryzae* interaction.

5.3 Materials and Methods

5.3.1 Inoculation and harvesting of plant tissue

The experimental set-up was as detailed in Section 2.5 with inoculation of the compatible and incompatible ecotypes Bd21 and ABR6 with *M. oryzae*. Plants were inoculated as described in Section 2.3 with the control and infected plants for each time point placed within the same plastic propagators. The plants remained under high humidity conditions throughout the experiment and were removed 1 hour prior to sampling. Metabolomic samples were harvested as described in Section 2.4 with two *B. distachyon* leaves placed in each eppendorf tube. For transcriptomic samples, 10 similarly sampled *B. distachyon* leaves were placed in each eppendorf tube.

5.3.2 Metabolite and RNA extraction

A global extraction protocol was used for the extraction of the metabolomic samples as described in Section 2.6. 1.4 ml of extraction solvent (CHCl_3 :MeOH:H₂O; 1:2.5:1; v:v:v) was added per sample. Samples for LC-HRMS analyses were further prepared as described in Section 2.6. RNA was extracted from samples for RNA-Seq analysis as described in Section 2.11

5.3.3 Mass spectral metabolomic analyses

FIE-HRMS analyses were conducted as described in Section 2.7. All 864 samples were distributed across 12 batches, each evenly randomised to ensure equal class distribution across the batches. For LC-HRMS analyses, 10 samples were randomly selected from each class. These samples were then pooled in pairs to give 5 replicates per class for each inoculation, resulting in 15 replicates for each treatment class. LC-HRMS analysis was conducted as described in Section 2.8 with each ecotype run separately across 3 batches each.

5.3.4 Metabolomic data mining

Pre-treatment and initial quality assessment of FIE-HRMS data was applied as described in Section 2.7. Random Forest classification and feature selection with an external validation resampling strategy were used to assess class discrimination as described in Section 4.3.3. To give overall binary comparison classification and feature selection results, the results were averaged across the independent experiment comparisons. A threshold of 1% was applied to selection frequency FPR to identify explanatory variables. Putative annotation of explanatory features was conducted as described in Section 2.10.1.

XCMS was used for signal processing of LC-HRMS data as described in Section 2.8. QC sample based variable filtering ($N = 12$) was applied as described in Section 2.8. Random Forest classification and feature selection resampling were conducted as described in Section 2.9. A threshold of 1% was applied to

selection frequency FPR to identify explanatory variables. Putative annotation of explanatory features was conducted as described in Section 2.10.2.

5.3.5 RNA-Seq and transcriptomic data mining

Library preparation and RNA-Seq was performed on RNA extracts as described in Section 2.11. QC, read alignment, transcriptome assembly, differential expression analysis and functional enrichment analysis of RNA-Seq reads was performed as described in Section 2.12.

5.4 Results and discussion

5.4.1 Omics-level differences between the *B. distachyon* ecotypes ABR6 and Bd21

Considerable geographic differences invariably increase the extent of genetic diversity within a species, especially over the ecological range of *B. distachyon*. The ecotypes used to study this interaction with *M. oryzae* are no exception. ABR6, whose origin is in Spain, shows substantial differences in some very important phenotypic traits such as growth habit and vernalisation requirements. Underlying these phenotypic differences will be considerable genotypic differences that are likely to affect a large proportions of the *B. distachyon* genome (Gordon et al. 2014).

As can be seen in Figure 5.1, substantial compositional differences were found in both the metabolomes and transcriptomes of ABR6 and Bd21 that are unrelated to the colonisation of *M. oryzae*. This observation is of no surprise considering the distances between the geographical origin of these ecotypes, with ABR6 originating in Spain and Bd21 in Iraq (Routledge et al. 2004; Mur et al. 2011).

An important point to make about the presence of these differences in both the metabolomes and transcriptomes of these ecotypes is that this is likely to confound any direct comparisons of their responses to *M. oryzae*. With this in

mind any comparisons that will be made between the compatible and incompatible interactions here will be qualitative and observational in nature to avoid the implications of this confounding factor.

5.4.2 Metabolomic changes during early phases of the *B. distachyon* and *M. oryzae* interaction

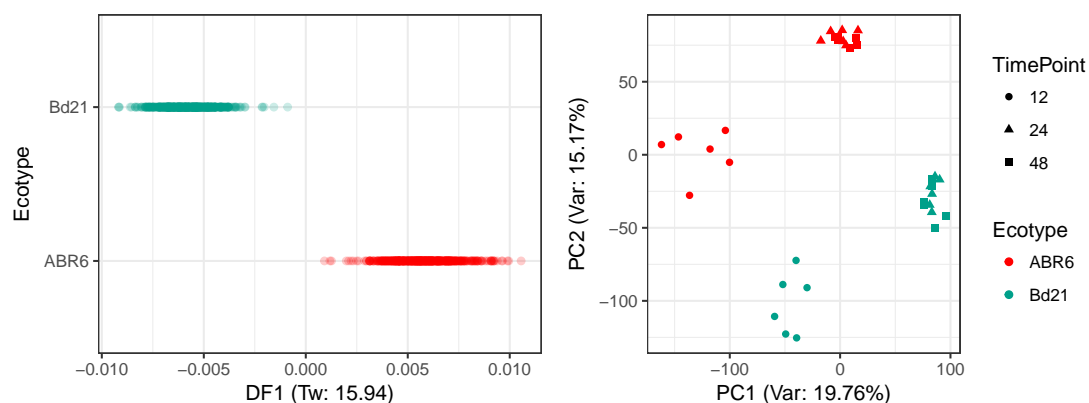
Both FIE-HRMS fingerprinting and LC-HRMS profiling were used to analyse metabolomic changes from 0-60 hpi. The following sections will assess the extent of discrimination between infected and control treatments at each time point as well as discuss the relationships between annotations of identified explanatory m/z features.

5.4.2.1 Discrimination between control and infected tissue during pre-symptomatic phases

Random Forest classification identified explanatory discrimination for *M. oryzae* infection at a number of time points in both the FIE-HRMS and LC-HRMS analyses (Figure 5.2). There was little discrimination found at 0 hpi in all analyses due to the appropriate control of the experiments as discussed in Chapter 4. Similar overall trends in random forest model margins were found in both the FIE-HRMS and LC-HRMS analyses; however, the model margins for the LC-HRMS were substantially more explanatory than the FIE-HRMS margins. There was little explanatory discrimination in positive mode of the FIE-HRMS analyses.

The extent of discrimination was dynamic over the time course of infection in both ecotypes. There was not a gradual linear progression of increasing discrimination as the time course progressed. Instead, there was a drop in discrimination at either 24 or 36 hpi in both ecotypes and in both metabolomic analyses; with discrimination increasing at time points subsequent to this (Figure 5.2). These time points represent the points at which primary and secondary cell invasion are occurring within the infection cycle. This reduction in discrimination could reflect a transitional phase between host recognition and the initiation of defences.

(a) PC-LDA of negative mode FIE-HRMS fingerprints (b) PCA of RNA-Seq gene expression profiles



(c) Base peak ion chromatograms of C_{18} LC-HRMS profiles

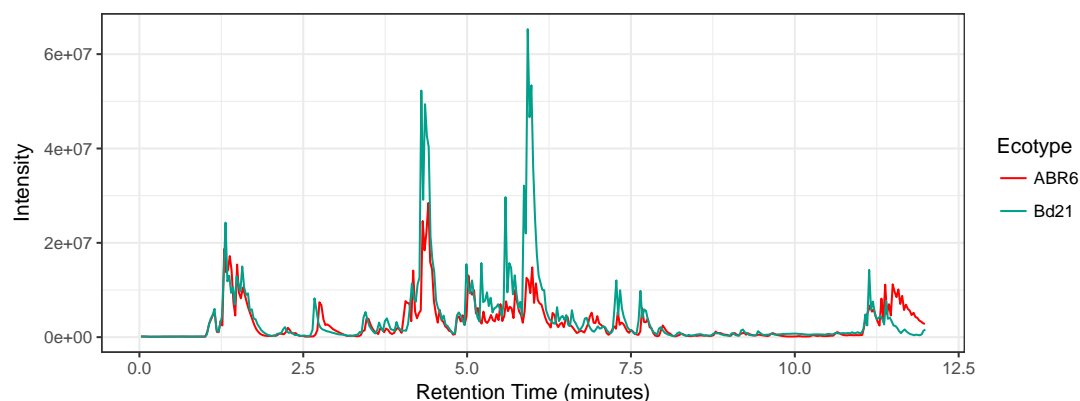


Figure 5.1: Metabolome and transcriptome differences between the *B. distachyon* ecotypes **ABR6** and **Bd21**. a) PC-LDA of FIE-HRMS data for a comparison between *ABR6* and *Bd21* across all infection time points and treatments with 1560 variables. Tw value > 2 is considered explanatory (Enot et al. 2008). b) PCA plot of Cufflinks assembled transcriptome using 28872 variables. c) Comparison of example C_{18} LC-HRMS QC chromatograms between *ABR6* and *Bd21*.

However, these results agree with the relative extent of discrimination found by Parker et al. (2009) between infected and healthy treatments of a compatible interaction in the ecotype ABR1, with little relative discrimination 1 day after inoculation and increasing discrimination up to 5 days after inoculation.

Interestingly, 12 hpi showed high explanatory discrimination in both ecotypes, especially in negative ion mode of the LC-HRMS analyses. This high discrimination is likely to have several sources. Within the host, initial recognition of the presence of the *M. oryzae* spores will be occurring by the detection of PAMPs along with the initiation of innate host defences. The *M. oryzae* spores will also have completed spore germination and would be undergoing appressorium development. This is known to cause major metabolic shifts within the spores. It is unclear as to which process would contribute most to this metabolic discrimination but it is likely that both processes are contributing to some degree.

The controlling of the experiment with the use of an inoculum containing inviable spores means that, if the discrimination identified at 12 hpi is caused by host response to the presence of the *M. oryzae* spores, these responses would have to be elicited by molecules secreted by the *M. oryzae* spores post germination. These responses could not be as a result of just the presence of the spores on the leaf surface as these have been controlled for. They would have to be as a result of secretions from the spores, post germination.

As mentioned in Section 1.3, *M. oryzae* spores secrete an adhesive that initially allows the spores to anchor themselves to the highly hydrophobic leaf surface. There is significant variability in the composition of adhesives used for attachment by fungal pathogens; with no evidence of a common adhesive compound in phytopathogenic fungal species (Tucker and Talbot 2001). *M. oryzae* spores also secrete cuticle and cell wall degrading enzymes during appressorium development; including cutinases cellulases and xylanases. These act to weaken the cell wall of the primary host cell prior to penetration (Howard and Valent 1996). The recognition of secretions such as these, that are likely to be specialized and particular to few fungal phytopathogens, could represent a more specialized

aspect of the PAMP triggered responses beyond that of the recognition of generic fungal constituents.

The most explanatory time point was found to be 60 hpi in both metabolomic analyses and in both ecotypes. This is to be expected as it is the point at which both the compatible and incompatible interactions are most developed in this time course. In the compatible interaction of Bd21, it is the time point at which *M. oryzae* would have invaded the greatest area of cells and therefore have the highest number of plant cells directly responding to it. In the incompatible interaction of ABR6, the host defences such as the hypersensitive response would be well under way which are known to have major metabolic consequences (López-Gresa et al. 2010). Systemic signals would also have had more time than at previous time points to reach the greatest number of cells and therefore have an additive affect on the host response to the *M. oryzae* infection.

5.4.2.2 Explanatory m/z features identified during pre-symptomatic infection phases

A total of 81 and 68 explanatory m/z were identified from both positive and negative mode FIE-HRMS analyses for comparisons between control and infected tissue across all time points in the incompatible interaction of ABR6 and the compatible interaction of Bd21 respectively. In LC-HRMS analyses, 176 and 139 were identified respectively.

Of the 116 unique explanatory features identified across the infection time course of both ecotypes in the FIE-HRMS analyses, 60% of these were assigned putative molecular formulas or metabolite annotations. In the LC-HRMS analyses, of the 275 unique explanatory features identified, 35% were assigned annotations. These are shown in Tables 5.1 and 5.2. Full tables of the explanatory features can be found in Appendix E. FIE-HRMS analyses identified a number of amino acids, dipeptides and fatty acids as well as mono and di-saccharides as explanatory features. The LC-HRMS analyses also identified fatty acids and dipeptides but also a number of purines, pyrimidines and nucleosides.

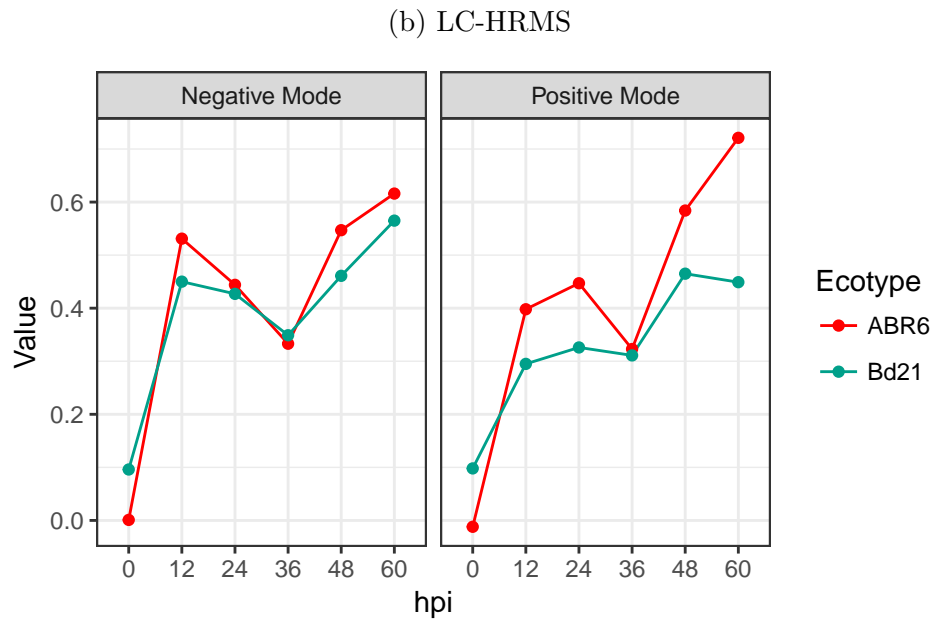
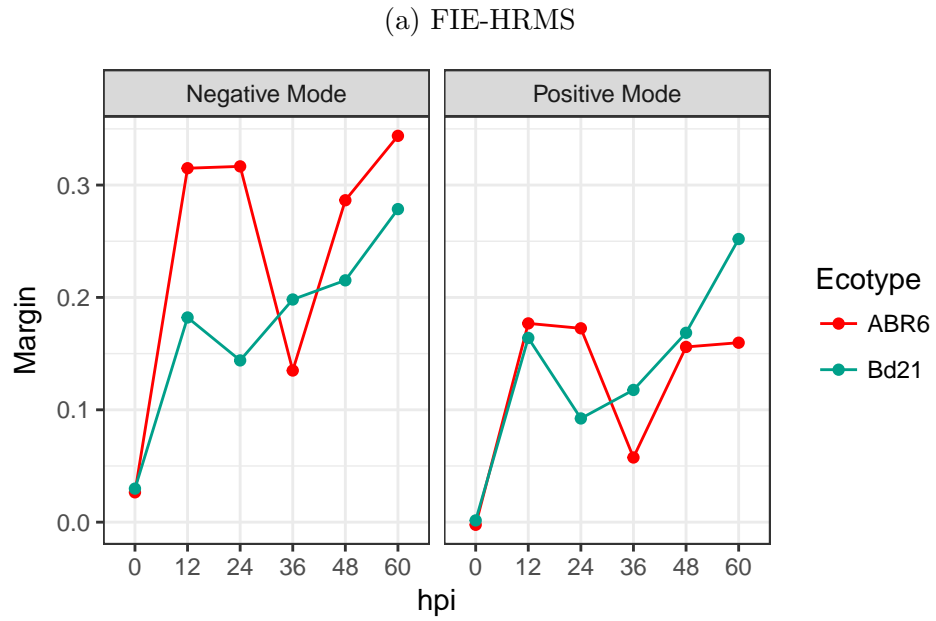


Figure 5.2: **Random Forest margins for control and infected tissue comparisons between 0 and 60 hpi for both ABR6 and Bd21. Margins > 0.2 were considered explanatory.**

In the FIE-HRMS analyses of ABR6, the very early phases (12 and 24 hpi) had the most explanatory features (Figure 5.3a). In the compatible interaction of Bd21, the later phases (48 and 60 hpi) have the most explanatory features in both the metabolomic analyses (Figures 5.3b & 5.4b). Few of the explanatory features were shared between the time points, with 60 hpi usually having the greatest proportion of unique features across all the analyses except for the LC-HRMS analyses in the incompatible interaction of ABR6 where 12 hpi had the greatest proportion of unique features (Figure 5.4b).

K means cluster analyses across both the interactions and metabolomic techniques revealed an even split between up regulated and down regulated clusters (Figures 5.5b, 5.6b & 5.7b). Except for the LC-HRMS analyses of the compatible Bd21 interaction where up-regulated clusters were dominant (Figure 5.8b).

Both metabolomic analyses identified large clusters of m/z that showed a decrease across all the time points from 12 hpi (Clusters 6, 6 and 2 in Figures 5.5a, 5.6a and 5.8a respectively). Other clusters (Clusters 1 and 2 in Figures 5.5a and 5.6a respectively) also showed a down regulation across all time points. Putative annotations within these clusters include mono and di-saccharides, amino acids as well as a number dipeptides, purines, pyrimidines and nucleosides (Tables 5.1 and 5.2). A reduction in these metabolites such as these is suggestive of their catabolism for energy production (further discussion in Section 6.4.1).

A reduction in the levels of sucrose in infected tissue compared to control tissue across all the time points in the compatible interaction of Bd21 does not agree with the trends found by Parker et al. (2009). Increases in sucrose were instead found at 48 and 72 hpi. This could be as a result the different control treatments that were used for the experiments with Parker et al. (2009) using a gelatin control treatment where as here a non-pathogenic inoculum has been used (see Section 4.4.1)

There were clusters identified in both ecotypes and metabolomic analyses that showed a likely influence of circadian rhythms with respect to the the responses to *M. oryzae* infection (Clusters 1 & 2 in Figure 5.5a; 1,2 & 3 in Figure 5.6a; 1

& 2 in 5.7a; 4 in 5.8a). The 12, 36 and 60 hpi time points were sampled 2 hours before the end of the light cycle whereas the 24 and 48 hpi time points were sampled 2 hours after the beginning of the light cycle. The trends explanatory m/z within these clusters all share a common ‘zig-zag’ feature, where the time points at each end of the light cycle share similar trends while those at opposite ends differ.

Primary metabolism in plants is heavily governed by the daily oscillation of light and temperature. Circadian regulation allow plants to pre-emptively regulate responses to these daily cycles. For instance, carbon assimilation will peak at the end of the light cycle as C_3 photosynthesis is unable to occur under darkness. This has important implications for energy availability at each end of the light cycle with carbon starvation and starch degradation occurring during and shortly after the dark period. Amino acids have also been found to peak at the end of the light cycle (Farré and Weise 2012).

The sensitivity of explanatory features in their response to not only *M. oryzae* but also to light cycle is of no surprise however it raises important points about the underlying biological system. Plant defence responses are highly dependent on the availability cellular energy resources. Cells with insufficient energy resources will not be able to mount defence responses such as oxidative bursts and the hypersensitive response. The coincidence of important events within the *M. oryzae* infection process with differing points within the light cycle will likely cause differences in the metabolic response that is achieved. Not only would this have implications for host defence responses, but also for the nutrient acquisition by *M. oryzae*. If metabolites such as sucrose or amino acids that are important carbon and nitrogen sources for the fungus are only in abundance at the end of the light cycle and if primary host cell colonisation (a point of nutrient starvation) occurs at the start of the light cycle, the effector response of *M. oryzae* could be affected.

In the LC-HRMS analyses of the incompatible ABR6 interaction, the largest cluster (Cluster 3 in Figure 5.7a) shows up-regulation at 12 hpi. With the high

discrimination identified in this time point from the classification results discussed in Section 5.4.2.1, many explanatory features would be expected. For many m/z to be clustering coherently suggests the presence of a highly co-ordinated response within either the plant or pathogen. This cluster contains many of the putatively annotated unsaturated fatty acids (Table 5.2). As discussed previously in Section 4.4.1.3, unsaturated fatty acids are known to be important signalling molecules in plant defence responses. The presence of a cluster in Bd21 (Cluster 5 in Figure 5.8a) containing similar fatty acids (Table 5.2) suggests that these metabolic changes could be as a result of defence signalling due to PAMP triggered host responses. Similar clusters are also present in the FIE-HRMS analyses (Clusters 4 and 5 in Figures 5.5a and 5.6a).

Alternatively, these fatty acid changes could be as a result of lipid catabolism for energy production within the pathogen spores during appressorium development. Fatty acid β -oxidation to produce acetyl-CoA from triglyceride degradation is essential for appressorium development in *M. oryzae*. This is used by the spores to fuel melanin and fungal cell wall biosynthesis synthesis as part of appressorium strengthening prior to turgor pressure generation (Wang et al. 2007).

The LC-HRMS analyses in the incompatible interaction of ABR6 identified a small cluster of m/z that show a sharp sudden increase in the infected tissue compared to the control tissue at 36 hpi. Also in the LC-HRMS analyses of the compatible interaction of Bd21 there is a similar cluster with a sharp increase at 24 hpi. The function of these clusters is unclear as the only putative annotations available is Hydroperoxy-octadecandienoate in Bd21.

Table 5.1: Putative annotations of FIE-HRMS explanatory m/z . Explanatory m/z were explanatory in infected and control comparisons of independent time points between 12 and 60 hpi.

Bin	Name	MF	Adduct	Theoretical m/z	ABR6 m/z Δ PPM	Cluster	Bd21 m/z Δ PPM	Cluster
p260.05	Acetyl-glucosamine	C8H15NO6	[M+K] ⁺	260.05310	260.05298	5	260.05304	2
p136.06	Adenine	C5H5N5	[M+H] ⁺	136.06177	136.06189	1	136.06178	4
n88.04	Alanine	C3H7NO2	[M-H] ⁻	88.04040	88.04002	6	88.04002	2
p246.16	Alanyl-arginine	C9H19N5O3	[M+H] ⁺	246.15607	246.15636	1	246.15605	6
p203.14	Alanyl-leucine	C9H18N2O3	[M+H] ⁺	203.13902	203.13916	1		
p288.2	Arginyl-leucine	C12H25N5O3	[M+H] ⁺	288.20302	288.20334	1	288.20334	2
p274.19	Arginyl-valine	C11H23N5O3	[M+H] ⁺	274.18737	274.18723	5		
n131.05	Asparagine	C4H8N2O3	[M-H] ⁻	131.04622	131.04587	3		
n132.03	Aspartate	C4H7NO4	[M-H] ⁻	132.03023	132.03001	6		
n133.03	Aspartate ¹³ C	C3iCH7NO4	[M-H] ⁻	133.03359	133.03334	6	132.03000	2
n134.03	Aspartate ¹⁸ O	C4H7NO3iO	[M-H] ⁻	134.03447	134.03426	6	133.03331	2
n277.03	C13H9O7	C13H9O7	[M+Cl] ⁻	277.03319	277.03467	1	134.03426	2
n318.06	C15H13NO7	C15H13NO7	[M-H] ⁻	318.06138	318.05927	5	277.03467	4
p193.98	C2H6NO6P	C2H6NO6P	[M+Na] ⁺	193.98250	193.98262	6	193.98277	2
p209.96	C2H6NO6P	C2H6NO6P	[M+K] ⁺	209.95644	209.95667	6	209.95667	2
p211.95	C2H6NO6P	C2H6NO6P	[M+K] ⁺	211.95455	211.95473	6		
n243.04	C6H12N2O6	C6H12N2O6	[M+Cl] ⁻	243.03894	243.03979	3		
p116.11	C6H13N1O1	C6H13N1O1	[M+H] ⁺	116.10699	116.10696	4		
n283.08	C8H16N2O9	C8H16N2O9	[M-H] ⁻	283.07831	283.08008	5	283.07977	6
n256.06	C9H11N5O2	C9H11N5O2	[M+Cl] ⁻	256.06068	256.06067	5	256.06085	6
n153.02	Dihydroxy-benzoate	C7H6O4	[M-H] ⁻	153.01933	153.01906	3		
p333.2	Dihydroxy-octadecatrienoate	C18H30O4	[M+Na] ⁺	333.20363	333.20364	4		
p349.18	Dihydroxy-octadecatrienoate	C18H30O4	[M+K] ⁺	349.17757	349.17770	4		
p350.18	Dihydroxy-octadecatrienoate ¹³ C	C17iCH30O4	[M+K] ⁺	350.18092	350.18097	4		
p86.06	GABA	C4H9NO2	[M+H-H2O] ⁺	86.06004			86.06013	1
p180.09	Glucosamine	C6H13NO5	[M+H] ⁺	180.08665	180.08688	5		
n215.03	Glucose	C6H12O6	[M+Cl] ⁻	215.03279	215.03343	1		
p219.03	Glucose	C6H12O6	[M+K] ⁺	219.02655	219.02641	1	219.02655	4
n217.03	Glucose ³⁷ Cl	C6H12O6	[M+Cl] ⁻	217.02983	217.03081	1		
n259.02	Glucose-phosphate	C6H13O9P	[M+Cl] ⁻	259.02245			259.02322	3
n145.06	Glutamine	C5H10N2O3	[M-H] ⁻	145.06187	145.06169	3		
n168.99	Glyceraldehyde phosphate	C3H7O6P	[M-H] ⁻	168.99075	168.99068	2		
n105.02	Glycerate	C3H6O4	[M-H] ⁻	105.01933	105.01900	3	168.99071	1
n236.08	Glycyl-glucosamine	C8H15NO7	[M-H] ⁻	236.07760	236.07877	5		
p232.14	Glycyl-arginine	C8H17N5O3	[M+H] ⁺	232.14042	232.14050	1		
n187.11	Glycyl-leucine	C8H16N2O3	[M-H] ⁻	187.10882	187.10883	5		
p143.12	Glycyl-leucine	C8H16N2O3	[M+H-FA] ⁺	143.11789	143.11806	5		
p189.12	Glycyl-leucine	C8H16N2O3	[M+H] ⁺	189.12337	189.12344	1		
p365.17	Hydroperoxy-octadecatrienoate	C18H30O5	[M+K] ⁺	365.17248	365.17267	4	365.17264	5
n131.07	Leucate	C6H12O3	[M-H] ⁻	131.07137	131.07115	4	131.07117	5
n279.23	Linoleate	C18H32O2	[M-H] ⁻	279.23295	279.23444	5		

Continued on next page...

Table 5.1 – Continued from previous page

Bin	Name	MF	Adduct	Theoretical m/z	ABR6			Bd21		
					m/z	Δ PPM	Cluster	m/z	Δ PPM	Cluster
n133.01	Malate	C4H6O5	[M-H] ⁻	133.01425				133.01398	2.02	3
n114.02	Maleamate	C4H5NO3	[M-H] ⁻	114.01967				114.01939	2.43	2
n147.07	Mevalonate	C6H12O4	[M-H] ⁻	147.06628		2.43	6	147.06619	0.64	5
n193.04	Pectate	C6H10O7	[M-H] ⁻	193.03538		-0.06	1	193.03539	-0.06	4
n184.99	Phosphoglycerate	C3H7O7P	[M-H] ⁻	184.98567		-0.61	1			
n222.94	Phosphoglycerate ¹³ C	C3H7O7P	[M+K-2H] ⁻	222.94155		-2.43	1			
n185.99	Phosphoglycerate	C2iCH7O7P	[M-H] ⁻	185.98902		1.83	1			
n191.06	Quinate	C7H12O6	[M-H] ⁻	191.05611		-0.04	3			
n229.01	Ribose-phosphate	C5H11O8P	[M-H] ⁻	229.01188				229.01292	-4.54	3
n289.03	Sedoheptulose-phosphate	C7H15O10P	[M-H] ⁻	289.03301				289.03476	6.05	1
n404.11	Sucrose	C12H22O11	[M+NO3] ⁻	404.10458		1.9	6			
p381.08	Sucrose	C12H22O11	[M+K] ⁺	381.07937		-0.25	6	381.07959	-0.57	6
n341.11	Sucrose	C12H22O11	[M-H] ⁻	341.10894				341.10977	-2.44	2
p723.2	Sucrose	C12H22O11	[2M+K] ⁺	723.19559		-0.33	6	723.19861	-4.18	6
p382.08	Sucrose ¹³ C	C11iCH22O11	[M+K] ⁺	382.08273				382.08289	-0.41	6
n342.11	Sucrose ¹³ C	C11iCH22O11	[M-H] ⁻	342.11174				342.11150	-0.7	6
n378.09	Sucrose ¹³ C	C11iCH22O11	[M+iCl] ⁻	378.08842				378.08750	-2.43	6
n405.11	Sucrose ¹³ C	C11iCH22O11	[M+NO3] ⁻	405.10794				405.11075	6.94	6
p384.08	Sucrose ¹³ C ⁴¹ K	C11iCH22O11	[M+iK] ⁺	384.08139				384.08121	-0.47	6
p383.08	Sucrose ⁴¹ K	C12H22O11	[M+iK] ⁺	383.07748		-0.41	6	383.07764	-0.41	6
n406.11	Sucrose ¹⁸ O	C11i2H22O10iO	[M+NO3] ⁻	406.10883		0.25	6	406.11121	5.86	6
n171.01	Threonate	C4H8O5	[M+Cl] ⁻	171.00658				171.00641	0.97	1
n327.22	Trihydroxy-octadecadienoate	C18H32O5	[M-H] ⁻	327.21770		0.06	4	327.21768		
p367.19	Trihydroxy-octadecadienoate	C18H32O5	[M+K] ⁺	367.18813		-0.18	4	367.18820	-0.26	5
p368.19	Trihydroxy-octadecadienoate ¹³ C	C17iCH32O5	[M+K] ⁺	368.19149		0.52	4	368.19168		
n203.08	Tryptophan	C11H12N2O2	[M-H] ⁻	203.08260		-1.86	2	203.08298		
n180.07	Tyrosine	C9H11NO3	[M-H] ⁻	180.06662		-0.02	2	180.06662		
n132.04	Ureidoglycine	C3H7N3O3	[M-H] ⁻	132.04147		2.01	2	132.04120		
p242.97	Xylionate	C5H10O6	[M+2K-H] ⁺	242.96678				242.96666	0.5	3

Table 5.2: Putative annotations LC-HRMS explanatory m/z . Explanatory m/z were explanatory in infected and control comparisons of independent time points between 12 and 60 hpi.

Name	MF	Adduct	Theoretical m/z	ABR6			Bd21			
				m/z	rt	Δ PPM	Cluster	m/z	rt	Δ PPM
Adenine	C5H5N5	[M+H] ⁺	136.06177	136.06170	1.42	-0.51	1			
Adenosine	C10H13N5O4	[M-H] ⁻	266.08948	266.08970	2.27	0.83	1			
Alanyl-arginine	C9H19N5O3	[M+H] ⁺	246.15607	246.15600	1.25	-0.28	2			
Alanyl-Leucine	C9H18N2O3	[M+H] ⁺	203.13900	203.13900	3.99	-0.49	2	203.13900	3.98	0
Alanyl-Leucine	C9H18N2O3	[M-H] ⁻	201.12450					201.12410	4.01	-1.99
Alanyl-Leucine	C9H18N2O3	[M+H] ⁺	203.13900					203.13900	2.69	0
Aspartate	C4H7NO4	[M-H] ⁻	132.03023	132.03010	1.23	-0.98	1			
Aspartate	C4H7NO4	[M+H] ⁺	134.04479	134.04480	1.37	0.07	1			
C10H17N2O2	C10H17N2O2	[M] ⁻	197.12955	197.12970	6.04	0.78	5			
C11H17N2O4	C11H17N2O4	[M] ⁻	241.11938	241.11970	6.03	1.31	2			
C15H23O5	C15H23O5	[M] ⁻	283.15509	283.15550	8.84	1.45	2			
C16H22O7	C16H22O7	[M+Na] ⁺	349.12578	349.12540	6.24	-1.09	6			
C16H26O3	C16H26O3	[M+Cl] ⁻	301.15760	301.15790	10.97	1	4	301.15700	11.17	-1.99
C17H24N3O16P2	C17H24N3O16P2	[M+H] ⁺	589.07046					589.06990	1.61	-0.95
C18H24O2	C18H24O2	[M+H] ⁺	273.18491	273.18460	8.71	-1.13	3			
C18H26O2	C18H26O2	[M+H-H2O] ⁺	257.18999					257.19000	8.86	0.04
C18H26O2	C18H26O2	[M+H] ⁺	275.20056					275.20050	8.86	-0.22
C18H28O9	C18H28O9	[M-H] ⁻	387.16606	387.16660	6.67	1.39	6			
C18CH30O9	C18CH30O9	[M-H] ⁻	402.18506					402.18430	8.07	-1.89
C19H16N3O2	C19H16N3O2	[M] ⁺	318.12372	318.12400	3.62	0.89	5			
C19H30O9	C19H30O9	[M-H] ⁻	401.18170					401.18100	8.07	-1.74
C19CH32O9	C19CH32O9	[M-H] ⁻	416.20071					416.20000	8.61	-1.72
C21H30O10	C21H30O10	[M-H] ⁻	441.17662					441.17590	8.13	-1.63
C21H32O5	C21H32O5	[M-H] ⁻	363.21770	363.21820	9.04	1.38	4			
C21H34O5	C21H34O5	[M+Cl] ⁻	401.21003					401.20920	10.87	-2.07
C21H38O7	C21H38O7	[M-H] ⁻	401.25448					401.25370	8.14	-1.94
C22H26O6	C22H26O6	[M+Na] ⁺	409.16216	409.16170	9.41	-1.12	1			
C22H38O10	C22H38O10	[M+Na] ⁺	485.23570	485.23500	3.91	-1.44	5			
C22H41O13	C22H41O13	[M] ⁻	513.25525	513.25600	6.29	1.46	3			
C23H36O7	C23H36O7	[M-H] ⁻	423.23883					423.23830	10.16	-1.25
C28H44O6	C28H44O6	[M-H] ⁻	475.30650	475.30715	10.21	1.37	4			
C39H50O7	C39H50O7	[M+Cl] ⁻	665.32506	665.32770	4.78	3.97	5			
C5CH11O3	C5CH11O3	[M] ⁻	132.07472	132.07460	6.28	-0.91	3			
C6H11O3	C6H11O3	[M] ⁻	131.07137	131.07120	6.28	-1.3	3	131.07080	6.27	-4.35
C9H18N2O4	C9H18N2O4	[M+H] ⁺	219.13393	219.13380	2.2	-0.59	2			
C9H18N2O4	C9H18N2O4	[M+H] ⁺	219.13393	219.13370	3.78	-1.05	2			
C9H9O3	C9H9O3	[M] ⁻	165.05572	165.05570	6.35	-0.12	3	165.05520	6.33	-3.15
Deoxyuridine	C9H12N2O5	[M-H] ⁻	227.06735	227.06740	2.01	0.22	1	227.06700	2.09	-1.54
Dihydroperoxy-eicosatetraenoate	C20H32O6	[M-H] ⁻	367.21261	367.21310	9.35	1.33	3	367.21200	9.49	-1.66
Dihydroxy-hexadecanoate	C16H32O4	[M-H] ⁻	287.22278	287.22320	9.57	1.46	3			
Diterpene Glycoside	C26H40O9	[M-H] ⁻	495.25995	495.25790	4.67	-4.14	1			

Continued on next page...

Continued on next page...

Table 5.2 – Continued from previous page

Name	MF	Adduct	Theoretical m/z	ABR6			Bd21				
				m/z	rt	Δ PPM	Cluster	m/z	rt	Δ PPM	Cluster
Diterpene Glycoside	C26H42O11	[M-H] ⁻	529.26544	529.26350	4.67	-3.67	1	511.21770	8.76	-1.56	6
Gibberellin A2 O-glucoside	C25H36O11	[M-H] ⁻	511.21850					512.22100	8.77	-1.68	6
Gibberellin A2 O-glucoside ¹³ C	C24iCH36O11	[M-H] ⁻	512.22186								
Glycyl-proline	C7H12N2O3	[M+H] ⁺	173.09207	173.09200	5.99	-0.4	5				
Guanine	C5H5N5O	[M+H] ⁺	152.05669					152.05670	1.44	0.07	1
Hexadecanedioate	C16H30O4	[M-H] ⁻	285.20713	285.20760	9.36	1.65	3	285.20670	9.35	-1.51	5
Hexadecanedioate	C16H30O4	[M-H] ⁻	285.20713					285.20680	10.16	-1.16	5
Homovanillate	C9H10O4	[M-H] ⁻	181.05063	181.05060	4.22	-0.17	3				
Hydroperoxy-octadecadienoate	C18H32O4	[M+Na-2H] ⁻	333.20473					333.20420	10.28	-1.59	3
Hydroperoxy-octadecatrienoate	C18H30O5	[M-H] ⁻	325.20205	325.20250	8.84	1.38	3	325.20150	8.81	-1.69	5
Hydroperoxy-octadecatrienoate	C18H30O5	[M+Na] ⁺	349.19855	349.19810	8.83	-1.29	3	349.19850	8.81	-0.14	5
Hydroperoxy-octadecatrienoate	C18H30O4	[M-H] ⁻	309.20713					309.20670	10.3	-1.39	4
Hydroperoxy-octadecatrienoate ¹³ C	C17iCH30O5	[M+Na] ⁺	350.20190	350.20150	8.83	-1.14	3				
Hydroperoxy-octadecatrienoate ¹³ C	C17iCH30O4	[M-H] ⁻	310.21048					310.21010	10.3	-1.22	4
Hydroperoxy-octadecatrienoate ¹³ C	C17iCH30O5	[M-H] ⁻	326.20540	326.20590	8.84	1.53	3	326.20490	8.81	-1.53	5
Hydroxy-decanoate	C10H18O3	[M-H] ⁻	185.11832	185.11830	8.14	-0.11	3				
Hydroxy-dodecanoate	C12H22O3	[M-H] ⁻	213.14962					213.14910	9.27	-2.44	4
Hydroxy-hexadecandioate	C16H30O5	[M-H] ⁻	301.20205					301.20150	8.2	-1.83	4
Hydroxy-octadecadienoate ¹³ C	C17iCH28O3	[M+H] ⁺	294.21448	294.21410	8.88	-1.29	3				
Hydroxy-tetradecanedioate	C14H26O5	[M-H] ⁻	273.17075	273.17110	7.99	1.28	3				
Inosine	C10H12N4O5	[M-H] ⁻	267.07349	267.07370	2.02	0.79	4				
Isopropylmalate	C7H12O5	[M-H] ⁻	175.06120	175.06120	5.21	0	5				
isopropylmalate ¹³ C	C6iC1H12O5	[M-H] ⁻	176.06455	176.06460	5.23	0.28	5				
Leucyl-Arginine	C12H25N5O3	[M+H] ⁺	288.20300	288.20260	1.48	-1.39	2	260.16040	2.09	-0.38	2
Leucyl-Glutamine	C11H21N3O4	[M+H] ⁺	260.16050	260.16020	2.1	-1.15	2	261.16380	2.09	-0.15	2
Leucyl-glutamine ¹³ C	C10iCH221N3O4	[M+H] ⁺	261.16384	261.16350	2.1	-1.3	2	357.10320	1.33	-1.82	4
Melibionate	C12H22O12	[M-H] ⁻	357.10385					347.22220	9.9	-1.67	4
Methyl-hydroperoxy-eicosapentaenoate	C21H32O4	[M-H] ⁻	347.22278								
Mevalonate	C6H12O4	[M-H] ⁻	147.06628	147.06610	3.44	-1.22	3				
Octadecene-diyenoate ¹³ C	C17iCH26O2	[M+H] ⁺	276.20391	276.20360	8.87	-1.12	3	291.19540	8.8	-0.24	5
Oxo-octadecatetraenoate	C18H26O3	[M+H] ⁺	291.19547					293.21100	8.85	-0.41	5
Oxo-octadecatetraenoate	C18H28O3	[M+H] ⁺	293.21112								
Phenylalanyl-Alanine	C12H16N2O3	[M-H] ⁻	235.10880	235.10900	3.75	0.85	2				
Phospho-pantothenate	C9H18NO8P	[M-H] ⁻	298.06973	298.06980	2.44	0.23	1	298.06930	2.52	-1.44	2
SerinyL-Tyrosine	C12H16N2O5	[M+H] ⁺	269.11320	269.11290	2.37	-1.11	1				
Subaphyllin	C14H20N2O3	[M-H] ⁻	263.14012	263.14060	5.33	1.82	2				
Sucrose ¹³ C	C11iCH22O11	[M+K] ⁺	382.08273					382.08260	1.36	-0.35	2
Sucrose ¹³ C	C11iCH22O11	[M-H] ⁻	342.11230					342.11170	1.41	-1.77	2
Sucrose ¹⁸ O	C12H22O10iO	[M-H] ⁻	343.11319					343.11250	1.41	-2.01	2
Thymidine	C10H14N2O5	[M-H] ⁻	241.08300	241.08310	3.3	0.41	1	241.08270	3.39	-1.24	2
Thymine	C5H6N2O2	[M+H] ⁺	127.05020					127.05030	3.33	0.79	2
Trihydroxy-eicosatetraenoate	C20H32O5	[M+Na] ⁺	375.21420	375.21380	10.65	-1.07	3				
Trihydroxy-octadecadienoate	C18H32O5	[M-H] ⁻	327.21770	327.21800	8.87	0.92	3	327.21710	8.84	-1.83	5
Trihydroxy-octadecadienoate	C18H32O5	[M+Na-2H] ⁻	349.19964	349.20000	8.87	1.03	3	349.19890	8.85	-2.12	5

Continued on next page...

Table 5.2 – Continued from previous page

Name	MF	Adduct	Theoretical m/z	ABR6				Bd21			
				m/z	rt	Δ PPM	Cluster	m/z	rt	Δ PPM	Cluster
Trihydroxy-octadecadienoate	C18H32O5	[M+Na] ⁺	351.21420	351.21380	8.87	-1.14	3	351.21410	8.85	-0.28	5
Trihydroxy-octadecadienoate ¹³ C	C16C2H32O5	[M-H] ⁻	329.22440	329.22480	8.87	1.23	3				
Trihydroxy-octadecadienoate ¹³ C	C17iCH32O5	[M-H] ⁻	328.22104	328.22140	8.87	1.1	3	328.22050	8.84	-1.64	5
Trihydroxy-octadecadienoate ¹³ C	C17iCH32O5	[M+Na] ⁺	352.21810					352.21740	8.86	-1.99	5
Trihydroxy-octadecenoate	C18H34O5	[M+Na] ⁺	353.22985	353.22930	9.06	-1.56	3				
Uridine	C9H12N2O6	[M-H] ⁻	243.06226	243.06250	1.69	0.99	1	243.06190	1.76	-1.48	2
Uridine	C9H12N2O6	[M+FA-H] ⁻	289.06770	289.06810	1.69	1.38	1				
Uridine ¹³ C	C8iCH12N2O6	[M-H] ⁻	244.06561	244.06580	1.69	0.77	1				
Valyl-leucine	C11H22N2O3	[M+H] ⁺	231.17030	231.17005	4.56	-1.08	2	231.17030	4.55	0	2
Valyl-valine	C10H20N2O3	[M-H] ⁻	215.14010					215.13970	3.27	-1.86	2

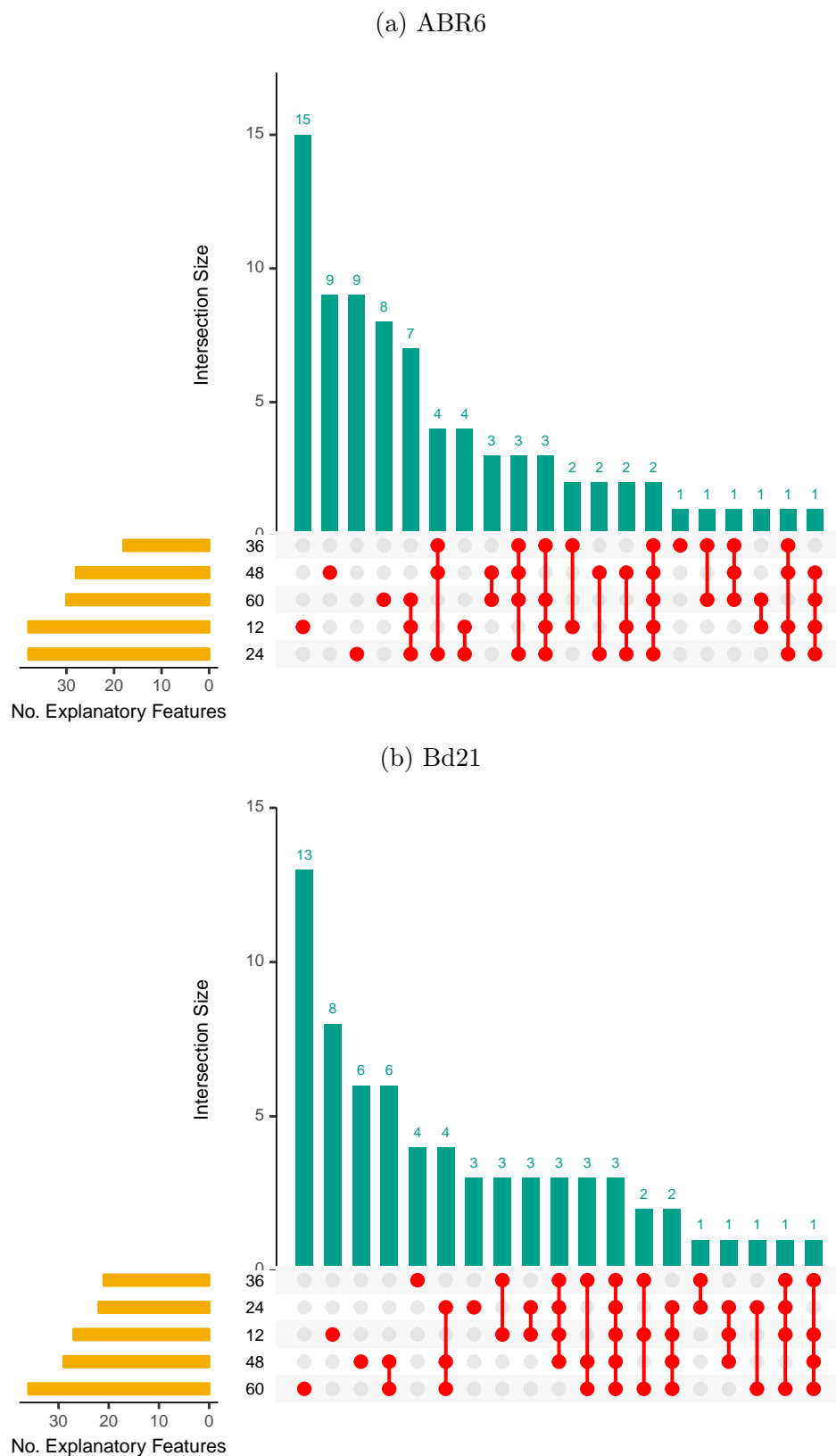


Figure 5.3: **Intersection plot of FIE-HRMS explanatory features identified by Random Forest in comparisons between control and infected tissue at each time point.** A selection frequency FPR of 1% was used for explanatory m/z feature thresholding.

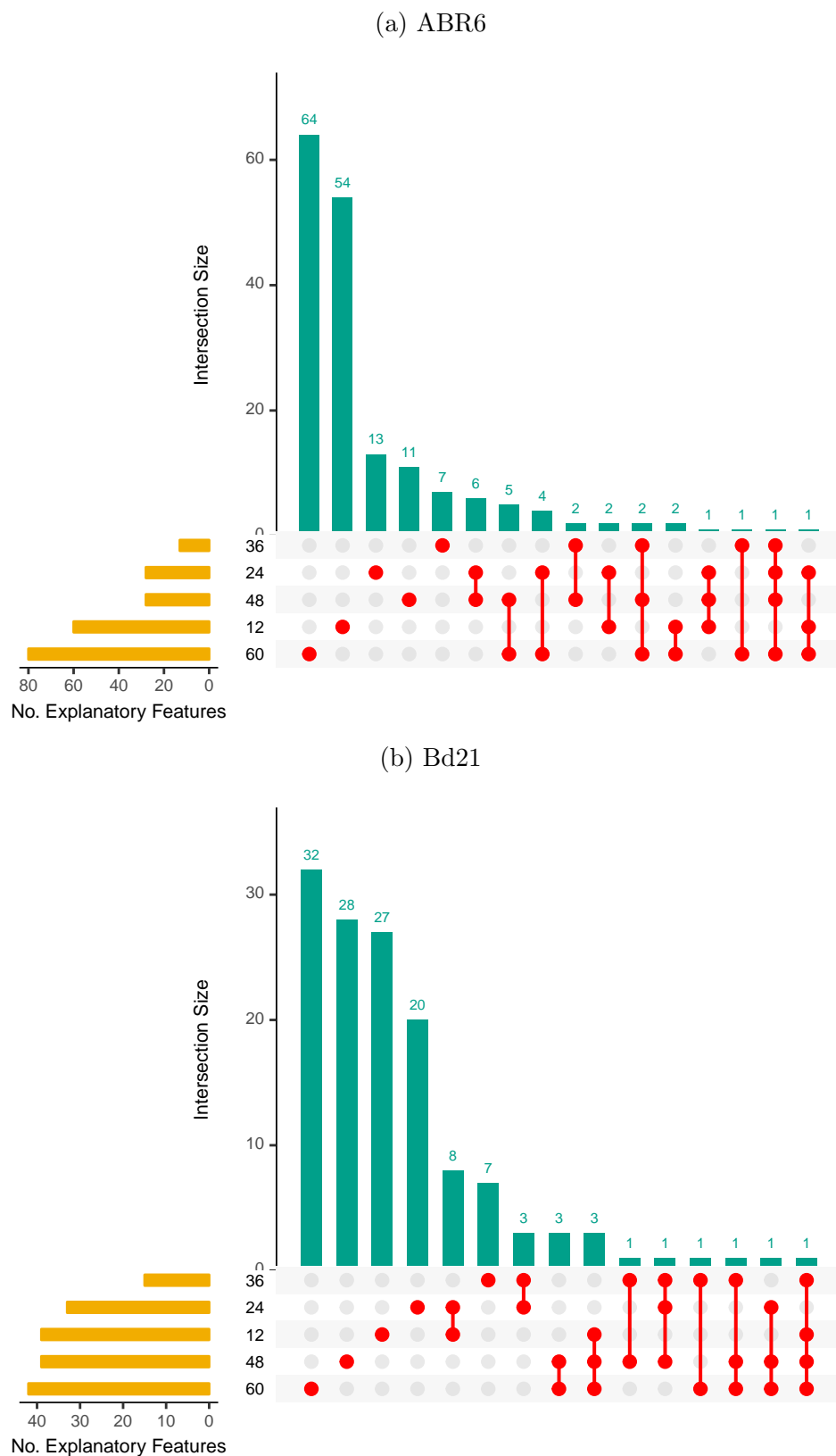


Figure 5.4: **Intersection plot of LC-HRMS explanatory m/z identified by Random Forest in comparisons between control and infected tissue at each time point.** A selection frequency FPR of 1% was used for explanatory m/z feature thresholding.

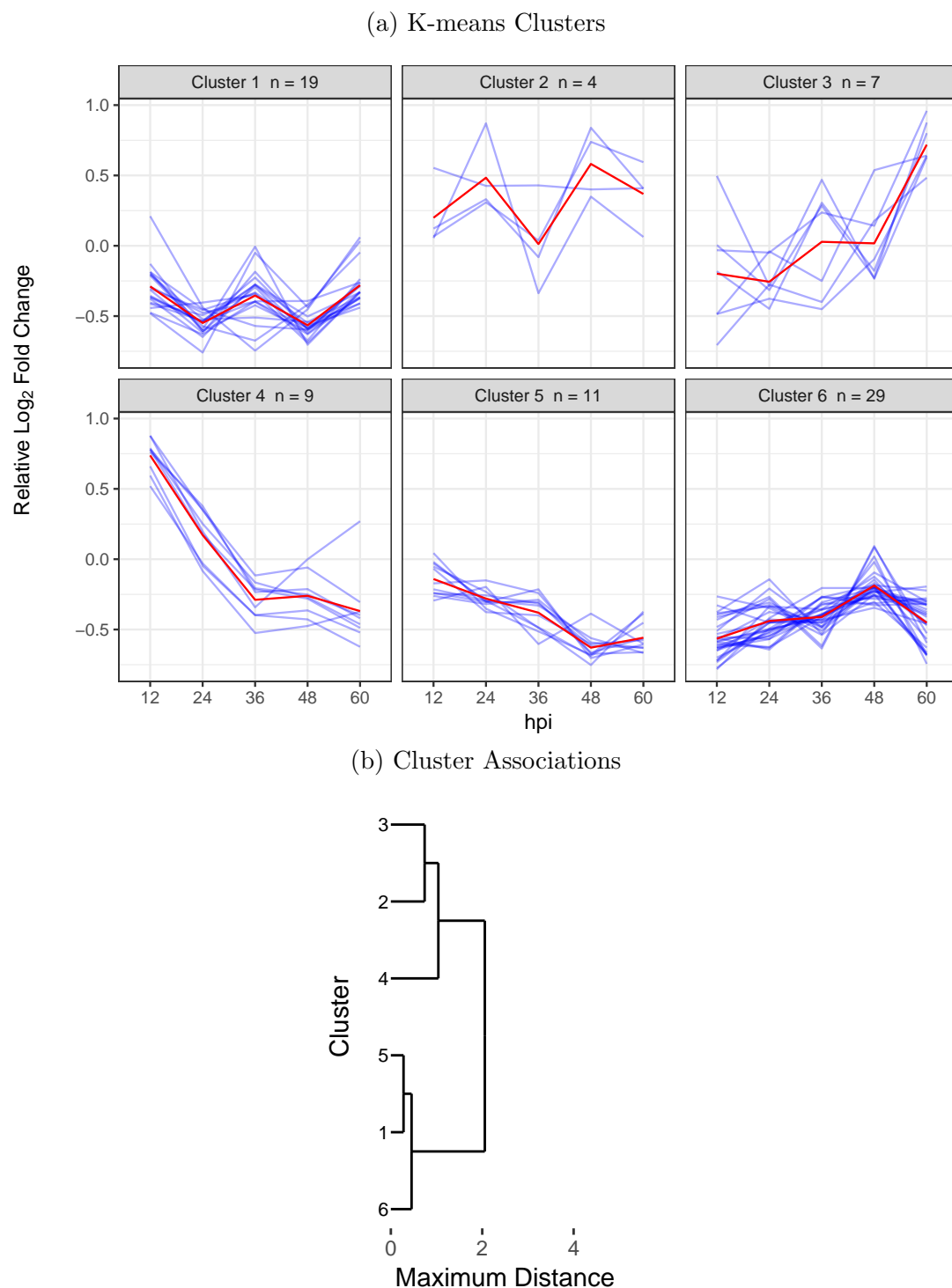


Figure 5.5: ***K*-means clustering of ABR6 explanatory m/z identified using FIE-HRMS.** a) Log₂ fold changes were normalised by the feature sum of squares. The red line denotes the cluster means. b) Maximum distance hierarchical clustering used to associate the cluster means identified in a).

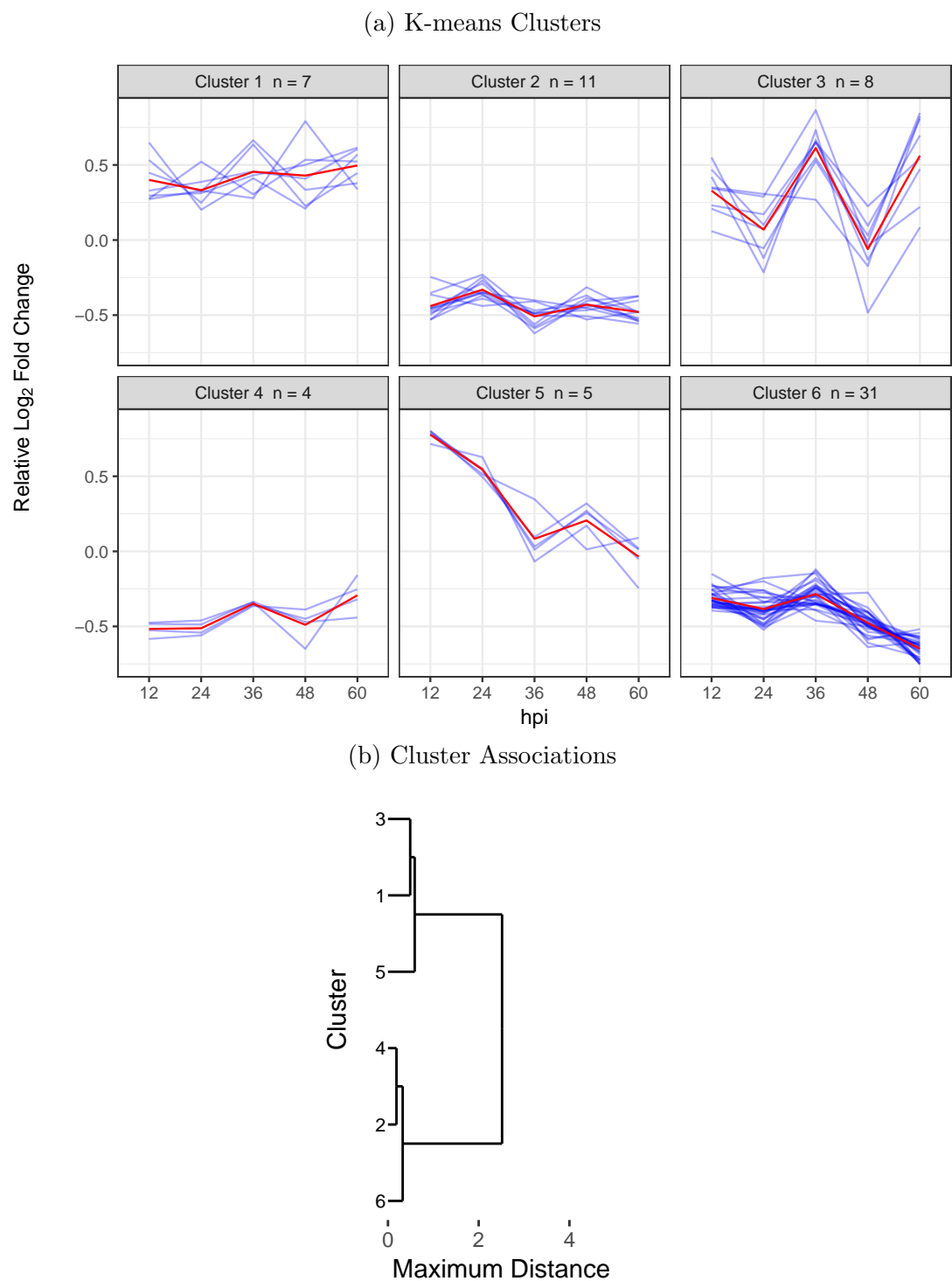


Figure 5.6: **K-means clustering of Bd21 explanatory m/z identified using FIE-HRMS.**a) Log₂ fold changes were normalised by the feature sum of squares. The red line denotes the cluster means. b) Maximum distance hierarchical clustering used to associate the cluster means identified in a).

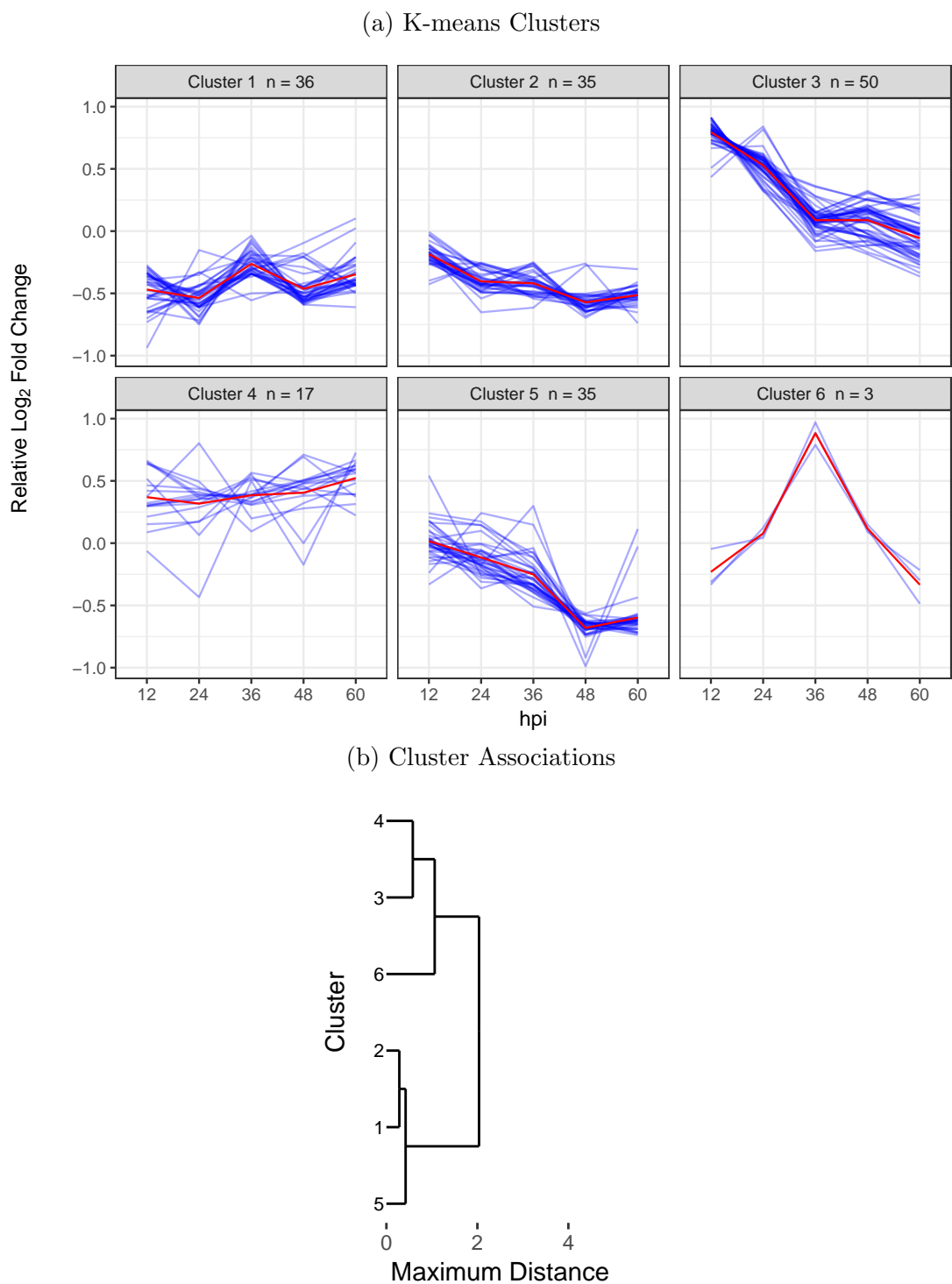


Figure 5.7: ***K*-means clustering of ABR6 explanatory m/z identified using LC-HRMS.**a) Log₂ fold changes were normalised by the feature sum of squares. The red line denotes the cluster means. b) Maximum distance hierarchical clustering used to associate the cluster means identified in a).

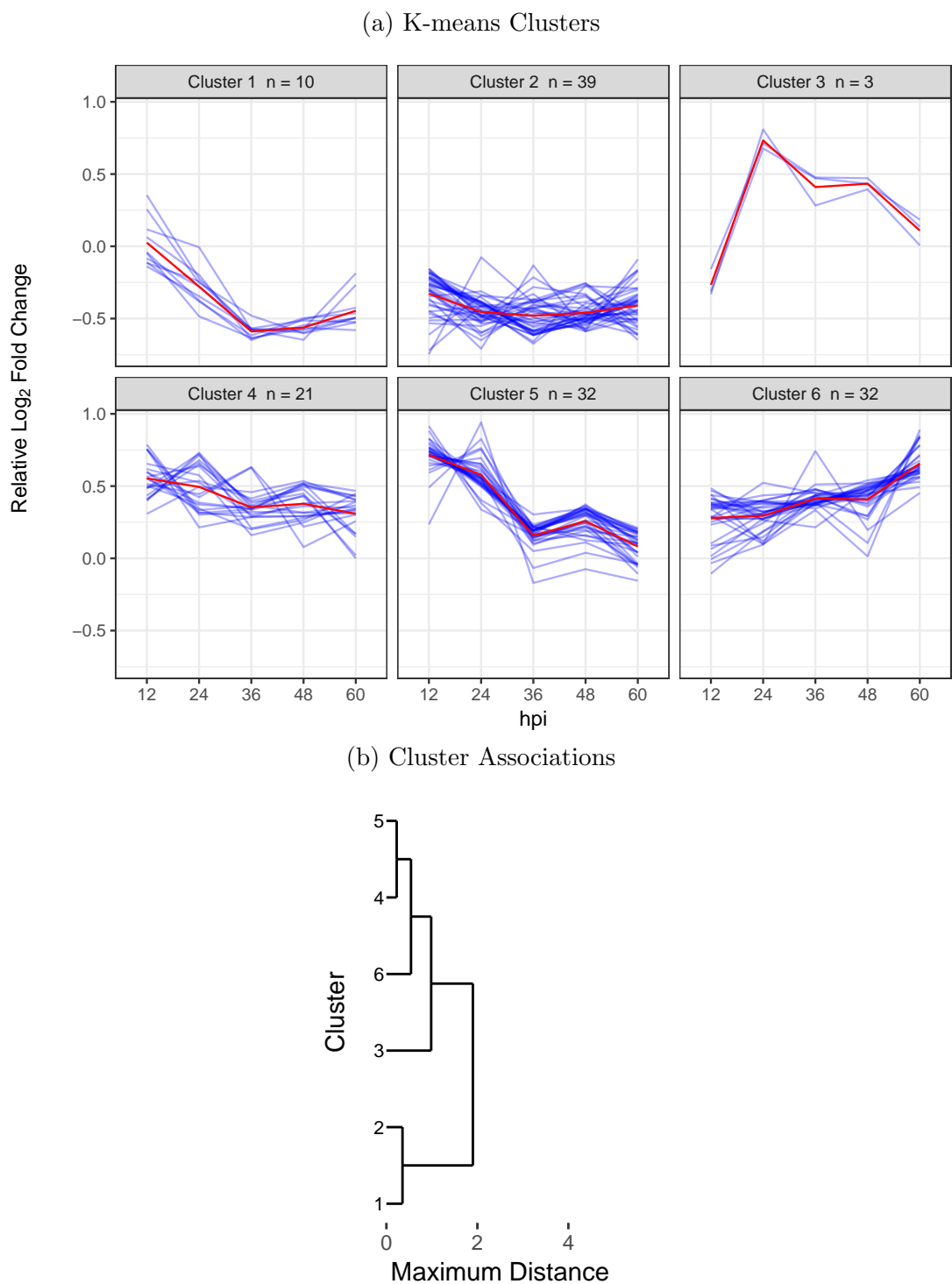


Figure 5.8: ***K*-means clustering of Bd21 explanatory m/z identified using LC-HRMS.**a) Log₂ fold changes were normalised by the feature sum of squares. The red line denotes the cluster means. b) Maximum distance hierarchical clustering used to associate the cluster means identified in a).

5.4.3 Transcriptomic changes during early phases of the *B. distachyon* and *M. oryzae* interaction

RNA-Seq transcriptomic analyses were used to analyse the transcriptomic changes at 12, 24 and 48 hpi using samples taken from the same inoculations as for the metabolomic analyses discussed in Section 5.4.2.

Due to cost limitations, not all of the six time points were able to be analysed. Therefore the three time points (12, 24 and 48 hpi) were chosen based on the results on the metabolomic analyses as well as their coverage of key time points. The time points 24 and 48 hpi are key microscopic time points where primary and secondary host cell invasion is occurring in the compatible interaction. By 48 hpi in the incompatible interaction, the initiation of host defences is well under way. 12 hpi was selected due to the extent of discrimination identified in the metabolomics analyses. An assessment of the underlying host transcriptome was seen as essential in understanding the source of this discrimination at such an early phase of the infection process.

The transcriptomic results presented here are that of only the host transcriptome. Due to the extent of multiplexing used for the RNA-Seq analysis, insufficient coverage was obtained in order to provide enough reads of *M. oryzae* origin for a transcriptome assembly. No more than 1% of the reads obtained for any of the infected samples in both interaction types could be aligned to the *M. oryzae* genome.

5.4.3.1 The *B. distachyon* interaction transcriptome during pre-symptomatic phases

A total of 33608 genes were detected across all the time points in both ecotypes. Of these 27549 and 27750 were expressed in ABR6 and Bd21 respectively. Also there was a total of 885 novel genes without previous annotations identified by Tophat.

5547 and 5798 differentially expressed gene (DEG)s were found in ABR6 and Bd21 respectively. Full lists of DEGs can be found in Appendix E. Similar trends

were also found between the ecotypes in terms of the numbers of DEGs that were found in comparisons between infected and control tissue across the three time points (Figure 5.9). Surprisingly, 12 hpi had substantially more DEGs compared to the other time points in both ecotypes and had the highest proportion of DEGs only present at that time point. 24 hpi had the lowest proportion of DEGs particular to that time point. Most of the genes that were differentially expressed at 24 hpi are also differentially expressed at either all three time points or at 12 hpi.

These results generally concur with the extent of discrimination that was found in the metabolomic analyses in that 12 hpi was highly explanatory, 24 hpi was showing relatively little discrimination and 48 hpi was somewhere in between. The extent of discrimination found in the transcriptomes of both the ecotypes at 12 hpi lends support to the idea that the origin metabolomic discrimination found at the same time point is as a result of PAMP triggered pathogen recognition responses within the host plant. This will further be discussed in Section 5.4.3.2.

5.4.3.2 Gene co-expression clusters and functional enrichment identified during pre-symptomatic infection phases

K means clustering was used to associate identified DEGs in co-expression clusters (Figures 5.10 & 5.11). In order to assess cluster function, enrichment analyses were used on Mapman bin ontology assignments to identify over-represented gene function within the co-expression clusters (Figures 5.12 & 5.13). Mapman and the Mapman bin ontology are discussed in Section 6.1.1. The clusters identified in ABR6 show an even split between up and down regulated clusters except for Cluster 6 that showed both up and down regulation across the time course (Figure 5.10b). Bd21 also showed an even split between up and down regulated cluster with Cluster 2 showing both up and down regulation over the time course (Figure 5.11b). The largest cluster in the incompatible interaction of ABR6 containing 878 DEGs (Cluster 4 in Figure 5.10a), showed up-regulation in infected compared to control tissue at all three time points. The largest cluster in the compatible

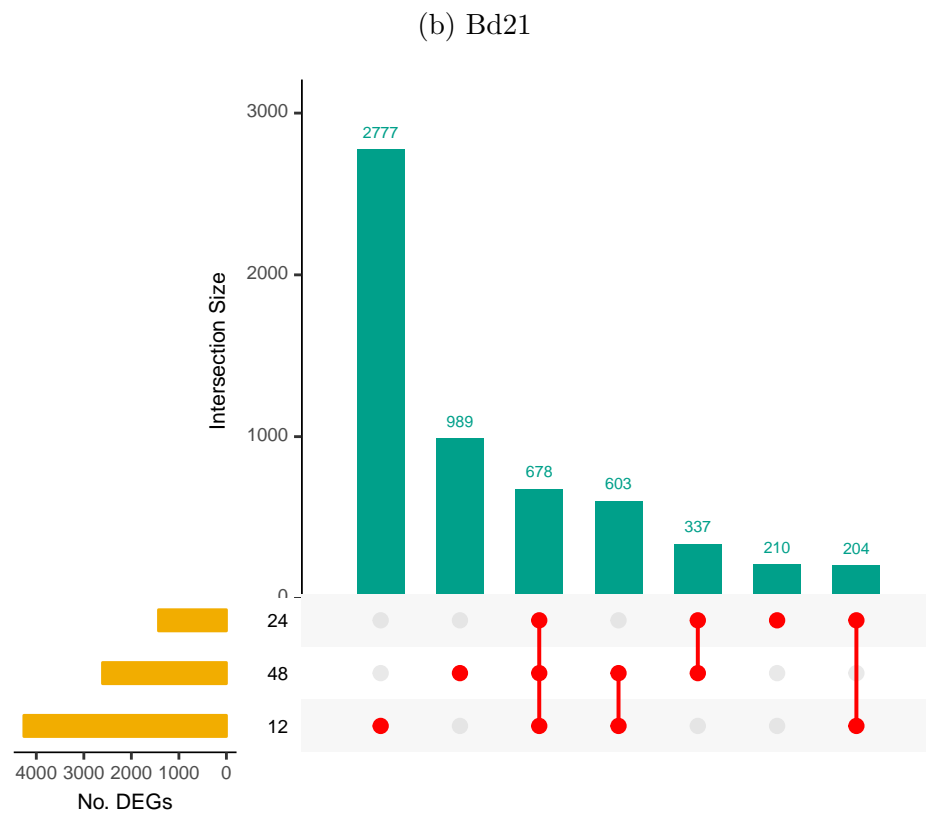
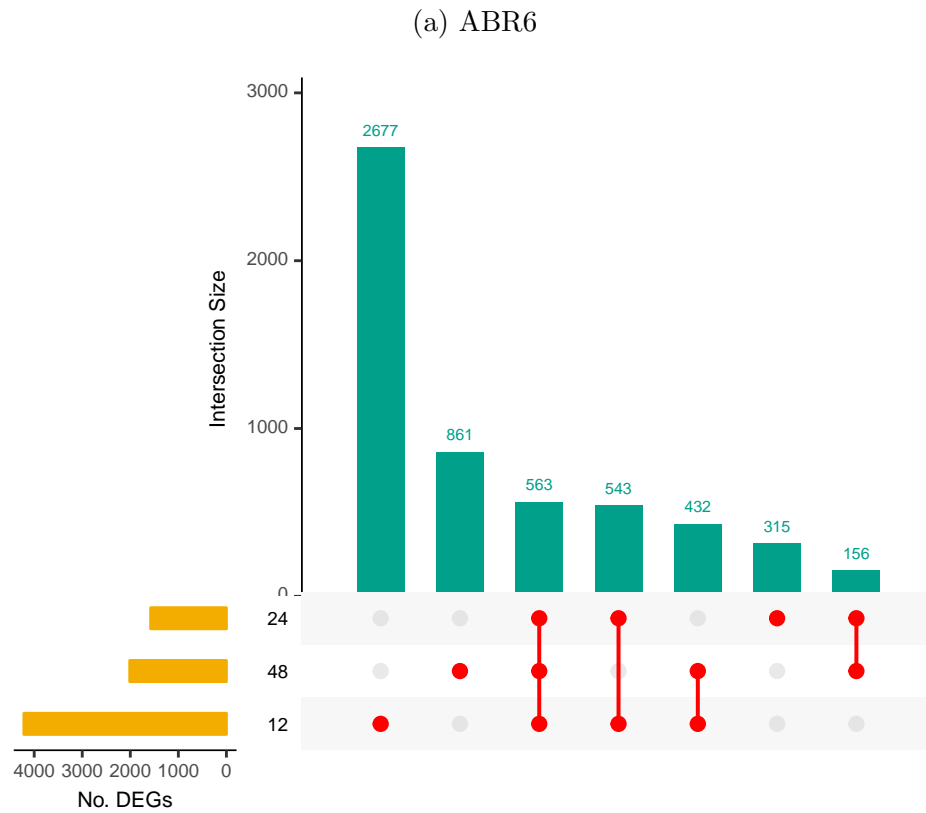


Figure 5.9: Intersection plot of RNAseq DEGs for in comparisons between control and infected tissue at each time point. A significance threshold of $p < 0.05$ was used for DEG identification.

interaction of Bd21 containing 825 DEGs (Cluster 1 in Figure 5.11a), showed up-regulation at 12 hpi with no difference in expression in infected compared to control tissue at the later time points.

There were substantially more over-represented Mapman bins found in the incompatible interaction of ABR6 compared to the compatible interaction of Bd21 (Figures 5.12 and 5.13). Also there were no over-represented bins found for Clusters 4 and 12 in Bd21. There was a high degree of similarity in the bins that were identified between the ecotypes. 99 and 57 Mapman bins were found to be over-represented in ABR6 and Bd21 respectively. Of those found in Bd21, 75% were also found in ABR6.

Further to the high numbers of DEGs found at 12 hpi, a number of clusters were found in both ecotypes that showed both up and down regulated differential responses at this time point (Figures 5.10 and 5.11). A notable functional response found within these clusters was an up regulation of photosynthesis related genes in both ecotypes at this time point (Clusters 6 and 10 in Figures 5.10a and 5.11a respectively). This is likely as a result of PAMP triggered immune responses. Photosynthetic productivity is an important part of the plants ability to effectively activate an oxidative burst (Bolton 2009). These changes in the expression of photosystem genes are likely to be in preparation of the production of an oxidative burst as part the defence response (discussed further in Section 6.4.3.1).

Phenylpropanoid metabolism genes were also found to be over-represented in clusters that were down regulated at 12 hpi in both ecotypes (Clusters 11 and 3 in Figures 5.10a and 5.11a respectively). There is little difference between expression in infected and control tissue at the subsequent time points. Interestingly, the increases in the phenylpropanoid metabolites found in the compatible interactions by Parker et al. (2009) were found at much later time points (> 24hpi). Phenylpropanoid metabolism is usually associated with the production of lignin monomers and phytoalexins during plant defence (Dixon et al. 2002). With this differential expression occurring very early in the infection time course, it is likely

that these changes are as a result of pathogen recognition responses. The reason for their down regulation however is unclear.

Based on the over-represented bins present in Cluster 12 of ABR6 (Figure 5.10a), there is evidence of an oxidative burst are under way within the *B. distachyon* leaf tissue by 48 hpi. This cluster is showing an increase of expression in infected tissue at 48 hpi and the over-representation of peroxidase encoding genes. There is also the presence of thioredoxin genes that are likely protectant measures against oxidative stress. ROS production encompassing a large number of cells around infection sites has previously been detected by 48 hpi in the *B. distachyon* ecotype ABR5 (Parker et al. 2009).

Genes relating to the synthesis of tryptophan were found to be over-represented in both ecotypes; although the trends of the clusters in which they were found differ (Clusters 12 and 2 in Figures 5.10a and 5.11a respectively). In the incompatible interaction of ABR6 Cluster 12 is up-regulated at 48 hpi whereas in the compatible interaction of Bd21 Cluster 2 is down regulated at 12 hpi. Tryptophan synthesis is essential for auxin production and auxins are key signalling molecules in plant defence alongside jasmonates and salicylic acid, although it shows variation in response to different pathogens (Kazan and Manners 2009). Auxins are antagonistic to salicylic acid signalling responses but share many commonalities with jasmonate responses. The cluster in Bd21 also contains salicylic acid synthesis related genes which is closely associated with chorismate production, the genes of which are also found in this cluster. However in ABR6, the purpose of this expression module is unclear and does not look to be associated with salicylic acid production. Interestingly, in both ecotypes, over-representation was found for genes involved in jasmonic acid synthesis (Clusters 9 and 11 in Figures 5.10a and Cluster 7 in Figure 5.11a). These clusters all show a strong down regulation at 12 hpi with little change at subsequent time points similar to the trends found for Cluster 2 in Bd21 in which salicylic acid synthesis over-representation was found. Doehlemann et al. (2008) found an up-regulation of jasmonic acid related genes during the pathogen penetration phases of the maize and *U. maydis* in-

teraction equivalent to 24 hpi in this interaction however these trends were not found here.

In ABR6 it is likely that this up-regulation of genes involved in the tryptophan synthetic pathway is linked to the initiation of host defence signalling in the hypersensitive response at 48 hpi. In Bd21 it's role may be more complex and unclear. Their down-regulation at 12 hpi may be as part of the initial host recognition response although this seems counter intuitive. A more obvious explanation would be that this is part of pathogen suppression of host defence responses. However, as it is happening so early in the infection time course, it seems unlikely that the *M. oryzae* spores would be developed enough to begin to secrete effectors that could elicit such a response.s

A potential example for subversion of host responses could be the presence of over-represented ubiquitin, and specifically the E3 ligases, genes in both ecotypes. However, the trends of the clusters in which they are present differ over the infection time course. In the incompatible interaction of ABR6 Cluster 10 shows up regulation at all three of the time points. In the compatible interaction of Bd21 Cluster 1 shows up regulation at 12 hpi, with little differential expression at subsequent time points. A diverse range of plant pathogens across kingdoms (bacteria, fungi and oomycetes) have been found to target ubiquitination processes in order to subvert host defences (Pritchard and Birch 2011). Ubiquitination is a key process in the initiation and execution of host defence responses. It is key in the regulation of the oxidative burst, hormone signalling and gene induction. The *M. oryzae* effector AvrPiz-t has been shown to target the RING E3 Ubiquitin ligase APIP6 rice. This effector was shown to be trans-located during the biotrophic phases of the infection process by 30 hpi. The initial up-regulation of these ubiquitin genes would occur as part of the initial pathogen recognition responses. The return of their expression levels to that of the control tissue in Bd21 would seem to agree with these observations. Disruption of ubiquitin E3 transcription occurs soon after the initiation of the host pathogen haustorial interface (24 hpi) in Bd21. Subsequently, the up-regulation of these genes ceases

in Bd21, while in ABR6 their up-regulation is maintained. The *B. distachyon* ortholog of APIP6 (Bradi2g35180) was not found to be differentially expressed in these experiments, which gives the potential for other effector targets to be present in this interaction.

Cluster 1 in ABR6 contains genes showing up-regulation at 24 hpi. Significant bins within this cluster suggest that substantial chromatin restructuring could be occurring within the ABR6 cells as this cluster contains genes for histone production. This is further discussed in Section 6.4.2.1

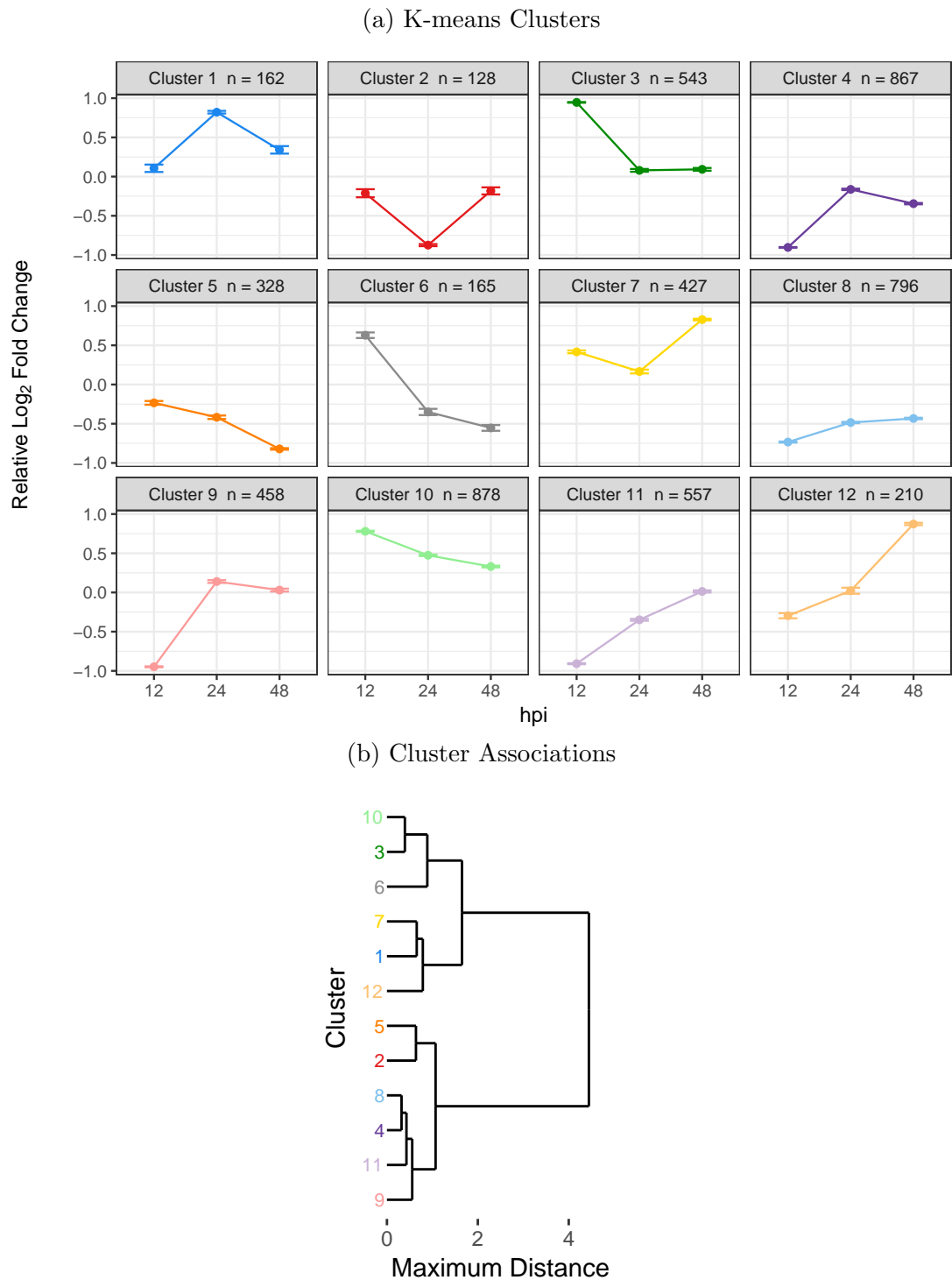


Figure 5.10: **K-means clustering to identify co-expression clusters of ABR6 DEGs.** a) Cluster means are plotted. Error bars are 95% confidence intervals, estimated using the t distribution. b) Maximum distance hierarchical clustering of cluster means identified in a). Cluster colours are retained between the plots.

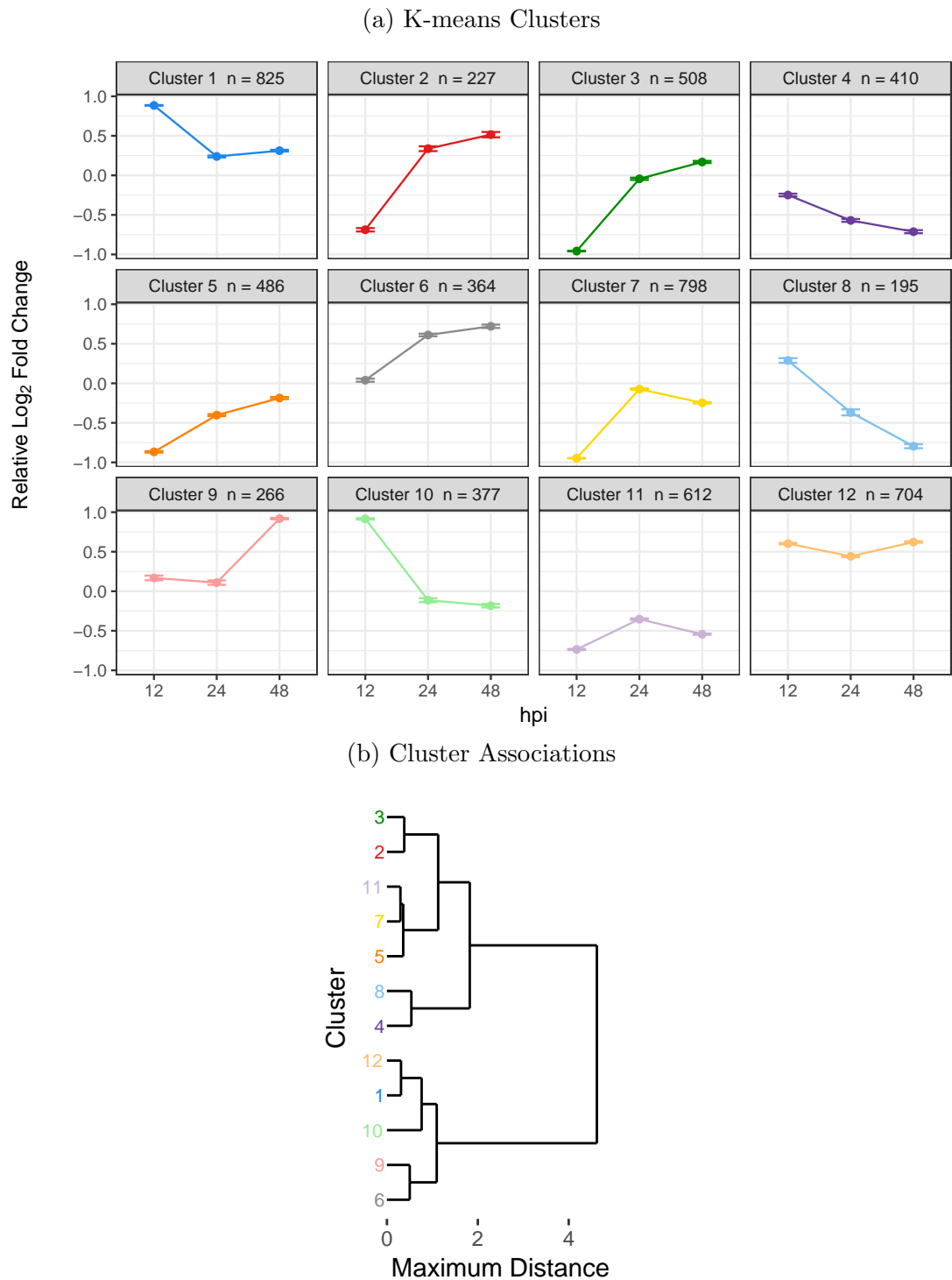


Figure 5.11: ***K*-means clustering to identify co-expression clusters of Bd21 DEGs.** a) Cluster means are plotted. Error bars are 95% confidence intervals, estimated using the *t* distribution. b) Maximum distance hierarchical clustering of cluster means identified in a). Cluster colours are retained between the plots.

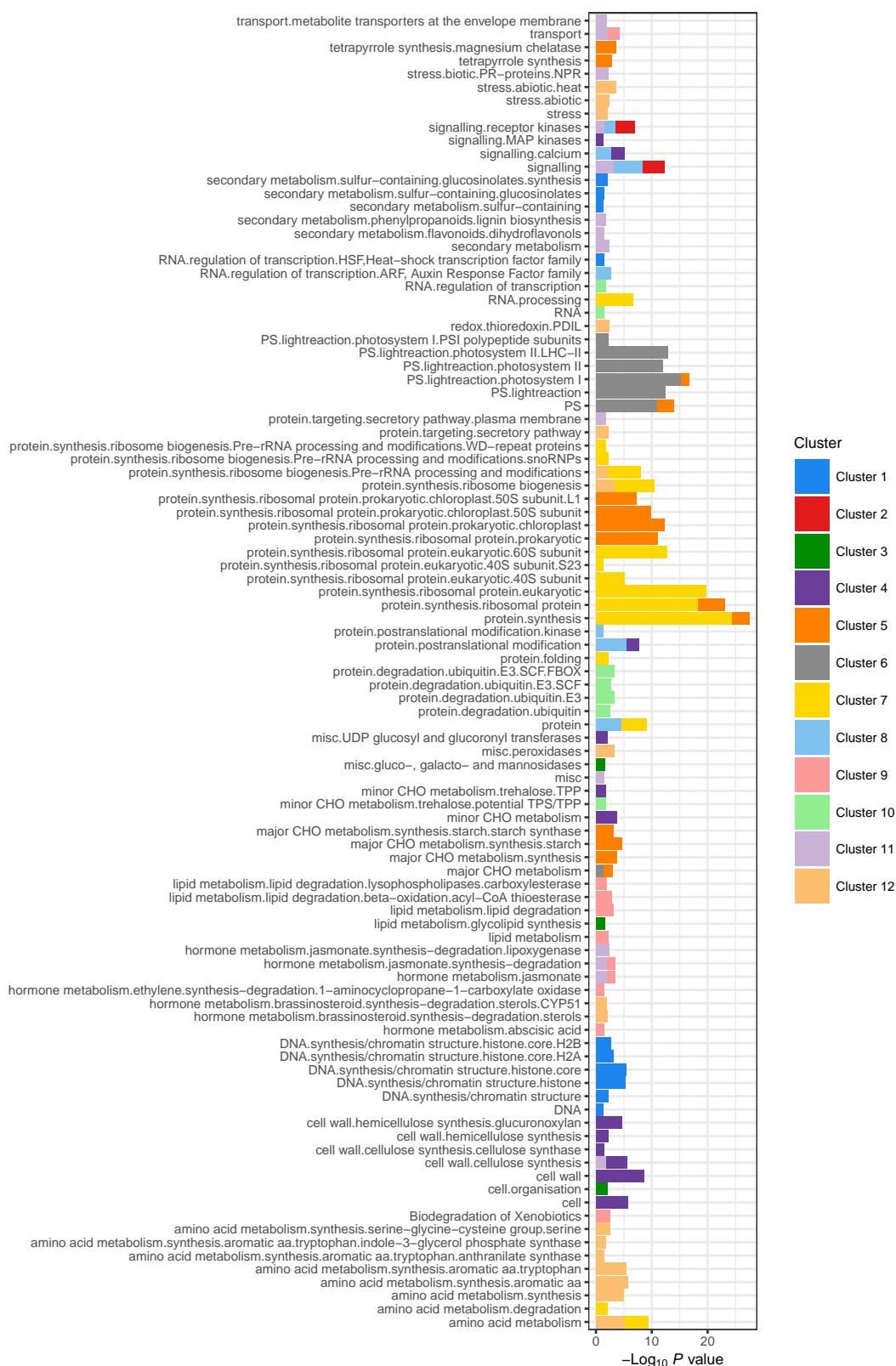


Figure 5.12: **Functional enrichment analysis of ABR6 co-expression clusters.** Clusters and cluster colours are based on those shown in 5.10. Mapman bins were used for functional ontology assignment. Fischers exact test was used to test for functional over-representation. An electronic version can be found in Appendix F.

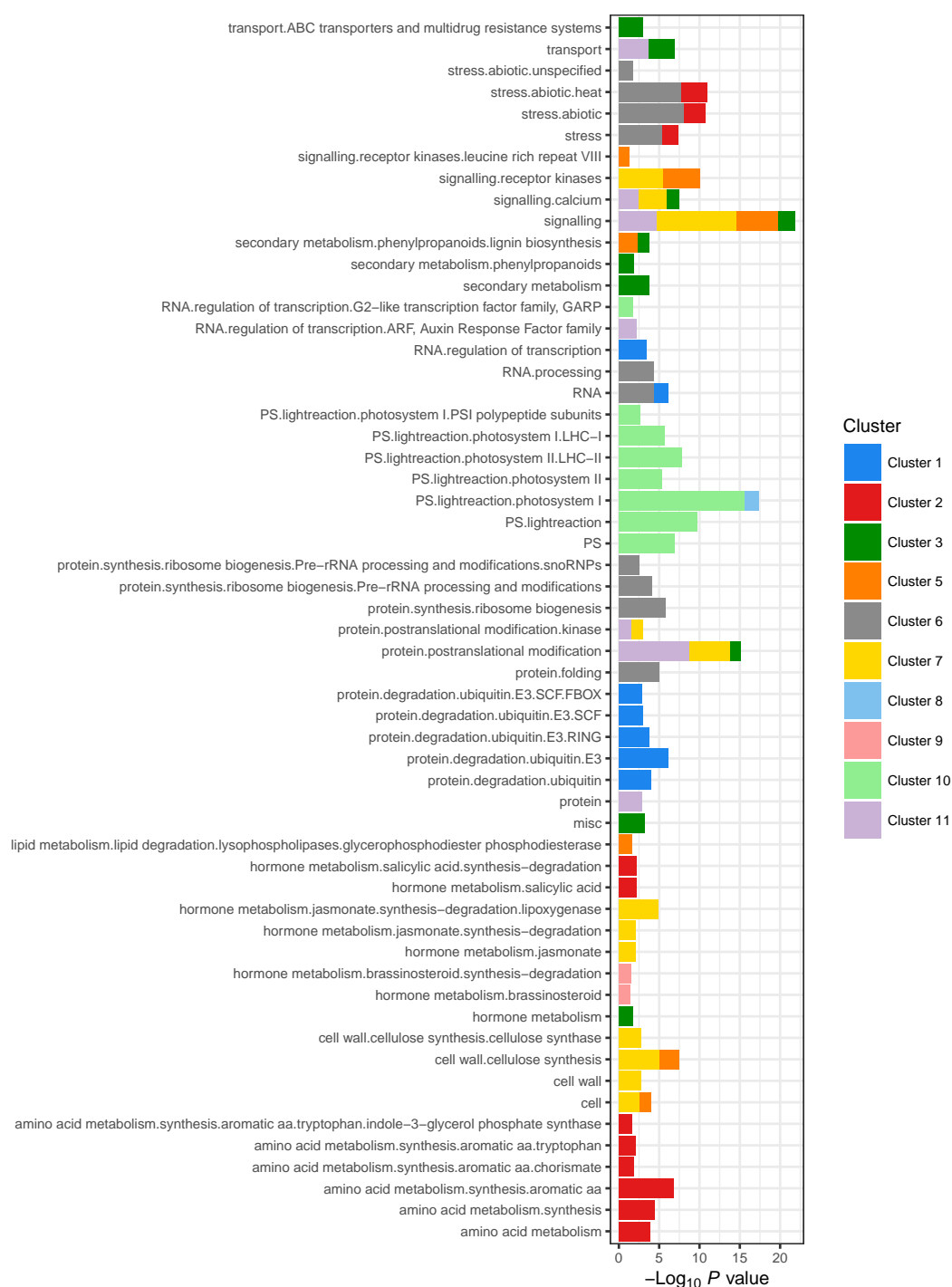


Figure 5.13: **Functional enrichment analysis of Bd21 co-expression clusters.** Clusters and cluster colours are based on those shown in 5.11. Mapman bins were used for functional ontology assignment. Fischers exact test was used to test for functional over-representation. An electronic version can be found in Appendix F.

5.5 Concluding remarks

These metabolomic and transcriptomic analyses have shown that infection by *M. oryzae* causes significant changes during the time course analysed in both compatible and incompatible interactions. However, the extent of these changes were dynamic over the course of infection and were not linear with disease progression.

Key microscopic events such as initial host cell penetration did not necessarily confer greater transcriptional or metabolic changes. In fact, initial host colonisation phases at 24 and 36 hpi were found to have the least discrimination in the metabolomic analyses and 24 hpi had the fewest identified DEGs. Interestingly 12 hpi was found to have substantial metabolomic and transcriptomic changes in both interactions. It was hypothesised that these responses were associated with PAMP triggered immune responses that would be caused by secreted compounds and enzymes, post spore germination. This is likely to be caused by the adhesive secretions that the spores use when attaching to the leaf surface as well as cell wall degrading enzymes during appressorium development.

There was the potential for disruption to ubiquitination by *M. oryzae* in the compatible interaction post 12 hpi. Also there was evidence for the activation of an oxidative response in the incompatible interaction by 48 hpi with the up-regulation of peroxidases and thioredoxins.

The overall objective of this chapter was to analyse each of the omics analyses individually to generate initial hypotheses with some qualitative comparisons between them. The following chapter will aim to directly integrate these data sets in order make more objective comparisons between these analyses. This will attempt to further inform the hypotheses generated in this chapter as well as potentially identify new interesting aspects of these interaction responses.

Chapter 6

Omics integration to elucidate key pathways in the *B.* *distachyon* and *M. oryzae* interaction

6.1 Introduction

Omics analyses provide systems level measurements that can cover a wide range of aspects of the cellular environment. One of the current challenges for investigators in all fields of omics application is the interpretation of these large-scale data sets with respect to the fundamental biological information that they contain. Effective analysis and interpretation of omics data is also tightly linked with its visualisation (Gehlenborg et al. 2010). Integrating of multiple omics data sets provides a solution to this problem, allowing investigators to broaden the view of the underlying biological system. However, this also creates it's own challenges, apart from further increasing the overall volume of available data.

The integration of omics data sets is especially important for studying dynamic processes such as plant-pathogen interactions. These are multi layered processes where molecular interactions are not restricted to a single level of the

cellular hierarchy. This includes both the subversion of host systems by fungal pathogens for nutritional gain as well as the initiation of the complex host defence responses.

6.1.1 Strategies for integrating omics data

Similar to the analysis and interpretation of individual omics levels, omics integration is centered upon identifying functional alterations and associations within biological systems as a result of experimental perturbation. The crucial difference is that information from multiple omics levels is included in these analyses. Strategies for the integration of omics data sets can be separated into data and knowledge driven approaches. The key distinction between these approaches is the premise on which the biological network will be constructed from the omics data. Data driven analyses attempt to identify all associations that may present in the data and use these to inform biological interpretation. Knowledge driven analyses use previously characterised pathways and associations and impose the data upon these to see how they are altered by a biological process. These strategies are not to be applied in isolation, with each able to inform the interpretation of the other (Cavill et al. 2015).

Prior to integration, each omics level data set will require its own pre-treatments to account for technical artifacts and biases that are particular to that analysis (Joyce and Palsson 2006). For instance, LC-HRMS metabolomics data will need substantially different processing requirements compared to RNA-Seq transcriptomics (Sections 5.1.1 & 5.1.4). The data will also require standardisation if potential cross-experiment variability is also going to be introduced. Omics data from differing levels are measured in completely differing magnitudes of scale. For instance, fold changes in gene expression do not necessarily have the same biological significance to fold changes of metabolite concentrations of the same magnitude. The relative trends with respect to the experimental perturbation are of more interest for integrated analyses rather than their absolute magnitude. Sum of squares normalisation of metabolite and transcript \log_2 fold changes has

previously been used to integrate metabolomic and transcriptomic data (Hirai et al. 2004).

Data driven omics integration uses the data to infer associations to represent the edges of the underlying biological network. These strategies rely upon the ‘guilt by association’ heuristic that can identify modules of highly connected nodes within the biological network. These are likely to have similar biological function within the context of the experimental perturbation (Wolfe, Kohane, and Butte 2005). Network associations can be calculated using correlations or by using clustering algorithms such as k means clustering or self organising maps (Gehlenborg et al. 2010). The addition of knowledge based functional ontology based enrichment analyses allows the identification of over-represented functional groups within the identified network modules. This can give biological context to the identified clusters and also inform further predictions of gene or metabolite function (Subramanian et al. 2005).

Knowledge driven omics integration analyses use previously established information collected about the biological system to derive network associations, upon which the newly collected data can be imposed. These can include metabolic pathways, signalling pathways or gene regulation pathways (Cavill et al. 2015). Knowledge based approaches can be targeted by focusing upon active modules identified using data driven approaches. This can then further enhance the interpretation of the function of these modules by giving them further functional context.

Crucial to these integration approaches is effective visualisation of the large volumes of omics data. Associative data driven approaches have different requirements to the pathway based knowledge approaches. Data driven approaches can be visualised through the use of heatmaps, dendrograms or interaction networks. Due to the volumes of data, these visualisation techniques can quickly become complex and overcrowded (Gehlenborg et al. 2010). Tools utilising these techniques therefore need to be interactive to allow investigators to identify areas of interest within the data.

There are a number of tools that have been developed for the visualisation of large and complex biological networks. Cytoscape allows interactive network visualisation for integrating interaction networks, expression profiles and phenotypes. It also extends with plug-ins that allow additional computational analyses (Cline et al. 2007). Another tool is VANTED that is tailored more for the visualisation and analysis of biological networks with related experimental data (Junker, Klukas, and Schreiber 2006).

Mapman is an extensive tool that allows the visualisation of experimental data in the context of metabolic pathways and biological processes in plants (Thimm et al. 2004). It is based on functional modules of plant metabolism known 'bins' that have > 200 hierarchical categories that include both genes and metabolites. These groupings can then be used to display experimental data onto pathway diagrams. Ontology maps of 'bins' are customisable. The web based Mercator tool can be used for the functional annotation of gene sequences into Mapman bins for previously uncharacterised organisms (Lohse et al. 2014). Due to the plant orientated nature of Mapman 'bin' ontologies, they have been found to outperform standard gene ontologies with respect to functional annotation (Klie and Nikoloski 2012).

6.1.2 Omics integration for plant stress responses

There has been a wide application of integrated omics investigations across the plant sciences. Many integrated analyses of plant stress have focused on abiotic stress mainly using transcriptomic and metabolomic data. Hirai et al. (2004) integrated FT-ICR-MS metabolomics and microarray transcriptomic data to investigate global responses to sulfur and nitrogen deficiency stress in *A. thaliana*. A data driven approach was used that incorporated PCA and batch-learning self-organising maps on relative \log_2 fold changes between treatment and control classes of the gene and metabolite profiles. This allowed them to associate the trends of genes and metabolites and suggested that general responses to these stresses were involved. In particular genes and metabolites involved in glucosino-

late metabolism were found to be coordinated.

There are few examples of the application of integrated omics analyses to investigate interactions between plants and their pathogens. Kumar et al. (2016) integrated NMR metabolomics and quantitative label-free proteomics to investigate the metabolic reprogramming of chickpea roots by *Fusarium oxysporum* by applying both data driven and knowledge driven integration approaches. Hierarchical clustering was used to associate proteins with ontology enrichment analyses applied to the identified clusters. Among these clusters, enrichment for proteins involved in lignin biosynthesis, protein degradation, and defence responses were found. Pathway mapping of proteins involved in lignin biosynthesis, phytoalexin synthesis, glycolysis and the TCA cycle revealed an up-regulation in the susceptible cultivar with down-regulation in the resistant cultivar. From the NMR metabolomics analyses, the phytoalexin luteolin was found to be decreased in the susceptible cultivar along with alterations to a number of amino acids as well as glucose and sucrose. Expression levels of a number of key genes were examined which were found to support many of the metabolomic and proteomic alterations already identified, particularly with respect to changes in carbon and nitrogen metabolism of the resistant cultivar.

Similarly, Gunnaiah et al. (2012) integrated LC-HRMS metabolomics and shotgun proteomic data to investigate resistance mechanisms of a QTL in Wheat against *Fusarium graminearum*. Resistance related metabolites were identified by correlating canonical discriminant vectors, used to classify the observations, with resistance phenotypes. These were then pathway mapped along with the proteomic profiles and identified metabolites involved in the phenylpropanoid pathway and enzymes contributing to cell wall thickening as important for functions of the QTL.

Other studies such as Doehlemann et al. (2008) and Voll (2011) have used both metabolic and transcriptional data to investigate plant pathogen interactions. However, they have used targeted metabolic analyses, focusing only on central metabolites, usually using the transcriptomic analyses to inform the targeting of

these metabolites. These are not strictly metabolomic techniques and cannot be counted as integrative omics analyses.

6.2 Aims

The aim of this chapter is to integrate the explanatory feature trends of metabolite and gene expression changes identified in Chapter 5. These changes can then be related to their likely molecular sources; both of pathogen origin (system manipulation and growth) and host defence responses (pathogen recognition and defence initiation).

Data integration will utilise both data and knowledge driven strategies. This gives the best chances of not only identifying changes that are directly linked to prior knowledge but also novel associations that would be otherwise be missed.

Once data has been integrated from the two omics levels, the system differences between the compatible and incompatible responses can be compared and contrasted. This will enable further identification of key areas of plant systems that pathogens perturb in order to suppress host defences during early colonisation phases.

This provides us with the following aims for this chapter:

- Integrate data from metabolomic analyses to identify areas of metabolism that are altered during the interaction.
- Integrate data from both metabolomic and transcriptomic analyses to identify associations indicative of pathogen perturbation or host defence responses.
- Identify key areas of metabolism that are altered during the interaction from which specific hypotheses can be developed and compare these between compatible and incompatible interactions.

6.3 Materials and Methods

6.3.1 Metabolomic and transcriptomic data preparation for integrative analyses

Metabolite and transcript profiles of explanatory m/z and differentially expressed genes were identified and annotated from FIE-HRMS, LC-HRMS and RNA-Seq of the pre-symptomatic phases of compatible and incompatible responses of the interaction between *B. distachyon* with *M. oryzae* described in Sections 5.3.4 and 2.12. Only explanatory m/z with putative annotations of at least a molecular formula were retained for integrative analyses with the removal of isotopic peaks. For m/z with more than one adduct present, the m/z with the most intense signal was retained. For metabolites that were represented in both the FIE-HRMS and LC-HRMS explanatory features, the profile from the LC-HRMS analyses were retained.

6.3.2 Integrative correlation network analysis and pathway mapping of metabolomic and transcriptomic data

For correlation network construction of data obtained from FIE-HRMS and LC-HRMS metabolomic analyses Pearson's correlations were calculated from the treatment and control means at each time point (12, 24, 36, 48, 60 hpi) for each of the ecotypes independently. Significant correlations were retained ($p < 0.05$) with non-significant correlations set to 0.

For correlation network construction for the integration of metabolite and transcript profiles, metabolite trends were first standardised for integration by taking \log_2 ratios between means of treatment and control classes at each time point (12, 24, 48 hpi) for each ecotype. The ratios of each feature was then normalised by its sum of squares to give relative fold changes. To reduce overcrowding of correlation network visualisations, the relative fold changes of cluster

means of gene co-expression clusters identified in Section 5.4.3 were used instead of individual transcript profiles. This allowed the association of general cluster functions identified by over-representation analysis of mapman bin ontologies (Sections 2.12 & 5.4.3).

Relative fold change of treatment and control means at each time point (12, 24, 48 hpi) for individual metabolite and transcript profiles were used for pathway mapping using and Mapman (version 3.5.1R2). Mapman bin ontologies were constructed using the Mercator tool as described in Section 2.12. Files for Mapman analyses can be found in Appendix G.

6.4 Results and Discussion

6.4.1 Integration of metabolomic analyses

The results of the metabolomic analyses described in Chapter 5 were integrated to both further their biological interpretation as well as ensure analytical agreement between the data sets. The trends of the explanatory m/z identified in Section 5.4.2.2 with the same accurate mass within 5 ppm of each other were compared between the FIE and LC-HRMS analyses.

As shown in Table 6.1, there were 8 and 4 m/z found with the same accurate mass in ABR6 and Bd21 respectively. For all of the FIE-HRMS m/z with matches there was only one corresponding LC-HRMS m/z . Of these, only two features (n327.22, p203.14) in ABR6 did not have significant correlations ($p < 0.05$). However in both cases, the significance of the correlations of these m/z were very close to the threshold.

The positive correlation of the explanatory m/z between the two metabolomics techniques allow two things. Firstly it allows the identification of FIE-HRMS bins that contain multiple isomers that have then been chromatographically resolved in the LC-HRMS analyses. This strategy can then allow duplicated features between analyses to be removed. Secondly the coherence of these m/z allows validation of the accuracy of the analytical approaches, signal processing and data

pre-treatments used to make these metabolomic measurements.

Correlation networks were constructed using putative annotations from both analyses for both ecotypes. For metabolites represented by multiple adducts, the most intense adduct was selected. Isotope annotations were also removed. The LC-HRMS m/z were retained for metabolites represented in both analyses. This identified 78 and 57 annotated metabolites in ABR6 and Bd21 respectively.

The metabolites found in the incompatible interaction of ABR6 showed a much greater diversity in trends over the infection time course compared to those found in the compatible interaction of Bd21. This is shown by the smaller cluster sizes in ABR6 in Figure 6.1 compared those of Bd21 in Figure 6.2. The metabolite network of Bd21 can broadly be separated into two main negatively correlated clusters.

It was identified in Section 5.4.2.2 that mono and di-saccharides and amino acids as well as a number dipeptides, purines, pyrimidines and nucleosides were present in clusters that were showing decreases in infected tissue compared to infected tissue. These metabolites show clear association in both ecotypes when the metabolite data sets are combined (Figures 6.1 & 6.2).

It is likely that the down-regulation of these metabolites is as a result of energy production. Sucrose and glucose are obvious direct energy sources through glycolysis. Plants are known to have the ability to degrade purines, pyrimidines and their nucleosides in order to boost energy production (Zrenner et al. 2006). Purine degradation can allow the production of glyoxylate via the production of allantoin. This glyoxylate can then be fed into the glyoxylate cycle to produce energy.

Pyrimidines can be catabolised to produce β alanine that can be used for co-enzyme A biosynthesis which can then in turn be used for energy production. Alanine has been putatively annotated in the FIE-HRMS analyses and is associated with the pyrimidines in both ecotypes. Phospho-pantothenate is similarly present in both ecotypes and in the same clusters, a metabolite present in the co-enzyme A biosynthesis pathway.

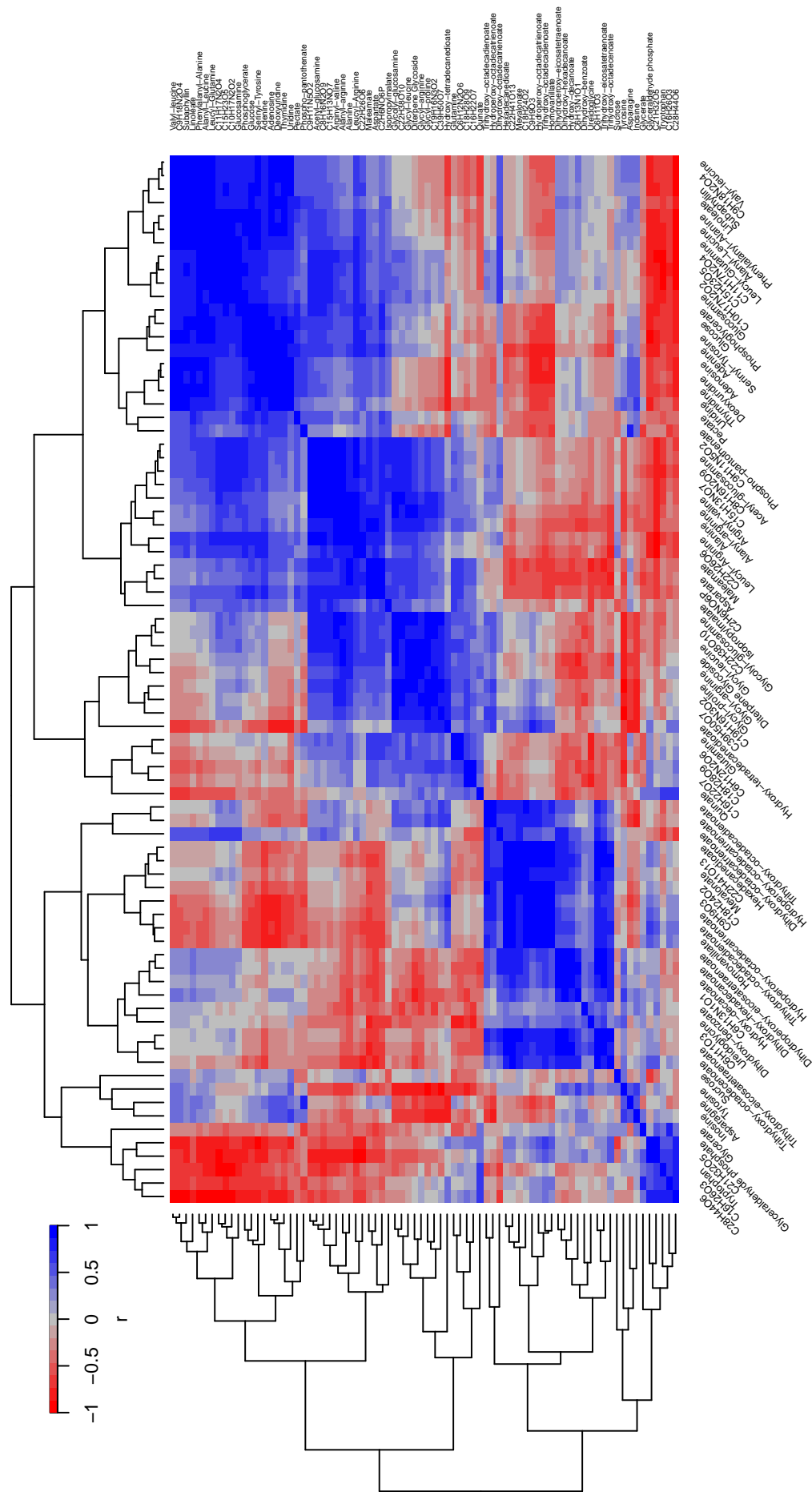
It is unclear as to which organism this catabolism is likely to be occurring as it can be argued that energy production would be important in both organisms during the entire time course of the infection. However, a point to add is that these changes could potentially originate from *M. oryzae* as the all the time points are similarly decreased instead of a gradual decrease over the time course. As discussed in Section 5.4.2.2, β -oxidation of fatty acids is essential for energy production in *M. oryzae* appressorium development and cell wall synthesis during spore germination. With energy production already highlighted as a requirement for primary cell penetration, these metabolites could represent other energy sources that are also utilised by *M. oryzae* during this phase.

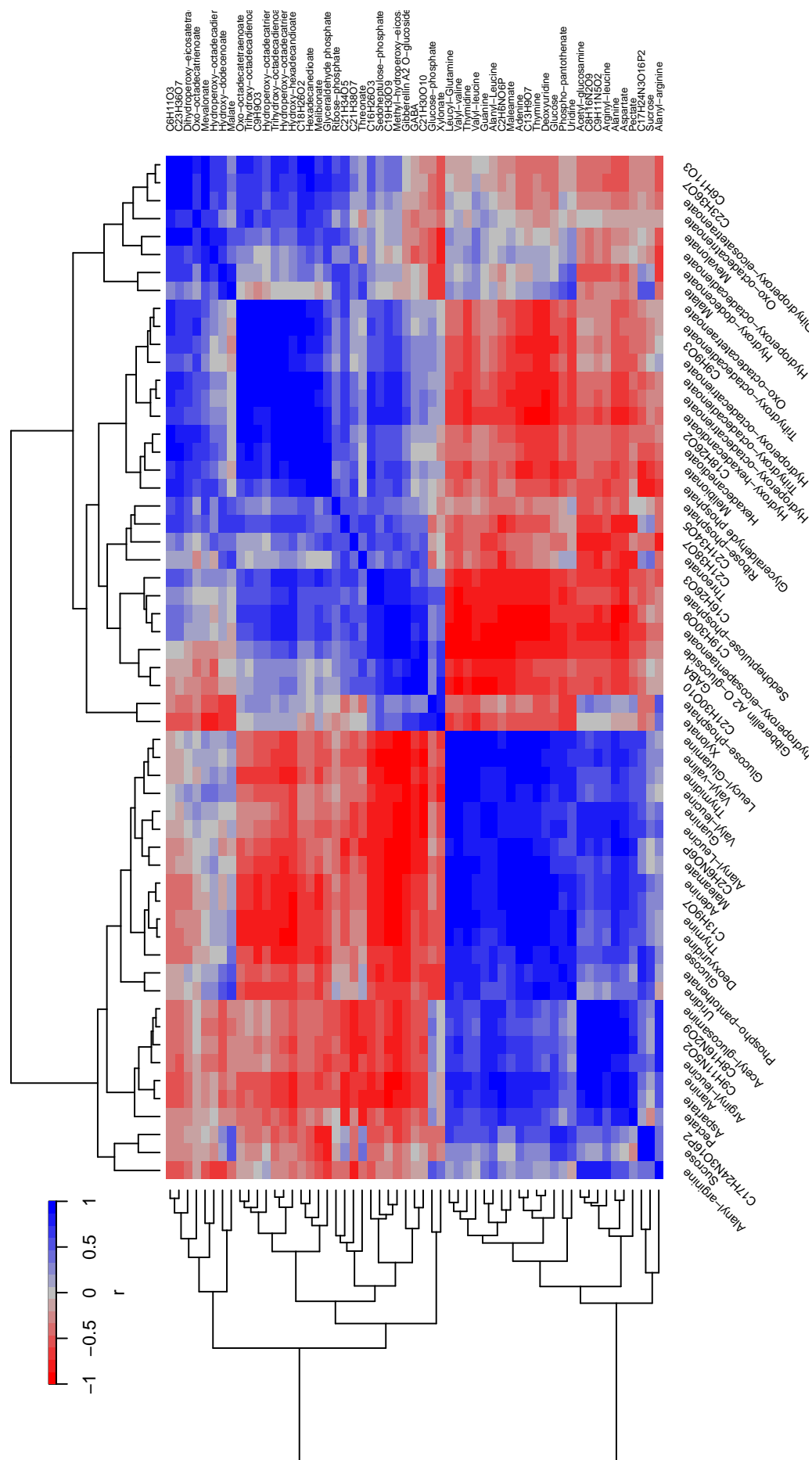
In Bd21 a number of Calvin-Benson cycle metabolites along with malate were also associated with the unsaturated fatty acids. The Calvin-Benson cycle metabolites include glyceraldehyde-phosphate, ribose-phosphate and sedoheptulose-phosphate. In ABR6 only glyceraldehyde phosphate and phospho-glycerate were identified. The role of these metabolites in the interaction will be further discussed in Section 6.4.3.1.

Unsaturated fatty acids were identified in both metabolomic analyses and were found to be associated together in both the ecotypes. These showed strong upregulation at 12 hpi when pathogen recognition responses will be occurring. As mentioned in Section 5.4.2.2, the likely functions of these include signalling for the initiation of host defence responses as well as the result of lipid catabolism for energy production. The role of these metabolites in the interaction will also be further discussed in Section 6.4.3.1.

Table 6.1: **Correlations of explanatory m/z found in both FIE-HRMS and LC-HRMS analyses.** Class means (N=10) for each time point and treatment for each ecotype. were used for pearson's correlation of trends between the metabolomic analyses excluding the 0 hpi samples. For ID's of both techniques, the lower case prefix refers to the ionisation acquisition mode and the following number is the m/z . For LC-HRMS IDs the following number gives the retention time of the m/z in minutes, prefixed by 'T'.

Ecotype	FIE-HRMS ID	LC-HRMS ID	r	p Value
ABR6	n131.07	nM131.07T6.28	0.96293	0.00004
	n132.03	nM132.03T1.23	0.97543	0.00001
	n277.03	nM277.03T1.29	0.79885	0.01672
	n327.22	nM327.22T8.87	0.68730	0.05618
	p136.06	pM136.06T1.42	0.98307	0.00000
	p203.14	pM203.14T3.99	0.61738	0.05720
	p246.16	pM246.16T1.25	0.97712	0.00001
	p288.2	pM288.2T1.48	0.94388	0.00016
Bd21	n131.07	nM131.07T6.27	0.95957	0.00004
	n342.11	nM342.11T1.41	0.92357	0.00027
	p382.08	pM382.08T1.36	0.94024	0.00016
	p517.01	pM517.01T1.36	0.91462	0.00027





6.4.2 Integration of metabolomic and transcriptomic analyses

As only three of the six presymptomatic time points that were analysed in the metabolomic analyses of the *B. distachyon* and *M. oryzae* interaction, only explanatory m/z present at 12, 24 and 48 hpi were used for integration with the transcriptomic analyses. As for integrating the metabolomics analyses, putatively annotated m/z that were not isotopic signals were carried for further analyses. For those m/z with multiple adducts present, the most intense ion was selected. The LC-HRMS m/z were selected for metabolites with annotations across both metabolomic analyses.

6.4.2.1 Network analysis of metabolomic and transcriptomic analyses

To aid in the visualisation and interpretation of metabolomic and transcriptomic networks, the cluster centers of the transcriptomic clusters identified in Figures 5.10 and 5.11 were used for integration with metabolite profiles. Contrary to the metabolomic networks identified across all five of the early phase time points, those constructed for the three transcriptomic and metabolomic time points showed a lower diversity of trends in ABR6 than in Bd21.

It was found that in ABR6 the amino acids aspartate, asparagine, alanine and glutamine were all closely associated with Cluster 12 of the transcriptomic data. As shown in Figure 5.12, Mapman bins related amino acid synthesis were found to be over-represented in Cluster 12. Also closely associated were the aromatic amino acids tyrosine and tryptophan. Functionally enriched subdivisions of the amino acid synthesis Mapman bins in Cluster 12 included transcripts involved directly in the synthesis of these aromatic amino acids as well as indole-3-glycerol phosphate synthesis. These amino acids were showing increases at 48 hpi, relative to the other time points. This indicates that these changes are likely as a result of the host defence responses. The amino acids aspartate, asparagine, alanine and glutamine are likely to reflect key changes in metabolism for the mobilisation of nitrogen resources.

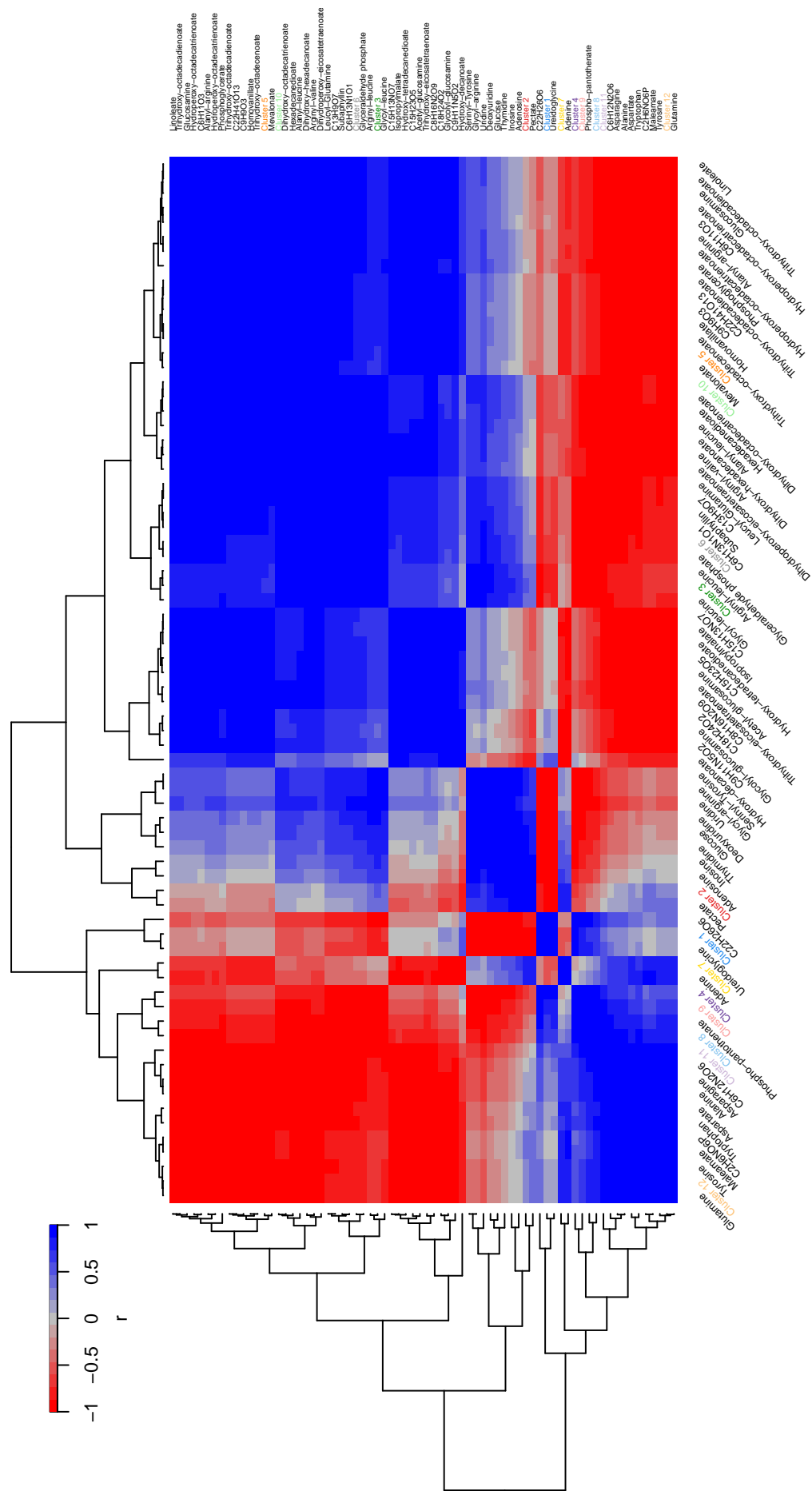
One of the key uses aromatic amino acid metabolism is in the biosynthesis of auxins. Auxins are key factors in the regulation of plant defences and show differential responses for different pathogens. These are antagonistic to salicylic acid signalling but commonalities have been found with jasmonic acid signalling (Kazan and Manners 2009). Transcripts relating to jasmonic acid synthesis were also found to be over-represented in the similarly associated Cluster 11. Brassinosteroid synthesis were found over-represented in Cluster 12 suggesting the likely importance that hormone signalling is having in the host resistance response.

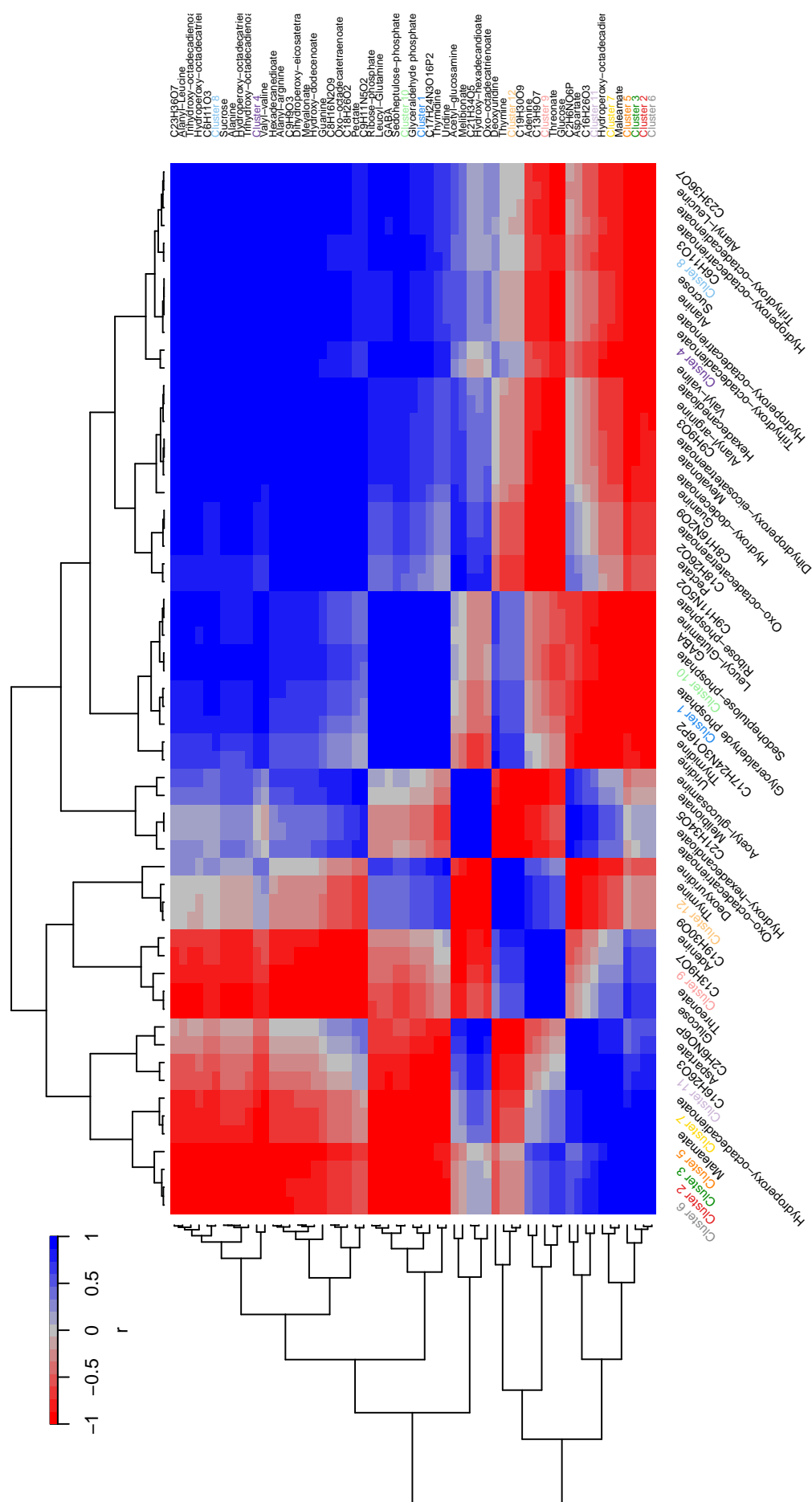
It was previously identified in Section 6.4.1 that purines, pyrimidines and nucleosides were associating with other metabolites involved in energy production. It was therefore inferred that changes in these metabolites could also be as a result of this. Interestingly in the incompatible interaction of ABR6, these metabolites are associated with genes in Cluster 1. This cluster was identified in Section 5.4.3.2 as being functionally over-represented for chromatin structure and showed strong upregulation at 24 hpi.

Priming of chromatin structure is important for ensuring fast and robust expression of defence related gene expression (Conrath 2011). For instance in *Arabidopsis*, ‘histone replacement’ of the H2A histone with the H2A.Z variant is crucial for appropriately regulating salicylic acid induced plant immunity. Mutants displayed phenotypes of hypersensitivity and spontaneous cell death (March-Díaz et al. 2008). This alteration to the expression of genes involved in histone production suggests that chromatin rearrangement is occurring at 24 hpi and that this could be pre-emptive of the substantial gene expression changes that would be needed to elicit defence responses such as the hypersensitive response. However, the link of this function with that of the cellular levels of purines, pyrimidines and nucleosides is unclear.

Calvin-Benson cycle metabolites have already been identified as important in both ecotypes as part of the host pathogen recognition responses at 12 hpi along with that of unsaturated fatty acids in Section 6.4.1. These metabolites were found to cluster with transcript Clusters 6 and 10 in ABR6 and Bd21 re-

spectively. Both of these gene clusters are functionally over-represented for gene involved in photosystems I and II. Interestingly, functional over-representation of lipoxygenase genes were found in Clusters 11 and 7 in ABR6 and Bd21 respectively. Both of these clusters are negatively associated with the clusters containing photosynthesis over-representation, Calvin-Benson cycle metabolites and unsaturated fatty acids. In ABR6, Cluster 5 was found to be associated with the unsaturated fatty acids. Cluster 5 included over-represented transcripts for chloroplastic ribosomal subunits. This implicates importance of the chloroplasts in the host defence responses with this likely representing alterations to protein synthesis occurring within the chloroplasts. These associations will be discussed further in Section 6.4.3.1.





6.4.2.2 Pathway mapping of metabolomic and transcriptomic analyses

Mapman was used to map both transcriptional and metabolic changes to primary and secondary metabolic pathways. This identified substantial differences between both the sampled time points and ecotypes (Figures 6.5, 6.6 & 6.7).

There was widespread down regulation of many genes found in cell wall, energy, lipid and secondary metabolism in Bd21 at 12 hpi. There were 5 clusters (Clusters 2,3,5,7, & 11) identified in Bd21 (Figure 5.11a) that showed a down-regulation at 12 hpi. These clusters contained a total of 2322 transcripts. However, enrichment analyses identified few over-represented function groups. Amongst these identified clusters there was variability in the trends of these transcripts at the later time points (24 and 48 hpi). Cell wall and lipid metabolism transcripts were over-represented in Cluster 7 as well as phenylpropanoid metabolism in Cluster 3, each being down regulated at 12 hpi.

These areas of metabolism form key areas for host defences and targets for the assimilation of nutrients by fungal pathogens. Subversion of these areas would be necessary for any successful pathogen. This extensive down-regulation could represent a subversion of host metabolism at a very early stage. Alternatively it could represent differential innate immune responses inherent in Bd21 compared to ABR6.

As discussed in Section 5.4.2.1, high discrimination between control and infected treatments is likely to be the result of PAMP triggered responses to spore secretions, post germination. Similar molecular responses would be expected between the ecotypes if these were generic innate responses to the presence of a fungal pathogen. The extensive differences in responses shown in Figure 6.5 suggest the contrary to this. However, it is unclear whether the extensive differences in these responses are as a result of the underlying genetic variation between the ecotypes in how they initially respond to the presence of fungal pathogens. Alternatively, these extensive differences could be directly linked to the recognition of the presence *M. oryzae* and the subsequent resistance response that is elicited in

ABR6. The initiation of defences that lead to incompatibility with ABR6 could be elicited as early as 12 hpi, prior to primary host cell penetration by *M. oryzae*.

There were relatively few transcriptional and metabolic changes occurring in Bd21 in the subsequent time points (Figures 6.6 & 6.7). In contrast in the incompatible interaction ABR6, there is both up and down regulation of genes distributed throughout the overview of metabolic pathways at both time points.

In Bd21 at 48 hpi there is down regulation of many genes involved in the light dependant photosynthetic reactions as well as an up regulation of some genes involved in amino acid synthesis. These areas of metabolism will be further discussed in Sections 6.4.3.1 and 6.4.3.2

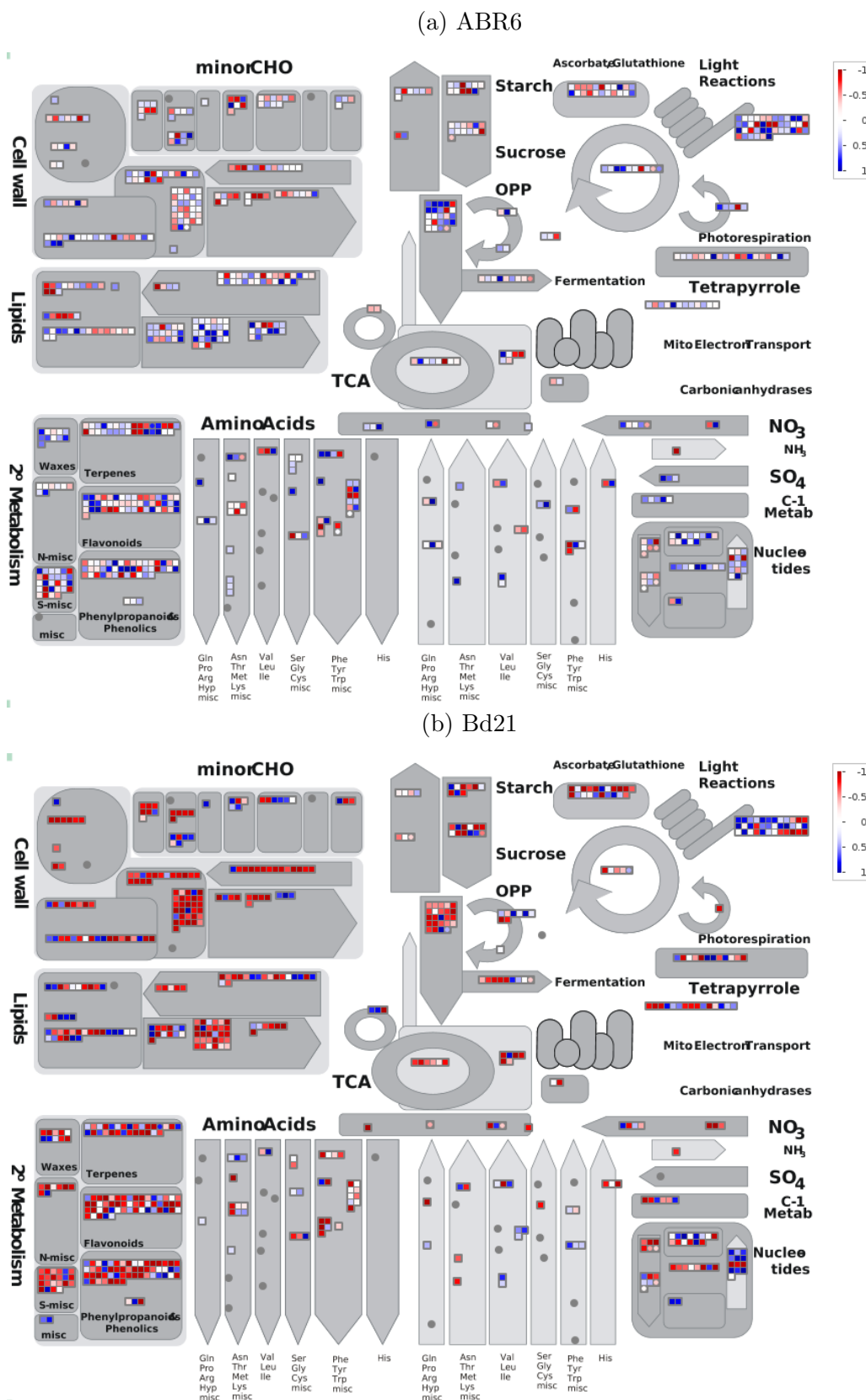


Figure 6.5: Mapman visualisations of metabolism for 12 hpi metabolite and gene expression changes. Trends are given as relative log₂ ratios between means of the treatment classes (N = 3). 820 and 754 data points are shown in (a) and (b) respectively of explanatory metabolites identified using FIE-HRMS or LC-HRMS and transcripts identified using RNA-Seq. Metabolites are represented as circles and transcripts represented as squares.

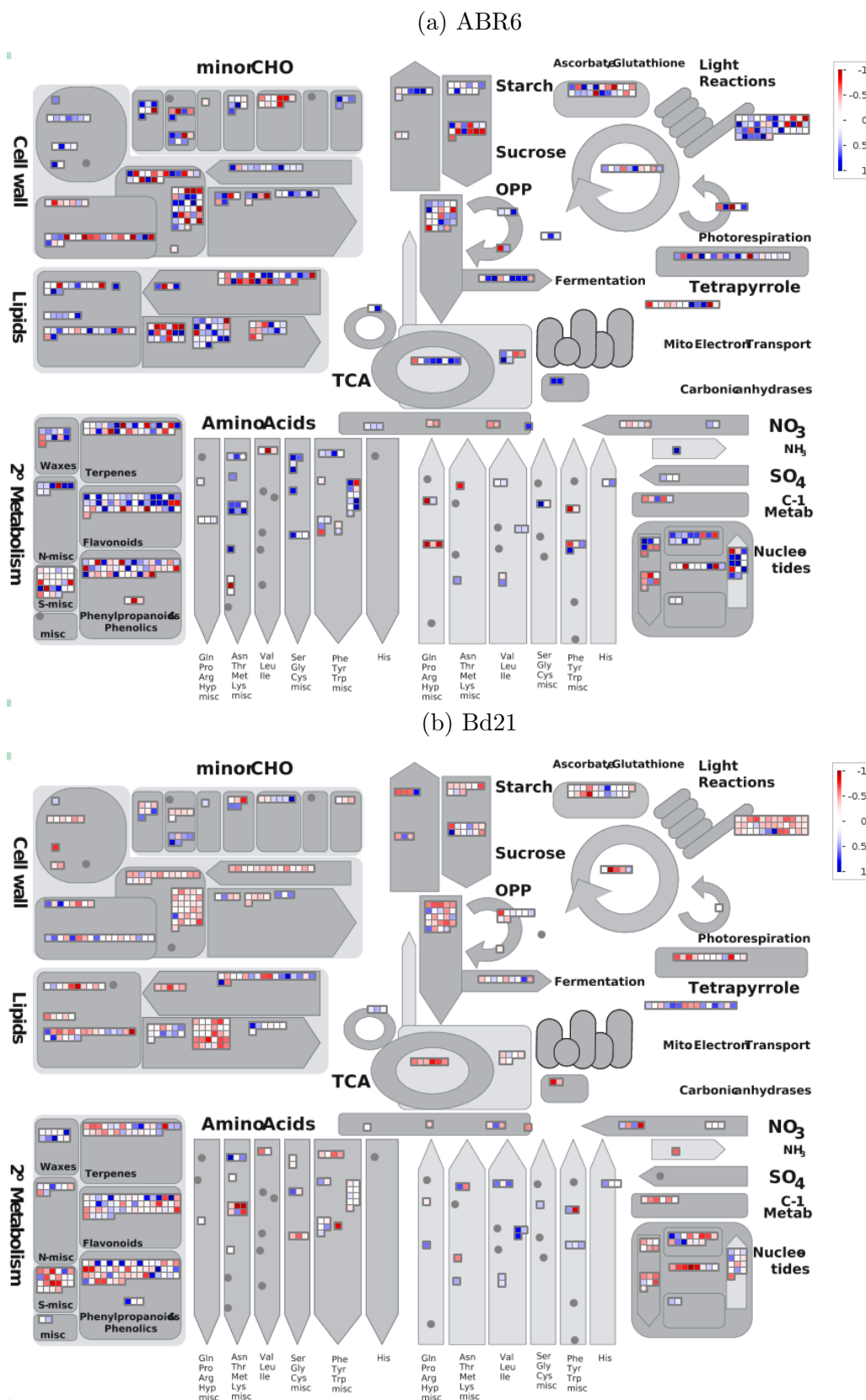


Figure 6.6: Mapman visualisations of metabolism for 24 hpi metabolite and gene expression changes. Trends are given as relative \log_2 ratios between means of the treatment classes ($N = 3$). 820 and 754 data points are shown in (a) and (b) respectively of explanatory metabolites identified using FIE-HRMS or LC-HRMS and transcripts identified using RNA-Seq. Metabolites are represented as circles and transcripts represented as squares.

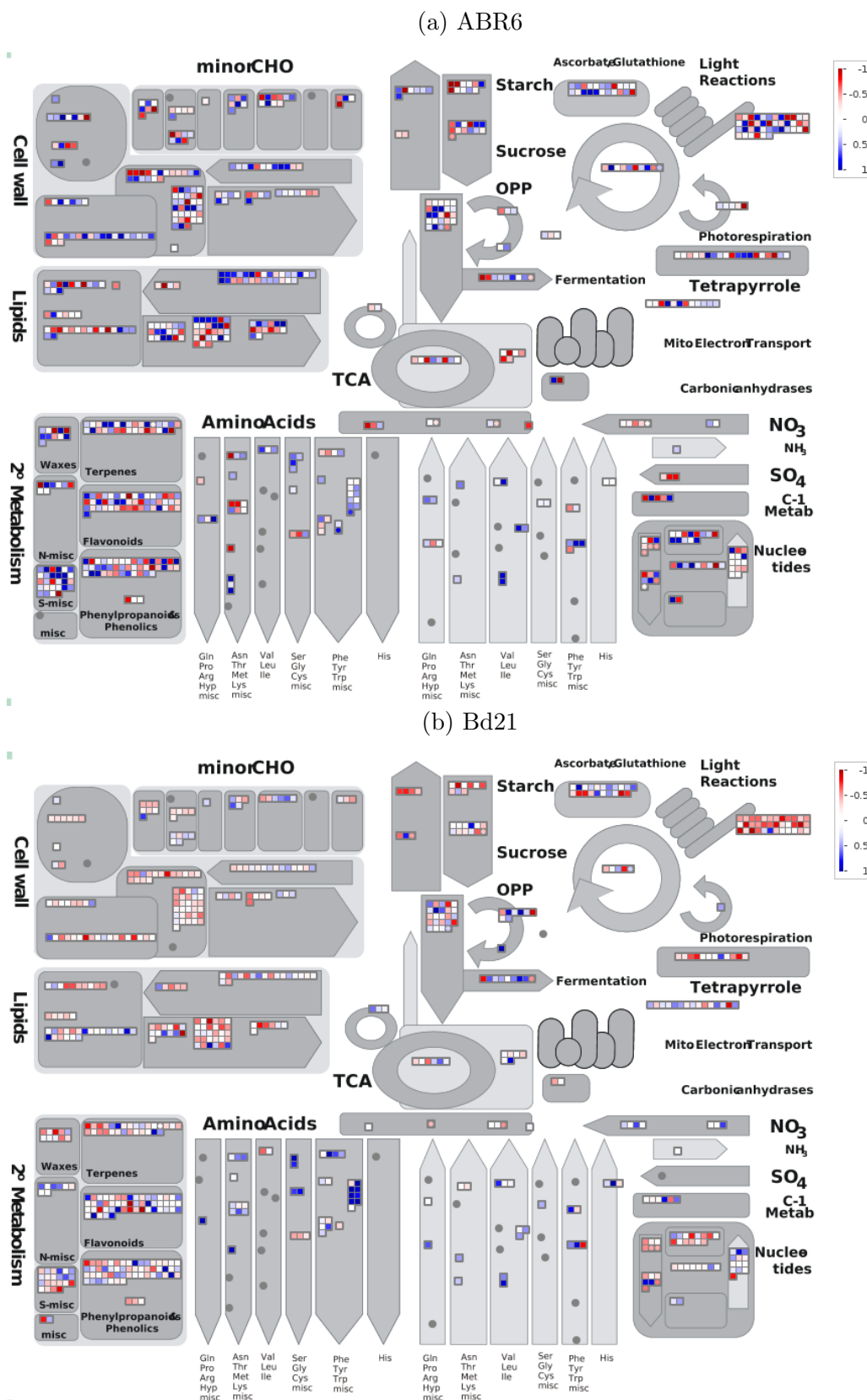


Figure 6.7: Mapman visualisations of metabolism for 48 hpi metabolite and gene expression changes. Trends are given as relative log₂ ratios between means of the treatment classes (N = 3). 820 and 754 data points are shown in (a) and (b) respectively of explanatory metabolites identified using FIE-HRMS or LC-HRMS and transcripts identified using RNA-Seq. Metabolites are represented as circles and transcripts represented as squares.

6.4.3 Key pathways in the early phases of the interaction between *M. oryzae* and *B. distachyon*

A number of areas of metabolism and signalling have already been identified in previous sections as be important during the pre-symptomatic phases of the *B. distachyon*-*M. oryzae* interaction. The previous analyses have focused on comparing the relative ratios of differences between infected and control treatments at each time point. This section will aim to highlight specific areas of metabolism and focus on the profiles of individual genes and metabolites in order to generate relevant hypotheses.

6.4.3.1 Chloroplasts are areas of metabolic and transcriptional change in *B. distachyon* during *M. oryzae* infection

Chloroplasts have long been recognised as centers of molecular alterations occurring during plant pathogen interactions. Not only do they function as carbon assimilation sites but are also important in the generation of oxidative bursts essential to the hypersensitive response. They therefore represent important targets for pathogen effectors for both nutrient assimilation and the suppression of host defences.

Upon pathogen recognition, plant cells will reduce carbon assimilation and transition from source into sink metabolism (Kangasjärvi et al. 2012). This is thought to be the result of a number of reasons. The production of defence related compounds would take priority and reduce the capacity of photosynthetic assimilation. Also a reduction in photosynthesis could reduce the compartmental damage caused by oxidative stress within chloroplasts.

Interestingly, a lack of correlation between photosynthetic rates and transcriptional repression has been observed in whole leaf analyses of the interaction between *Arabidopsis* and *P. syringae* (Bonfig et al. 2006). Targeted analyses of just the infection sites for transcriptional analysis revealed that this is likely due to a dilution of the transcriptional response when it is averaged across the whole leaf. Correlation has also been found in the transcriptional response of

ribulose-bisphosphate carboxylase (RuBisCo) and chlorophyll a/b-binding genes (Swarbrick, Schulze-Lefert, and Scholes 2006).

The trend of phosphoglycerate levels in Figure 6.8a could reflect a reduction in photosynthetic activity were levels are attenuated in the infected tissue compared to the control tissue. Glycerate-3-phosphate is the product of the carboxylase reaction between ribulose-1,5-bisphosphate catalysed by RuBisCo. There was no transcriptional change found in genes encoding for the RuBisCo subunits; however the trend of phosphoglycerate shows a reduction in levels at 12 and 24 hpi in both ecotypes. The trend is continued at 48 hpi in the incompatible interaction of ABR6 but in the compatible interaction of Bd21 levels have returned to that seen in the control. This return in Bd21 could reflect subversion of host photosynthetic activity by *M. oryzae* in order to increase carbon assimilation for nutrient acquisition. The trends of the genes shown in Figure 6.8b all show a down-regulation in ABR6 with few differences in Bd21 except for Ubiquinol oxidase. This trend could also be associated with the down regulation of photosynthesis during the incompatible interaction with *M. oryzae*.

Chloroplasts as the generators of ROS are important in the elicitation of the hypersensitive response. ROS can be formed from excess excitation energy in photosystems I and II, depending on the light conditions. This is quickly scavenged by both the stromal and thylakoidal scavenging systems (Kangasjärvi et al. 2012). The generation of ROS can also be achieved from NADPH oxidases, independent of light conditions. However light-driven generation from chloroplasts provides a less metabolically costly form of ROS generation. ROS production can increase under high light conditions and it has been shown that this can enhance the hypersensitive response in response to plant pathogens (Liu et al. 2007). Investigations by Parker (2006) revealed the *B. distachyon* ecotype ABR5, that exhibits an incompatible response to *M. oryzae*, could be induced to produce a compatible interaction if incubated under dark conditions for 56 hours post inoculation. This suggests the likely importance of light dependant ROS production in the *B. distachyon*-*M. oryzae* interaction. The reduction of Ubiquinol

oxidase expression observed in both ecotypes could be considered as a response to allowing an increase in ROS within the chloroplast.

The production of polyunsaturated fatty acids from chloroplastic membranes has been found to be important for the production of ROS in the *Arabidopsis* and *P. syringae* (Yaeno, Matsuda, and Iba 2004). Polyunsaturated fatty acids are produced by fatty acid desaturases and are able to induce NADPH oxidase activity to produce ROS. Trienoic unsaturated fatty acids of chloroplastic origin such as hexadecatrienoic acid and linolenic acid have been found to increase in *Arabidopsis* within just a few hours of exposure to avirulent strains of *Pseudomonas syringae*. Mutants showed that they were also essential for effective reactive oxygen species generation (Yaeno, Matsuda, and Iba 2004). Oxylipins can be produced by the addition of oxygen to polyunsaturated fatty acids. These are known to be potent signalling molecules in plant responses to disease. In rice, enhanced resistance to *Magnaporthe grisea* was found with the suppression of ω -3 fatty acid desaturases and a deficiency of 18:3 derived oxylipins (Yara et al. 2008). They have also been found to be involved in the activation of salicylic acid signalling and systemic responses (Ongena et al. 2004).

Lipoxygenase genes were found to show the same trends in both ecotypes with up regulation in control tissue at 12 hpi (Figure 6.9a). This was also true for the NADPH oxidase genes (Figure 6.9b). The ω -6-fatty acid desaturase gene was decreased in infected tissue at all the time points in both the compatible and incompatible interactions (Figure 6.9a). There was little relationship found between the trends of the fatty acids levels and the genes involved in their production as well as NADPH oxidases (Figure 6.9c). This likely reflects the regulation of NADPH oxidases and the metabolism of these fatty acids for plant defence at the enzyme level rather than that of the transcriptional level.

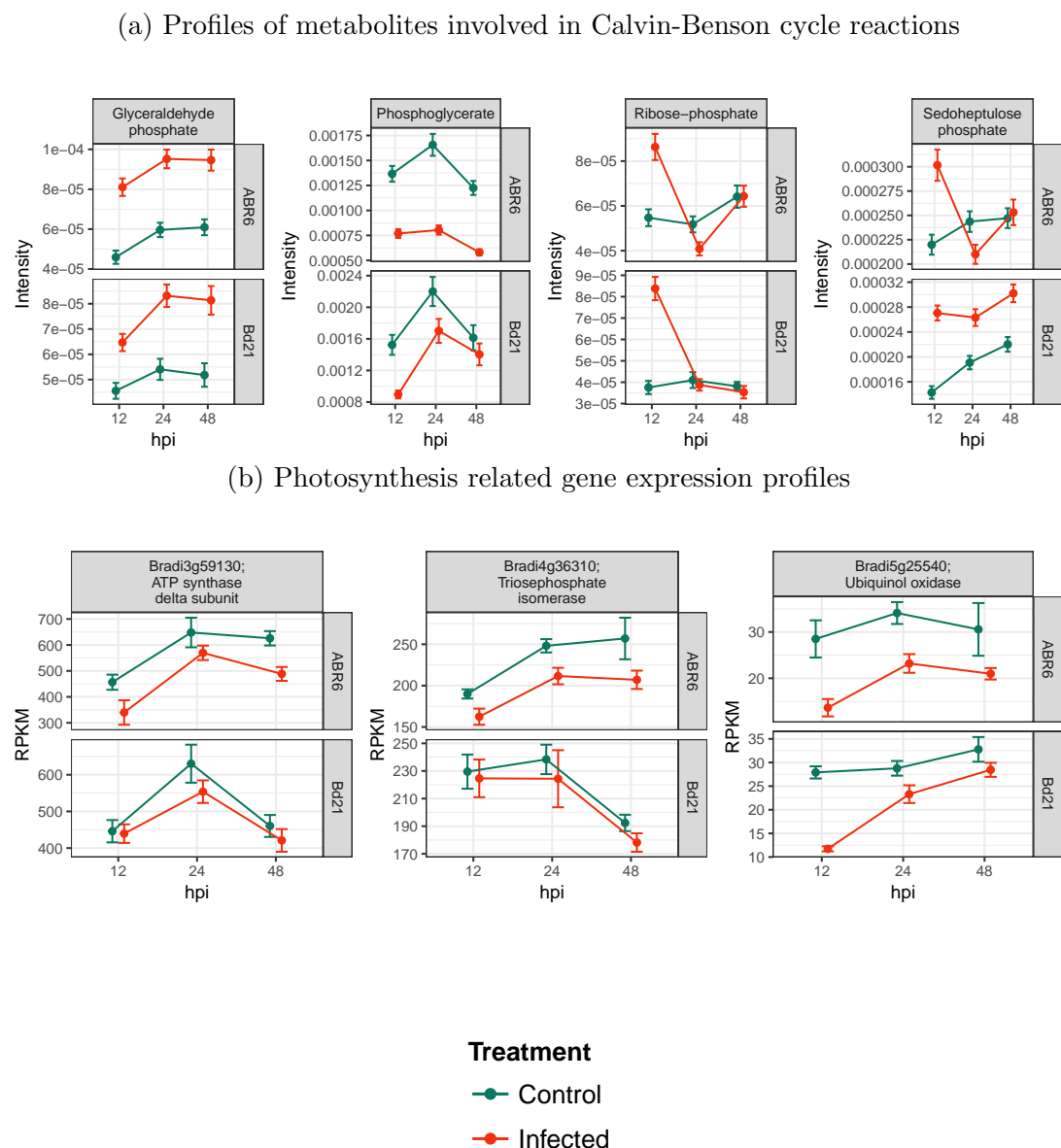


Figure 6.8: **Profiles of explanatory photosynthesis related metabolites and transcripts.** Points show treatment means ($N = 3$ for transcripts and $N = 10$ for metabolites) and error bars are one standard error of the mean. Metabolites profiles were measured using either FIE-HRMS or LC-HRMS and transcript profiles using RNA-Seq.

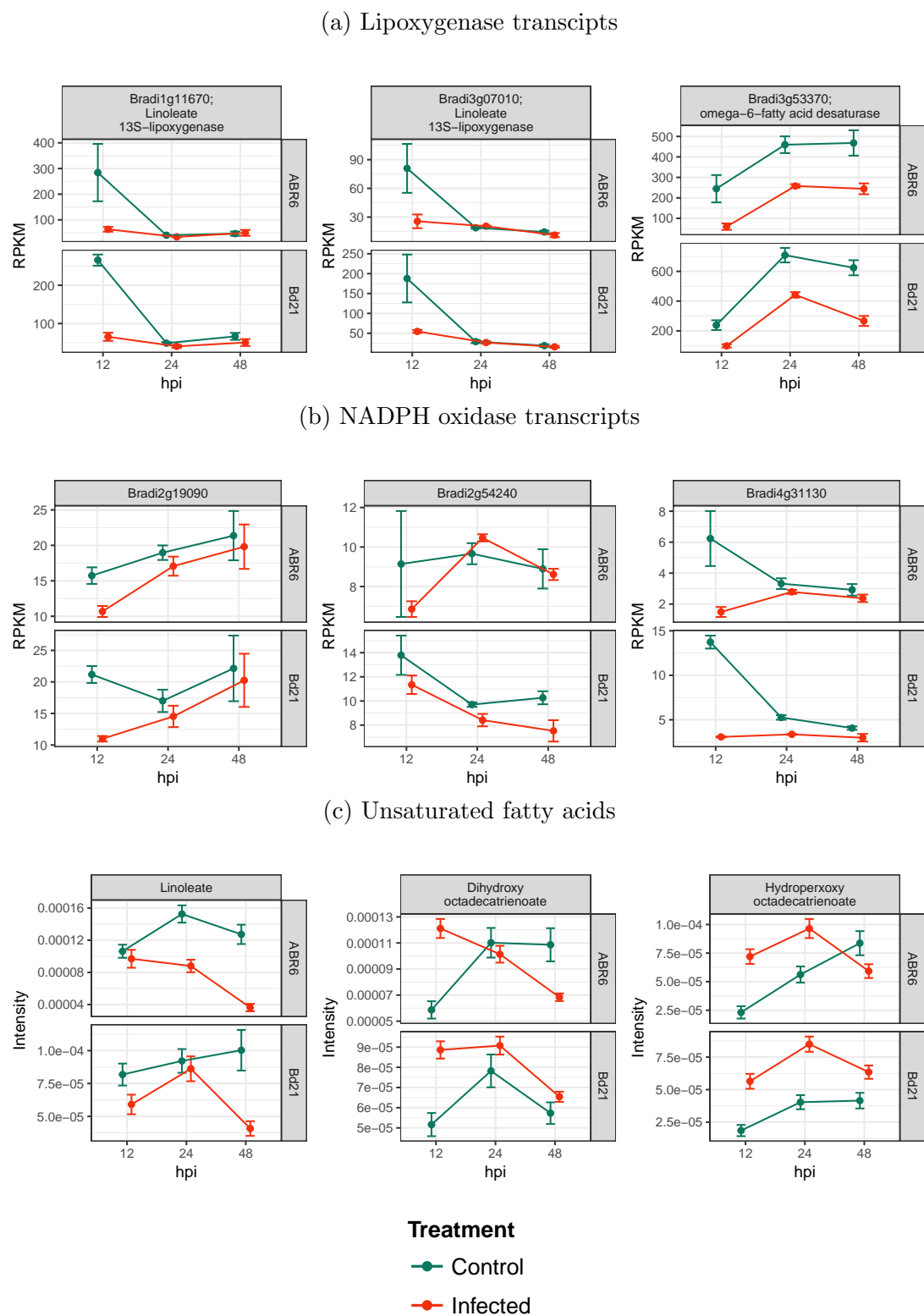


Figure 6.9: **Profiles of explanatory lipoxygenases, NADPH oxidases and unsaturated fatty acids.** Points show treatment means (N = 3 for transcripts and N = 10 for metabolites) and error bars are one standard error of the mean. Metabolites profiles were measured using either FIE-HRMS or LC-HRMS and transcript profiles using RNA-Seq.

6.4.3.2 Nitrogen metabolism is altered in both compatible and incompatible responses to *M. oryzae* infection

The amino acids glutamate, glutamine, asparagine and aspartate are important nitrogen transport and storage molecules within plant tissues. The energy requirements of plant defence can lead to the shuttling of these amino acids into pathways for energy production (Bolton 2009). The GABA shunt is one such pathway which allows carbon from glutamate to be fed directly into the TCA cycle. The nitrogen status of plant tissue is highly linked to the success of disease development in many interactions. Nitrogen deficiency in leaf tissue can encourage disease development with cells less able mount an effective disease response (Solomon, Kar-chun, and Oliver 2003). Nitrogen is also involved in the production of reactive nitrogen species that, together with ROS, are able to trigger the hypersensitive response.

Plant tissues will mobilise nitrogen away from infection sites during pathogen infection. This is hypothesised to be an attempt to deprive the pathogen of vital nutrients for growth (Tavernier et al. 2007). Asparagine is the preferred amino acid for nitrogen transport in plants due to its high nitrogen to carbon ratio. Disease resistance to bacterial and oomycete pathogens is enhanced by the conversion of aspartate to asparagine. The induction of asparagine synthetase 1 was associated with the infection of pepper by *Xanthomonas campestris*. Induction was also found to be induced by salicylic acid and jasmonic acid treatment (Hwang, An, and Hwang 2011).

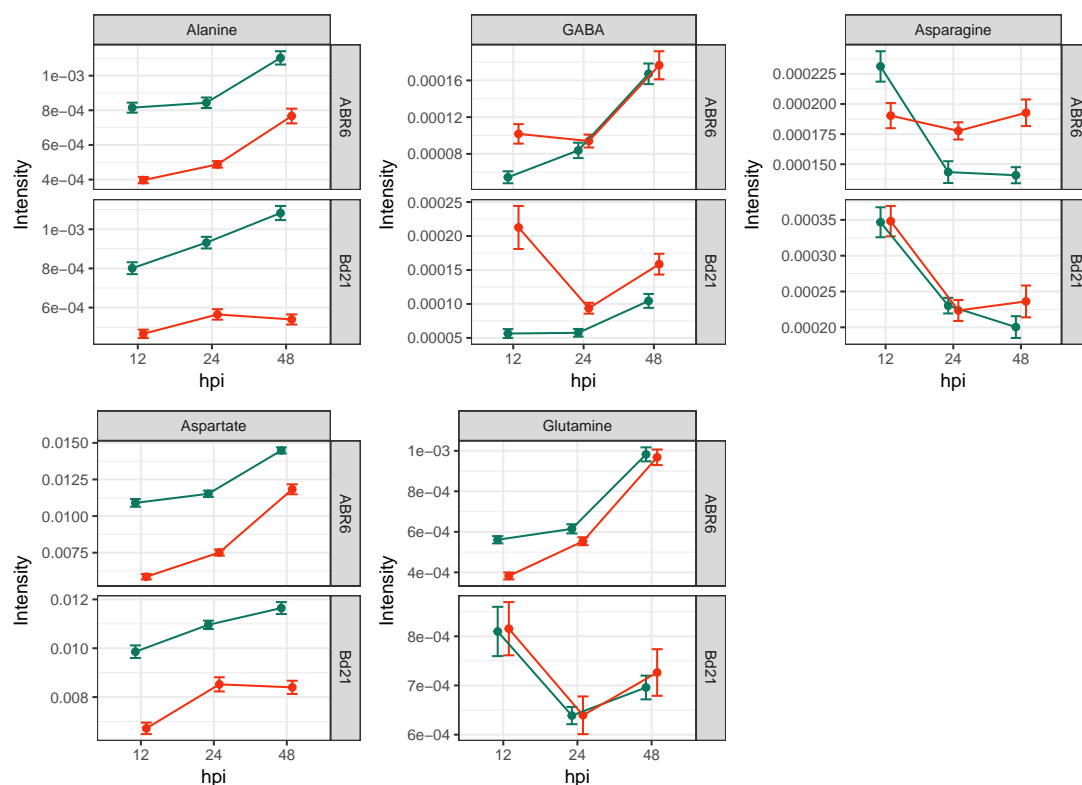
In the interaction between *B. distachyon* and *M. oryzae* an increase in asparagine was found at both 24 and 48 hpi in infected tissue in ABR6. Little difference was found in Bd21 (Figure 6.10a). Interestingly, Asparagine synthetase 1 was found to be down-regulated in infected tissue at 12 hpi with little difference at later time points. Glutamine levels were also found to be reduced in infected tissue of ABR6 at 12 hpi. It is the amide group of glutamine that is donated during the formation of asparagine from aspartate by asparagine synthetase, suggesting that these trends could be linked.

Both alanine and aspartate levels were found to be reduced at all time points (Figure 6.10a). Alanine transaminase catalyses the transamination between alanine and α -ketoglutarate to produce pyruvate and glutamate. It was found to be up regulated at 12 hpi suggesting its responsiveness during pathogen recognition phases.

GABA was found to be reduced in infected tissue in both ecotypes at 12 hpi. However, the GABA transaminase gene expression was found to be increased in control tissue at this time point. This suggests that the changes in GABA level are as a result of a decrease in control tissue rather than an increase in infected tissue.

The results suggest that the alteration of nitrogen metabolism is important during the initial pathogen recognition phases of this interaction with a number of other alterations such as that of asparagine occurring at later phases. However, there is also discordance between gene transcriptional changes and metabolite changes associated with the levels of aspartate and alanine. Nitrogen metabolism is heavily involved in the spatial aspects of plant disease responses, plant cells responding differently depending on their proximity to an infection site. Therefore the lack of spatial resolution in this data could be masking key differential responses of cells in terms of their nitrogen metabolism.

(a) Amino acids



(b) Nitrogen metabolism transcripts

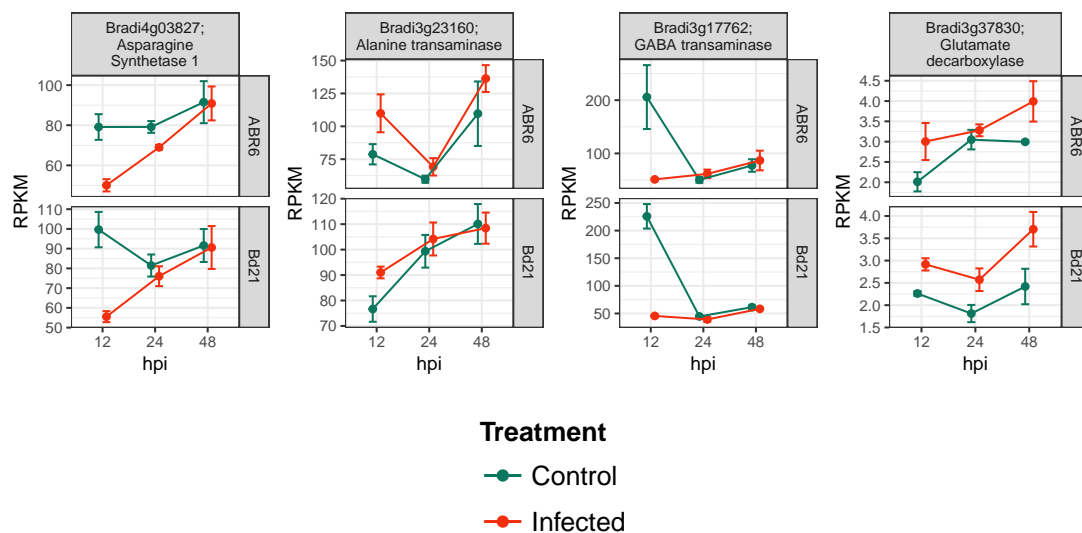


Figure 6.10: **Profiles of explanatory nitrogen metabolism transcripts and metabolites.** Points show treatment means ($N = 3$ for transcripts and $N = 10$ for metabolites) and error bars are one standard error of the mean. Metabolites profiles were measured using either FIE-HRMS or LC-HRMS and transcript profiles using RNA-Seq.

6.5 Concluding Remarks

With few previous examples of omics integration in the investigation of plant-pathogen interactions, the results presented in this chapter represent the first example of the integration of large scale transcriptomic and metabolomic analyses in the pre-symptomatic phases of the *B. distachyon* and *M. oryzae* interaction. Data integration revealed a strong correlation of explanatory features identified in both metabolomic analyses confirmed analytical agreement between the analyses. Network analyses of these data revealed the association of metabolically related metabolites across the metabolomic techniques such as fatty acids, purines and pyrimidines. Energy metabolism was hypothesised as being altered in response to *M. oryzae* infection in both compatible and incompatible interactions. However, it was unclear as to which of the organisms in the interaction to which these changes could be attributed, as both have high energy requirements for responses during these phases.

The integration of data from both transcriptomic and metabolomic analyses identified key association between alterations in gene expression and metabolism during *M. oryzae* infection beyond that of their individual analysis. This included links between amino acid trends with changes in amino acid synthesis of which tryptophan synthesis was important. Alteration to genes involved in histone production were also associated with the trends of purines and pyrimidines.

Pathway mapping identified that there was widespread transcriptional and metabolic down-regulation in primary and secondary metabolism occurring in Bd21 at 12 hpi, with reduced changes occurring at subsequent time points compared to ABR6. With extensive transcriptional differences in the recognition responses at 12 hpi it may be that these are associated with the initiation of the incompatible responses of ABR6 rather than only that of general innate responses to the presence of a fungal pathogen.

Network analysis and pathway mapping indicated that the chloroplasts were important centers for responses to *M. oryzae* infection. Alterations to Calvin-Benson cycle metabolites and genes associated with photosystems I and II were

identified. It was hypothesised that these were related to alterations to photosynthetic activity and that this could be associated with light dependant generation of ROS production. Also changes to poly unsaturated fatty acids and oxylipins that are thought to be of chloroplastic origin were hypothesised to be related to light independent generation of ROS. There was discordance between transcriptional and metabolic changes that suggested these roles in response to *M. oryzae* are strongly influenced by regulation at the enzymatic level.

Nitrogen metabolism was also identified to be altered in both compatible and incompatible interactions. Changes in the levels of amino acids and genes involved in their synthesis were identified to be changing during pathogen recognition phases. It was hypothesised that these changes could be as the result of energy production needs as well as nitrogen transport. However due to the spatial nature of the role of this metabolism in plant defence responses it was difficult to interpret alterations only in their temporal context.

Chapter 7

Identification of the Rbr1 disease resistance locus using computer vision based phenotyping

7.1 Introduction

There has been considerable research effort expended to understand the genetic bases of plant resistance to disease. As plant diseases contribute substantially to annual global crop losses, this importantly stimulates the breeding of resistant genotypes. In model interactions, such as that between *B. distachyon* and *M. oryzae*, identifying the genes responsible for diversity in pathogen responses is essential in elucidating the molecular mechanisms involved. However, the ability to accurately identify resistant genotypes is dependant on precise and robust phenotyping of plant responses to disease. Computer vision based techniques are becoming the standard approaches for quantitatively assessing plant responses to disease.

7.1.1 Quantitative plant phenotyping using computer vision

The wide spread application of omics and NGS technologies has spurred the need for automated, high throughput, quantitative plant phenotyping solutions. Alongside this has also begun the development of the field of phenomics for which computer vision is becoming ever important. As manual methods of plant phenotyping are often time consuming, computer vision improves throughput and reduces the potential for operator error. Computer vision encompasses the use of computers in the acquisition, processing and analysis of digital images. A general outline of the elements of computer vision analyses are shown in Figure 7.1.

Methods for image acquisition are highly varied within computer vision and are dependant on the experimental question and the biological scale of interest. This can be from individual cells to field-wide phenotyping. Both standard and confocal microscopy have been applied to enable image acquisition at the cellular level for investigations such as monitoring meristem or hypocotyl growth in *Arabidopsis* (Sozzani et al. 2014). Digital cameras are most commonly used in image acquisition due to their low cost and easy availability. This allows images to be taken non-destructively at the whole plant or plant organ levels (Bock et al. 2010). The use of images taken from differing points of view can also be combined to produce 3D data sets (Eliceiri et al. 2012). It is not only sensors that rely upon the visible regions of the spectrum that have been applied to plant phenotyping situations. Both X-ray computed tomography and NMR imaging have been used for root phenotyping problems where easy access to the plant organ is difficult (Metzner et al. 2015). Computer vision can also be applied at the field scale where digital cameras or multispectral imaging equipment can be mounted to unmanned aerial vehicles allowing entire experimental field plots to be imaged aerially (Araus and Cairns 2014).

Irrespective of the method of image acquisition, acquired images require processing in order to extract measurements from features of interest upon which statistical analyses can be made. Image segmentation methods allow the isola-

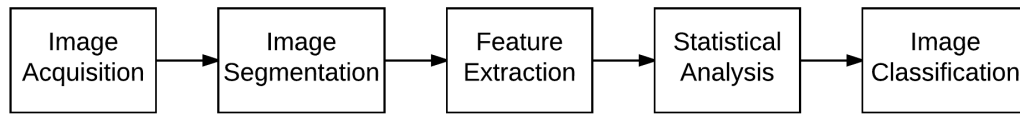


Figure 7.1: **A general workflow in computer vision analyses.**

tion of features of interest. There are many different methods for doing so, the most common of which is simple thresholding where pixel intensities above or below a set threshold are excluded. This can be further enhanced by the use of colour space transformations such as from RGB to HSV that allow colour thresholding as opposed to the thresholding of individual grey scale channels. Edge detection allows segmentation through the detection of sharp changes in pixel intensity around features of interest. Commonly 1st and 2nd derivatives are used for detection. However these techniques are not trivial and are often computationally intensive (Chen and Leung 2004). Clustering algorithms such as *K*-means clustering can also be used for segmentation however they require that there is a predefined set of features of interest (Zhang, Fritts, and Goldman 2008).

Once features of interest have been isolated, quantitative measures can then be extracted. These measures will be dependant on the experimental question but can commonly include area, colour and texture measures. Quantitative measures of image features can then allow classification or statistical analyses on either a whole image or individual feature basis.

With respect to the use of computer vision for characterising plant disease, applications have included disease detection, classification of different disease symptoms as well as disease symptom quantification (Barbedo 2013). Phadikar and Sil (2008) used neural networks for identifying disease in rice that included rice blast and brown spot. Peressotti et al. (2011) applied a semi-automated computer vision workflow for quantifying the sporulation of downy mildew on grape for resistance phenotyping.

7.1.2 The genetic basis of plant resistance to disease

The plant immune system consists of a host of defence responses that can be used to repel invading pathogens. These include both innate immune responses that allow plants to resist non-specialized pathogens and specific responses for specialized pathogens that have more sophisticated strategies in host colonisation. The resistance of plants to disease causing pathogens can be divided into three main categories; non-host, vertical and horizontal resistance.

Non-host resistance refers to a plants ability to resist pathogens for which it is not a host. This can be due to physical barriers such as thick cuticular layers on a leaf's surface that prevent pathogen infection. It may also be that the a plant does not provide a suitable environment on which the pathogen can grow and extract the necessary nutrition it needs in order to reproduce or simply that the pathogen cannot recognise the plant and initiate its colonisation.

Vertical resistance in its most basic form consists of individual gene-for-gene interactions. A gene product in the plant can recognise a gene product from the pathogen during the interaction which initiates defence responses leading to host resistance. These pathogen gene products are known as effectors and are secreted with the purpose of host defence suppression and subversion of cellular processes for nutrient acquisition (Jones and Dangl 2006). Plant resistance gene products are known as R proteins. Cell surface pattern-recognition receptors monitor the extra cellular environment and are activated by recognising highly conserved PAMPs. These are often leucine rich repeat or lysine motif kinases and contribute mainly to innate immune responses. Intracellularly plants detect pathogen colonisation using NB-LRR receptors. These constitute the majority of R proteins present within plants (Eitas and Dangl 2010).

The evolutionary arms race that inevitably occurs between a pathogen and its host has led to a diversity in the molecular interactions that occur between effectors and R genes, beyond that of the gene-for-gene concept. There are three main mechanisms of interaction; direct, guard and bait (Dodds and Rathjen 2010). In direct recognition, the effector binds directly to the NB-LRR receptor.

Guard recognition requires an accessory protein that could either be the effector target protein or a mimic, is modified by the effector which is then sensed by the NB-LRR. The bait mechanism also requires an accessory protein which interacts with the effector and is directly sensed by the NB-LRR. In other cases, several NB-LRRs are required to mediate defence responses (Eitas and Dangl 2010). One of the NB-LRRs will act as a sensor and is activated by the effector, the other as a helper that is required for function.

NB-LRR receptors are multi-domain structures that allow them to act as sensors, switches and response factors. They consist of two primary domains; the nucleotide-binding site (NBS) and the carboxy-terminal LRRs. This class of receptors can be divided based on the type of amino-terminal domain. These are the toll interleukin 1 and coiled-coil domain containing NB-LRRs. Those containing toll interleukin 1 regions are rarely found in grasses (Tan and Wu 2012).

Many R genes have been putatively identified in plant genomes. In rice, 85 complete rice blast resistance genes have been identified of which 80% are NB-LRRs (Ballini et al. 2008). There is considerable diversity in NBS disease resistance genes between maize, sorghum, *Brachypodium* and rice with only 3.83% of 496 ancestral NBS families showing conservation between species (Li et al. 2010a). This reflects the highly specialised nature of these genes and a strong effect of natural selection in ensuring that a species can adapt to evolving pathogen strategies. In *Brachypodium*, 239 nucleotide-binding site disease resistance genes have been putatively identified (Tan and Wu 2012).

In interactions involving horizontal resistance, a spectrum of responses from fully compatible to fully incompatible can be observed. This is often referred to as partial or quantitative resistance. It is the result of genetic diversity across many gene loci within both the plant and pathogen that regulate many layers of plant defence and pathogen colonisation strategies (Poland et al. 2009).

Genes of many different functions have been related to horizontal resistance and can be related to both constitutive and induced resistance responses. In rice,

it has been found that increased expression of constitutively expressed defence genes can increase resistance to *M. oryzae* (Schaffrath et al. 2000). It is hypothesised that many genes involved in horizontal resistance are involved in defence responses such as cell wall thickening, production of cytotoxins and the hypersensitive response. Oxalate oxidase-like proteins that are likely to be involved in the production of oxidative bursts have been identified in rice as enhancers of quantitative resistance (Walz et al. 2008). Influences of environmental factors such as temperature, humidity and host nutrient status can also greatly affect the outcome of quantitative interactions. Horizontal resistance is much less well understood than vertical resistance due to its greater complexity and the number of interacting gene loci.

Crop breeding programs strive for durable disease resistance in the field. Single R gene based resistance has had limited success due to pathogen adaptation within 2-3 seasons. One exception to this is the *mlo* gene for powdery mildew resistance in barley that has remained durable for more than 40 years, still being widely used in across Europe (Brown 2015). Breeding models for durable disease resistance rely on using multiple R genes as well as genes that enhance a number of different aspects of the disease response. This greatly reduces the potential for a pathogen to adapt and overcome resistant varieties.

7.1.3 Linking phenotype to genotype: QTL mapping to identify disease resistance loci

A QTL is a gene locus or region that is responsible for a particular quantitative phenotype of interest. These can encompass any number of linked genes within the same QTL region or multiple QTLs linked to the same phenotypic trait. QTL mapping encompasses the suite of techniques used to associate quantitative phenotypes to genetic loci. These include the generation of mapping populations, the use of genetic markers to construct linkage maps and QTL analysis for linking markers to phenotype.

In order to link a phenotype to a genotype, firstly a population of individuals

is required that is showing diversity or segregating for the trait of interest such as a disease response (Figure 7.2). In self-pollinating species such as *B. distachyon*, the generation of mapping populations is simple compared to cross pollinating species. These can be crosses between inbred parental lines from which the F₁ hybrids are further inbred to construct recombinant inbred lines (RIL).

With the generation of a population suitable for QTL mapping, genetic markers or polymorphisms need to be identified to map the genetic differences between the individuals in the population, both physically and genetically. There are a number of techniques that can be used to identify genetic markers. These can be hybridization-based, polymerase chain reaction (PCR)-based or DNA sequence based. Markers can be visualised by gel electrophoresis, staining or the addition of radioactive or colourimetric probes (Collard et al. 2005). They can generally be classed as dominant or co-dominant. Co-dominant markers can discriminate between homozygotes and heterozygotes.

Genetic markers can be assembled into linkage maps that indicate their chromosomal positions and the relative genetic distance between markers. They are based on the principle that markers segregate via recombination during meiosis. Tightly linked markers are more likely to be transmitted together from parent to progeny. As there will be a mixture of parental and recombinant genotypes present in a segregating population, the frequency of recombinant genotypes allows the calculation of the genetic linkage or distance between markers (Collard et al. 2005). Mapping functions can be used for calculating genetic distance. Most commonly used are the Kosambi and Haldane functions, for which the units of genetic distance is given in centiMorgans (cM). Genetic distance is not necessarily related to physical distance, the relationship of which can vary across a chromosome due to the presence of recombination ‘hot spots’ (Paris, Haen, and Gill 2000).

The ABR6 x Bd21 mapping population consists of 155 RIL from a cross between parental *B. distachyon* ecotypes ABR6 and Bd21 (Bettgenhaeuser et al. 2016). The founding F₂ population was generated by self pollination of three

individuals confirmed as hybrids in the F_1 population, from the initial cross of the parental lines. These lines were advanced by single seed decent to the F_4 stage. The genetic map was constructed by selecting single nucleotide polymorphism (SNP) markers every 10 cM, based on the previously characterized Bd21 x Bd3-1 F_2 genetic map (Barbieri et al. 2012). This ensured an even distribution of markers for both physical and genetic distance. At the F_4 stage 115 lines were genotyped for a total of 252 markers. The linkage map is shown in Figure 7.3.

QTL analysis can be used to detect the presence of QTLs for a particular phenotypic trait within a mapped segregating population. There are three common techniques for detecting QTLs. Firstly, single-marker analyses use univariate statistical techniques such as analysis of variance (ANOVA) to test for the presence of associations between the phenotypic trait and genetic markers on a single marker basis (Tanksley 1993). Simple interval mapping (SIM) analyses the intervals between adjacent markers and is able to compensate for recombination events between markers. This is more statistically powerful than single-marker analyses (Lander and Botstein 1989). Composite interval mapping (CIM) combines SIM with a linear regression strategy, with additional genetic markers included in the model. This allows the detection of multiple linked loci for a given phenotypic trait. It is more sensitive for detecting QTLs than single-marker and SIM analyses (Kao, Zeng, and Teasdale 1999). SIM and CIM analyses produce profiles across each chromosome of logarithmic of odds (LOD) scores. Significance thresholds for these scores can be calculated by permutation testing (Churchill and Doerge 1994).

QTL analyses have been applied to many crop species for a very wide range of phenotypic traits from disease resistance, vernalisation requirements and yield. In *B. distachyon* three QTLs have been identified for resistance to false brome rust that were dynamic across developmental stages (Barbieri et al. 2012).

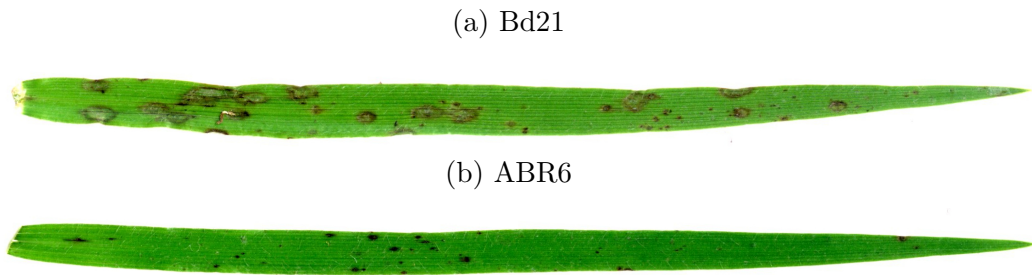


Figure 7.2: **Rice blast disease response phenotypes in the *B. distachyon* ecotypes Bd21 and ABR6 6 days post inoculation.** Bd21 shows a compatible interaction and ABR6 an incompatible interaction.

7.2 Aims

One of the main limitations in the omics analyses of this interaction in the preceding chapters is that the direct comparison of compatible and incompatible interaction responses are confounded by the high discrimination between the two ecotypes (Section 5.4.1). One way that this confounding factor could be overcome in the future is to identify the gene loci that are responsible of resistance in ABR6. Bd21 could then be genetically transformed so that it could elicit an incompatible response to *M. oryzae* pathogenesis. This would then allow compatible and incompatible interactions to be directly compared, without the confounding effects of the genetic diversity between ecotypes.

The central aim of this chapter is to identify gene loci that are potentially responsible for the differential phenotypes in ABR6 and Bd21 in response to *M. oryzae* pathogenesis. This requires the application of QTL mapping techniques using a population of RILs from a cross between ABR6 and Bd21. In turn this requires reliable and precise phenotyping of large numbers of individual plants. Quality and reproducibility in phenotyping can be achieved using computer vision based image analysis techniques over manual scoring. This chapter will also apply computer vision based image analysis to quantitatively assess *M. oryzae* response phenotypes. This provides us with the following aims for this chapter:

- Use computer vision based image analysis to quantitatively assess RIL response phenotypes to *M. oryzae* pathogenesis.

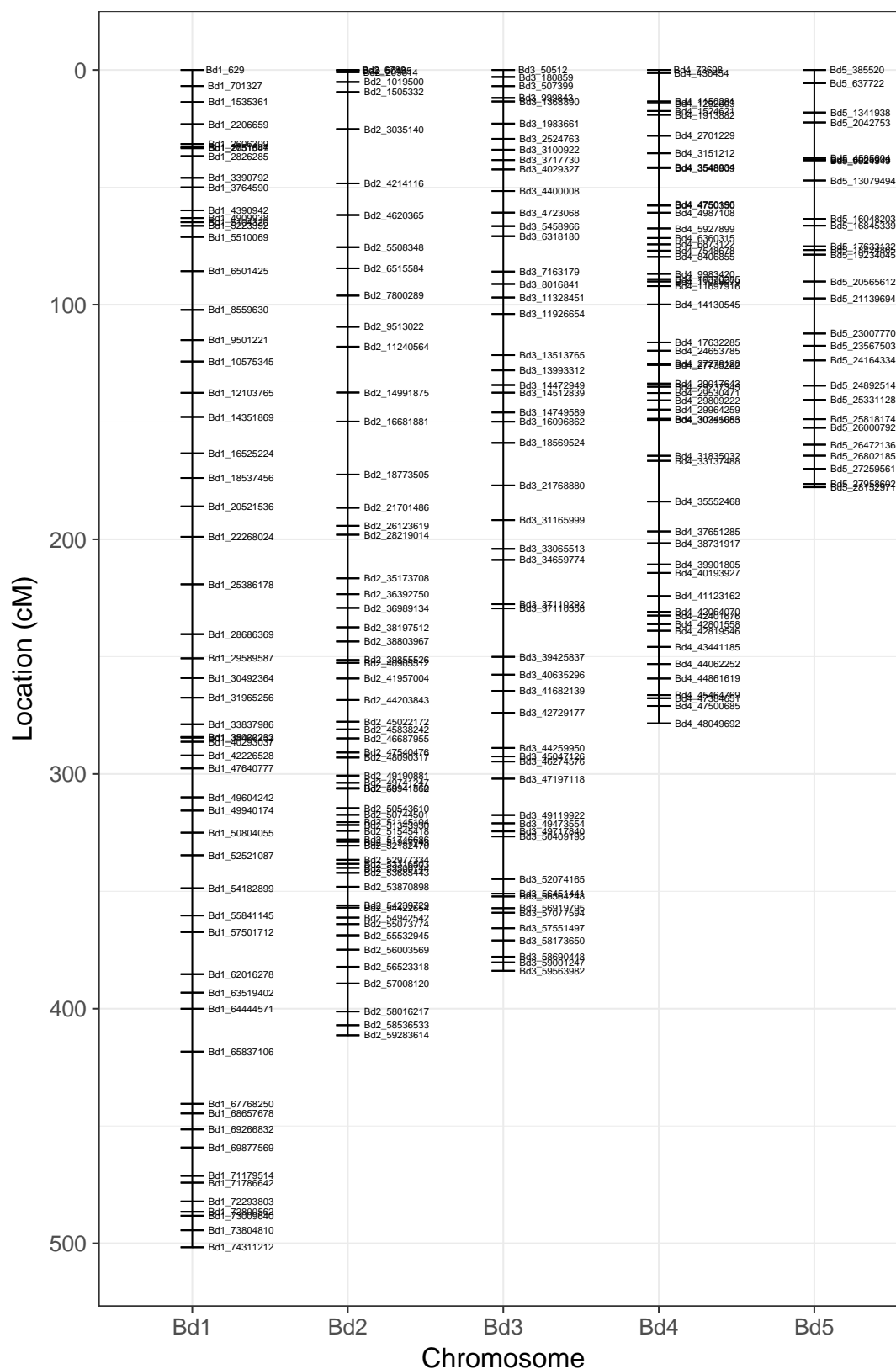


Figure 7.3: The ABR6 x Bd21 F₄ genetic map.

-
- Apply QTL analysis to identify gene loci responsible for ABR6 resistance to *M. oryzae* pathogenesis.
 - Identify potential candidate resistance genes within linked gene loci.

7.3 Materials and Methods

7.3.1 Inoculation of F_{4:5} ABR6 x Bd21 population with rice blast

115 lines and parental ABR6 and Bd21 ecotypes were grown as described in Section 2.1 with 5-6 plants of each line per module. The plants were inoculated at 21 days old as described in Section 2.3. Propagator lids were removed two days post inoculation.

7.3.2 Image acquisition and processing

At 6 days post inoculation the 3rd leaf from the base of each plant was detached and scanned using an Epson GT-12000 flat-bed scanner at 800 dots per inch. The acquired images were processed using a custom script in R (Appendix D). Details of the image segmentation and feature extraction of rice blast disease responses are described and explained in Section 7.4.1.






7.3.3 Manual scoring and QTL analyses

Manual scoring of rice blast disease response images were based on the observations of Routledge et al. (2004). The images were blindly scored three times and the scores averaged for each line. Scoring used a scale from 0-4 based on the extent of disease response scores; with 0 representing a highly resistant response and scores of 4 representing a highly susceptible response and scores of 2 being intermediate. Examples are shown in Table 7.1.

Analyses for QTL detection (simple interval mapping (SIM) and composite interval mapping (CIM)) were performed using QTL Cartographer (Version 1.17j;

<http://statgen.ncsu.edu/qtlcart/>). CIM was performed under an additive model ($H_0:H_1$) with the selection of five background markers at a walking speed of 2 cM and a window size of 10 cM. 1,000 permutations with reselection of background markers were performed for determining the statistical significance of QTLs. SIM was used for estimating the 2-LOD support intervals.

Table 7.1: Manual scoring of rice blast disease responses in the ABR6 x Bd21 RIL population.

Image	Score	Comment
	0	Few, very small highly localized necrotic flecks
	1	Numerous, small highly localized necrotic flecks, some larger black lesions present
	2	Intermediate to scores 0 and 4 containing both necrotic flecks and large spreading lesions
	3	Large spreading lesions brown to black in colour
	4	Large, coalescing lesions, brown to black in colour.

7.4 Results and Discussion

7.4.1 Computer vision based analysis for quantitatively assessing *B.distachyon* and *M. oryzae* interaction phenotypes

The effective development of a computer vision based algorithm for assessing plant-pathogen interaction phenotypes requires a logistically suitable and reproducible method of image acquisition. Image aspect, quality and scale are all important factors when deciding upon the most suitable method of image acquisition to ensure that measured features are comparable between images (Bock et al. 2010). It is also important to consider aspects related to the disease response such as whether plants need to be imaged multiple times as well as at which time points in the disease progression are most suitable.

Here it was decided that imaging the plants once, 6 days post inoculation would provide disease symptoms most optimal for image acquisition. This gave the best compromise for lesion size, without lesions coalescing, which would affect the accuracy of lesion size and density measurements. It was also decided that imaging detached leaves using a flat-bed scanner would provide a high degree of reproducibility in aspect and scale between images that would be difficult with whole plant imaging. The interest here was in accurate assessment of the disease response rather than how the disease developed over time, allowing destructive sampling to become a viable option.

7.4.1.1 Image segmentation of Rice Blast disease symptoms

Prior to extracting quantitative information from acquired images of disease phenotypes, the images needed to be partitioned to isolate ROI. In the case of rice blast disease symptoms, the ROI are the disease legions associated with both the compatible and incompatible responses (Figure 7.2).

The partitioning of rice blast disease related ROI was a two part process. Firstly, the leaf needs to be segmented from the image background, to remove

any unrelated noise that may have been acquired. Then the disease lesions can be segmented from the rest of the healthy tissue. At each stage of this segmentation, a binary black and white reference image can be produced that defines the boundaries of the ROI.

As can be seen in Figure 7.4, all the channels provide very clear differences between the leaf and background. The histogram of the blue channel in Figure 7.4e shows how clearly the background can be defined; the background having pixel intensities close to 100%, the leaf with values close to 0%. A threshold pixel intensity of 35% in the blue channel was found to suitably segment the leaf the image background.

Segmentation of the disease lesions from the healthy leaf tissue was found to be more difficult than segmenting the leaf from the image background. This was mainly due to the lighter colour found at the centre of compatible disease lesions (Figure 7.5a). As can be seen in Figures 7.5b & c, the lesion information can be seen in both the red and green channels, although the red channel is very subtle. The green channel was suitable for directly thresholding the incompatible lesions as there is a high pixel intensity difference between the lesions and surrounding healthy tissue (Figure 7.4c). However, thresholding of the compatible lesions directly on both the red and green channel was found to be ineffective. This was due to the low pixel intensity difference between the lesion information and the healthy leaf tissue. Colour space transformations such as HSV and LAV were also found to be ineffective at providing suitable intensity differences to segment entire compatible lesions. Subtracting the red channel from the green channel to produce a pseudo-image was found to be an effective method for combining the information held within these two channels, allowing the compatible lesions to be effectively segmented (Figure 7.5d).

It was also found that the colour of healthy leaf tissue varied across all the acquired images. The effect of this was to cause the relative frequencies of pixel intensities to fluctuate and so making direct thresholding of the subtracted image ineffective. Using the pixel intensity with the highest density on the image, it

was found that subtracting 20% from this value gave a suitable threshold for segmenting the incompatible lesions in the green channel (Figure 7.5e).

A similar strategy was used for thresholding the subtracted image for the compatible lesions. Instead of subtracting a value from the maximum density pixel intensity, 50% of this value was found to be a suitable threshold (Figure 7.5f). Using a dynamic thresholding approach for the lesion segmentation allows for the individual RIL to vary in colour without compromising the partitioning of disease symptom segmentation.

With the use of dual thresholding strategies for the compatible and incompatible lesion types that, while being separate, are also able to threshold lesions with some overlap, means that the entire spectrum of lesion types can be captured. Figure 7.6 shows the results of the segmentation techniques described here for a *B. distachyon* leaf showing a variety of lesion types. It can be seen that both leaf and lesion boundaries have been accurately partitioned across the entire spectrum of lesion types.

7.4.1.2 Extraction of disease related features

The segmentation of disease response related ROI, allows quantitative measurements to be extracted. In relation to rice blast disease these features include measures relating to lesion density, size, shape, colour and texture (Camargo and Smith 2009). Measurements are extracted on an individual lesion basis which can then be averaged to give the overall lesion measures for individual leaves.

The initial segmentation of the entire leaf allows the total leaf area to be measured and therefore the lesion density. Size measures are extracted as pixel counts but can be calibrated to ‘real world’ units by measuring the size of a single pixel. For the acquired images, a single pixel is equivalent to $8.94 \times 10^{-4} \text{ mm}^2$. The number of lesions segmented on an individual leaf can be divided by the leaf area to give the lesion density.

Lesion area and eccentricity give measures of lesion size and shape. Similar to leaf area, lesion area is also given in pixel counts that can be calibrated to mm^2 .

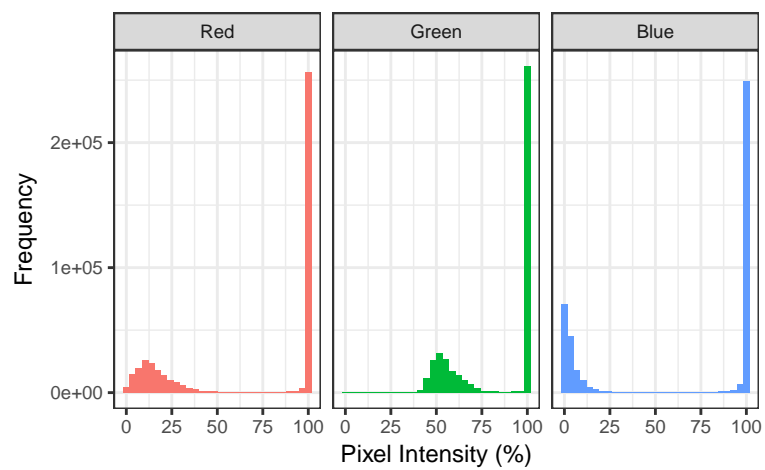
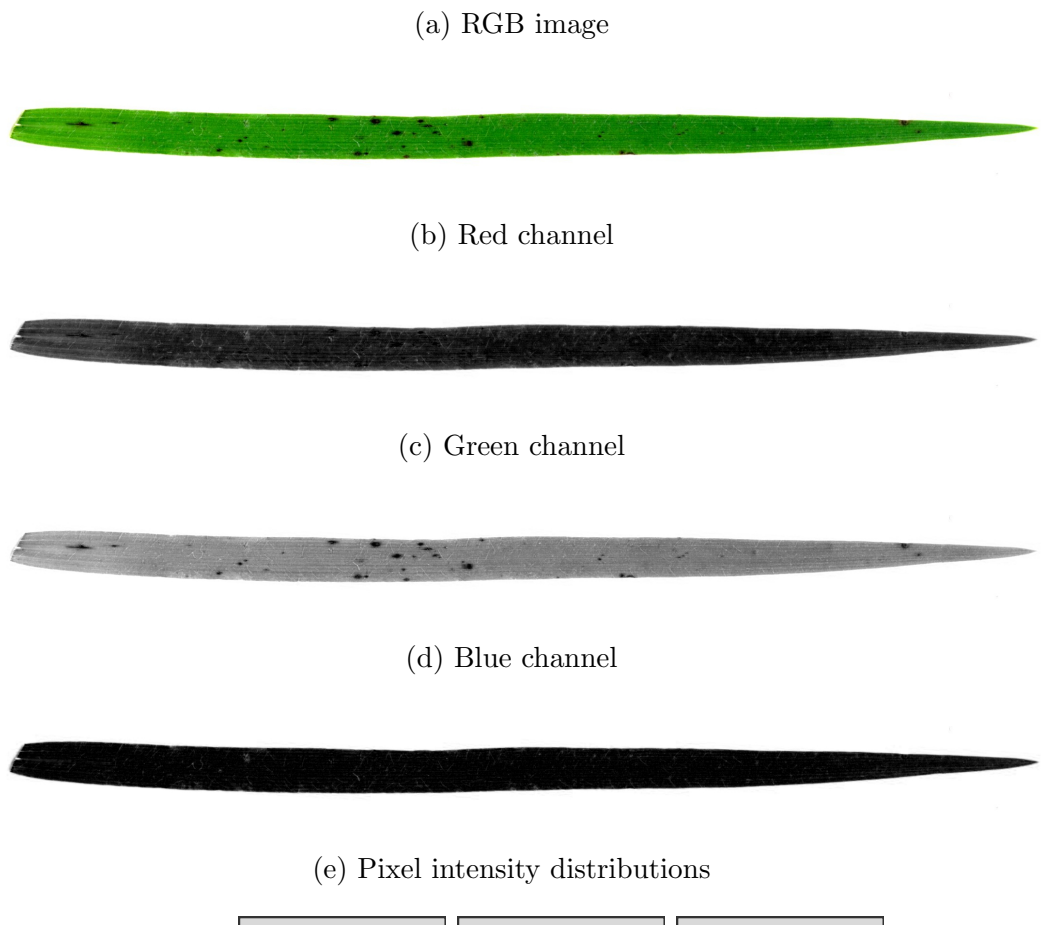


Figure 7.4: RGB image channels of an example ABR6 leaf showing incompatible disease symptoms

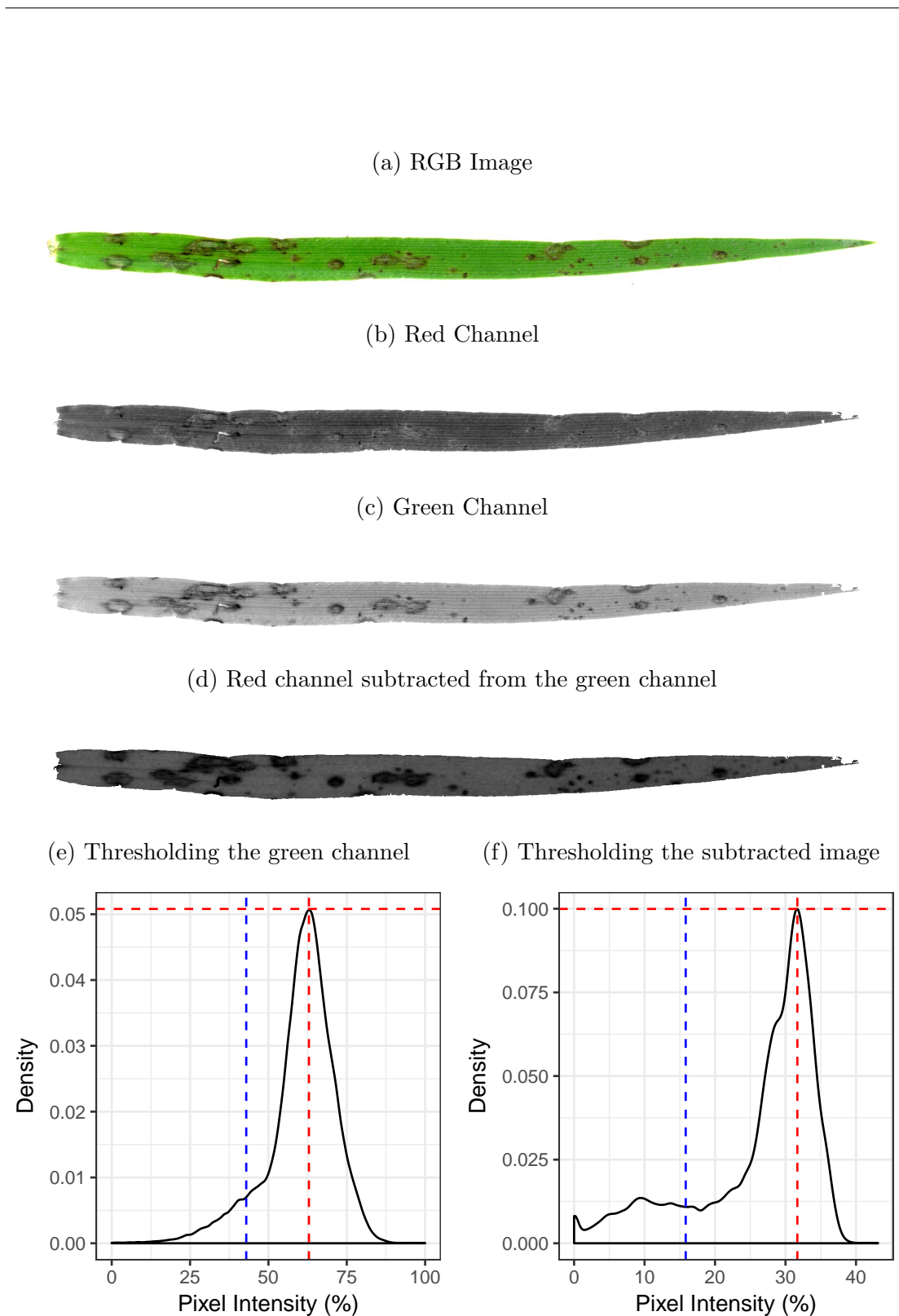


Figure 7.5: **Lesion segmentation using image subtraction and thresholding.** e & f are density plots of c & d respectively. The red lines indicate the pixel intensity with the highest density. The blue lines indicate the thresholding values used.

(a) Original Image



(b) Segmented leaf



(c) Segmented lesions



(d) Outlined segmented lesions



Figure 7.6: Segmentation results of Rice Blast disease response phenotypes.

Eccentricity measures how elliptical a lesion is with circles having a value of 0 and straight lines a value of 1.

Lesion colour measures can be calculated by averaging pixel intensities across the lesion for each RGB colour channel. Also the standard deviation can be calculated which can indicate the extent of variability in a lesions colour.

Features relating to image texture can be used describe the spatial arrangement of pixel intensities across a lesion. There are numerous measures that are calculated based on co-occurrence matrices. These features include angular second moment, contrast, correlation and entropy (Haralick, Shanmugan, and Dinstein 1973).

As can be seen from Figures 7.7, 7.8 and 7.9, most of the extracted features show little difference between the differential response phenotypes of Bd21 and ABR6. The only feature that is showing a significant difference between the ecotypes is lesion area. The compatible lesions of Bd21 are 5 times larger in area than the incompatible lesion of ABR6 with averages of 0.43 and 0.09 mm² respectively (Figure 7.7b).

As would be expected, lesion density did not show any difference between the ecotypes (Figure 7.7a). This more likely to be affected by differences caused during inoculation and the coalescence of lesions. It could potentially be a useful measure when developing inoculation techniques as a way of checking for uniformity in inoculation across a tray of plants. Interestingly, there was no appreciable differences in the eccentricity between compatible and incompatible lesions (Figure 7.7c).

There was no difference found between compatible and incompatible lesions in both their colour and texture features in both the red and green channels.

7.4.1.3 Validation of computer vision disease measurements

Important in the use of computer vision based image analysis for plant phenotyping is the validation of segmentation and feature extraction results relevant to the biological question. This is especially important for quantifying plant disease

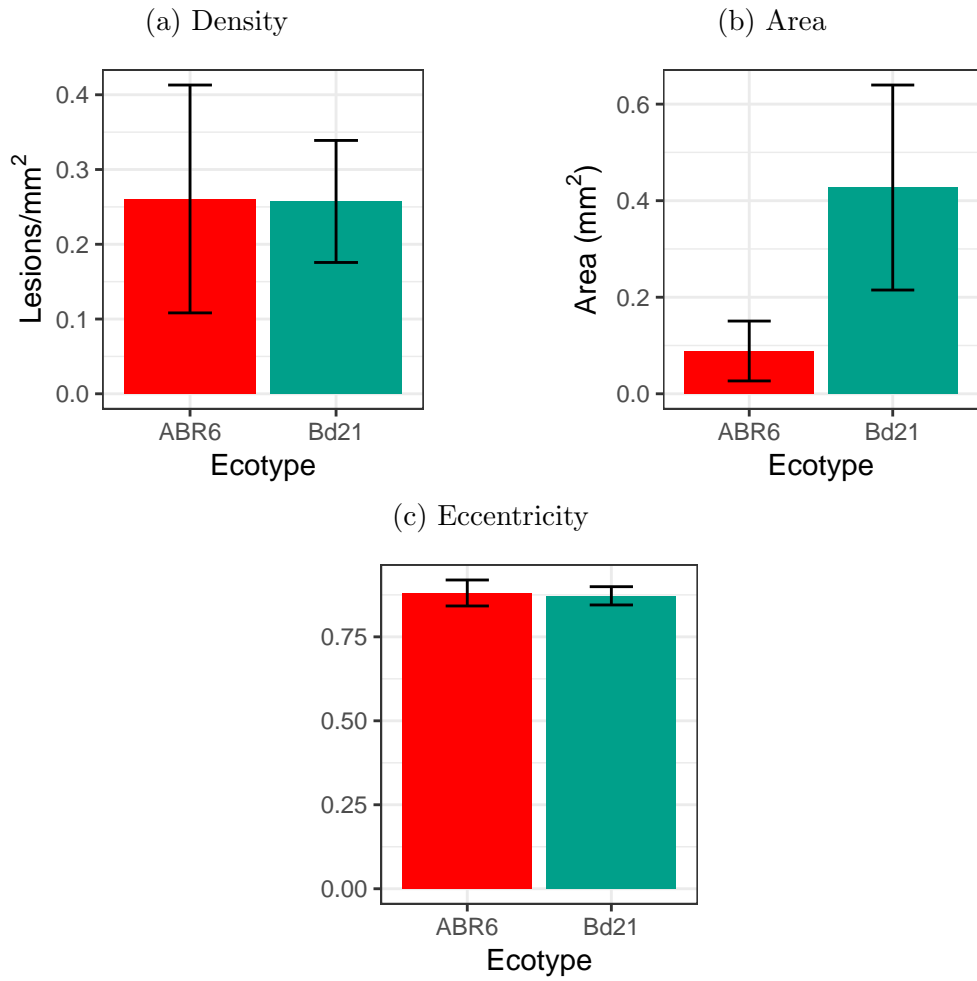


Figure 7.7: **Ecotype comparisons of density, size and shape features extracted from leaf images.** Elliptical eccentricity is calculated by $\sqrt{1 - \text{minoraxis}^2 / \text{majoraxis}^2}$ with a circle having a value of 1 and a straight line a value of 0. Error bars show 95% confidence intervals estimate using the t distribution.

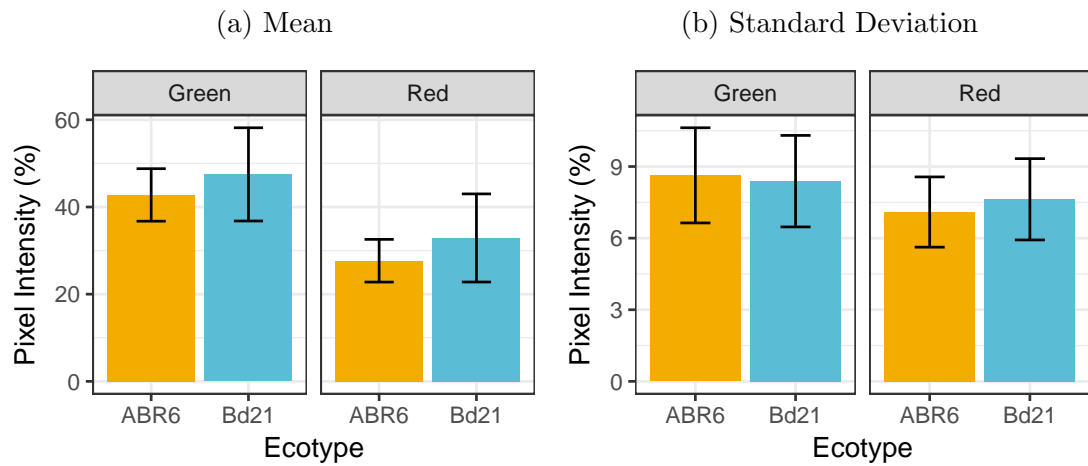


Figure 7.8: **Ecotype comparisons of colour features extracted from leaf images.** Individual graphs based on green and red RGB channels. Error bars show 95% confidence intervals estimate using the t distribution.

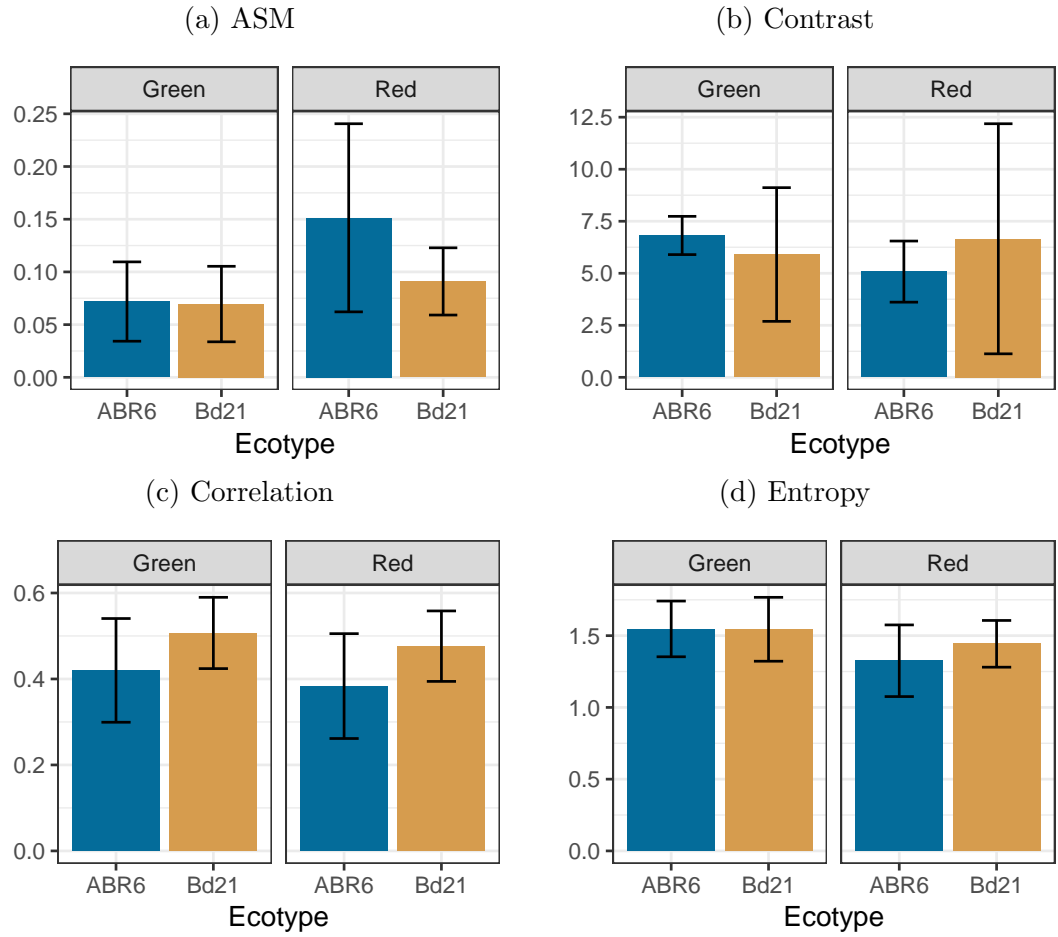


Figure 7.9: **Ecotype comparisons of texture based features extracted from leaf images.** Texture features based on those described in (ref). Individual graphs based on green and red RGB channels. Error bars show 95% confidence intervals estimate using the t distribution.

response phenotypes as these are complex phenotypes that involve a number of factors that contribute to their classification. The main goal of using computer vision for plant disease response quantification is that it can remove bias, subjectivity and inconsistency that scoring based phenotyping can introduce (Bock, Hotchkiss, and Wood 2016). Scoring based phenotyping approaches also require a certain level of expertise and experience in identifying the correct phenotype. This becomes even more complex when response phenotypes are quantitative in nature. However, response scoring has application in this context in ensuring the validity and relevance of image extracted measurements to the disease response phenotype.

In the previous section (Section 7.4.1.2) it was identified that lesion area showed a significant difference between ABR6 and Bd21. To confirm this association, lesion area was compared to response scores of individuals from the entire ABR6 x Bd21 mapping population (Figure 7.10). An exponential relationship was identified with an R^2 of 0.87 ($p < 0.001$) suggesting that lesion size contributes to a substantial proportion to the characterisation of rice blast disease response phenotypes.

7.4.2 The identification of linked loci for rice blast disease resistance *B. distachyon*

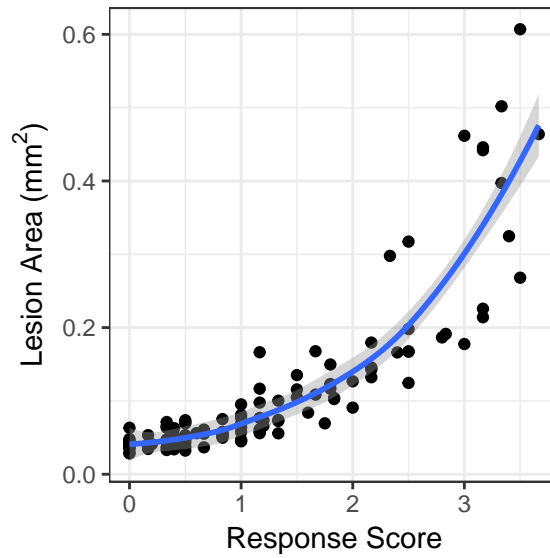
The extraction of lesion area measurements, suitable as phenotypic indicators of rice blast response between ABR6 and Bd21, allows their analysis for QTL linkage. This section focuses on QTL analyses for QTLs related to rice blast resistance and the identification of candidate NB-LRR resistance genes that may be present in these regions.

7.4.2.1 QTL mapping to identify the Rbr1 disease resistance locus

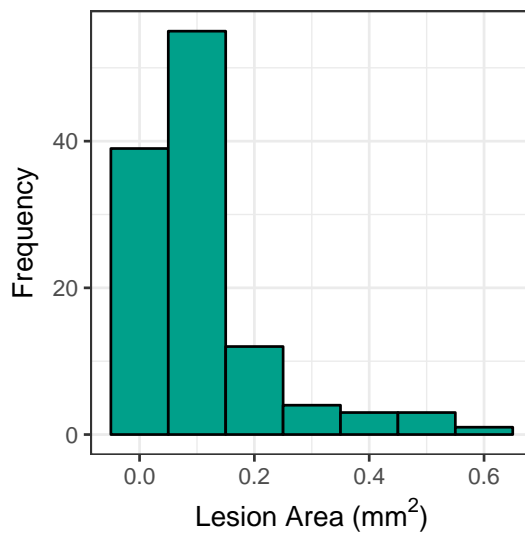
CIM analyses of average lesion area from the $F_{4:5}$ population against the F_4 linkage map identified two significantly linked markers on chromosome Bd4 (Figure 7.11a). This corresponds to markers Bd4_8406855 and Bd4_11697916. A third

(a) Lesion area versus response score

```
## 'geom_smooth()' using method = 'loess'
```



(b) Lesion area distribution



(c) Response score distribution

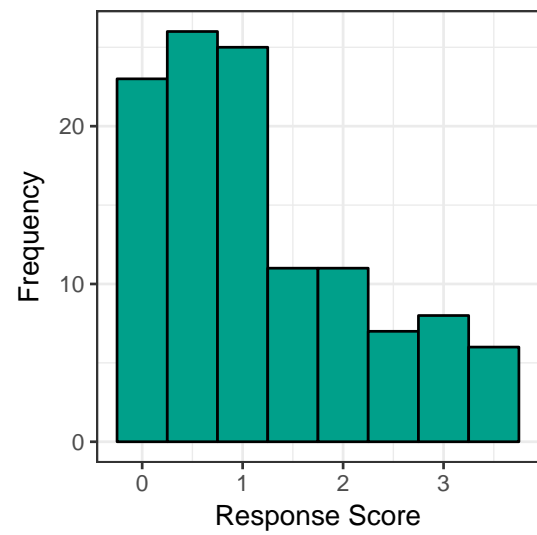


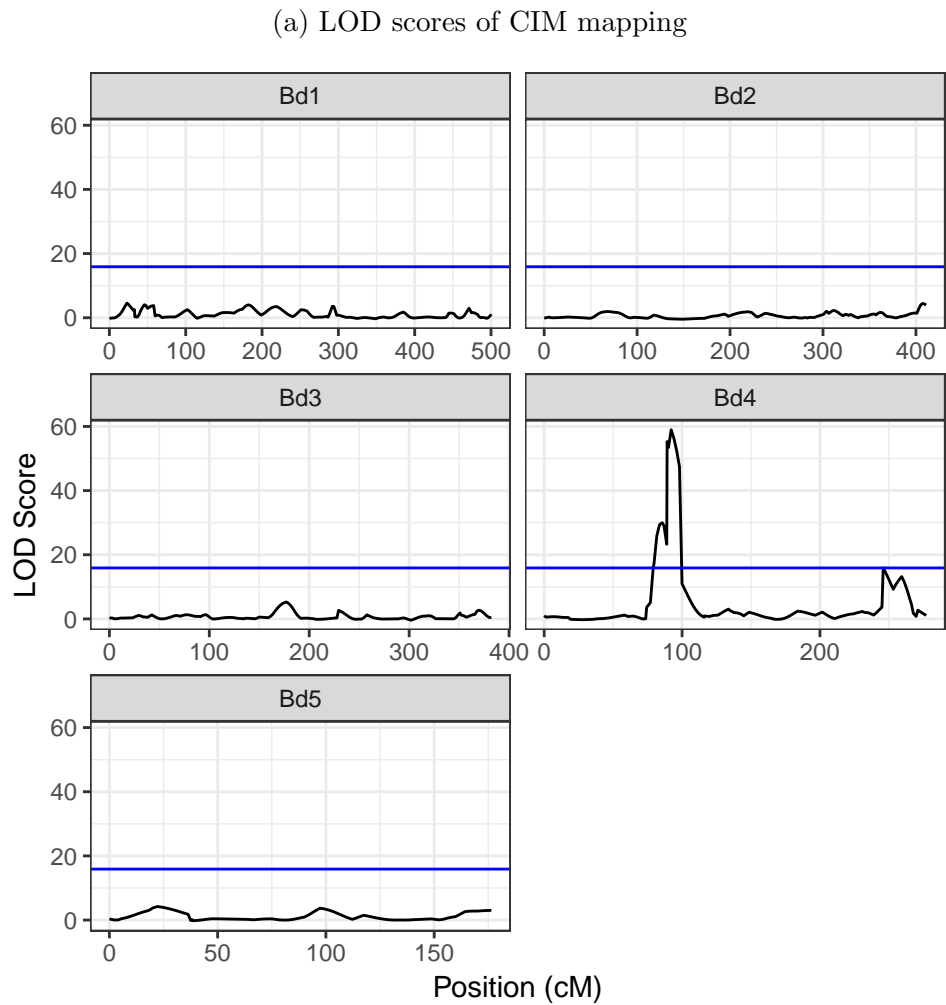
Figure 7.10: Validation of lesion area measurements using response scores in the ABR6 x Bd21 mapping population.

locus was identified (Bd4_43441185) that was just below the permuted LOD significance threshold.

Figure 7.11b shows genotype-phenotype plots for each of the markers identified. Both markers Bd4_8406855 and Bd4_11697916 show clear linkage with homozygotes for the ABR6 allele and heterozygotes showing lesions areas $< 0.2 \text{ mm}^2$ corresponding to a resistant disease response phenotype. The homozygotes for the Bd21 allele (BB) are showing larger lesion areas corresponding to a susceptible response phenotype. However, there are a number of homozygous lines that are also showing smaller mean lesion areas. This could be as a result of under inoculation of those lines or the result of other environmental factors. In marker Bd4_8406855 showed a few dominant homozygotes (AA) that had an average lesion area above 0.2 mm^2 . With the close proximity of these loci and insufficient genetic resolution within the ABR6xBd21 genetic map, they were named as a single locus, Rice Blast Resistance 1 (Rbr1).

7.4.2.2 Candidate NB-LRR genes at the Rbr1 locus

LOD support intervals from SIM were used to identify candidate NB-LRR genes within the identified locus. Figure 7.12 shows the respective LOD support interval of markers Bd4_8406855 and Bd4_11697916 with respect to their physical distance. The intervals of these markers clearly overlap showing their close proximity. There were a total of 102 found between the 2-LOD support interval on *B. distachyon* chromosome 4. Of these genes, 8 have putative annotations as NB-LRRs resistance genes (Tan and Wu 2012). These are shown in Table 7.2. No previous association of these genes has been found with disease resistance. However, interactions of Bradi4g10030, Bradi4g10060, Bradi4g10190 and Bradi4g10220 with micro RNAs have been previously identified in response to *Fusarium culmorum* infection (Lucas, Bata, and Budak 2014).



(b) Genotype vs phenotype plots of significant markers

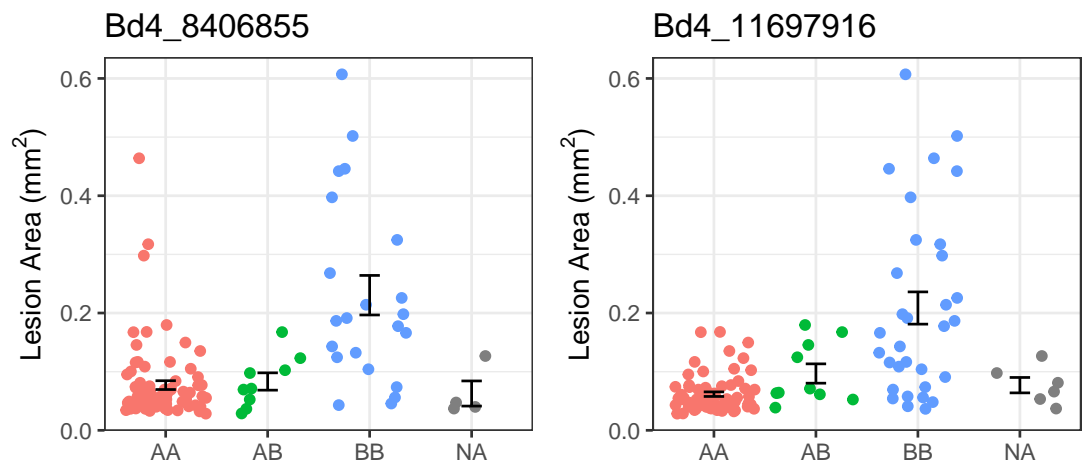


Figure 7.11: **QTL analysis of lesion area data using the ABR6 x Bd21 F_4 genetic map.** a) Lesion area CIM results. The blue line denotes the permuted LOD significance threshold based on 1000 permutations. b) Plots of markers found to be above the LOD significance threshold.

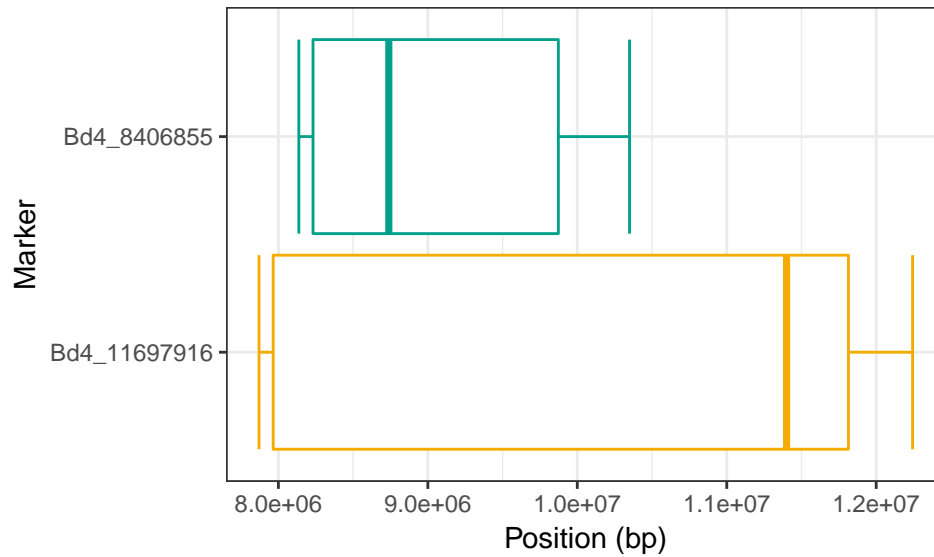


Figure 7.12: **LOD support intervals of significant markers found in the Rbr1 locus.** Bold line denotes the physical position of the marker. The box denotes the 1st LOD support interval. The whiskers denote the 2nd LOD support interval.

Table 7.2: **NB-LRRs identified within the 2-LOD support interval of the Rbr1 locus.**

Gene	Start	Stop	Length	Annotation
Bradi4g10030	9657274	9665856	8582	NBS LRR
Bradi4g10060	9703405	9712329	8924	NBS LRR
Bradi4g10180	9836900	9846084	9184	NBS LRR
Bradi4g10190	9872684	9878529	5845	NBS LRR
Bradi4g10220	9901509	9904547	3038	NBS LRR
Bradi4g11920	11867878	11871924	4046	NBS LRR
Bradi4g11930	11878021	11880671	2650	NBS LRR
Bradi4g11940	11900832	11903917	3085	NBS LRR

7.5 Concluding remarks

This chapter has shown the applicability of computer vision approaches for accurately phenotyping plant disease responses for the identification of linked gene loci. A computer vision method was developed that was able to differentiate lesion area between the compatible and incompatible responses of Bd21 and ABR6. This method of image processing could potentially be applied to other plant pathogen interactions with the adjustment of thresholding parameters for lesion segmentation. Also this would be dependant on the similarity of symptoms to those of the interaction between *B. distachyon* and *M. oryzae*.

Lesion area measurements using computer vision were able to identify a significantly linked genetic locus (Rbr1), responsible for resistance and susceptibility in ABR6 and Bd21. Within the Rbr1 locus, 8 candidate NB-LRR genes were identified.

Chapter 8

General conclusions and future work

8.1 Summary of general conclusions

The central aim of this thesis was to integrate omics analyses to investigate the pre-symptomatic phases of both compatible and incompatible interactions of *M. oryzae* and *B. distachyon*.

Accomplishing this required the development of a spectral binning method for FIE-HRMS metabolome fingerprinting in **Chapter 3**. The software package binneR was developed as a solution for applying spectral binning to FIE-HRMS data. A bin width of 0.01 amu was found to be optimal for spectral binning; this retained resolution, without introducing substantial amounts of missing data or artifacts.

Noise and variability and missing data within FIE-HRMS fingerprints requires rigorous data pre-treatment and filtering prior to statistical analysis. Suitable normalisation, scaling and variable filtering need to be applied that are dependent on the experiment and sample context. The added resolution of FIE-HRMS fingerprints also allows increased confidence in putative metabolite identification compared to previous nominal mass techniques. This has the potential for substantially improving the quality of hypothesis generation when applying

metabolome fingerprinting that can then better inform further metabolomic analyses and experimentation.

The development of FIE-HRMS metabolome fingerprinting in **Chapter 3** allowed key aspects of experimental control and robustness of the interaction between *B. distachyon* and *M. oryzae* to be investigated in **Chapter 4**. It was identified that inoculum constituents other than pathogenic fungal spores were able to elicit substantial metabolomic responses in *B. distachyon*, unrelated to pathogenesis of *M. oryzae*. A method of suitably controlling for these responses, by the use of a non-pathogenic control inoculum, was developed using cycles of deep freezing and vortexing to neutralise *M. oryzae* spores.

An external validation re-sampling approach was applied to investigate the robustness of metabolome changes between independent inoculations in the pre-symptomatic phases of both compatible and incompatible interaction of *B. distachyon* and *M. oryzae*. It was found that metabolome changes do vary between inoculations and that this variability is dynamic between time points with some phases of pathogenesis being more reproducible than others. This highlighted the importance of taking into account the variability in inoculations when investigating plant-pathogen interactions using omics analyses.

The investigation of elements key of omics analyses of plant-pathogen interactions in **Chapter 4** allowed the application of both metabolomic analyses and transcriptomics analysis of the pre-symptomatic phases of both compatible and incompatible interactions between *B. distachyon* and *M. oryzae*. **Chapter 5** identified a range of metabolites that were found to be altered during these initial interaction phases. These included amino acids, fatty acids, purines and pyrimidines in both compatible and incompatible interactions using FIE-HRMS and LC-HRMS analyses. Gene co-expression clusters were also identified from RNA-Seq transcriptomic analyses that included functional enrichment for transcripts related to photosynthesis, amino acid synthesis and hormone metabolism in both compatible and incompatible interactions. It was identified that the extent of metabolomic and transcriptomic changes were not found to be linear with

disease progression with *M. oryzae* spore germination phases being more explanatory than its penetration and initial colonisation phases in both the compatible and incompatible interactions.

With the genes and metabolites identified as explanatory during the interactions between *B. distachyon* and *M. oryzae* in **Chapter 5**, **Chapter 6** attempted to directly integrate these omics data. Pathway and network analyses revealed that chloroplasts were important centres in response to *M.oryzae* infection and that nitrogen metabolism was altered in both compatible and incompatible responses. There were alterations to metabolites involved in light independent photosynthetic reactions and genes involved in photosystems I and II. It was hypothesised that this was in response to reductions in photosynthetic rates during the initial host recognition phases. Alterations to lipoxygenase genes expression changes and poly unsaturated fatty acids were also hypothesised to be related to light independent generation of ROS. Changes in amino acids levels and genes involved in their synthesis indicated that there was alteration to nitrogen mobilisation. However, the interpretation of these changes were difficult due to the likely spatial differentiation of these responses.

There was a substantial difference in the extent of down regulation in primary and secondary metabolism related genes and metabolites in the compatible compared to the incompatible interaction during the spore germination phases. It was hypothesised that these differences in pathogen recognition responses could be related to the initiation of the incompatible responses rather than those of only innate responses present in both the ecotypes.

It was identified in **Chapter 5** that there were substantial transcriptome and metabolome differences between the ecotypes used in the investigations of compatible and incompatible interaction with *M. oryzae*. These would be likely to confound direct comparison of compatible and incompatible responses. A solution to this would be the use of transformed isogenic lines differing only in genes responsible for responses to *M. oryzae*. To initiate investigations towards this goal, a computer vision based approach was used to quantify disease responses for

QTL analysis of RILs from an ABR6 x Bd21 mapping population. A significantly linked genetic locus, subsequently named Rbr1, was found on chromosome 4 of *B. distachyon*. There were 8 candidate NB-LRR disease resistance genes identified within this locus.

The results presented here are a novel example of the integration of metabolomic and transcriptomic analyses to investigate a plant-pathogen interaction. They lay a foundation for substantial further investigation into elucidating key *M. oryzae* effector targets within the *B. distachyon* system as well as pathways involved in the resistant responses initiated by *B. distachyon*.

8.2 Future work

Immediately furthering these analyses presented in this thesis would require the confirmation of putative metabolite annotations using MS/MSⁿ analyses, comparing fragmentation patterns with that of chemical standards. The transcriptional profiles of RNA-Seq analyses would also require confirmation by qPCR analyses.

These integrative analyses could be enhanced by the application of further omics levels such as proteomic analyses. As many of the molecular interactions involved during plant-pathogen interactions are that of protein-protein interactions and enzymatic disruption, the use of shotgun proteomic analyses could provide valuable information in addition the transcriptional and metabolite associations identified here (Haynes and Roberts 2007).

There are a number of technical issues in analysing this interaction using omics analyses. The further characterisation of the Rbr1 locus and the eventual transformation of Bd21 to yield a genotype resistant to *M. oryzae* would allow the removal of genetic variability and direct comparisons between compatible and incompatible interactions to be made. This could allow substantial characterisation of the recognition pathways and response pathways involved in the incompatible *M. oryzae* response as well as the identification of potential effector targets involved in the suppression of these responses.

One of the limitations of the metabolomic analyses presented in this thesis

is confidently assigning which organism is the source of the metabolite changes. Unlike the transcriptional changes, where the alignment of reads to an organisms genome allows automatic assignment, many primary metabolites will be shared between the organisms.

In the interaction between *A. thaliana* and *P. syringae*, a dual metabolomics approach has been developed utilising plant cell-pathogen co-cultures (Allwood et al. 2010). The host and pathogen cells can be separated by differential filtering and centrifugation, allowing metabolomic analyses to be conducted separately on each organism. Unlike bacterial pathogens that colonise the plant apoplast, biotrophic fungal pathogens establish intimate interfaces with their hosts (O’Connell et al. 2012). This would make the separation of plant cell-pathogen co-cultures difficult. The utility of the use plant cell cultures is also questionable for studying plant pathogen interactions when they are not under the physiological constraints that would be found *in planta* (Allwood et al. 2010).

Stable isotopic labelling of the host and/or pathogen species could provide an alternative means of assigning organism specific metabolic changes. This would allow the differentiation of mass spectral signals resultant from either the pathogen or the host (Godin, Fay, and Hopfgartner 2007). However, the changes resolvable using this method are likely to be short lived as there will a dilution of the isotope, Conducting the interaction would require the exposure of one or other of the labelled organisms to an unlabelled environment. Horst et al. (2010) used stable isotope labelled ^{15}N to investigate alterations in nitrogen allocation during the interaction between *Zea mays* and *U. maydis*, identifying that *U. maydis* induced tumours showed a reduce assimilation of soil-derived $^{15}\text{NO}_3^-$ becoming strong nitrogen sinks.

Single cell metabolomics could provide another solution to organism specific assignment. Cellular contents can be removed using a metal-coated microcappillary under video-microscopy observation. This can be directly fed into a mass spectrometer using nanospray ionisation. Ion mobility separation can be used to provide further separation post ionisation (Fujii et al. 2015). Cellular contents of

both the plant and fungal cells could be sampled independently giving organism specific metabolite profiles. This could also be extended to single cell transcriptomics, applying RNA-Seq to single cell mRNA profiles to investigate the gene expression of the individual cells (Tang et al. 2009).

Closely linked to the issue of organism assignment of metabolite profiles is the spatial changes occurring within plant tissues during pathogenesis. The results of this thesis deal only with the temporal changes; however, as discussed in Sections 6.4.3.1 and 6.4.3.2, many cellular responses are spatially differentiated during pathogenesis such as photosynthesis or nitrogen mobilisation responses. Resolution of the spatial responses occurring across the leaf tissue would also provide information as to the likely organism origin of those changes.

Applying omics analyses to investigate the spatial changes during plant pathogen interactions would be reliant on appropriately sampling and analysing separate regions of the plant tissue. There are a number of techniques that could be used in order to achieve this. The single cell sampling by microcapillary and microscopic observation discussed previously would provide one method that could allow both transcriptomic and metabolomic analyses. Micro dissection of leaf regions would also allow this (Kueger et al. 2012).

There are few metabolomic techniques that would allow spatial distributions to be investigated. These include NMR spectroscopy that can be used to produce metabolomic fingerprints of intact plant organs (Kim, Choi, and Verpoorte 2011). Mass spectral imaging techniques utilising desorption electrospray ionisation and matrix assisted laser desorption ionisation coupled to MS, can provide spatial information on a wide range of molecules. They use solvent droplets or lasers respectively to ionise molecules from the surface of intact plant tissues, however they are other limited by necessary dehydration tissue pre-treatments prior to analysis (Muller et al. 2011; Kaspar et al. 2011). A further method is the use of laser ablation electrospray ionisation that can be used on fresh living plant tissue and has been used to investigate the depletion of α -Tomatine at interaction sites in tomato during *Cladosporium fulvum* infection (Etalo et al. 2015).

Applying these suite of techniques would be essential in successfully elucidating the complex and dynamic system alterations that occur during *B. distachyon* and *M. oryzae* interaction.

References

- Ahuja, I., R. Kissen, and A. M. Bones (2012). “Phytoalexins in defense against pathogens”. In: *Trends in Plant Science* 17.2, pp. 73–90.
- Allwood, J. W., D. I. Ellis, and R. Goodacre (2008). “Metabolomic technologies and their application to the study of plants and plant-host interactions”. In: *Physiologia Plantarum* 132.2, pp. 117–135.
- Allwood, W. J., E. David, J. Heald, R. Goodacre, and L. a. J. Mur (2006). “Metabolomic approaches reveal that phosphatidic and phosphatidyl glycerol phospholipids are major discriminatory non-polar metabolites in responses by *Brachypodium distachyon* to challenge by *Magnaporthe grisea*”. In: *Plant Journal* 46.3, pp. 351–368.
- Allwood, W. J., A. Clarke, R. Goodacre, and L. A. J. Mur (2010). “Dual metabolomics: A novel approach to understanding plant-pathogen interactions”. In: *Phytochemistry* 71.5-6, pp. 590–597.
- Araus, L. and J. E. Cairns (2014). “Field high-throughput phenotyping : the new crop breeding frontier”. In: *Trends in Plant Science* 19.1, pp. 52–61.
- Baena-González, E. and J. Sheen (2008). “Convergent energy and stress signaling”. In: *Trends in Plant Science* 13.9, pp. 474–482.
- Bagnaresi, P., C. Biselli, L. Orru, S. Urso, L. Crispino, P. Abbruscato, P. Piffanelli, E. Lupotto, L. Cattivelli, and G. Val?? (2012). “Comparative Transcriptome Profiling of the Early Response to *Magnaporthe oryzae* in Durable Resistant vs Susceptible Rice (*Oryza sativa* L.) Genotypes”. In: *PLoS ONE* 7.12, pp. 1–26.

-
- Ballini, E., J.-B. Morel, G. Droc, A. Price, B. Courtois, J.-L. Notteghem, and D. Tharreau (2008). “A Genome-Wide Meta-Analysis of Rice Blast Resistance Genes and Quantitative Trait Loci Provides New Insights into Partial and Complete Resistance”. In: *Molecular Plant-Microbe Interactions* 21.7, p. 859.
- Barabási, A.-L. and Z. N. Oltvai (2004). “Network biology: understanding the cell’s functional organization.” In: *Nature reviews. Genetics* 5.2, pp. 101–113.
- Barbedo, J. G. (2013). “Digital image processing techniques for detecting, quantifying and classifying plant diseases.” In: *SpringerPlus* 2.1, pp. 660–671.
- Barbieri, M., T. C. Marcel, R. E. Niks, E. Francia, M. Pasquariello, V. Mazzamurro, D. F. Garvin, N. Pecchioni, and a.E. Van Deynze (2012). “QTLs for resistance to the false brome rust *Puccinia brachypodii* in the model grass *Brachypodium distachyon* L.” In: *Genome* 55.2, pp. 152–163.
- Beckmann, M., D. P. Enot, D. P. Overy, and J. Draper (2007). “Representation, comparison, and interpretation of metabolome fingerprint data for total composition analysis and quality trait investigation in potato cultivars”. In: *Journal of Agricultural and Food Chemistry* 55.9, pp. 3444–3451.
- Beckmann, M., D. Parker, D. P. Enot, E. Duval, and J. Draper (2008). “High-throughput, nontargeted metabolite fingerprinting using nominal mass flow injection electrospray mass spectrometry.” In: *Nature protocols* 3.3, pp. 486–504.
- Ben-David, A. (2008). “Comparison of classification accuracy using Cohen’s Weighted Kappa”. In: *Expert Systems with Applications* 34.2, pp. 825–832.
- Bettgenhaeuser, J., F. M. Corke, M. Opanowicz, P. Green, I. Hernandez-Pinzon, J. H. Doonan, and M. J. Moscou (2016). “Natural variation in *Brachypodium distachyon* links VRN2 and FT loci as major flowering determinants”. In: *Under review*.
- Bock, C. H., M. W. Hotchkiss, and B. W. Wood (2016). “Assessing disease severity: Accuracy and reliability of rater estimates in relation to number of diagrams in a standard area diagram set”. In: *Plant Pathology* 65.2, pp. 261–272.

-
- Bock, C. H., G. H. Poole, P. E. Parker, and T. R. Gottwald (2010). “Plant Disease Severity Estimated Visually, by Digital Photography and Image Analysis, and by Hyperspectral Imaging”. In: *Critical Reviews in Plant Sciences* 29.2, pp. 59–107.
- Bohnert, H. U. (2004). “A Putative Polyketide Synthase/Peptide Synthetase from *Magnaporthe grisea* Signals Pathogen Attack to Resistant Rice”. In: *the Plant Cell Online* 16.9, pp. 2499–2513.
- Bolger, A. M., M. Lohse, and B. Usadel (2014). “Trimmomatic: A flexible trimmer for Illumina sequence data”. In: *Bioinformatics* 30.15, pp. 2114–2120.
- Bolton, M. D. (2009). “Primary metabolism and plant defense—fuel for the fire.” In: *Molecular plant-microbe interactions : MPMI* 22.5, pp. 487–497.
- Bonfig, K. B., U. Schreiber, A. Gabler, T. Roitsch, and S. Berger (2006). “Infection with virulent and avirulent *P. syringae* strains differentially affects photosynthesis and sink metabolism in *Arabidopsis* leaves”. In: *Planta* 225.1, pp. 1–12.
- Boulesteix, A. L. and M. Slawski (2009). “Stability and aggregation of ranked gene lists”. In: *Briefings in Bioinformatics* 10.5, pp. 556–568.
- Brachypodium Initiative, T. I. (2010). “Genome sequencing and analysis of the model grass *Brachypodium distachyon*”. In: *Nature* 463.7282, pp. 763–768.
- Braga, G. U., D. E. Rangel, v. K. Fernandes, S. D. Flint, and D. W. Roberts (2015). “Molecular and physiological effects of environmental UV radiation on fungal conidia”. In: *Current Genetics* 61.3, pp. 405–425.
- Breiman, L. (1996). “Bagging Predictors”. In: *Machine Learning* 24.421, pp. 123–140.
- (2001). “Random Forests”. In: *Machine Learning* 45, pp. 5–32.
- Broadhurst, D. I. and D. B. Kell (2006). “Statistical strategies for avoiding false discoveries in metabolomics and related experiments”. In: *Metabolomics* 2.4, pp. 171–196.
- Brougham, D. F., G. Ivanova, M. Gottschalk, D. M. Collins, A. J. Eustace, R. O’Connor, and J. Havel (2011). “Artificial neural networks for classification

-
- in metabolomic studies of whole cells using ^1H Nuclear Magnetic Resonance”. In: *Journal of Biomedicine and Biotechnology* 2011, pp. 1–8.
- Brown, J. K. (2015). “Durable Resistance of Crops to Disease: A Darwinian Perspective”. In: *Annual Review of Phytopathology* 53.1, pp. 513–539.
- Camargo, a. and J. S. Smith (2009). “Image pattern classification for the identification of disease causing agents in plants”. In: *Computers and Electronics in Agriculture* 66.2, pp. 121–125.
- Castaldi, P. J., I. J. Dahabreh, and J. P. A. Ioannidis (2011). “An empirical assessment of validation practices for molecular classifiers”. In: *Briefings in Bioinformatics* 12.3, pp. 189–202.
- Cavill, R., D. Jennen, J. Kleinjans, and J. J. Briedé (2015). “Transcriptomic and metabolomic data integration.” In: *Briefings in bioinformatics* August, pp. 1–11.
- Chang, W., J. Cheng, J. Allaire, Y. Xie, and J. McPherson (2016). “shiny: Web Application Framework for R”. In: URL: project.org/package=shiny.
- Chen, S. and H. Leung (2004). “Survey over image thresholding techniques and quantitative performance evaluation”. In: *Journal of Electronic Imaging* 13.1, p. 220.
- Chinchilla, D., Z. Bauer, M. Regenass, T. Boller, and G. Felix (2006). “The Arabidopsis receptor kinase FLS2 binds flg22 and determines the specificity of flagellin perception.” In: *The Plant cell* 18.February, pp. 465–476.
- Chisholm, S. T., G. Coaker, B. Day, and B. J. Staskawicz (2006). “Host-microbe interactions: Shaping the evolution of the plant immune response”. In: *Cell* 124.4, pp. 803–814.
- Churchill, G. A. and R. W. Doerge (1994). “Empirical threshold values for quantitative trait mapping”. In: *Genetics* 138.3, pp. 963–971.
- Cline, S. M., M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A. R. Pico, A. Vailaya, P.-L. Wang, A. Adler, B. R. Conklin, L. Hood, M. Kuiper,

-
- C. Sander, I. Schmulevich, B. Schwikowski, G. J. Warner, T. Ideker, and G. D. Bader (2007). “Integration of biological networks and gene expression data using Cytoscape”. In: *Nature protocols* 2.10, pp. 2366–2382.
- Collard, B. C. Y., M. Z. Z. Jahufer, J. B. Brouwer, and E. C. K. Pang (2005). “An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts”. In: *Euphytica* 142.1-2, pp. 169–196.
- Collins, G. S., J. a. de Groot, S. Dutton, O. Omar, M. Shanyinde, A. Tajar, M. Voysey, R. Wharton, L.-M. Yu, K. G. Moons, and D. G. Altman (2014). “External validation of multivariable prediction models: a systematic review of methodological conduct and reporting.” In: *BMC medical research methodology* 14.1, p. 40.
- Conrath, U. (2011). “Molecular aspects of defence priming”. In: *Trends in Plant Science* 16.10, pp. 524–531.
- Couch, B. C. and L. M. Kohn (2002). “A multilocus gene genealogy concordant with host preference indicates segregation of a new species, *Magnaporthe oryzae*, from *M. grisea*.” In: *Mycologia* 94.4, pp. 683–693.
- Crawley, M. J. (2013). *The R Book-Second Edition*.
- Davis, C. A., F. Gerick, V. Hintermair, C. C. Friedel, K. Fundel, R. Küffner, and R. Zimmer (2006). “Reliable gene signatures for microarray classification: Assessment of stability and performance”. In: *Bioinformatics* 22.19, pp. 2356–2363.
- De Jong, J., B. McCormack, N. Smirnoff, and N. Talbot (1997). “Glycerol generates turgor in rice blast”. In: *Nature* 389.September, p. 244.
- De Vos, R. C. H., S. Moco, A. Lommen, J. J. B. Keurentjes, R. J. Bino, and R. D. Hall (2007). “Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry.” In: *Nature protocols* 2.4, pp. 778–91.
- Dean, R. a., N. J. Talbot, D. J. Ebbole, M. L. Farman, T. K. Mitchell, M. J. Orbach, M. Thon, R. Kulkarni, J.-R. Xu, H. Pan, N. D. Read, Y.-H. Lee,

-
- I. Carbone, D. Brown, Y. Y. Oh, N. Donofrio, J. S. Jeong, D. M. Soanes, S. Djonovic, E. Kolomiets, C. Rehmeier, W. Li, M. Harding, S. Kim, M.-H. Lebrun, H. Bohnert, S. Coughlan, J. Butler, S. Calvo, L.-J. Ma, R. Nicol, S. Purcell, C. Nusbaum, J. E. Galagan, and B. W. Birren (2005). “The genome sequence of the rice blast fungus *Magnaporthe grisea*.” In: *Nature* 434.7036, pp. 980–986.
- Dell’Endice, F., J. Nieke, B. Koetz, M. E. Schaepman, and K. Itten (2009). “Improving radiometry of imaging spectrometers by using programmable spectral regions of interest”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 64.6, pp. 632–639.
- Desikan, R, S. J. Neill, and J. T. Hancock (2000). “Hydrogen peroxide-induced gene expression in *Arabidopsis thaliana*”. In: *Free radical biology & medicine* 28.5, pp. 773–778.
- Dillies, M. A., A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, N. S. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloë, C. Le Gall, B. Schaëffer, S. Le Crom, M. Guedj, and F. Jaffrézic (2013). “A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis”. In: *Briefings in Bioinformatics* 14.6, pp. 671–683.
- Ding, L., H. Xu, H. Yi, L. Yang, Z. Kong, L. Zhang, S. Xue, H. Jia, and Z. Ma (2011). “Resistance to hemi-biotrophic *f. graminearum* infection is associated with coordinated and ordered expression of diverse defense signaling pathways”. In: *PLoS ONE* 6.4.
- Dixon, R. A., L. Achnine, P. Kota, C. J. Liu, M. S. S. Reddy, and L. Wang (2002). “The phenylpropanoid pathway and plant defence - A genomics perspective”. In: *Molecular Plant Pathology* 3.5, pp. 371–390.
- Djamei, A., K. Schipper, F. Rabe, A. Ghosh, V. Vincon, J. Kahnt, S. Osorio, T. Tohge, A. R. Fernie, I. Feussner, K. Feussner, P. Meinicke, Y.-D. Stierhof, H. Schwarz, B. Macek, M. Mann, and R. Kahmann (2011). “Metabolic priming by a secreted fungal effector”. In: *Nature* 478.7369, pp. 395–398.

-
- Dodds, P. N. and J. P. Rathjen (2010). “Plant immunity: towards an integrated view of plant-pathogen interactions.” In: *Nature reviews. Genetics* 11.8, pp. 539–548.
- Doehlemann, G., R. Wahl, R. J. Horst, L. M. Voll, B. Usadel, F. Poree, M. Stitt, J. Pons-Kühnemann, U. Sonnewald, R. Kahmann, and J. Kämper (2008). “Re-programming a maize plant: Transcriptional and metabolic changes induced by the fungal biotroph *Ustilago maydis*”. In: *Plant Journal* 56.2, pp. 181–195.
- Draper, J., L. a.J. Mur, G. Jenkins, G. C. Ghosh-Biswas, P. Bablak, R. Hasterok, and A. P. Routledge (2001). “*Brachypodium distachyon*. A New Model System for Functional Genomics in Grasses¹”. In: *Plant physiology* 127.4, pp. 1539–1555.
- Draper, J., D. P. Enot, D. Parker, M. Beckmann, S. Snowdon, W. Lin, and H. Zubair (2009). “Metabolite signal identification in accurate mass metabolomics data with MZedDB, an interactive m/z annotation tool utilising predicted ionisation behaviour ‘rules’.” In: *BMC bioinformatics* 10, p. 227.
- Draper, J., A. J. Lloyd, R. Goodacre, and M. Beckmann (2013). “Flow infusion electrospray ionisation mass spectrometry for high throughput, non-targeted metabolite fingerprinting: A review”. In: *Metabolomics* 9.SUPPL.1, pp. 4–29.
- Dunn, W. B., N. J. C. Bailey, and H. E. Johnson (2005). “Measuring the metabolome: current analytical technologies.” In: *The Analyst* 130.5, pp. 606–625.
- Dunn, W. B., I. D. Wilson, A. W. Nicholls, and D. Broadhurst (2012). “The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans.” In: *Bioanalysis* 4.18, pp. 2249–64.
- Eitas, T. K. and J. L. Dangl (2010). “NB-LRR proteins: Pairs, pieces, perception, partners, and pathways”. In: *Current Opinion in Plant Biology* 13.4, pp. 472–477.
- Eliceiri, K. W., M. R. Berthold, I. G. Goldberg, L. Ibáñez, B. S. Manjunath, M. E. Martone, R. F. Murphy, H. Peng, A. L. Plant, B. Roysam, N. Stuurman,

-
- J. R. Swedlow, P. Tomancak, and A. E. Carpenter (2012). “Biological imaging software tools”. In: *Nature Methods* 9.7, pp. 697–710.
- Enot, D. P. and J. Draper (2007). “Statistical measures for validating plant genotype similarity assessments following multivariate analysis of metabolome fingerprint data”. In: *Metabolomics* 3.3, pp. 349–355.
- Enot, D. P., W. Lin, M. Beckmann, D. Parker, D. P. Overy, and J. Draper (2008). “Preprocessing, classification modeling and feature selection using flow injection electrospray mass spectrometry metabolite fingerprint data.” In: *Nature protocols* 3.3, pp. 446–470.
- Etalo, D. W., R. C. H. De Vos, M. Joosten, and R. D. Hall (2015). “Spatially resolved plant metabolomics: Some potentials and limitations of laser-ablation electrospray ionization mass spectrometry metabolite imaging”. In: *Plant Physiology* 169.3, pp. 1424–1435.
- Farré, E. M. and S. E. Weise (2012). “The interactions between the circadian clock and primary metabolism”. In: *Current Opinion in Plant Biology* 15.3, pp. 293–300.
- Fernández, V. and P. H. Brown (2013). “From plant surface to plant metabolism: the uncertain fate of foliar-applied nutrients”. In: *Frontiers in Plant Science* 4.July, p. 289.
- Fiehn, O., J. Kopka, P. Dörmann, T. Altmann, R. N. Trethewey, and L. Willmitzer (2000). “Metabolite profiling for plant functional genomics”. In: *Nature Biotechnology* 18.11, pp. 1157–1161.
- Fisher, M. C., D. a. Henk, C. J. Briggs, J. S. Brownstein, L. C. Madoff, S. L. McCraw, and S. J. Gurr (2012). “Emerging fungal threats to animal, plant and ecosystem health.” In: *Nature* 484.7393, pp. 186–94.
- Fujii, T, S Matsuda, M. L. Tejedor, T Esaki, I Sakane, H Mizuno, N Tsuyama, and T Masujima (2015). “Direct metabolomics for plant cells by live single-cell mass spectrometry”. In: *Nat Protoc* 10.9, pp. 1445–1456.
- Gehlenborg, N., S. I. O’donoghue, N. S. Baliga, A. Goesmann, M. A. Hibbs, H. Kitano, O. Kohlbacher, H. Neuweyer, R. Schneider, D. Tenenbaum, and A.-

-
- C. Gavin (2010). “Visualization of omics data for systems biology”. In: *Nat Methods* 7.3 Suppl, S56–68.
- Giraldo, M. C., Y. F. Dagdas, Y. K. Gupta, T. a. Mentlak, M. Yi, A. L. Martinez-Rocha, H. Saitoh, R. Terauchi, N. J. Talbot, and B. Valent (2013). “Two distinct secretion systems facilitate tissue invasion by the rice blast fungus *Magnaporthe oryzae*.” In: *Nature communications* 4, May, p. 1996.
- Godin, J.-P., L.-B. Fay, and G. Hopfgartner (2007). “LIQUID CHROMATOGRAPHY COMBINED WITH MASS SPECTROMETRY FOR 13 C ISOTOPIC ANALYSIS IN LIFE SCIENCE RESEARCH”. In: *Mass Spectrometry Reviews* 26, pp. 751–774.
- Gonzalez-dugo, V., J.-l. Durand, V. Gonzalez-dugo, and J.-l. Durand (2010). “Water deficit and nitrogen nutrition of crops. A review”. In: *Agronomy for Sustainable Development* 30, pp. 529–544.
- Goodacre, R., S. Vaidyanathan, W. B. Dunn, G. G. Harrigan, and D. B. Kell (2004). “Metabolomics by numbers: Acquiring and understanding global metabolite data”. In: *Trends in Biotechnology* 22.5, pp. 245–252.
- Gordon, S. P., H. Priest, D. L. Des Marais, W. Schackwitz, M. Figueroa, J. Martin, J. N. Bragg, L. Tyler, C. R. Lee, D. Bryant, W. Wang, J. Messing, A. J. Manzaneda, K. Barry, D. F. Garvin, H. Budak, M. Tuna, T. Mitchell-Olds, W. F. Pfender, T. E. Juenger, T. C. Mockler, and J. P. Vogel (2014). “Genome diversity in *Brachypodium distachyon*: Deep sequencing of highly diverse inbred lines”. In: *Plant Journal* 79.3, pp. 361–374.
- Grata, E., J. Boccard, G. Glauser, P. A. Carrupt, E. E. Farmer, J. L. Wolfender, and S. Rudaz (2007). “Development of a two-step screening ESI-TOF-MS method for rapid determination of significant stress-induced metabolome modifications in plant leaf extracts: The wound response in *Arabidopsis thaliana* as a case study”. In: *Journal of Separation Science* 30.14, pp. 2268–2278.
- Gromski, P. S., H. Muhamadali, D. I. Ellis, Y. Xu, E. Correa, M. L. Turner, and R. Goodacre (2015a). “A tutorial review: Metabolomics and partial least squares-

-
- discriminant analysis - a marriage of convenience or a shotgun wedding”. In: *Analytica Chimica Acta* 879, pp. 10–23.
- Gromski, P. S., Y. Xu, K. A. Hollywood, M. L. Turner, and R. Goodacre (2015b). “The influence of scaling metabolomics data on model classification accuracy”. In: *Metabolomics* 11.3, pp. 684–695.
- Gunnaiah, R., A. C. Kushalappa, R. Duggavathi, S. Fox, and D. J. Somers (2012). “Integrated metabolo-proteomic approach to decipher the mechanisms by which wheat qtl (Fhb1) contributes to resistance against *Fusarium graminearum*”. In: *PLoS ONE* 7.7, e40695.
- Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber, M. D. Macmanes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C. N. Dewey, R. Henschel, R. D. Leduc, N. Friedman, and A. Regev (2013). “De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis.” In: *Nature protocols* 8.8, pp. 1494–1512.
- Hagel, J. M. and P. J. Facchini (2008). “Plant metabolomics: Analytical platforms and integration with functional genomics”. In: *Phytochemistry Reviews* 7.3, pp. 479–497.
- Hamer, J. E., R. J. Howard, F. G. Chumley, and B. Valent (1988). “A mechanism for surface attachment in spores of a plant pathogenic fungus.” In: *Science* 239.4837, pp. 288–290.
- Hanson, J. and S. Smeeckens (2009). “Sugar perception and signaling - an update”. In: *Current Opinion in Plant Biology* 12.5, pp. 562–567.
- Haralick, R., K. Shanmugan, and I. Dinstein (1973). *Textural features for image classification*.
- Hardman, M. and A. A. Makarov (2003). “Interfacing the orbitrap mass analyzer to an electrospray ion source”. In: *Analytical Chemistry* 75.7, pp. 1699–1705.
- Hasegawa, M., I. Mitsuhashi, S. Seo, T. Imai, J. Koga, K. Okada, H. Yamane, and Y. Ohashi (2010). “Phytoalexin accumulation in the interaction between rice

-
- and the blast fungus.” In: *Molecular plant-microbe interactions : MPMI* 23.8, pp. 1000–1011.
- Haynes, P. A. and T. H. Roberts (2007). “Subcellular shotgun proteomics in plants: Looking beyond the usual suspects”. In: *Proteomics* 7.16, pp. 2963–2975.
- He, Z. and W. Yu (2010). “Stable feature selection for biomarker discovery”. In: *Computational Biology and Chemistry* 34.4, pp. 215–225.
- Hématy, K., C. Cherk, and S. Somerville (2009). “Host-pathogen warfare at the plant cell wall”. In: *Current Opinion in Plant Biology* 12.4, pp. 406–413.
- Hikosaka, K., K. Ishikawa, A. Borjigidai, O. Muller, and Y. Onoda (2006). “Temperature acclimation of photosynthesis: Mechanisms involved in the changes in temperature dependence of photosynthetic rate”. In: *Journal of Experimental Botany* 57.2, pp. 291–302.
- Hirai, M. Y., M. Yano, D. B. Goodenowe, S. Kanaya, T. Kimura, M. Awazuhara, M. Arita, T. Fujiwara, and K. Saito (2004). “Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*.” In: *Proceedings of the National Academy of Sciences of the United States of America* 101.27, pp. 10205–10210.
- Horst, R. J., G. Doehlemann, R. Wahl, J. Hofmann, A. Schmiedl, R. Kahmann, J. Kämper, U. Sonnewald, and L. M. Voll (2010). “*Ustilago maydis* infection strongly alters organic nitrogen allocation in maize and stimulates productivity of systemic source leaves.” In: *Plant physiology* 152.1, pp. 293–308.
- Hou, Q., G. Ufer, and D. Bartels (2015). “Lipid signalling in plant responses to abiotic stress.” In: *Plant, cell & environment*.
- Howard, R. J. and B. Valent (1996). “BREAKING AND ENTERING: Host Penetration by the Fungal Rice Blast Pathogen *Magnaporthe grisea*”. In: *Annual Review of Microbiology* 50.1, pp. 491–512.
- Hrydziusko, O. and M. R. Viant (2012). “Missing values in mass spectrometry based metabolomics: An undervalued step in the data processing pipeline”. In: *Metabolomics* 8, pp. 161–174.

-
- Hu, Q., R. J. Noll, H. Li, A. Makarov, M. Hardman, and R. G. Cooks (2005). "The Orbitrap: A new mass spectrometer". In: *Journal of Mass Spectrometry* 40.4, pp. 430–443.
- Huber, W, V. J. Carey, R Gentleman, S Anders, M Carlson, B. S. Carvalho, H. C. Bravo, S Davis, L Gatto, T Girke, R Gottardo, F Hahne, K. D. Hansen, R. A. Irizarry, M Lawrence, M. I. Love, J MacDonald, V Obenchain, A. K. Oles, H Pages, A Reyes, P Shannon, G. K. Smyth, D Tenenbaum, L Waldron, and M Morgan (2015). "Orchestrating high-throughput genomic analysis with Bioconductor". In: *Nat Methods* 12.2, pp. 115–121.
- Hwang, I. S., S. H. An, and B. K. Hwang (2011). "Pepper asparagine synthetase 1 (CaAS1) is required for plant nitrogen assimilation and defense responses to microbial pathogens". In: *Plant Journal* 67.5, pp. 749–762.
- Ingle, R. a., M. Carstens, and K. J. Denby (2006). "PAMP recognition and the plant-pathogen arms race". In: *BioEssays* 28.9, pp. 880–889.
- Ioannidis, J. P. a. and M. J. Khoury (2011). "Improving validation practices in "omics" research." In: *Science* 334.6060, pp. 1230–1232.
- Jones, J. D. G. and J. L. Dangl (2006). "The plant immune system." In: *Nature* 444.7117, pp. 323–329.
- Jones, O. A. H., M. L. Maguire, J. L. Griffin, Y. H. Jung, J. Shibato, R. Rakwal, G. K. Agrawal, and N. S. Jwa (2011). "Using metabolic profiling to assess plant-pathogen interactions: An example using rice (*Oryza sativa*) and the blast pathogen *Magnaporthe grisea*". In: *European Journal of Plant Pathology* 129.4, pp. 539–554.
- Jörnsten, R., H.-Y. Wang, W. J. Welsh, and M. Ouyang (2005). "DNA microarray data imputation and significance analysis of differential expression." In: *Bioinformatics (Oxford, England)* 21.22, pp. 4155–4161.
- Joyce, A. R. and B. Ø. Palsson (2006). "The model organism as a system: integrating 'omics' data sets." In: *Nature reviews. Molecular cell biology* 7.3, pp. 198–210.

-
- Junker, B. H., C. Klukas, and F. Schreiber (2006). “VANTED: a system for advanced data analysis and visualization in the context of biological networks.” In: *BMC bioinformatics* 7, p. 109.
- Jurman, G., S. Merler, A. Barla, S. Paoli, A. Galea, and C. Furlanello (2008). “Algebraic stability indicators for ranked lists in molecular profiling”. In: *Bioinformatics* 24.2, pp. 258–264.
- Kachroo, A. and P. Kachroo (2009). “Fatty AcidDerived Signals in Plant Defense”. In: *Annual Review of Phytopathology* 47.1, pp. 153–176.
- Kalousis, A., J. Prados, and M. Hilario (2007). “Stability of feature selection algorithms: a study on high dimensional spaces”. In: *Knowledge and Information Systems* 12, pp. 95–116.
- Kangasjärvi, S., J. Neukermans, S. Li, E. M. Aro, and G. Noctor (2012). “Photosynthesis, photorespiration, and light signalling in defence responses”. In: *Journal of Experimental Botany* 63.4, pp. 1619–1636.
- Kao, C. H., Z. B. Zeng, and R. D. Teasdale (1999). “Multiple interval mapping for quantitative trait loci.” In: *Genetics* 152.3, pp. 1203–16.
- Kaspar, S., M. Peukert, A. Svatos, A. Matros, and H. P. Mock (2011). “MALDI-imaging mass spectrometry - An emerging technique in plant biology”. In: *Proteomics* 11.9, pp. 1840–1850.
- Kawahara, Y., Y. Oono, H. Kanamori, T. Matsumoto, T. Itoh, and E. Minami (2012). “Simultaneous RNA-seq analysis of a mixed transcriptome of rice and blast fungus interaction.” In: *PloS one* 7.11, e49423.
- Kazan, K. and J. M. Manners (2009). “Linking development to defense: auxin in plant-pathogen interactions”. In: *Trends in Plant Science* 14.7, pp. 373–382.
- Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg (2013). “TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.” In: *Genome biology* 14.4, R36.
- Kim, H. K., Y. H. Choi, and R. Verpoorte (2011). “NMR-based plant metabolomics: Where do we stand, where do we go?” In: *Trends in Biotechnology* 29.6, pp. 267–275.

-
- Kind, T. and O. Fiehn (2007). “Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry.” In: *BMC bioinformatics* 8, p. 105.
- Klie, S. and Z. Nikoloski (2012). “The choice between MapMan and Gene ontology for automated gene function prediction in plant science”. In: *Frontiers in Genetics* 3:JUN, pp. 1–14.
- Konukoglu, E. and M. Ganz (2014). “Approximate False Positive Rate Control in Selection Frequency for Random Forest”. In: *arXiv.org* cs.LG, p. 26.
- Kueger, S., D. Steinhauser, L. Willmitzer, and P. Giavalisco (2012). “High-resolution plant metabolomics: From mass spectral features to metabolites and from whole-cell analysis to subcellular metabolite distributions”. In: *Plant Journal* 70.1, pp. 39–50.
- Kuhl, C., R. Tautenhahn, C. Böttcher, T. R. Larson, and S. Neumann (2012). “CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets”. In: *Analytical Chemistry* 84.1, pp. 283–289.
- Kumar, Y., L. Zhang, P. Panigrahi, B. B. Dholakia, V. Dewangan, S. G. Chavan, S. M. Kunjir, X. Wu, N. Li, P. R. Rajmohan, N. Y. Kadoo, A. P. Giri, H. Tang, and V. S. Gupta (2016). “Fusarium oxysporum mediates systems metabolic reprogramming of chickpea roots as revealed by a combination of proteomics and metabolomics”. In: *Plant Biotechnology Journal*, pp. 1589–1603.
- Lander, E. S. and S. Botstein (1989). “Mapping mendelian factors underlying quantitative traits using RFLP linkage maps”. In: *Genetics* 121.1, p. 185.
- Langmead, B. and S. L. Salzberg (2012). “Fast gapped-read alignment with Bowtie 2”. In: *Nat Methods* 9.4, pp. 357–359.
- Li, J., J. Ding, W. Zhang, Y. Zhang, P. Tang, J. Q. Chen, D. Tian, and S. Yang (2010a). “Unique evolutionary pattern of numbers of gramineous NBS-LRR genes”. In: *Molecular Genetics and Genomics* 283.5, pp. 427–438.

-
- Li, P., L. Ponnala, N. Gandotra, L. Wang, Y. Si, S. L. Tausta, T. H. Kebrom, N. Provart, R. Patel, C. R. Myers, E. J. Reidel, R. Turgeon, P. Liu, Q. Sun, T. Nelson, and T. P. Brutnell (2010b). “The developmental dynamics of the maize leaf transcriptome.” In: *Nature genetics* 42.12, pp. 1060–1067.
- Li, Y., W. Uddin, and J. E. Kaminski (2014). “Effects of relative humidity on infection, colonization and conidiation of *Magnaporthe oryzae* on perennial ryegrass”. In: *Plant Pathology* 63.3, pp. 590–597.
- Liu, W., J. Liu, Y. Ning, B. Ding, X. Wang, Z. Wang, and G.-L. Wang (2013). “Recent progress in understanding PAMP- and effector-triggered immunity against the rice blast fungus *Magnaporthe oryzae*.” In: *Molecular plant* 6.3, pp. 605–20.
- Liu, Y., D. Ren, S. Pike, S. Pallardy, W. Gassmann, and S. Zhang (2007). “Chloroplast-generated reactive oxygen species are involved in hypersensitive response-like cell death mediated by a mitogen-activated protein kinase cascade”. In: *Plant Journal* 51.6, pp. 941–954.
- Lo Presti, L., D. Lanver, G. Schweizer, S. Tanaka, L. Liang, M. Tollot, A. Zucaro, S. Reissmann, and R. Kahmann (2015). “Fungal Effectors and Plant Susceptibility.” In: *Annual review of plant biology* 66, pp. 513–545.
- Lohse, M., A. Nagel, T. Herter, P. May, M. Schroda, R. Zrenner, T. Tohge, A. R. Fernie, M. Stitt, and B. Usadel (2014). “Mercator: A fast and simple web server for genome scale functional annotation of plant sequence data”. In: *Plant, Cell and Environment* 37.5, pp. 1250–1258.
- López-Gresa, M. P., F. Maltese, J. M. Bellés, V. Conejero, H. K. Kim, Y. H. Choi, and R. Verpoorte (2010). “Metabolic response of tomato leaves upon different plant-pathogen interactions”. In: *Phytochemical Analysis* 21.1, pp. 89–94.
- Lucas, S. J., K. Bata, and H. Budak (2014). “Exploring the interaction between small RNAs and R genes during *Brachypodium* response to *Fusarium culmorum* infection”. In: *Gene* 536.2, pp. 254–264.

-
- Mahadevan, S, S. L. Shah, T. J. Marrie, and C. M. Slupsky (2008). “Analysis of metabolomic data using support vector machines”. In: *Anal Chem* 80.19, pp. 7562–7570.
- Makarov, A., E. Denisov, A. Kholomeev, W. Balschun, O. Lange, K. Strupat, and S. Horning (2006). “Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer”. In: *Analytical Chemistry* 78.7, pp. 2113–2120.
- Mandadi, K. K. and K.-B. G. Scholthof (2015). “Genome-wide analysis of alternative splicing landscapes modulated during plant-virus interactions in *Brachypodium distachyon*.” In: *The Plant cell* 27.1, pp. 71–85.
- March-Díaz, R., M. García-Domínguez, J. Lozano-Juste, J. León, F. J. Florencio, and J. C. Reyes (2008). “Histone H2A.Z and homologues of components of the SWR1 complex are required to control immunity in *Arabidopsis*”. In: *Plant Journal* 53.3, pp. 475–487.
- Marshall, A. G. and C. L. Hendrickson (2008). “High-Resolution Mass Spectrometers”. In: *Annual Review of Analytical Chemistry* 1.1, pp. 579–599.
- Martens, L., M. Chambers, M. Sturm, D. Kessner, F. Levander, J. Shofstahl, W. H. Tang, A. Rompp, S. Neumann, A. D. Pizarro, L. Montecchi-Palazzi, N. Tasman, M. Coleman, F. Reisinger, P. Souda, H. Hermjakob, P.-A. Binz, and E. W. Deutsch (2011). “mzML—a Community Standard for Mass Spectrometry Data”. In: *Molecular & Cellular Proteomics* 10.1, R110.000133.
- Martin, L. B. B., Z. Fei, J. J. Giovannoni, and J. K. C. Rose (2013). “Catalyzing plant science research with RNA-seq.” In: *Frontiers in plant science* 4.April, p. 66.
- McDougall, G., I. Martinussen, and D. Stewart (2008). “Towards fruitful metabolomics: High throughput analyses of polyphenol composition in berries using direct infusion mass spectrometry”. In: *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences* 871.2, pp. 362–369.
- Mellersh, D. G., I. V. Foulds, V. J. Higgins, and M. C. Heath (2002). “H₂O₂ plays different roles in determining penetration failure in three diverse plant - fungal interactions”. In: *Science* 29, pp. 257–268.

-
- Meng, X. and S. Zhang (2013). “MAPK cascades in plant disease resistance signaling.” In: *Annual review of phytopathology* 51, pp. 245–66.
- Mentlak, T. a., A. Kombrink, T. Shinya, L. S. Ryder, I. Otomo, H. Saitoh, R. Terauchi, Y. Nishizawa, N. Shibuya, B. P. H. J. Thomma, and N. J. Talbot (2012). “Effector-Mediated Suppression of Chitin-Triggered Immunity by *Magnaporthe oryzae* Is Necessary for Rice Blast Disease”. In: *The Plant Cell* 24.1, pp. 322–335.
- Mercier, J. and S. E. Lindow (2000). “Role of leaf surface sugars in colonization of plants by bacterial epiphytes”. In: *Applied and Environmental Microbiology* 66.1, pp. 369–374.
- Metzner, R., A. Eggert, D. V. Dusschoten, D. Pflugfelder, S. Gerth, U. Schurr, N. Uhlmann, and S. Jahnke (2015). “Direct comparison of MRI and X-ray CT technologies for 3D imaging of root systems in soil : potential and challenges for root trait quantification”. In: *Plant Methods* 11.17, pp. 1–11.
- Meyer, D., S. Pajonk, C. Micali, R. O’Connell, and P. Schulze-Lefert (2009). “Extracellular transport and integration of plant secretory proteins into pathogen-induced cell wall compartments”. In: *Plant Journal* 57.6, pp. 986–999.
- Mochida, K. and K. Shinozaki (2011). “Advances in omics and bioinformatics tools for systems analyses of plant functions”. In: *Plant and Cell Physiology* 52.12, pp. 2017–2038.
- Moghaddam, M. R. B. and W. Van Den Ende (2012). “Sugars and plant innate immunity”. In: *Journal of Experimental Botany* 63.11, pp. 3989–3998.
- Moorthy, K., M. Mohamad, and S. Deris (2014). “A Review on Missing Value Imputation Algorithms for Microarray Gene Expression Data”. In: *Current Bioinformatics* 9.1, pp. 18–22.
- Morozova, O., M. Hirst, and M. a. Marra (2009). “Applications of new sequencing technologies for transcriptome analysis.” In: *Annual review of genomics and human genetics* 10, pp. 135–151.
- Mullineaux, P. M. (2006). “Spatial Dependence for Hydrogen Peroxide-Directed Signaling in Light-Stressed Plants”. In: *Plant Physiology* 141.2, pp. 346–350.

-
- Mur, L. A. J., P. Kenton, A. J. Lloyd, H. Ougham, and E. Prats (2008). “The hypersensitive response; The centenary is upon us but how much do we know?” In: *Journal of Experimental Botany* 59.3, pp. 501–520.
- Mur, L. A. J., J. Allainguillaume, P. Catalan, R. Hasterok, G. Jenkins, K. Lesniewska, I. Thomas, and J. Vogel (2011). “Exploiting the brachypodium tool box in cereal and grass research”. In: *New Phytologist* 191.2, pp. 334–347.
- Murray, D. B., M. Beckmann, and H. Kitano (2007). “Regulation of yeast oscillatory dynamics.” In: *Proceedings of the National Academy of Sciences of the United States of America* 104.7, pp. 2241–6.
- Muller, T., S. Oradu, D. R. Ifa, R. G. Cooks, and B. Krautler (2011). “Direct Plant Tissue Analysis and Imprint Imaging by Desorption Electrospray Ionization Mass Spectrometry”. In: *Analytical Chemistry* 83.14, pp. 5754–5761.
- Nagalakshmi, U., Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder (2008). “The transcriptional landscape of the yeast genome defined by RNA sequencing.” In: *Science* 320.5881, pp. 1344–9.
- O’Brien, J. a., A. Daudi, V. S. Butt, and G. P. Bolwell (2012). “Reactive oxygen species and their role in plant defence and cell wall metabolism”. In: *Planta* 236.3, pp. 765–779.
- Obuchowski, N. A., M. L. Lieber, and F. H. Wians (2004). “ROC curves in Clinical Chemistry: Uses, misuses, and possible solutions”. In: *Clinical Chemistry* 50.7, pp. 1118–1125.
- O’Connell, R. J. and R. Panstruga (2006). “Tete a tete inside a plant cell: Establishing compatibility between plants and biotrophic fungi and oomycetes”. In: *New Phytologist* 171.4, pp. 699–718.
- O’Connell, R. J., M. R. Thon, S. Hacquard, S. G. Amyotte, J. Kleemann, M. F. Torres, U. Damm, E. a. Buiate, L. Epstein, N. Alkan, J. Altmüller, L. Alvarado-Balderrama, C. a. Bauser, C. Becker, B. W. Birren, Z. Chen, J. Choi, J. A. Crouch, J. P. Duvick, M. a. Farman, P. Gan, D. Heiman, B. Henrissat, R. J. Howard, M. Kabbage, C. Koch, B. Kracher, Y. Kubo, A. D. Law, M.-H. Lebrun, Y.-H. Lee, I. Miyara, N. Moore, U. Neumann, K. Nordström, D.

-
- G. Panaccione, R. Panstruga, M. Place, R. H. Proctor, D. Prusky, G. Rech, R. Reinhardt, J. a. Rollins, S. Rounsley, C. L. Schardl, D. C. Schwartz, N. Shenoy, K. Shirasu, U. R. Sikhakolli, K. Stüber, S. a. Sukno, J. a. Sweigard, Y. Takano, H. Takahara, F. Trail, H. C. van der Does, L. M. Voll, I. Will, S. Young, Q. Zeng, J. Zhang, S. Zhou, M. B. Dickman, P. Schulze-Lefert, E. Ver Loren van Themaat, L.-J. Ma, and L. J. Vaillancourt (2012). “Lifestyle transitions in plant pathogenic *Colletotrichum* fungi deciphered by genome and transcriptome analyses.” In: *Nature genetics* 44.9, pp. 1060–5.
- O’Driscoll, A., J. Daugelaite, and R. D. Sleator (2013). “‘Big data’, Hadoop and cloud computing in genomics”. In: *Journal of Biomedical Informatics* 46.5, pp. 774–781.
- Ongena, M., F. Duby, F. Rossignol, M.-L. Fauconnier, J. Dommes, and P. Thonart (2004). “Stimulation of the lipoxygenase pathway is associated with systemic resistance induced in bean by a nonpathogenic *Pseudomonas* strain.” In: *Molecular plant-microbe interactions : MPMI* 17.9, pp. 1009–1018.
- Opanowicz, M., P. Vain, J. Draper, D. Parker, and J. H. Doonan (2008). “Brachypodium distachyon: making hay with a wild grass”. In: *Trends in Plant Science* 13.4, pp. 172–177.
- Overy, D. P., D. P. Enot, K. Tailliant, H. Jenkins, D. Parker, M. Beckmann, and J. Draper (2008). “Explanatory signal interpretation and metabolite identification strategies for nominal mass FIE-MS metabolite fingerprints”. In: *Nat Protoc* 3.3, pp. 471–485. ISSN: 1754-2189. DOI: 10.1038/nprot.2007.512. URL: <http://www.ncbi.nlm.nih.gov/pubmed/18323817>{\%}5Cn<http://www.nature.com/nprot/journal/v3/n3/pdf/nprot.2007.512.pdf>.
- Paris, J. D., K. M. Haen, and B. S. Gill (2000). “Saturation mapping of a gene-rich recombination hot spot region in wheat”. In: *Genetics* 154.2, pp. 823–835.
- Park, C.-H., S. Chen, G. Shirsekar, B. Zhou, C. H. Khang, P. Songkumarn, A. J. Afzal, Y. Ning, R. Wang, M. Bellizzi, B. Valent, and G.-L. Wang (2012). “The *Magnaporthe oryzae* effector AvrPiz-t targets the RING E3 ubiquitin ligase

-
- APIP6 to suppress pathogen-associated molecular pattern-triggered immunity in rice.” In: *The Plant cell* 24.11, pp. 4748–62.
- Parker, D. (2006). “A metabolomic approach to investigate host pathogen interactions”. PhD thesis. Aberystwyth University.
- Parker, D., M. Beckmann, D. P. Enot, D. P. Overy, Z. C. Rios, M. Gilbert, N. Talbot, and J. Draper (2008). “Rice blast infection of *Brachypodium distachyon* as a model system to study dynamic host/pathogen interactions.” In: *Nature protocols* 3.3, pp. 435–445.
- Parker, D., M. Beckmann, H. Zubair, D. P. Enot, Z. Caracuel-Rios, D. P. Overy, S. Snowdon, N. J. Talbot, and J. Draper (2009). “Metabolomic analysis reveals a common pattern of metabolic re-programming during invasion of three host plant species by *Magnaporthe grisea*”. In: *Plant Journal* 59.5, pp. 723–737.
- Pedrioli, P. G. a., J. K. Eng, R. Hubley, M. Vogelzang, E. W. Deutsch, B. Raught, B. Pratt, E. Nilsson, R. H. Angeletti, R. Apweiler, K. Cheung, C. E. Costello, H. Hermjakob, S. Huang, R. K. Julian, E. Kapp, M. E. McComb, S. G. Oliver, G. Omenn, N. W. Paton, R. Simpson, R. Smith, C. F. Taylor, W. Zhu, and R. Aebersold (2004). “A common open representation of mass spectrometry data and its application to proteomics research.” In: *Nature biotechnology* 22.11, pp. 1459–1466.
- Peressotti, E., E. Duchêne, D. Merdinoglu, and P. Mestre (2011). “A semi-automatic non-destructive method to quantify grapevine downy mildew sporulation”. In: *Journal of Microbiological Methods* 84.2, pp. 265–271.
- Phadikar, S. and J. Sil (2008). “Rice disease identification using pattern recognition techniques”. In: *2008 11th International Conference on Computer and Information Technology* Iccit, pp. 25–27.
- Pieterse, C. M., D. Van der Does, C. Zamioudis, A. Leon-Reyes, and S. C. Van Wees (2012). “Hormonal Modulation of Plant Immunity”. In: *Annual Review of Cell and Developmental Biology* 28.1, pp. 489–521.

-
- Pignocchi, C. and C. H. Foyer (2003). “Apoplastic ascorbate metabolism and its role in the regulation of cell signalling”. In: *Current Opinion in Plant Biology* 6.4, pp. 379–389.
- Poland, J. A., P. J. Balint-Kurti, R. J. Wisser, R. C. Pratt, and R. J. Nelson (2009). “Shades of gray: the world of quantitative disease resistance”. In: *Trends in Plant Science* 14.1, pp. 21–29.
- Prince, J. T. and E. M. Marcotte (2006). “Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping”. In: *Analytical Chemistry* 78.17, pp. 6140–6152.
- Pritchard, L. and P. Birch (2011). “A systems biology perspective on plant-microbe interactions: Biochemical and structural targets of pathogen effectors”. In: *Plant Science* 180.4, pp. 584–603.
- Rapaport, F., R. Khanin, Y. Liang, M. Pirun, A. Krek, P. Zumbo, C. E. Mason, N. D. Socci, and D. Betel (2013). “Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data.” In: *Genome Biol* 14.9, R95.
- Roden, L. C. and R. A. Ingle (2009). “Lights, rhythms, infection: the role of light and the circadian clock in determining the outcome of plant-pathogen interactions.” In: *The Plant cell* 21.9, pp. 2546–52.
- Rolland, F., E. Baena-Gonzalez, and J. Sheen (2006). “SUGAR SENSING AND SIGNALING IN PLANTS: Conserved and Novel Mechanisms”. In: *Annual Review of Plant Biology* 57.1, pp. 675–709.
- Routledge, a. P. M., G. Shelley, J. V. Smith, N. J. Talbot, J. Draper, and L. a. J. Mur (2004). “Magnaporthe grisea interactions with the model grass *Brachypodium distachyon* closely resemble those with rice (*Oryza sativa*)”. In: *Molecular Plant Pathology* 5.4, pp. 253–265.
- Saeyns, Y., T. Abeel, and Y. Peer (2008). “Robust Feature Selection Using Ensemble Feature Selection Techniques”. In: *European conference on Machine Learning and Knowledge Discovery in Databases* 5212, pp. 313–325.

-
- Sandri, M. and P. Zuccolotto (2006). “Variable Selection using Random Forests”. In: *Pattern Recognitions Letters* 31 31.14, pp. 2225–2236.
- Sattelmacher, B. (2001). “The apoplast and its significance for plant mineral nutrition”. In: *New Phytologist* 149.2, pp. 167–192.
- Schaffrath, U, F Mauch, E Freydl, P Schweizer, and R Dudler (2000). “Constitutive expression of the defense-related Rir1b gene in transgenic rice plants confers enhanced resistance to the rice blast fungus *Magnaporthe grisea*.” In: *Plant molecular biology* 43.1, pp. 59–66.
- Schönherr, J. (2006). “Characterization of aqueous pores in plant cuticles and permeation of ionic solutes”. In: *Journal of Experimental Botany* 57.11, pp. 2471–2491.
- Scott, I. M., W Lin, M Liakata, J. E. Wood, C. P. Vermeer, D Allaway, J. L. Ward, J Draper, M. H. Beale, D. I. Corol, J. M. Baker, and R. D. King (2013). “Merits of random forests emerge in evaluation of chemometric classifiers by external validation”. In: *Analytica Chimica Acta* 801, pp. 22–33.
- Sesma, A. and A. E. Osbourn (2004). “The rice leaf blast pathogen undergoes developmental processes typical of root-infecting fungi.” In: *Nature* 431.7008, pp. 582–586.
- Smith, C. A., E. J. Want, G. O’Maille, R. Abagyan, and G. Siuzdak (2006). “XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification”. In: *Analytical Chemistry* 78.3, pp. 779–787.
- Smith, R., D. Ventura, and J. T. Prince (2013). “Controlling for confounding variables in MS-omics protocol: Why modularity matters”. In: *Briefings in Bioinformatics* 15.5, pp. 768–770.
- Soanes, D. M., A. Chakrabarti, K. H. Paszkiewicz, A. L. Dawe, and N. J. Talbot (2012). “Genome-wide transcriptional profiling of appressorium development by the rice blast fungus *Magnaporthe oryzae*”. In: *PLoS Pathogens* 8.2, e1002514.

-
- Sokolova, M. and G. Lapalme (2009). “A systematic analysis of performance measures for classification tasks”. In: *Information Processing and Management* 45.4, pp. 427–437.
- Solfanelli, C., A. Poggi, E. Loreti, A. Alpi, P. Perata, A. Biotechnology, C. Nazionale, and S. Anna (2006). “Sucrose-Specific Induction of the Anthocyanin Biosynthetic Pathway in Arabidopsis”. In: *Society* 140.February, pp. 637–646.
- Solomon, R. J., T. Kar-chun, and R. P. Oliver (2003). “The nutrient supply of pathogenic fungi ; a fertile field for study”. In: *Molecular Plant Pathology* 4, pp. 203–210.
- Southam, A. D., T. G. Payne, H. J. Cooper, T. N. Arvanitis, and M. R. Viant (2007). “Dynamic range and mass accuracy of wide-scan direct infusion nano-electrospray fourier transform ion cyclotron resonance mass spectrometry-based metabolomics increased by the spectral stitching method”. In: *Analytical Chemistry* 79.12, pp. 4595–4602.
- Sozzani, R., W. Busch, E. P. Spalding, and P. N. Benfey (2014). “Advanced imaging techniques for the study of plant growth and development”. In: *Trends in Plant Science* 19.5, pp. 304–310.
- Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, and B. L. Ebert (2005). “Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.43, pp. 15545–15550.
- Swarbrick, P. J., P. Schulze-Lefert, and J. D. Scholes (2006). “Metabolic consequences of susceptibility and resistance (race-specific and broad-spectrum) in barley leaves challenged with powdery mildew”. In: *Plant, Cell and Environment* 29.6, pp. 1061–1076.
- Talbot, N. J. (1995). “Having a blast: exploring the pathogenicity of *Magnaporthe grisea*”. In: *Trends in Microbiology* 3.1, pp. 9–16.

-
- Tan, S. and S. Wu (2012). “Genome wide analysis of nucleotide-binding site disease resistance genes in *Brachypodium distachyon*”. In: *Comparative and Functional Genomics 2012*, pp. 1–12.
- Tang, F., C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao, and M. A. Surani (2009). “mRNA-Seq whole-transcriptome analysis of a single cell”. In: *Nature Methods* 6.5, pp. 377–382.
- Tanksley, S. D. (1993). “Mapping polygenes”. In:
- Tautenhahn, R., C. Böttcher, and S. Neumann (2008). “Highly sensitive feature detection for high resolution LC/MS”. In: *BMC Bioinformatics* 9.1, p. 504.
- Tautenhahn, R., G. J. Patti, D. Rinehart, and G. Siuzdak (2012). “XCMS online: A web-based platform to process untargeted metabolomic data”. In: *Analytical Chemistry* 84.11, pp. 5035–5039.
- Tavernier, V., S. Cadiou, K. Pageau, R. Laugé, M. Reisdorf-Cren, T. Langin, and C. Masclaux-Daubresse (2007). “The plant nitrogen mobilization promoted by *Colletotrichum lindemuthianum* in *Phaseolus* leaves depends on fungus pathogenicity”. In: *Journal of Experimental Botany* 58.12, pp. 3351–3360.
- Thimm, O., O. Bläsing, Y. Gibon, A. Nagel, S. Meyer, P. Krüger, J. Selbig, L. a. Müller, S. Y. Rhee, and M. Stitt (2004). “MAPMAN: A user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes”. In: *Plant Journal* 37.6, pp. 914–939.
- Tolstikov, V. V. and O. Fiehn (2002). “Analysis of Highly Polar Compounds of Plant Origin : Combination of Hydrophilic Interaction Chromatography and Electrospray Ion Trap Mass Spectrometry”. In: 307, pp. 298–307.
- Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter (2012). “Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.” In: *Nature protocols* 7.3, pp. 562–78.

-
- Trapnell, C., D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter (2013). “Differential analysis of gene regulation at transcript resolution with RNA-seq.” In: *Nature biotechnology* 31.1, pp. 46–53.
- Trouvelot, S., M.-C. Héloir, B. Poinssot, A. Gauthier, F. Paris, C. Guillier, M. Combier, L. Trdá, X. Daire, and M. Adrian (2014). “Carbohydrates in plant immunity and plant protection: roles and potential application as foliar sprays.” In: *Frontiers in plant science* 5.November, p. 592.
- Tucker, S. L. and N. J. Talbot (2001). “Surface Attachment and Pre-Penetration Stage Development by Plant Pathogenic Fungi”. In: *Annual review of phytopathology* 39, pp. 385–417.
- Van den Berg, R. a., H. C. J. Hoefsloot, J. a. Westerhuis, A. K. Smilde, and M. J. van der Werf (2006). “Centering, scaling, and transformations: improving the biological information content of metabolomics data.” In: *BMC genomics* 7, p. 142.
- Veneault-Fourrey, C., M. Barooah, M. Egan, G. Wakley, and N. J. Talbot (2006). “Autophagic fungal cell death is necessary for infection by the rice blast fungus.” In: *Science* 312.5773, pp. 580–583.
- Vogel, J. P., D. F. Garvin, O. M. Leong, and D. M. Hayden (2006). “Agrobacterium-mediated transformation and inbred line development in the model grass *Brachypodium distachyon*”. In: *Plant Cell, Tissue and Organ Culture* 84.2, pp. 199–211.
- Voll, L. (2011). “Common motifs in the response of cereal primary metabolism to fungal pathogens are not based on similar transcriptional reprogramming”. In: *Frontiers in Plant Science* 2.August, pp. 1–17.
- Walz, A., I. Zingen-Sell, M. Loeffler, and M. Sauer (2008). “Expression of an oxalate oxidase gene in tomato and severity of disease caused by *Botrytis cinerea* and *Sclerotinia sclerotiorum*”. In: *Plant Pathology* 57.3, pp. 453–458.
- Wang, Z. Y., D. M. Soanes, M. J. Kershaw, and N. J. Talbot (2007). “Functional analysis of lipid metabolism in the rice blast fungus *Magnaporthe grisea* re-

-
- veals a role for peroxisomal [beta]-oxidation in appressorium-mediated plant infection". In: *Mol. Plant Microbe Interact.* 20.5, pp. 475–491.
- Wang, Z., M. Gerstein, and M. Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics." In: *Nature reviews. Genetics* 10.1, pp. 57–63.
- Want, E. J., I. D. Wilson, H. Gika, G. Theodoridis, R. S. Plumb, J. Shockcor, E. Holmes, and J. K. Nicholson (2010). "Global metabolic profiling procedures for urine using UPLC-MS." In: *Nature protocols* 5.6, pp. 1005–1018.
- Ward, J. L., S. Forcat, M. Beckmann, M. Bennett, S. J. Miller, J. M. Baker, N. D. Hawkins, C. P. Vermeer, C. Lu, W. Lin, W. M. Truman, M. H. Beale, J. Draper, J. W. Mansfield, and M. Grant (2010). "The metabolic transition during disease following infection of *Arabidopsis thaliana* by *Pseudomonas syringae* pv. tomato". In: *Plant Journal* 63.3, pp. 443–457.
- Weber, R. J. M., A. D. Southam, U. Sommer, and M. R. Viant (2011). "Characterization of isotopic abundance measurements in high resolution FT-ICR and Orbitrap mass spectra for improved confidence of metabolite identification". In: *Analytical Chemistry* 83.10, pp. 3737–3743.
- Weitbrecht, K., K. Müller, and G. Leubner-Metzger (2011). "First off the mark: Early seed germination". In: *Journal of Experimental Botany* 62.10, pp. 3289–3309.
- Wishart, D. S. (2008). "Quantitative metabolomics using NMR". In: *TrAC - Trends in Analytical Chemistry* 27.3, pp. 228–237.
- Wolfe, C. J., I. S. Kohane, and A. J. Butte (2005). "Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks." In: *BMC bioinformatics* 6, p. 227.
- Xian, F., C. L. Hendrickson, and A. G. Marshall (2012). "High resolution mass spectrometry". In: *Analytical Chemistry* 84.2, pp. 708–719.
- Xu, J. R. and J. E. Hamer (1996). "MAP kinase and cAMP signaling regulate infection structure formation and pathogenic growth in the rice blast fungus *Magnaporthe grisea*". In: *Genes Dev* 10, pp. 2696–2706.

-
- Yaeno, T., O. Matsuda, and K. Iba (2004). “Role of chloroplast trienoic fatty acids in plant disease defense responses”. In: *Plant Journal* 40.6, pp. 931–941.
- Yara, A., T. Yaeno, J. L. Montillet, M. Hasegawa, S. Seo, K. Kusumi, and K. Iba (2008). “Enhancement of disease resistance to *Magnaporthe grisea* in rice by accumulation of hydroxy linoleic acid”. In: *Biochemical and Biophysical Research Communications* 370.2, pp. 344–347.
- Yuan, J. S., D. W. Galbraith, S. Y. Dai, P. Griffin, and C. N. Stewart (2008). “Plant systems biology comes of age”. In: *Trends in Plant Science* 13.4, pp. 165–171.
- Zhang, H., J. E. Fritts, and S. A. Goldman (2008). “Image segmentation evaluation: A survey of unsupervised methods”. In: *Computer Vision and Image Understanding* 110.2, pp. 260–280.
- Zrenner, R., M. Stitt, U. Sonnewald, and R. Boldt (2006). “Pyrimidine and Purine Biosynthesis and Degradation in Plants”. In: *Annual Review of Plant Biology* 57.1, pp. 805–836.
- Zubair, H. (2014). “Metabolic reprogramming of *Brachypodium distachyon* challenged with *Magnaporthe grisea*”. PhD thesis. Aberystwyth University.