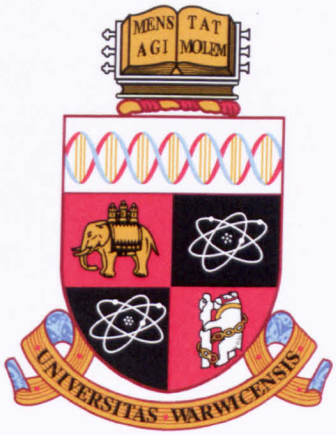


Using systems biology approaches to elucidate gene regulatory networks controlling the plant defence response

by

Steven John Kiddle

Thesis

Submitted to the University of Warwick
for the degree of

Doctor of Philosophy
in Systems Biology

Department of Systems Biology

June 2012

THE UNIVERSITY OF
WARWICK

Contents

List of Tables

List of Figures

Acknowledgments

I would like to gratefully acknowledge the following people, without whom this PhD would not have been possible. This is not comprehensive, and sorry to the many people I will inevitably have forgotten.

Many thanks to Katherine Denby and Sach Mukherjee for their enthusiastic, supportive and ambitious supervision. Many thanks to my family: Pete, Sue and Emma, whose support has been invaluable. Many thanks to my bandmates: Laurie Ainley, Rowan Gifford, Stephen Henry and Nick Dugdale, without whom I may not have even considered a career in research. Many thanks to the Systems and MOAC cohorts for fun times, especially to Tyson and Hickman. Many many thanks to the two postdocs, Claire Hill and Volkan Cevik, that could coax biological results out of a mathematician! You both have the patience of a saint, and I will miss working with both of you. Many thanks to my advisory committee: Isabelle Carre, Jim Beynon and Julia Brettschneider. In no particular order, many thanks to: Peijun Zhang, Alison Jackson, Oliver Windram, Stuart McHattie, Nicki Adams, Zennia Paniwnyk, Youn-Sung Kim, Steve Hill, Chris Oates, Chris Penfold, Dafyd Jenkins, Justyna Prusinska, Emma Cooke, Jo Rhodes, Alex Jironken, Jens Steinbrenner, Daniel Tome, Laura Lewis, Sascha Ott, Vicky Buchanan-Wollaston and all the rest of PRESTA. Many thanks to Sarah Harvey and Alison Eyres for lab chat.

Dedicated to Carl Blakey, may he rest in peace.

This thesis was typeset with LaTeX 2_ε[1] by the author.

[1] LaTeX 2_ε is an extension of LaTeX. LaTeX is a collection of macros for TeX. TeX is a trademark of the American Mathematical Society. The style package *warwickthesis* was used.

Declarations

This thesis is presented in accordance with the regulations for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. The work in this thesis has been undertaken by myself except where otherwise stated.

PAGINATED
BLANK PAGES
ARE SCANNED AS
FOUND IN
ORIGINAL
THESIS

NO
INFORMATION
MISSING

Abstract

Transcriptional regulation controlling pathogen-responsive gene expression in Arabidopsis is believed to underlie the plant defence response, which confers partial immunity of Arabidopsis to infection by *Botrytis cinerea*. In this thesis networks of transcriptional regulation mediating the defence response are studied in various ways.

First transcriptional regulation was predicted for all genes differentially expressed during *B. cinerea* infection by development of a novel clustering approach, Temporal Clustering by Affinity Propagation (TCAP). This approach finds groups of genes whose expression profile time series have strong time-delayed correlation, a measure that is demonstrated to be more predictive of transcriptional regulation than conventionally used similarity measures. TCAP predicts the known regulation of *GI* by *LHY*, and co-clusters *ORA59* and some of its downstream targets. Predicted novel regulators of pathogen-responsive gene expression were then studied in a reverse genetics screen, which discovered several novel but weakly altered susceptibility phenotypes. Comparison of predicted targets to known targets was complicated by the sparsity of mutant versus wildtype gene expression experiments performed during *B. cinerea* infections in the literature.

To explore the context-dependence of transcriptional regulation, evidence of transcriptional regulation in different contexts was collected. This was compiled to generate a qualitative model of transcriptional regulation during the defence response. This model was validated and extended by experimental analysis of transcription factor-promoter binding in Yeast and transcriptional activation *in planta*. Comparative transcriptomics showed that downstream genes of some of these regulators — TGA3, ARF2, ERF1 and ANAC072 — are over-represented in the list of genes differentially expressed during *B. cinerea* infection, which is consistent with these targets being regulated by them during *B. cinerea* infection.

Finally this qualitative model was used as prior information and was used along with gene expression time series to infer quantitative models of the gene regulatory network mediating the defence response. Some known regulation was predicted, and additionally ANAC055 was predicted to be a central regulator of pathogen-responsive gene expression.

Resulting publications and those in preparation

- S J Kiddle, O P F Windram, S McHattie, A Mead, J Beynon, V Buchanan-Wollaston, K J Denby and S Mukherjee (2010) *Temporal clustering by affinity propagation reveals transcriptional modules in Arabidopsis thaliana.* Bioinformatics, Vol. 26, No. 3, pp. 355-362.

- E Breeze, E Harrison, S McHattie, L Hughes, R Hickman, C Hill, S Kiddle, Y-S Kim, C A Penfold, D Jenkins, C Zhang, K Morris, C Jenner, S Jackson, B Thomas, A Tabrett, R Legaie, J D Moore, D L Wild, S Ott, D Rand, J Beynon, K Denby, A Mead and Vicky Buchanan-Wollaston. *High-Resolution Temporal Profiling of Transcripts during Arabidopsis Leaf Senescence Reveals a Distinct Chronology of Processes and Regulation* Plant Cell, Vol. 23, pp. 873-894.

- O Windram, P Madhou, S J Kiddle, C Hill, R Hickman, E Cooke, S McHattie, D Jenkins, C Penfold, Y-S Kim, L Baxter, C Zhang, A Tabrett, J Moore, D L Wild, A Mead, D Rand, J Beynon, S Ott, V Buchanan-Wollaston and K Denby. *Defence against Botrytis cinerea: chronology and regulation deciphered by high-resolution temporal transcriptomic analysis.* In preparation.

- It is likely that Chapter 3 will be adapted into paper form.

Talks and poster presentations

- Demonstrated my method, TCAP, to European postdocs and PhD students. Plant systems biology summer school, Palazzo Pesaro-Papafava, Venice. September 2011.

- Poster presented. 10th conference on Machine Learning in Systems Biology, Edinburgh University, October 2010.

- Poster presented. Systems Biology: Global regulation of gene expression, Cold Spring Harbor Laboratory, USA, March 2010.

- Poster and 20-minute talk. 5th Utrecht University PhD summer school in environmental signaling, Netherlands, August 2009.

TCAP Software

The TCAP software package can be downloaded from:

http://www.wsbc.warwick.ac.uk/stevenkiddle/tcap.html

Abbreviations

0.0.1 Acronyms

ABA = Abscisic acid

ANOVA = ANalysis Of VAriance

AP = Affinity Propagation

APPLES = Analysis of Plant Promoter Linked Elements

AP2-ERE BP = APETALA2-Ethylene Responsive Element Binding Proteins (a transcription factor family)

ARF = Auxin Response Factor (a transcription factor family)

bp = base-pairs

BN = Bayesian Networks

CATMA = Complete Arabidopsis Transcriptome Microarray

ChIP = Chromatin Immuno-Precipitation

cDNA = complementary Deoxyribose Nucleic Acid

DAMP = Damage (or Danger) Associated Molecular Pattern

DNA = Deoxyribonucleic Acid

EM = Expectation-Maximisation

EMS = Ethylmethanesulfonate

EMSA = Electrophoretic Mobility Shift Assay

ERE = Ethylene Responsive Element

ET = Ethylene

GAL4 AD = The activation domain of the DNA-binding Yeast protein GAL4

GFP = Green Fluorescent Protein

GO = Gene Ontology

GRN = Gene Regulatory Network

GUS = beta-glucuronidase

hpi = hours post infection

JA = Jasmonic acid

JAZ = Jasmonate ZIM-domain

MAANOVA = MicroArray ANalysis Of VAriance

MAPK = Mitogen Activated Protein Kinase

MAST = Motif Alignment and Search Tool

MeJA = Methyl Jasmonate

MEME = Multiple EM for Motif Elicitation

mRNA = messenger Ribonucleic Acid

MS = Murashige and Skoog medium

MWW = Mann-Whitney-Wilcoxon

NAC = NAM/ATAF/CUC transcription factor family

NINJA = Novel Interactor of JAZ

ODEs = Ordinary Differential Equations

OG = Oligogalacturnide

PAM = Partitioning Around Medoids

PAMP = Pathogen Associated Molecular Pattern

PCC = Pearson's Correlation Coefficient

PCR = Polymerase Chain Reaction

PEG = Polyethylene Glycol

PGIPs = Polygalacturonase Inhibiting Proteins

PRESTA = Plant Response to Environmental Stress in Arabidopsis

PSSMs = Position Specific Scoring Matrices

PTI = Pathogen Triggered Immunity

QDR = Quantitative Disease Resistance

QTL = Quantitative Trait Loci

RNA = Ribonucleic Acid

ROC = Receiver Operator Characteristic

RT-PCR = Reverse Transcriptase-Polymerase Chain Reaction

SA = Salicylic acid

SD-L = SD minus Leucine

SD-T = SD minus Tryptophan

SD-LT = SD minus Leucine and Tryptophan

SD-LTH = SD minus Leucine, Tryptophan and Hisitidine

SSM = State Space Model

TAIL-PCR = Thermal Asymmetric Interlaced-Polymerase Chain Reaction

TAIR = The Arabidopsis Information Resource

TCP = TB1/CYC/PCF transcription factor family

TCAP = Temporal Clustering by Affinity Propagation

TGA = Transcription factor family with the characteristic binding sequence TGACG

TF = Transcription Factor

Ti = Tumor inducing

T-DNA = Transfer-Deoxyribonucleic Acid

TSS = Transcriptional Start Site

UTR = Untranslated Region

VBSSM = Variational Bayesian State Space Model

WRKY = Transcription factor family with the characteristic amino acid sequence WRKYGQK

Y1H = Yeast-one-hybrid

3AT = 3-Amino-1,2,4-triazole

0.0.2 Gene/Protein/Knockout naming conventions

In this report the standard Arabidopsis naming conventions are used. Genes appear in italicised capitals (e.g. *WRKY33*), whereas proteins appear as non-italicised capitals (e.g. WRKY33). Gene names in lower case represent knockout mutants of that gene (e.g. *wrky33*), i.e. plants that do not express that gene and so cannot make that protein.

0.0.3 Gene reference

The table provided on the next page can be used to translate between the gene name as given in the literature and the corresponding TAIR (The Arabidopsis Information Resource) unique identifier.

Table 1: Gene name reference table

Gene Name	AGI	Gene Name	AGI	Gene Name	AGI
ABI4	At2g40220	*ERF6*	At4g17490	*PGIP1*	At5g06860
ANAC019	At1g52890	*ERF71*	At2g47520	*PGIP2*	At5g06870
ANAC055	At3g15500	*GI*	At1g22770	*PIF7*	At5g61270
ANAC072	At4g27410	*HSFC*	At3g24520	*PRR7*	At5g02810
ANAC092	At5g39610	*HUB1*	At2g44950	*PRR9*	At2g46790
ARF1	At1g59750	*JAR1*	At2g46370	*RAP2.6*	At1g43160
ARF2	At5g62000	*JAZ1*	At1g19180	*RAP2.6L*	At5g13330
ARF5	At1g19850	*JAZ6*	At1g72450	*RGL1*	At1g66350
ATAF1	At1g01720	*JAZ7*	At2g34600	*RST1*	At3g27670
ATERF1	At4g17500	*JAZ8*	At1g30135	*SDG8*	At1g77300
ATG18a	At3g62770	*JAZ9*	At1g70700	*SPL4*	At1g53160
ATML1	At4g21750	*JAZ10*	At5g13220	*SYD*	At2g28290
BAP1	At3g61190	*JAZ12*	At5g20900	*TCP1*	At1g67260
BDL	At1g04550	*LACS2*	At1g49430	*TCP14*	At3g47620
BIK1	At2g39660	*LBD41*	At3g02550	*TCP15*	At1g69690
BES1	At1g19350	*LHY*	At1g01060	*TCP16*	At3g45150
BHLH100	At2g41240	*MAPKKKα*	At1g63700	*TCP20*	At3g27010
CAMTA3	At2g22300	*MEKK1*	At4g08500	*TCP3*	At1g53230
CCA1	At2g46830	*MKK4*	At1g51660	*TCP4*	At3g15030
CDF	At5g62430	*MKK5*	At3g21220	*TCP8*	At1g58100
CERK1	At3g21630	*MKS1*	At3g18690	*TGA2*	At5g06950
CHE	At5g08330	*MPK3*	At3g45640	*TGA3*	At1g22070
CHIB	At3g12500	*MPK4*	At4g01370	*TGA5*	At5g06960
COI1	At2g39940	*MPK6*	At2g43790	*TGA6*	At3g12250
EFR	At5g20480	*MYB108*	At3g06490	*TOC1*	At5g61380
DREB2A	At5g05410	*MYB15*	At3g23250	*TOPLESS*	At1g15750
EIN2	At5g03280	*MYB49*	At5g54230	*WAK1*	At1g21250
EIN3	At3g20770	*MYBL2*	At1g71030	*WRKY20*	At4g26640
ERF1	At3g23240	*MYC2*	At1g32640	*WRKY25*	At2g30250
ERF10	At1g03800	*NF-YB5*	At2g47810	*WRKY31*	At4g22070
ERF11	At1g28370	*NINJA*	At4g28910	*WRKY33*	At2g38470
ERF13	At2g44840	*NPR1*	At1g64280	*WRKY39*	At3g04670
ERF14	At1g04370	*NUB*	At1g13400	*WRKY48*	At5g49520
ERF15	At2g31230	*OCP3*	At5g11270	*WRKY60*	At2g25000
ERF2	At5g47220	*ORA59*	At1g06160	*WRKY70*	At3g56400
ERF4	At3g15210	*PAD3*	At3g26830	*ZFAR1*	At2g40140
ERF5	At5g47230	*PDF1.2*	At5g44420		

Chapter 1

Introduction

1.1 Motivation

The human food chain ultimately depends, either directly or through other animals, on the availability of plant biomass. The availability of edible plant biomass is therefore a strong determinant of the availability of the food that humans consume. Because of supply and demand, reductions in the quantity of food that can be produced, or increases in our demand for food, raise the price of food. Our demand for food is increasing because of the size of our population, which has been growing since 1350 (Biraben, 1980). Our ability to produce large quantities of food can depend on many factors such as the availability of high-yield crop varieties, the cost of fertilisers, current climate conditions and the prevalence of plant diseases. For example, improvements in these factors have been credited with the 'green revolution' that led to substantial increases in the production of food between 1960-2000 (reviewed in Evenson and Gollin, 2003).

In the United Nations Millennium Development Goals, goal one is to "end poverty and hunger". While the availability of food is only one of many factors affecting world hunger, it has been shown to be a key cause (reviewed in Bowbrick, 1986). Increased production of food can also have societal impacts, such as the increase in food security (reviewed in Rosegrant and Cline, 2003). Food security is a rising concern, due to the increasing world population and changing climatic conditions. For these reasons it is important to develop new methods that can increase the yield of crops.

The ability of crops to resist infection by pathogens is a major factor affecting their yield. This is most noticeable when a plant is introduced to a disease to which it has not evolved resistance to. For example, the Irish Potato Famine was caused by the introduction of the oomycete pathogen *Phytopthora infestans* to Ireland (reviewed

in Ristain, 2002).

1.2 Disease resistance in plants

Animals have an adaptive immune system that allows them to develop immunity by exposure to most diseases; successful infections occur when pathogens evolve new mechanisms to evade detection. The immune system of plants, which is referred to as the defence response, is innate, i.e. resistance to a pathogen evolves over many generations rather than during the course of each plant's life (reviewed in Jones and Dangl, 2006). This means that the key to decreasing disease related yield loss in crops is to use varieties that are resistant to the pathogens that they are likely to encounter.

The resistance of crop varieties can be improved by selective breeding, but this is time consuming and is limited by the genetic diversity of the source varieties (reviewed in Roane, 1973). Greater knowledge of the molecular mechanisms that confer resistance to pathogens can be used to develop resistant varieties more rapidly. A similar approach has been successfully used to develop greater tolerance to flooding in rice. The isolation of a specific gene responsible for flood-resistance in a wild variety allowed the rapid development of flood-resistant versions of high-yield rice varieties (Xu et al., 2006a). This was achieved by marker assisted selection (reviewed in Ribault and Hoisington, 1998).

1.2.1 The plant defence response

Study of the defence response of plants to infection by pathogens has revealed two broad defence mechanisms: Pathogen Triggered Immunity (PTI) and Effector Triggered Immunity (reviewed in Chisholm et al., 2006; Jones and Dangl, 2006). PTI results from the recognition of elicitors of the defence response (reviewed in Boller and Feli, 2009). Elicitors of the defence response that are recognisably of pathogenic (i.e. non-self) origin are called Pathogen Associated Molecular Patterns (PAMPs; reviewed in Janeway and Medzhito, 2002); whereas endogenous (i.e. self) elicitors are called Danger/Damage Associated Molecular Patterns (DAMPs; reviewed in Matzinger, 2007).

PAMPs are often highly conserved protein domains required by the pathogen for pathogenicity. A good example of a PAMP is bacterial flagellin, a protein required for bacterial mobility. Recognition by the plant *Arabidopsis thaliana* (Arabidopsis) of a peptide representing the most conserved domain of flagellin, flg22, results in activation of the expression of the Arabidopsis gene *WRKY29*; expression of *WRKY29*

leads to resistance of Arabidopsis to infection by bacterial pathogens such as *Pseudomonas syringae* (Asai et al., 2001). i.e. recognition of the PAMP flg22 by Arabidopsis leads to PTI, which protects it from infection by *P. syringae*. Some PAMPs are not proteins; for example, chitin is a polymer of N-acetylglucosamine and forms the major component of fungal cell walls (reviewed in Bartnicki-Garcia, 1968). In a study by Ramonell et al. (2005) recognition of chitin by Arabidopsis led to the differential expression of many genes; knockout mutants of 9 of these genes were screened for altered susceptibility to the fungal pathogen *Erysiphe cichoracearum*. Three of the mutants were found to be more susceptible to infection, suggesting the differential expression of the corresponding genes was responsible for the activation of PTI of Arabidopsis to infection by *E. cichoracearum*.

DAMPs are molecules produced by the plant that are damaged in characteristic ways by pathogens and therefore elicit the defence response (reviewed in Matzinger, 2007). For example, Cervone et al. (1989) have shown that oligogalacturinides (OGs) are DAMPs formed by the interaction of a *Botrytis cinerea* protein, Polygalaturnase, with the plant cell wall. Ferrari et al. (2007) have shown that Arabidopsis treated with OGs are more resistant to infection by *B. cinerea*, showing that OGs act as an elicitor of the plant defence response.

PTI resulting from the detection of PAMPs or DAMPs can successfully lead to either partial or full immunity of plants to infection by certain pathogens. However, some pathogens have evolved to overcome PTI by secreting effector proteins into the plant; these proteins typically block recognition of PAMPs by the plant (reviewed in Abramovitch et al., 2006). Susceptibility of a plant to infection by a pathogen, which secretes effectors to blocks its recognition by the plant immune system, has been called Effector Triggered Susceptibility (reviewed in Chisholm et al., 2006; Jones and Dangl, 2006). For example, He et al. (2006) have shown that the bacterial pathogen *P. syringae* secretes an effector called AvrPto into the plant, which suppresses the detection of flagellin. In response, some plants have evolved Effector Triggered Immunity, which corresponds to the detection by the plant of these pathogen secreted effector proteins (reviewed in Chisholm et al., 2006; Jones and Dangl, 2006).

Plant pathogens have a range of lifestyles; biotrophic pathogens reduce yield by developing a parasitic relationship, whereas necrotrophic pathogens necrotise/kill plant tissue to extract nutrients. Whereas PTI has been demonstrated for both biotrophic and necrotrophic pathogens, secretion of effectors into the plant by necro- trophic pathogens to suppress PTI has not yet been found. One defence response deployed

against biotrophic pathogens is localised cell death; Govrin and Levine (2000) have shown that cell death is an effective defence response against biotrophic pathogens like *P. syringae*, but is beneficial to necrotrophic pathogens like *B. cinerea*. The need for different responses to different types of pathogen is reflected by the use of different hormones in defence signalling (reviewed in McDowell and Dangl, 2000). Glazebrook et al. (2003) have shown that the balance between defence responses to biotrophic and necrotrophic pathogens partly relies on crosstalk between the salicylic acid (SA) and jasmonic acid (JA) hormone pathways, which are responsible for differential regulation of the expression of partially overlapping lists of genes. In general, defence responses to biotrophic pathogens rely on the SA pathway; whereas necrotrophic pathogens rely on the JA and ethylene (ET) pathways, which repress the SA pathway and therefore suppress the cell death response (reviewed in McDowell and Dangl, 2000; Glazebrook, 2005; Pieterse et al., 2009).

Defence responses against biotrophic pathogens are usually reported as qualitative in nature; the defence response against necrotrophic pathogens seems to be more variable, with different levels of partial resistance observed. This has been termed Quantitative Disease Resistance (QDR; reviewed in Poland et al., 2008). Historically, resistance to biotrophic pathogens has been more comprehensively studied, which is why this thesis concentrates on resistance to the necrotrophic pathogen *B. cinerea*. The generation of varieties with greater disease-resistance to necrotrophic pathogens is desirable as they could increase the yield of economically important crops. It is hoped that increased knowledge of the defence response of plants to infection by necrotrophic pathogens will allow disease reistant varieties to be developed. In the case of crop resistance to infection by *B. cinerea*, crop varieties with increased resistance are required because the pathogen is rapidly developing resistance to previously effective chemical controls (reviewed in Rosslenbroich and Stuebler, 2000).

1.2.2 The Arabidopsis-Botrytis pathosystem as a tractable experimental model of the plant defence response to infection by necrotrophic pathogens

B. cinerea is a fungal plant pathogen that is able to infect over 200 different plants, including many grapes, vegetables, berries, stone fruits and Arabidopsis (Williamson et al., 2007; Jarvis, 1977; Koch and Slusarenko, 1990). Its effect pre- and post-harvest causes substantial reductions in yield for economically important crops (Droby and Lichter, 2004; Williamson et al., 2007). It is fortunate for researchers that *B. cinerea* can infect the model plant Arabidopsis (Koch and Slusarenko, 1990); while not economically important in itself, Arabidopsis is well studied, experimen-

tally convenient (reviewed in Fink, 1998) and is related to hosts of *B. cinerea* that are important for the human food chain.

Arabidopsis is infected by B. cinerea spores/conidia, which germinate and then form appressoria, described by van Kan (2006) as infection structures that differentiate on the surface and form a penetration peg that breaches the cuticle. The exact timing of penetration of Arabidopsis cells by *B. cinerea* appressoria after attachment of B. cienerea spores/conidia is not known, but conidia have been shown to germinate after six hours in water (Hawker and Hendy, 1963). Additionally, fast growth in B. cienera biomass on Arabidopsis leaves occurs, presumably during spore/conidia germination and hyphae formation (branching infection structure), up until a lag phase which occurs 20-28 hours post infection (hpi). This lag phase appears to correspond to the formation of dark lesions, presumably caused by penetration of plant cells by *B. cinerea*, which then grow (both in fungal biomass and lesion size) from 36 hpi onwards (Windram et al., in preparation).

It is hoped that by studying the defence response of Arabidopsis, resistance mechanisms will be highlighted that could be targeted by future controls, or modified by breeding or genetic engineering in a relevant host to decrease yield losses. The Arabidopsis genome has been sequenced (Bevan and Walsh, 2005), methods exist to disrupt or insert new sequences into its genome (reviewed in Krysan et al., 1999), and microarrays have been produced which can monitor the expression of its genes (Allemeersch et al., 2005). These developments should allow the role of Arabidopsis genes in resistance against infection by *B. cinerea* to be characterised relatively rapidly. This makes the Arabidopsis-Botrytis pathosystem a comparatively tractable model of plant defence response to necrotrophic pathogens. For example, Denby et al. (2004) have shown that quantitative resistance of Arabidopsis to infection by *B. cinerea* varies with Arabidopsis ecotype and *B. cinerea* isolate. Both Denby et al. (2004) and Rowe and Kliebenstein (2008) have shown that quantitative resistance of Arabidopsis to infection by *B. cinerea* is under complex genetic control, suggesting that it is the interaction of a set of genes that controls resistance, i.e. QDR of Arabidopsis to infection by *B. cinerea* is a polygenic trait.

The differential expression of genes appears to be a key aspect of the defence response of Arabidopsis to infection by *B. cinerea*; Ferrari et al. (2007) have shown that 4,813 Arabidopsis genes are differentially expressed 48 hours post infection (hpi) with *B. cinerea* relative to uninfected leaves. i.e. the levels of mRNA corresponding to 4,813 Arabidopsis genes have been found to be statistically significantly different between infected and uninfected leaves. This corresponds to approximately 20% of

the genes which have corresponding probes on the microarrays used. Other studies have found smaller, but partially overlapping, lists of genes differentially expressed in Arabidopsis leaves during *B. cinerea* infection relative to uninfected leaves (for example AbuQamar et al., 2006; Rowe et al., 2010; Mulema and Denby, 2012).

Mutants of some known regulators of gene expression have been shown to have altered susceptibility to infection by *B. cinerea* (15 such mutants are reviewed in Birkenbihl and Somssich, 2011). The number of known regulators of gene expression that have an effect on the defence response, and the scale of *B. cinerea*-responsive changes in Arabidopsis gene expression, suggest the existence of a gene regulatory network (GRN) that underpins the defence response of Arabidopsis to infection by *B. cinerea*. Even though some regulators are known, knowledge of the genes they regulate during *B. cinerea* infection is currently very sparse. If the structure of the GRN underpinning the defence response of Arabidopsis to infection by *B. cinerea* was known, it could be manipulated to increase resistance. For this reason it is highly desirable to increase knowledge of the structure of the GRN underpinning the defence response and to develop models that could predict the effect of genetic perturbations on the ability of Arabidopsis to resist infection by *B. cinerea*. To achieve this, novel modelling and experimental approaches will have to be developed.

1.3 Gene regulatory networks

While the aim of this thesis is to identify and model GRNs underpinning the defence response of Arabidopsis to infection by *B. cinerea*, novel modelling and experimental approaches may also prove to be useful in the study of gene regulation in other contexts. This provides additional motivation because of the importance of gene regulation in many different contexts, both in plants and in other organisms. Gene regulation is the mechanism controlling the spatial and temporal expression of genes; changes in spatial gene expression have been shown to be instrumental in the development of embryos (Davidson and Erwin, 2006), and in the developmental of different cell types from the same original stem cells in plants (for example Espinosa-Sotoa et al. (2004), other examples are reviewed in Pu and Brady (2009)) and animals (reviewed in Graf and Enver, 2009). Temporal changes of gene expression have been found to control developmental (Breeze et al., 2011), seasonal (Aikawa et al., 2010) and diurnal (Locke et al., 2006; McClung, 2008) changes in plants. Gene regulation has also been shown to be critical for the differences between species, for example, many human-specific traits have been linked to the loss of DNA sequences controlling the expression of genes (McLean et al., 2011). This helps to

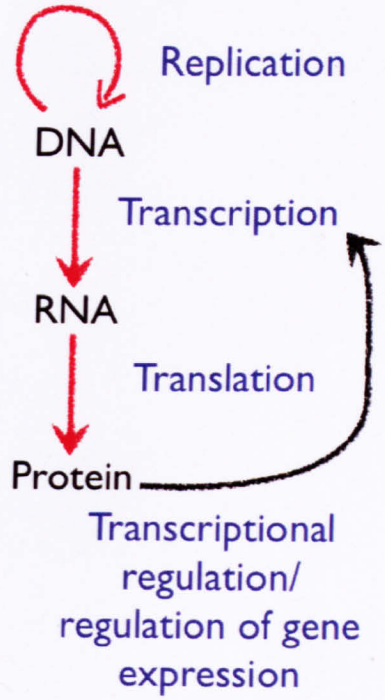
explain why many different species have highly conserved gene-coding sequences, despite highly divergent appearance and lifestyles.

1.3.1 The regulation of gene expression

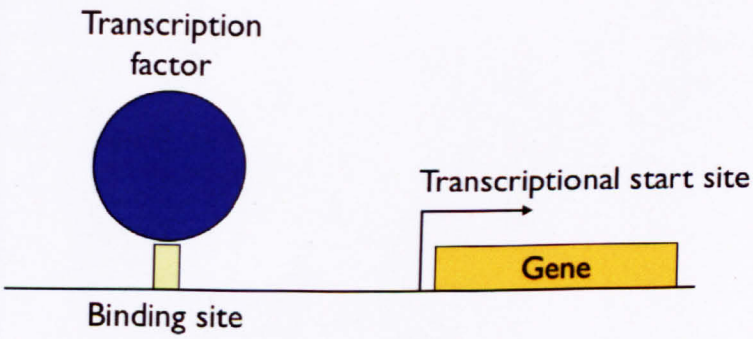
The central dogma of molecular biology is that DNA (Deoxyribose Nucleic Acid) can be transcribed into RNA (Ribose Nucleic Acid), and that RNA can be translated into protein (as illustrated in Figure 1.1(a)). A gene, stored as a sequence of DNA bases, encodes mRNA (messenger RNA) which is produced from the gene by transcription. Transcription is the process of generating mRNA from a DNA template, which is performed by the enzyme RNA polymerase. Gene expression/transcription is regulated by transcription factors (TFs), which are proteins that can affect the transcription rate of genes by binding upstream of the gene and encouraging or blocking the recruitment of RNA polymerase to the transcriptional start site (as illustrated in Figure 1.1(b) and reviewed in Dynan and Tjian (1985)). Such upstream sequences containing binding motifs of TFs are referred to as promoters, as they allow TFs to promote or inhibit transcription, and therefore expression, of the gene downstream of the transcriptional start site. This is called transcriptional regulation and is summarised conceptually in Figure 1.1(a). After transcription, mRNA is degraded (Beelman and Parker, 1995). The concentration of mRNA, which is affected by both the transcription and degradation rate, is often referred to as gene expression. The mRNA in turn encodes a protein, which is produced from the mRNA by translation.

Transcriptional regulation is an important component of regulation within a cell, however it is by no means the only form of regulation. Other types of regulation include: epigenetics, mRNA splicing, translational regulation, post-translational modifications and protein-protein interactions. Many of these are able to regulate transcription indirectly, for example Mazzucotelli et al. (2008) have reviewed post-transcriptional and post-translational regulation controlling transcription in the plant defence response to abiotic stress.

Plant TF families have been reviewed in Meshi and Iwabuchi (1995); they have been shown to form large families that have similar protein structures and DNA-binding specificities. Knowledge of the function of Arabidopsis TFs, which account for approximately 5-10% of the Arabidopsis genome, is currently sparse (reviewed in Riechmann and Ratcliffe, 2000). 'Omic' technologies are showing promise as a way to assign function to plant TFs (reviewed in Mitsuda and Ohme-Takagi, 2009).



(a) Central dogma of molecular biology and the role of transcriptional regulation



(b) Transcriptional regulation

Figure 1.1: The role and regulation of transcription. (a) The central dogma of molecular biology (red arrows) and the role of transcriptional regulation (black arrow). (b) Binding of transcription factors (specific proteins) upstream of the transcriptional start site affect the recruitment of RNA Polymerase II and therefore the rate of mRNA production.

1.3.2 The role of DNA binding and activation/repression domains in transcriptional regulation

TFs are defined to be proteins that can bind to DNA and regulate gene expression. The function of TFs can be understood by studying their DNA-binding specificity and their effect on gene expression in certain contexts.

DNA-binding domains recruit TFs to sequence motifs

TFs were first identified by *in vitro* DNA-binding assays, which also demonstrated the variety of DNA-binding specificity of different DNA-binding domains (reviewed in Mitchell and Tjian, 1989). TFs bind to short DNA sequences called motifs, which are approximately 5–8 base pairs (bp) long (reviewed in Wray et al., 2003). Typically a TF can only bind to specific variations of one or more core binding motifs (Badis et al., 2009). The DNA-binding specificity of a TF is determined by the sequence of its DNA interacting domain, which is typically 60-100 amino acids long (reviewed in Mitchell and Tjian, 1989).

The ability of a TF to bind to a promoter can be experimentally tested in a number of ways, such as by electrophoretic mobility shift assays (EMSAs), Yeast one-hybrid (Y1H) or chromatin immunoprecipitation (ChIP). EMSAs measures the ability of a TF to bind to a given DNA sequence *in vitro*, by observing whether the protein reduces the speed at which the DNA can move through a gel during electrophoresis (Garner and Revzin, 1981). Y1H is a method to test the interaction of a TF and a DNA sequence in Yeast, this will be properly introduced in Chapter 3. ChIP is a method to immunoprecipitate DNA to which a protein is bound *in planta*. Immunoprecipitated DNA can then be amplified by PCR, recognised by microarray or sequenced to reveal which sequences are bound by the protein (Collas, 2010). EMSAs are a comparatively low-throughput approach to study DNA-binding proteins; Y1H (performed against TF libraries) and ChIP (combined with microarrays or sequencing) are 'interactomic' approaches that can be used to characterise many TF-DNA interactions in high-throughput.

Activation/repression domains regulate gene expression

The ability of TFs to bind to DNA has been shown to be necessary, but not sufficient, for regulation of gene expression (reviewed in Mitchell and Tjian (1989)). This is why TF binding does not necessarily lead to transcriptional regulation; the function performed by the binding of TFs is to allow other domains to be brought into play. Activation/repression of gene expression by a TF requires an activation/repression domain which is separate from the DNA-binding domain. These do-

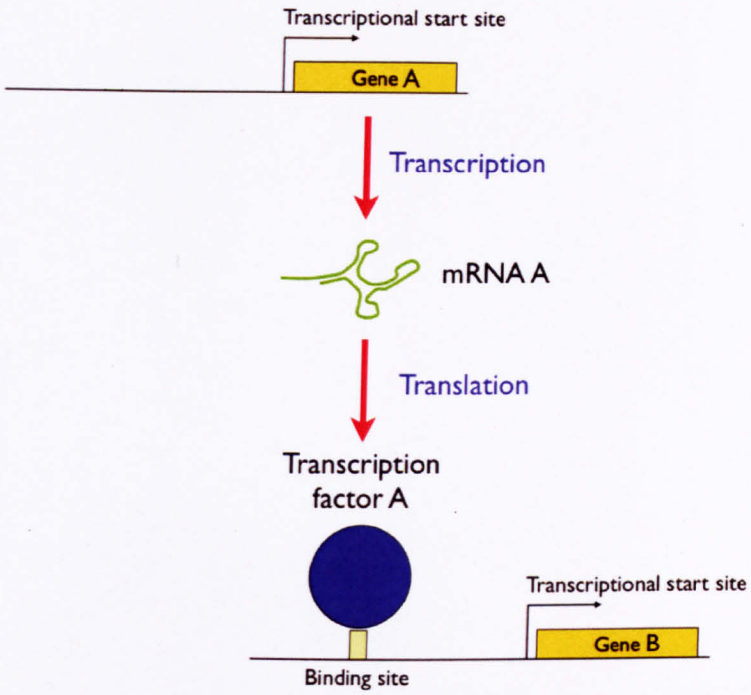
mains are typically 30-100 amino acids long (reviewed in Mitchell and Tjian (1989)). The independent function of DNA-binding and transcriptional activation/repression domains has been demonstrated by chimeric proteins combining domains from different TFs, which are able to regulate gene expression (reviewed in Ptashne (1988)).

To find genes regulated by a TF, the expression of genes in plants which have mutations in that TF's DNA sequence can be measured; this is called a mutant versus wildtype gene expression experiment. Typically knockout mutants, that cannot express a certain gene, or over-expressor mutants, that highly express a certain gene, are used. Genes regulated by a TF will usually have altered expression in mutants of it; these can be revealed by reverse transcriptase-PCR (RT-PCR), microarray or sequencing. Alternatively, transcriptional regulation of a gene can be studied by fusing a reporter to the genes promoter; such a promoter-reporter fusion has the transcriptional activaty of the original gene (minus the effect of local chromatin conformation), but the mRNA and protein degradation of the reporter gene. A commonly used reporter gene in plants encodes the beta-glucuronidase (GUS) protein from *Escherichia coli* (Jefferson et al., 1987). The expression of the reporter, in over-expressors or knockouts of certain TFs, can be used to identify transcriptional regulators. RT-PCR or promoter-reporter experiments are usually low-throughput, whereas microarray or sequencing can be used to identify genes whose expression is regulated by a TF in high-throughput.

1.3.3 Gene Regulatory Networks

Traditionally, molecular biology has been studied using reductionist approaches, which attempt to study biological processes by concentrating on single components. The reductionist approach has been enormously successful, and is responsible for much of our current understanding of molecular biology. However, many complicated biological processes can only be understood at a systems-level, where many different components interact in a manner that is too complex to understand if studied in a reductionist way (reviewed in Regenmortel, 2004). For example, gene expression can be controlled by the products of genes themselves, because of this gene regulation can form networks whose interconnectivity, feedback and feed-forward can produce complex dynamics of gene expression (reviewed in Alon, 2007).

Direct transcriptional regulation is a directed pairwise relationship (A,B) between genes, where gene A encodes a TF that binds to the promoter of gene B and affects its expression; this is illustrated in Figure 1.2(a). Given a set of genes, the GRN describing the interconnectivity of transcriptional regulation in a specific biological context can be illustrated as a graph, where nodes represent genes and directed edges



(a) Transcriptional regulation

(b) Simplified diagram of transcriptional regulation

Figure 1.2: Diagrams of transcriptional regulation. (a) TFs are proteins, which are produced by translation from the mRNA of a gene, and are capable of binding to DNA. Some TFs are also capable of transcriptional activation or inhibition by changing the binding affinity of RNA Polymerase to the transcriptional start site, when this occurs it is called transcriptional regulation. (b) Transcriptional regulation can be represented as a simplified network diagram.

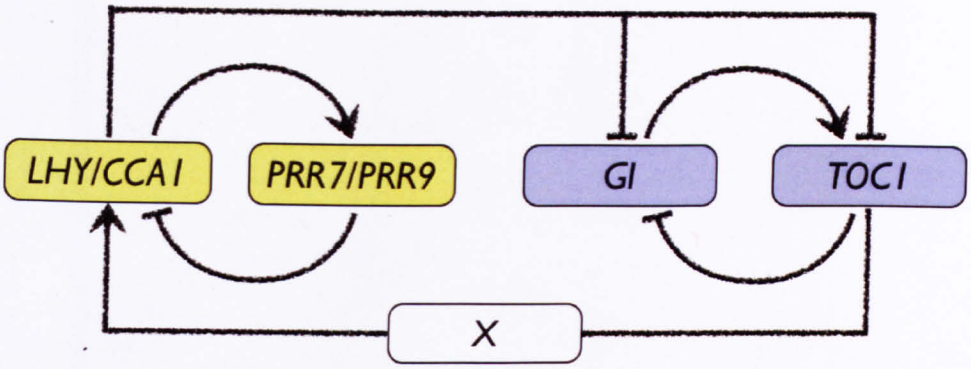


Figure 1.3: The interconnectivity of the Arabidopsis circadian clock GRN as known in 2006, redrawn from Locke et al. (2006).

(i.e. arrows) represent transcriptional regulation. For example the transcriptional regulation depicted in Figure 1.2(a) can be summarised as a graph, as depicted in Figure 1.2(b). The dynamics of gene expression depend on GRN structure and kinetics, which can only be properly understood within the context of the whole GRN. For example, the circadian clock in the model plant Arabidopsis is an example of a GRN, in which periodic expression of genes is achieved by multiple feedback loops of gene regulation (Locke et al., 2006; McClung, 2008); this can be illustrated as a graph, as shown in Figure 1.3. It is obvious that mutations in any of these genes could have a strong impact on the dynamics of the expression of the other genes; this would not be obvious if the interconnectivity of the circadian clock GRN was not taken into account.

Because thousands of Arabidopsis genes are known to be differentially expressed during *B. cinerea* infection (Ferrari et al., 2007) and many TFs are known to affect the susceptibility of Arabidopsis to infection by *B. cinerea* (reviewed in Birkenbihl and Somssich, 2011), gene expression during *B. cinerea* infection is likely to be regulated combinatorially by an interconnected GRN. For example, over-expression of the TFs *ERF1* or *ORA59* reduces the susceptibility of Arabidopsis to infection by *B. cinerea* (Berrocal-Lobo et al., 2002; Pré et al., 2008). Both *ERF1* and *ORA59* are up-regulated during *B. cinerea* infection (Berrocal-Lobo et al., 2002; Pré et al., 2008). *CHIB* is a gene which is up-regulated during *B. cinerea* infection (Pré et al., 2008); it is also known to be up-regulated by over-expression of either *ERF1* (Solano et al., 1998) or *ORA59* (Pré et al., 2008). *ERF1* itself is known to be upregulated by the TF EIN3 (Solano et al., 1998). This shows that TFs regulating *B. cinerea* responsive gene expression can combinatorially regulate the expression of other genes and can themselves be regulated transcriptionally by other TFs. Knowledge of the interconnectivity of the GRN, underpinning the defence response of Arabidopsis to

infection by *B. cinerea*, is likely to be required before the effect of genetic perturbations on disease resistance can be predicted.

Because current knowledge of the structure of this GRN is sparse, there are four major goals: identify the TFs responsible for controlling differential gene expression during *B. cinerea* infection; gather together current knowledge of the structure of the GRN linking these TFs and their targets; validate/extend the knowledge of the structure of the GRN; and develop a quantitative model of the regulation of Arabidopsis gene expression during *B. cinerea* infection. These are the approaches pursued in this thesis.

1.4 Transcriptional regulation controlling susceptibility of Arabidopsis to infection by Botrytis

Many TFs have knockout or over-expresssor mutants which have been found to have altered susceptibility to *B. cinerea* (reviewed in Birkenbihl and Somssich, 2011). For example, mutants of the following TFs have been found to have altered susceptibility to infection by *B. cinerea*: *ARF2, MYC2, ANAC019, ANAC055, ANAC092, TGA3, EIN3, ERF1, MYB46, MYB108, ZFAR1, WRKY70, WRKY33, ORA59, CAMTA3* and *ATAF1* (Youn-Sung Kim et al., in preparation; Lorenzo et al., 2004; Bu et al., 2008; Windram, 2010; Zhu et al., 2011; Berrocal-Lobo et al., 2002; Ramirez et al., 2011; Mengiste et al., 2003; AbuQamar et al., 2006; Zheng et al., 2006; Pré et al., 2008; Galon et al., 2008; Wang et al., 2009). Of these, *ANAC019, ATAF1, ERF1, MYB108, MYC2, WRKY70* and *ZFAR1* have have been shown to be differentially expressed during *B. cinerea* infection, in the microarray experiments of AbuQamar et al. (2006) and Ferrari et al. (2007). Additionally, *CAMTA3, ORA59* and *WRKY33* have been shown to be differentially expressed during *B. cinerea* infection by Reverse Transcriptase-Polymerase Chain Reaction (RT-PCR) experiments (Galon et al., 2008; Pré et al., 2008; Zheng et al., 2006).

Gene expression can be controlled by TFs, and also by chromatin modifiers. Chromatin describes both DNA and the proteins called histones that bind to DNA; histones are responsible for 'packing' the DNA into a smaller volume, called a closed chromatin state. A closed chromatin state reduces or blocks transcription by reducing accessibility of the promoter and transcriptional start site (TSS). Chromatin modifiers are proteins which can alter the chromatin state of DNA (reviewed in Saha et al., 2006); when they alter the chromatin state around DNA regions that contain a gene, they can allow transcription to be controlled. For example, knockout mutants of *HUB1*, which encodes a RING E3 ligase which monoubiquitinates histone

H2B (Liu et al., 2007), have been shown to have increased susceptibility to infection by *B. cinerea* (Dhawan et al., 2009). Another chromatin modifier, SDG8, a histone methyltransferase, has been found to regulate defence related gene expression, and therefore its loss of function mutant *sdg8_1* shows increased susceptibility to *B. cinerea* (Berr et al., 2010). Reporters fused to the promoters of *HUB1* and *SDG8* have been shown to be up-regulated locally during *B. cinerea* infection (Dhawan et al., 2009; Berr et al., 2010). Additionally, a knockout mutant of *SPLAYED (SYD)*, a gene encoding a SWI/SNF class chromatin remodeling ATPase, has been shown to be more susceptible to infection by *B. cinerea* (Walley et al., 2008). The altered susceptibility of mutants of TFs and chromatin modifiers is evidence of the important role of gene regulation in resistance of Arabidopsis to infection by *B. cinerea*.

1.4.1 Inputs to the GRN controlling the defence response of Arabidopsis to infection by Botrytis

Before the expression of genes can change in response to *B. cinerea*, infection must be detected by the plant. Pathogen perception typically takes place at the cell membrane, whereas transcriptional regulation takes place within the nucleus. For example, OGs are believed to be produced by the interaction of fungal polygalacturnases with the plant cell wall (reviewed in Cervone et al., 1989) and OG treatment has been shown to increase the resistance of Arabidopsis to infection by *B. cinerea* (Ferrari et al., 2007). The gene *WAK1* has been shown to encode the receptor for OGs. *WAK1* encodes a protein which contains an extracytoplasmic domain (ectodomain) which can interact with OGs *in vitro* (Decreux et al., 2006). Brutus et al. (2010) have shown that this domain is able to detect OGs; this was achieved by the production of a chimeric receptor which contained both the ectodomain of WAK1 and the kinase domain of the EFR receptor. The WAK1 ectodomain was found to activate the EFR kinase domain in the chimeric protein following OG treatment, resulting in the activation of known EFR kinase regulated genes. They also showed that over-expression of *WAK1* decreases the susceptibility of Arabidopsis to infection by *B. cinerea*, suggesting that WAK1 allows the plant to detect OGs produced during fungal attack and therefore to activate appropriate defence responses. Ferrari et al. (2007) have shown that detection of OGs by Arabidopsis results in the differential expression of 1,854 genes; 953 of which are also differentially expressed during *B. cinerea* infection. Some of the genes differentially expressed during OG treatment and *B. cinerea* infection encode TFs, whose mutants have altered susceptibility to *B. cinerea*, e.g. *CAMTA3, ERF1, TGA3* and *WRKY33* (introduced in Section 1.4). This shows that elicitors corresponding to *B. cinerea* infection can be detected by receptors located at the cell membrane leading to: differential expression of TFs

that control the defence response; and an enhanced resistance to infection by *B. cinerea*.

Although chitin treatment has not been tested for its ability to activate the defence response of Arabidopsis to infection by *B. cinerea*, treatment with its deacetylated derivate chitosan has been shown to reduce the susceptibility of Arabidopsis to infection by *B. cinerea* (Povero et al., 2011). Additionally the TFs *WRKY33* and *WRKY70*, which both have knockout mutants with increased susceptibility to infection by *B. cinerea* (Zheng et al., 2006; AbuQamar et al., 2006), are differentially expressed both during chitin treatment (Libault et al., 2007) and *B. cinerea* infection (AbuQamar et al., 2006). This suggests that the PAMP chitin, like OGs (which are perceived as DAMPs), are elicitors of the defence response of Arabidopsis against infection by *B. cinerea*. Both chitin and OGs are known to alter the expression of known regulators of the defence response and are therefore inputs to the GRN mediating the defence response of Arabidopsis to infection by *B. cinerea*.

After the detection of elicitors of the defence response, some molecule or molecules must be responsible for signalling to regulators in the nucleus that *B. cinerea* infection has been detected. These signalling molecules may be small molecules or proteins, which couple pathogen perception to regulation of gene expression. As such these signalling mechanisms must function as inputs to the GRN controlling *B. cinerea* responsive changes in gene expression. Many signalling molecules and proteins have been implicated in the defence response of Arabidopsis to infection by *B. cinerea*, most of which are either hormones or protein kinases.

Hormones have been implicated in the regulation of gene expression during *B. cinerea* infection, by experiments which have shown that mutants that are insensitive to hormone treatment also have altered susceptibility to *B. cinerea* infection. For example, Feys et al. (1994) have shown that the knockout mutant *coi1* is insensitive to treatment with Methyl Jasmonate (MeJA), a derivative of JA. Subsequently *coi1* was shown to be more susceptible to infection by *B. cinerea* (Thomma et al., 1998; Ferrari et al., 2003a). Similarly Guzman and Ecker (1990) have shown that the knockout mutant *ein2* is insensitive to ET treatment. Thomma et al. (1999) and Ferrari et al. (2003a) have shown that the *ein2* mutant has increased susceptibility to infection by *B. cinerea*. A mutant, *npr1* (Cao et al., 1997), involved in signalling downstream of SA has not been found to have altered susceptibility to infection by *B. cinerea* (Thomma et al., 1999; Ferrari et al., 2003a); whereas a mutant, *NahG*, which expresses a bacterial enzyme, salicylate hydroxlyase, making it unable to accumulate SA (Delaney et al., 1994), has been shown to be more susceptible to in-

fection by *B. cinerea* (Govrin and Levine, 2002; Ferrari et al., 2003a). The difference in susceptibility between the *nahG* and *npr1* mutants fits with the proposal that SA is involved in the defence response through it's effect on other signalling pathways. These results show that the JA, ET and SA hormones are involved in signalling in response to *B. cinerea* infection, which fits with the differential expression of some *B. cinerea* responsive genes in mutants of these genes, as shown in Glazebrook et al. (2003).

Although JA, ET and SA signalling pathways are important for controlling the expression of *B. cinerea* responsive genes, other hormone signalling pathways seem to be involved as well. Two additional hormones implicated in the defence response of Arabidopsis to infection by *B. cinerea* are auxin and abscisic acid (ABA). Three mutants, *axr1, axr2* and *axr6*, that affect auxin signalling have also been shown to have increased susceptibility to *B. cinerea* (Llorente et al., 2008). Auxin is believed to regulate gene expression by modifying the degradation of Aux/IAA proteins which bind to Auxin Response Factor (ARF) TFs (Tiwari et al., 2003; Reed, 2001). Similarly, ABA has been shown to affect susceptibility of Tomato to infection by *B. cinerea* (Audenaert et al., 2002). Additionally, a knockout mutant of an Arabidopsis TF, *zfar1*, is has been shown to be more susceptible to infection by *B. cinerea* and more sensitive to ABA treatment during germination (AbuQamar et al., 2006).

As well as signalling of pathogen perception by hormones, protein kinases have been shown to be responsible for *B. cinerea* responsive changes in gene expression. For example, Ferrari et al. (2007) have shown that resistance to *B. cinerea* induced by treatment with OGs – which are perceived by the plant as DAMPs – is independent of signalling by the JA, ET or SA hormone pathways. This was shown by inducing resistance to *B. cinerea* in the *coi1, ein2, nahG* and *npr1* lines with OG treatment. However, resistance induced by OG perception was abolished by a knockout mutant of *PAD3* (Ferrari et al., 2007), this meant that OG induced resistance of Arabidopsis to infection by *B. cinerea* depended on *PAD3*. Ren et al. (2008) have shown that induction of *PAD3* expression following *B. cinerea* infection depends on a protein kinase cascade involving MAPKKKα/MEKK1-MKK4/MKK5-MPK3/MPK6. They also showed that a knockout of *MPK3*, which encodes a protein kinase in that cascade, is more susceptible to infection by *B. cinerea*.

Recently the TF WRKY33 has been shown to be involved in the induction of *PAD3* expression by MPK3/MPK6 during *B. cinerea* infection (Mao et al., 2011). More recently Lai et al. (2011a) have shown that *PAD3* expression can still be induced in a knockout mutant of *WRKY33*, which suggests that *PAD3* expression may be under

combinatorial transcriptional regulation. Other kinases, such as the OG receptor WAK1 and Botrytis Induced Kinase 1 (BIK1) have also been shown to affect the susceptibility of Arabidopsis to infection by *B. cinerea*, implicating them in defence related signalling (Brutus et al., 2010; Veronese et al., 2006).

In summary, the elicitors chitin and OGs, PAMPs and DAMPs respectively, are believed to be recognised by Arabidopsis leading to increased resistance to *B. cinerea*. Signalling involving hormones (JA, ET, SA, ABA and auxin), as well as protein kinases (WAK1, BIK1 and the MPK3 kinase cascade), have been shown to affect the ability of Arabidopsis to resist infection by *B. cinerea*. Both pathogen perception and signalling have been shown to change the expression of many genes, some of which are known to encode regulators of differential gene expression during *B. cinerea* infection. This suggests that pathogen perception by detection of elicitors and subsequent signalling cascades are inputs to the GRN controlling the defence response of Arabidopsis to infection by *B. cinerea*.

1.4.2 Physiological outputs of the GRN controlling the defence response of Arabidopsis to infection by Botrytis

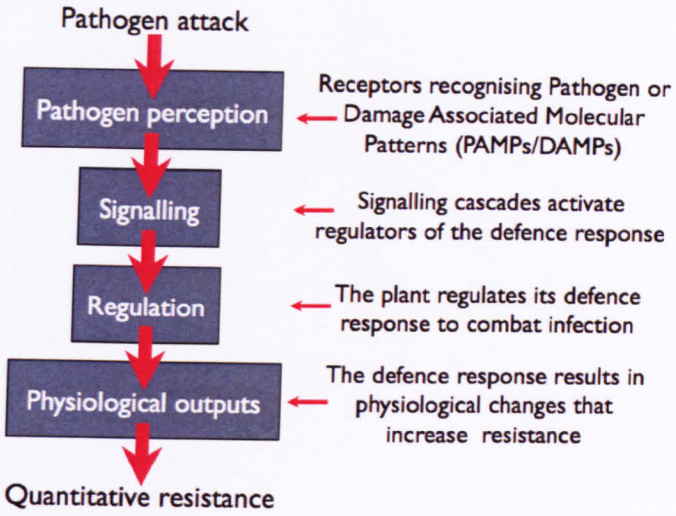
It is expected that the changes in gene expression are partly an active response that improves the resistance of Arabidopsis to infection by *B. cinerea*. This is supported by the TF mutants which have altered susceptibility to infection by *B. cinerea*. It is also supported by the literature linking TFs to regulation of 'physiological outputs', i.e. enzymes directly mediating resistance of Arabidopsis to infection by *B. cinerea*. In the previous section an example was given of pathogen perception leading to the induction of *PAD3* expression by the TF WRKY33. *pad3* was found in a mutant screen to be incapable of producing the anti-microbial compound camalexin in response to infection by *P. syringae* (Glazebrook and Ausubel, 1994). Additionally *PAD3* has been found to encode an enzyme, CYP71B15, that catalyses the final step in camalexin biosynthesis (Schuhegger et al., 2006). Camalexin concentration has been found to negatively correlate with susceptibility in various ecotypes of Arabidopsis (Denby et al., 2004). Therefore it appears that *PAD3* is a 'physiological output' of resistance, that is transcriptionally regulated by the TF WRKY33 during infection by *B. cinerea* to increase the production of camalexin and therefore increase the resistance of Arabidopsis to infection by *B. cinerea*. The role of the transcriptional regulator WRKY33 in the chain of events leading from OG treatment to *PAD3* dependent resistance can be summarised as: OGs (elicitor) → WAK1 (pathogen perception receptor) → MPK3 cascade (signalling) → WRKY33 (transcriptional regulator) → PAD3 (physiological output) → camelexin (anti-fungal molecule). The process of pathogen perception, signalling, regulation

and manipulation of physiological outputs, as is seen with the regulation of *PAD3*, is illustrated in Figure 1.4.

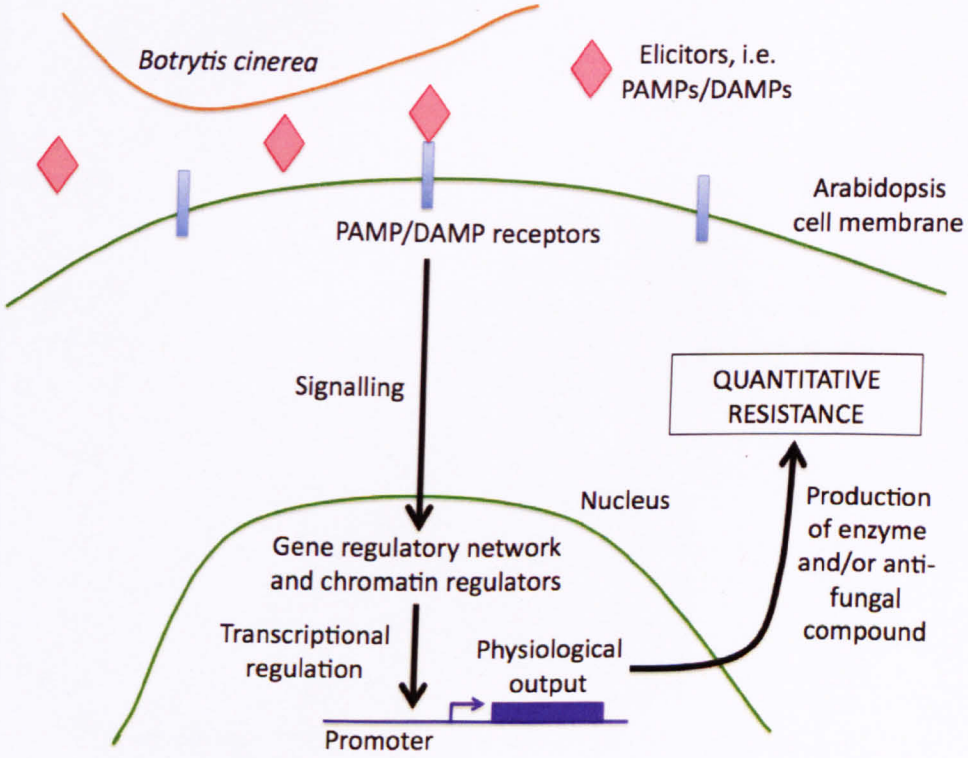
While *PAD3* is the best characterised physiological output, controlling resistance of Arabidopsis to infection by *B. cinerea* in response to transcriptional regulation, other good candidates exist. For example, Polygalacturonase Inhibiting Proteins (PGIPs) are known to be locally up-regulated during *B. cinerea* infection, both are also known to be up-regulated after OG treatment (Ferrari et al., 2003b). Over-expressors of *PGIP1* and *PGIP2* have been shown to have decreased susceptibility to *B. cinerea* infection; PGIPs reduce susceptibility of Arabidopsis to infection by *B. cinerea* by inhibiting fungal polygalacturonase, which would otherwise damage plant cells (Ferrari et al., 2003b). *PGIP2* expression requires *COI1*, suggesting that it is downstream of the JA hormone signalling pathway (Ferrari et al., 2003b). Because PGIPs are up-regulated during *B. cinerea* infection, and can lead to resistance by inhibiting fungal proteins, they are good candidates for a physiological output regulated by the GRN controlling *B. cinerea* responsive gene expression. Both *PAD3* and *PGIP1* expression were decreased in a knockout of *ARF2* in seedlings of Arabidopsis (Vert et al., 2008), suggesting that ARF2 is able to repress their expression (possibly indirectly).

For both *PAD3* and *PGIP1* there is a clear mechanism by which they directly contribute to resistance of Arabidopsis to infection by *B. cinerea*, mutants of both have been shown to affect resistance, and their differential expression during infection suggests transcriptional regulation and some potential transcriptional regulators are known. In the future other potential physiological outputs will hopefully be characterised to that level. However, no other candidates were found in the literature with all of the above attributes. Some candidates fell short at only one of these attributes, for example chitinases and *LACS2*.

The fungal PAMP chitin is hydrolysed by Arabidopsis chitinases. There are 22 Arabidopsis genes which are believed to encode chitinases, many of which are known to be induced by the interaction of Arabidopsis with pathogens (reviewed in Passarinho and Vries, 2002). It is believed that chitinases have a direct effect on resistance of Arabidopsis to *B. cinerea* infection, i.e. by inhibiting fungal growth by damaging the fungal cell wall (reviewed in Schlumbaum et al., 1986). However no literature was found showing that mutants of Arabidopsis genes encoding chitinases had altered susceptibility to infection by *B. cinerea*. For example, a study which used anti-sense RNA to suppress the level of a chitinase in Arabidopsis produced inconclusive results; while increased susceptibility was observed it was not found to be statistically



(a) Conceptual overview of the plant defence response



(b) Illustration of the role of the gene regulatory network underpinning the defence response of Arabidopsis to infection by *B. cinerea*

Figure 1.4: The defence response of Arabidopsis to infection by *B. cinerea*. (a) Summarised conceptually. (b) Summarised by illustration.

significant because of high variability of susceptibility between replicates (Samac and Shah, 1994). However, expression of a rice chitinase gene in cucumber has been shown to lead to reduced susceptibility to *B. cinerea* infection (Tabei et al., 1998), demonstrating that chitinases are able to inhibit growth of *B. cinerea* on plants. Upregulation of chitinases have been seen following *B. cinerea* infection (AbuQamar et al., 2006), as well as after chitin (in rice, Nishizawa et al. (1999)), MeJA (Thomma et al., 1998) or ET treatment (McGrath et al., 2005). As briefly introduced in Section 1.3.3, some TFs such as *ERF1* and *ORA59* are known to control the expression of the chitinase *CHIB*. Additionally, *CHIB* has been shown to be differentially expressed in a knockout of the TF *MYC2* (Dombrecht et al., 2007). Except for the lack of a published Arabidosis chitinase mutant with altered susceptibility, chitinases, and especially *CHIB* as transcriptional regulators are known, are good candidates for physiological outputs, controlling resistance of Arabidopsis to infection by *B. cinerea* in response to transcriptional regulation.

Mutants with morphological traits, such as cuticle defects, have also been found to have altered susceptibility to *B. cinerea*. For example, the *lacs2-2* mutant has a higher cuticle permeability and therefore is better able to secrete antimicrobial compounds, making it very resistant to infection by *B. cinerea* (Bessire et al., 2007). It is not known whether *LACS2* enhances resistance in a developmental or a pathogen responsive way, but it has been found to be differentially expressed during *B. cinerea* infection at 48 hpi suggesting that it is transcriptionally regulated in response to infection (Ferrari et al., 2007). Except for the lack of known transcriptional regulators, *LACS2* is a good candidate for a physiological ouput, controlling resistance of Arabidopsis to infection by *B. cinerea* in response to transcriptional regulation.

More speculatively, *BAP1*, which is known to be an inhibitor of programmed cell death that plays a role in the Arabidopsis defence response (Yang et al., 2007), has been found to be strongly up-regulated in the area close to the *B. cinerea* infection site (Mulema and Denby, 2012). This suggests that *BAP1* may be involved in the defence response against *B. cinerea*, by preventing cell death which would benefit *B. cinerea*. *BAP1* is known to be up-regulated in response to treatment by OGs (Ferrari et al., 2007), and also in knockouts of the TFs *ARF2* and *MYC2* which are also known to have altered susceptibility to infection by *B. cinerea* (as introduced in Section 1.4) (Vert et al., 2008; Dombrecht et al., 2007). No literature could be found in which mutants of *BAP1* were screened for altered susceptibility to *B. cinerea*; additionally the 'directness' of BAP1 in inhibiting programmed cell death is not known. For these two reasons a considerable amount of work is required to test whether *BAP1* is actually a physiological output, controlling resistance of Arabidopsis to

infection by *B. cinerea* in response to transcriptional regulation. However, *BAP1* highlights the potential for inhibition of programmed cell death pathways to control resistance of Arabidopsis to infection by *B. cinerea*. This is plausible as *B. cinerea* has been shown to manipulate the defence response of tomato to encourage programmed cell death, which increases the susceptibility of tomato to infection by *B. cinerea* (Oirdi et al., 2011).

These potential physiological outputs, and probably many more, control the resistance of Arabidopsis to infection by *B. cinerea*. This means that their differential expression, as controlled by the GRN underpinning the defence response, is required for Arabidopsis to successfully resist *B. cinerea* infection. For this reason it would be desirable to elucidate the network of regulators responsible for their differential expression, as manipulation of that GRN should allow the resistance of Arabidopsis to infection by *B. cinerea* to be modified.

1.5 Modelling approaches for gene regulatory networks

Qualitative models, like graphs of GRN structure, can usefully summarise experimental findings, but they do not describe the dynamics of the regulation in a particular context. For example, a graph of a subnetwork of the Arabidosis circadian clock GRN was shown in Figure 1.3; this qualitative model does not necessarily imply that the expression of these genes will repeat every 24 hours. These dynamic details emerge from a quantitative model of the regulation, and the values of parameters of the model. An example of this is the model suggested in Locke et al. (2006) describing the dynamics of the Arabidopsis circadian clock GRN. By modelling the expression levels of the components of a GRN, quantitative models allow complex hypotheses to be made which can be tested by comparison of modelled and experimentally observed gene expression. This is why a quantitative model of the GRN controlling differential expression of genes in response to *B. cinerea* infection is highly desirable. It would allow a cycle of hypotheses and experimental validation which could reveal the transcriptional regulation controlling the defence response of Arabidopsis to infection by *B. cinerea*.

1.5.1 Difficulties with using gene expression measurements to predict gene regulation

Modelling the effect of transcriptional regulation on gene expression is non-trivial. The expression of a gene may be regulated in a non-linear fashion, and may also be controlled simultaneously by many different transcription factors, as well as by epigenetic regulators such as chromatin state and DNA methylation (for example

Mazzucotelli et al., 2008). Furthermore, protein concentration may not correlate with RNA concentration for a given gene (Gygi et al., 1999). Gene expression is also a proxy measure of the balance between the mRNA transcription and RNA degradation rates (Beelman and Parker, 1995), only the former being controlled by transcriptional regulation. TFs may also require post-transcriptional activation before regulating direct targets (for example Mao et al., 2011). In fact, master regulators are likely to be regulated non-transcriptionally, as they will need to respond to environmental cues, e.g. TFs activated by kinase cascades, which are in turn activated by receptors that recognise pathogens. It is therefore clear that accurately predicting transcriptional regulation from gene expression is challenging, made harder by measuring gene expression from mixed populations of cell types.

There are also mathematical problems that affect the accuracy of such predictions, resulting from the large number of variables observed relative to the amount of observations made upon them (reviewed in Bellman, 1961; Cleaskens and Hjort, 2008). As a general rule the more complex the model of regulation, the fewer genes can be modelled together accurately given a finite amount of data. This results in a two-fold approach to modelling transcriptional regulation: determine which genes should be modelled together; and then use a modelling approach that is most suitable given the amount of genes and existing knowledge about the complexity of their regulation.

1.5.2 Determining which genes should be modelled together

The method used to chose genes to model together will affect the usefulness of that model. Also, as discussed above, the fewer genes modelled together the easier it is to fit the models accurately.

Genes whose mutants have altered phenotypes relevant to specific processes

Gene function can be revealed by observing the effect of altered expression. This can be performed using altered expression mutants such as gene knockouts and overexpressors. If a specific biological process is being studied, then genes whose altered expression affects the process are described as displaying an altered phenotype. If these genes are TFs then it is likely that the altered phenotype is caused because the TFs no longer regulate the expression of genes in the way they did in the wildtype organism, i.e. TFs whose mutants have altered phenotypes are likely to be important components of the GRNs that coordinate the biological process. Therefore it makes sense to model together genes which display altered phenotypes in the same (or similar) biological processes as they are likely to regulate, or be regulated by, each

other. For example, it would make sense to model the expression of genes that have mutants which have altered susceptibility to *B. cinerea*, as they would be likely to be involved in the defence response of Arabidopsis to infection by *B. cinerea*.

Experimental evidence suggesting transcriptional regulation

Experimental approaches can be used to study transcriptional regulation, as introduced in Section 1.3.2. Different experimental methods test different aspects of transcriptional regulation, such as the binding of TFs upstream of a gene or the effect of a TF on the expression of its target genes. Often these methods show likely transcriptional regulation rather validated transcriptional regulation, this can be because the experiment was performed either: *in vitro*; in Yeast; *in planta* but in a different context; with ectopic levels of the transcription factor; or in a way that fails to discriminate between direct and indirect regulation. It can be interesting to model the expression of these genes to see which of these likely regulatory events are supported by the expression data. This approach will be applied in Chapter 4.

Genes differentially expressed during specific processes

Both of the above methods are heavily biased towards genes which are already heavily studied, meaning that important unknown regulators are ignored. This can be a problem when key transcriptional regulators of a biological process have yet to be found. Transcriptomics can be used to find genes that are differentially expressed during specific biological processes, these are genes whose mRNA levels are being differentially regulated during the biological process. This regulation can either occur epigenetically, transcriptionally or by regulation of mRNA degradation. If some of the differentially expressed genes are TFs then it is likely that they are involved in transcriptional regulation during that biological process, so it makes sense to model them with the other differentially expressed genes.

1.5.3 Information theory

Information theory approaches that statistically model gene regulation, are applicable to large groups of genes. They typically work by calculating the 'similarity' of gene expression profiles over time or conditions. It is hoped that genes whose profiles are more 'similar' are more likely to be involved together in gene regulation. This approach is only able to infer undirected transcriptional regulation. That is, it infers pairs of genes in which the expression of one of the genes is affected by regulation by the protein encoded by the other, but not which of these genes is the regulator. An example of this is the paper by Carrera et al. (2009), which uses a continuous generalisation of mutual information introduced by Daub et al. (2004)

as a similarity measure and applies it to a compendium of microarray data. It is important in such approaches to demonstrate that the similarity measure can distinguish between regulation and its absence, at least better than guesswork and ideally better than commonly used similarity measures. An example of this kind of benchmarking of similarity measures can be seen in Yona et al. (2006).

Other commonly used measures of similarity are negative Euclidean distance, Pearson's correlation coefficient (PCC) and Spearman's rank correlation coefficient. A similarity measure made for gene expression time series was introduced by Qian et al. (2001) that can find time-delayed correlation; this has been shown to correctly predict some known regulation. This will be introduced fully and applied in Chapter 2.

Information theory approaches to model gene regulation have the advantage that they can be applied to large groups of gene expression profiles. However there are a number of important disadvantages:

1. Modelling large groups of genes means that many profiles will appear similar by chance alone.

2. Similarity of gene expression profiles is more often caused by co-regulation than by regulation.

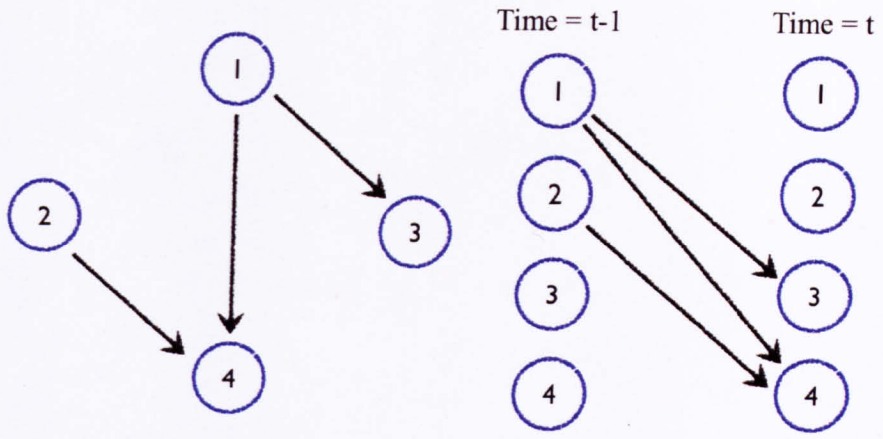
3. Similarity measures are by definition pairwise, meaning that combinatorial regulation would be undetectable.

4. It is not possible to determine the direction of regulation if a symmetric similarity measure is used

5. Even if a similarity measure predicts transcriptional regulation well, its false positive rate may still be too high to be of practical use in a given setting. While better than guesswork, the success rate may not be high enough to justify the use of researcher time and project resources required to validate the predictions.

1.5.4 Graphical models

Graphical models can be used to model the regulation of medium sized groups of genes, depending on the size of the gene expression dataset and the complexity of the modelling approach. A graph, G, is defined as $G := (V, E)$, where $V := \{V_1, V_2, \cdots V_n\}$ is a set of vertices and $E := \{E_1, E_2, \cdots E_m\}$ is a set of edges. For example, the graph in Figure 1.5(a) can be defined by $V := \{1, 2, 3, 4\}$ and



(a) Simple graph

(b) Simple bipartite graph representing dependence in a dynamic Bayesian network equivalent of the simple graph

Figure 1.5: An example of a graph, which can be used to represent networks of transcriptional regulation. Blue circles are vertices, which can represent genes, and black arrows are edges, which can represent transcriptional regulation. (a) A simple graph. (b) A simple bipartite graph representing the same relationships with a Markovian dependence.

$E := \{(1,3),(1,4),(2,4)\}$. When modelling gene regulation with a graphical model, each vertex can represent a gene and each edge can represent transcriptional regulation, i.e. (A,B) or A $\rightarrow$ B can represent gene A transcriptionally regulating gene B, as in Figure 1.2.

To connect gene expression data to a graph that represents its regulation, a probabilistic model must be defined. A commonly used approach is to model gene regulation using a Bayesian network (BN). A BN is defined as a directed acyclic graph, a directed graph with no loops, with edges representing the conditional dependence between random variables represented by the vertices. The distribution of random variable, A, given data, X, is defined as follows:

$$P(A|X) := \frac{P(X|A)P(A)}{P(X)} \quad \text{(Bayes theorem)} \tag{1.1}$$

$$\tag{1.2}$$

Where $P(X|A)$ is the distribution of X given A, $P(A)$ is a prior distribution over A and $P(X) := \int_A P(X|A)dA$ is the marginal distribution of X. $P(X|A)$ is called the likelihood, $P(A)$ is called the prior and $P(X)$ is called the marginal likelihood.

Gene regulation over time can be modelled using dynamic Bayesian networks (BNs), where two nodes exist for each gene, representing gene expression levels at current and previous time points respectively (Murphy and Mian, 1999), for example Figure 1.5(b). Dynamic BNs can model networks with loops by "unravelling the network in time", e.g. auto-regulation can be taken into account by allowing the random variable to depend on its value at the previous time point, this can be represented by a directed acyclic bipartite graph.

BN models can be used to learn network structure *de novo* by working out the Bayesian distribution of graph structures given the data and prior distributions. If prior knowledge of the network structure exists, it can be integrated with the data using the prior (Heckerman et al., 1995). Graphical approaches to model gene regulation have the advantage that they can be applied to medium size groups of genes. They can also model combinatorial regulation, i.e. the regulation of a gene's expression by several different TFs simultaneously. However there are a number of important disadvantages:

1. Given a dataset with a realistic number of observations, it can be hard to model non-linear regulation.

2. The number of different possible graphs grows super-exponentially with the number of vertices (Robinson, 1973), making it increasingly hard to distinguish between each model given a limited number of observations.

3. If prior information is used, prior weights must be set.

4. As with information theory approaches, graphical modelling methods must have a sufficiently low false positive rate to generate predictions likely to be useful to experimental biologists.

Literature on benchmarking graphical model inference approaches on gene regulatory networks

A wide range of different graphical model based network inference approaches have been applied to the prediction of the structure of GRNs. Methods that use time-series of gene expression have been reviewed in Sima et al. (2009), which introduces the methods and then discusses the need for comparative studies on 'ground truth' networks, i.e. GRNs of known structure. Unfortunately such networks are rare, which has limited researchers ability to perform comparative analyses of the accuracy of methods that predict network structure. Recent progress has been made

by using simulated data, synthetic *in vivo* GRNs or GRNs whose structures are partially known.

Simulated data has been used, for example in the DREAM challenges, where the 'ground truth' was only revealed to predictors after predictions have been made. Assessment was also made 'blind' to prevent conscious or subconscious bias in the comparison of the methods (Marbach et al., 2009; Prill et al., 2010; Marbach et al., 2010).

In Cantone et al. (2009) a synthetic biology approach was applied to benchmark various network modelling methods. A GRN was constructed in *S. cerevisiae* to be minimally affected by external influence, except by the switch from one type of growth medium to another. Because the GRN was constructed its exact network structure was known, which allowed the accuracy of predictions made by different methods to be compared. Because the synthetic network was generated *in vivo* the dataset produced is likely to be more realistic than simulated data.

An alternative approach is to take a naturally occurring GRN that is reasonably well characterised, and to take literature knowledge to be an 'approximate ground truth' against which predictions can be compared. This approach has been taken by both Morrissey et al. (2010) and Penfold and Wild (2011), using a time series of Arabidopsis gene expression to study the circadian clock GRN.

In all benchmarking studies performed so far, correct prediction of network structure has been shown to be incredibly hard, i.e. predictions of transcriptional regulation typically have non-trivial false positive rates. Incorrect predictions can be made by using insufficiently informative data, by making incorrect model assumptions (correct assumptions aren't fully known), and because of the complexity of combinatorial regulation. Because of this, accurate inference of GRN structure from gene expression data is an unsolved problem (Marbach et al., 2010).

VBSSM

One class of graphical model is the state space model (SSM)/Linear Gaussian Model/Kalman filter model, which is itself a dynamic BN in which some variables are considered to be unobserved (Brown and Hwang, 1997; Roweis and Ghahramani, 1999). In Beal et al. (2005) a SSM was applied to gene expression time series, where the unobserved variables can represent relevant unmeasured quantities, such as: genes whose expression has not been measured; protein levels; and the effect of degradation. In this section, the Variational Bayesian SSM (VBSSM) approach

introduced in Beal et al. (2005) is briefly introduced, following that paper.

In a generic SSM the measured variables are represented as a vector $\underline{y}_t$ at each time point, and so the time series of observations is a sequence $y_. := (\underline{y}_1, \cdots, \underline{y}_T)$, where T is the number of time-points. It is assumed that the value of $\underline{y}_t$ depends on the values of the unobserved/hidden state variables $\underline{x}_t$ at the same time point. It is also assumed that the value of the hidden states, $\underline{x}_t$, depends on the value, $\underline{x}_{t-1}$, of the hidden states at the previous time point, which is called a Markov Process assumption. This means that a sequence, $x_. := (\underline{x}_1, \cdots, \underline{x}_T)$, of hidden states, $\underline{x}_t$, exist. This results in the following joint probability for a given $x_.$ and $y_.$:

$$P(x_., y_.) = P(\underline{x}_1)P(\underline{y}_1|\underline{x}_1)\prod_{t=2}^{T} P(\underline{x}_t|\underline{x}_{t-1})P(\underline{y}_t|\underline{x}_t) \tag{1.3}$$

$\underline{x}_1$ is assumed to be normally distributed, and then dependencies are assumed to be linear and normally distributed, resulting in the following state space model:

$$\underline{x}_t = A\underline{x}_{t-1} + \underline{w}_t, \qquad \underline{w}_t \sim N(0, Q) \tag{1.4}$$

$$\underline{y}_t = C\underline{x}_t + \underline{v}_t, \qquad \underline{v}_t \sim N(0, R) \tag{1.5}$$

A contains the linear coefficients describing the dependency of the hidden state, $\underline{x}_t$, on its value at the previous time point, $\underline{x}_{t-1}$. C contains the linear coefficients describing the dependence of the observed (sometimes referred to as the emitted) state, $\underline{y}_t$, on the hidden state, $\underline{x}_t$, at the same time point. Q and R define the covariance of the noise vectors, $\underline{w}_t$ and $\underline{v}_t$, of the hidden states and observed variables respectively.

In the application of this SSM to gene expression data, $\underline{y}_t$ is taken to be a vector of gene expression measurements at time point t. To allow a dependence between the expression of genes this generic normal-linear SSM is extended in the following way:

$$\underline{x}_t = A\underline{x}_{t-1} + B\underline{y}_{t-1} + \underline{w}_t, \qquad \underline{w}_t \sim N(0, Q) \tag{1.6}$$

$$\underline{y}_t = C\underline{x}_t + D\underline{y}_{t-1} + \underline{v}_t, \qquad \underline{v}_t \sim N(0, R) \tag{1.7}$$

B contains the linear coefficients describing the dependency of, $\underline{x}_t$, the value of the hidden states at a time point, on the value, $\underline{y}_{t-1}$, of expression of all genes at the previous time point. D contains the linear coefficients describing the dependency of, $\underline{y}_t$, the expression of all genes at a time point on the value, $\underline{y}_{t-1}$, of expression

of all genes at the previous time point. In this way the hidden states and gene expression can all be allowed to affect each other at the next time point, providing a model of gene regulation and external influences over time. The equation for the gene expression data can be rewritten, by substitution of Equation (1.6), as:

$$\underline{y}_t = C(A\underline{x}_{t-1} + B\underline{y}_{t-1} + \underline{w}_t) + D\underline{y}_{t-1} + \underline{v}_t \quad (1.8)$$

$$\Rightarrow \underline{y}_t = CA\underline{x}_{t-1} + CB\underline{y}_{t-1} + C\underline{w}_t + D\underline{y}_{t-1} + \underline{v}_t \quad (1.9)$$

$$\Rightarrow \underline{y}_t = CA\underline{x}_{t-1} + (CB + D)\underline{y}_{t-1} + C\underline{w}_t + \underline{v}_t \quad (1.10)$$

And so to a first order $(CB + D)$ represents the effect of the expression of each gene on the expression of all genes at the next time point. A threshold can be applied to this matrix to give an inferred network structure, approximately representing the dependency of each gene on the expression of other genes over time. Because transcriptional regulation is one way in which the expression of one gene can affect the expression of another, this SSM can be used to model GRNs.

To predict the structure of a GRN, the distribution over parameters of the SSM is inferred given a gene expression time series dataset and prior distributions over the parameters. The distribution of these parameters can be inferred using a variational Bayesian expectation-maximisation (EM), as detailed in Beal et al. (2005).

1.5.5 Differential equation models

Differential equations can be used to model transcriptional regulation by relating the rate of change of gene expression to the concentration of the TFs that transcriptionally regulate that gene. An Ordinary Differential Equation (ODE) is defined as:

$$\frac{dx(t)}{dt} = f(x, t) \quad (1.11)$$

Where $\frac{dx(t)}{dt}$ is the rate of change of $x(t)$ with respect to time. Here $x(t)$ could be the expression level of a gene. For example $f(x, t)$ can be defined to take into account basal transcription, B, transcriptional activation, α, and mRNA degradation, δ (Adapted from Honkela et al., 2010):

$$\frac{dx(t)}{dt} = B + \alpha(t) - \delta(t)x(t) \quad (1.12)$$

A system of ODEs is the vector generalisation of an ODE, I.e.

$$\frac{d\underline{x}(t)}{dt} = \underline{f}(\underline{x}, t) \tag{1.13}$$

Where $\underline{x} = [x_1, x_2, \cdots, x_n]$. Systems of ODEs can be used to model both transcription and translation. They can also be used to model networks of several genes regulating each other's transcription. They are well suited for modelling non-linear combinatorial regulation of transcription, translation and degradation of mRNA or proteins. However, the more complex the model the more parameters are needed. Where these parameters have not been determined experimentally, they will have to be inferred and the more parameters to infer the more samples are need. Here again more complex models become harder to fit given limited data. This typically limits ODE modelling approaches to small sets of genes, ideally with the network of regulation already known.

Because the components and structure of the GRNs underpinning the defence response are insufficiently well known, information theory and graphical modelling approaches, rather than ODE modelling approaches, are applied in this thesis to predict the structure and dynamics of gene regulation during the defence response.

1.5.6 Literature on the prediction of gene regulation from gene expression in Arabidopsis

Recently several authors have attempted to use published microarray data to infer Arabidopsis GRNs using either information theory (Carrera et al., 2009) or discrete BNs (Needham et al., 2009). The work of Carrera et al. (2009) applies a continuous generalization of mutual information to data on the expression of approximately 20,000 genes from a compendium of 1,436 microarrays. This approach is not able to infer the direction of transcriptional regulation. The work of Needham et al. (2009) is better in two respects: it applies BN inference to find probable causal networks, allowing the direction of transcriptional regulation to be inferred; and it applies network modelling to small sets of genes at a time, which may allow an improvement in accuracy. Two shortcomings with this approach are that discretised gene expression data was used, and the genotypes of the biological samples were not taken into account, i.e. some samples are from mutants such as knockouts, which may have subtly different GRN structures due to the lack of a particular regulator.

A different approach would be to use a time series of gene expression and to infer a continuous state space model (SSM) as described briefly in Section 1.5.4 and fully in Beal et al. (2005). This approach has been applied to predict the structure

of a GRN controlling senescence in Arabidopsis (Breeze et al., 2011).

1.6 Aims and objectives

It is clear that a GRN is involved in the regulation of the defence response of Arabidopsis to infection by *B. cinerea*. However, knowledge of its components and structure are currently sparse. We aimed to find novel components of this GRN, such as TFs regulating the defence response, and to characterise the GRN structure linking them by transcriptional regulation. We also aimed to predict the components and structure of this GRN in a way that is data-driven and relatively unbiased with respect to the literature, e.g. not just investigating TFs which are already suspected to play a role in the defence response. This will allow us to reveal novel aspects of the defence response. We also aimed to develop a quantitative model of the the GRN, to allow predictions about gene regulation to be made and tested experimentally to refine the model.

1.7 Chapter outlines

In Chapter 2 a time series of gene expression during *B. cinerea* infection is introduced and used to predict regulators of differential expression in response to infection. Novel computational approaches to predict transcriptional regulation from gene expression time series are introduced, and predicted regulators tested in a reverse genetics screen.

In Chapter 3 experimental evidence of transcriptional regulation, between TFs and physiological outputs involved in the defence response, is compiled from the literature. This is used to generate a qualitative model of the GRN underpinning the defence response of Arabidopsis to infection by *B. cinerea*. This qualitative model is tested and extended by cloned TF library yeast one-hybrid and transient transactivation assays. Context-dependence of transcriptional regulation is studied by comparative transcriptomics.

In Chapter 4 VBSSM is used to predict the dynamics of gene regulation between the genes suggested to regulate each other in the qualitative model of the previous chapter. VBSSM is first used with an uninformative prior, and then with informative priors based on experimental evidence of gene regulation from the literature and the previous chapter. Predictions are compared to the temporal precedence of differential expression to highlight the most plausible predictions.

In Chapter 5 the results of the previous chapters are discussed and conclusions made.

Chapter 2

Genome-wide inference of transcriptional regulation from gene expression time series and phenotype screening of inferred regulators of the defence response

The aim of this chapter is predict specific transcriptional regulation between genes differentially expressed during *B. cinerea* infection. Knowledge of pathogen-responsive gene regulation would allow regulators to be targeted by genetic perturbations to alter the expression of downstream targets, and therefore the ability of Arabidopsis to resist infection. In this way, strategies to increase the resistance of Arabidopsis to infection by *B. cinerea* may be revealed. These strategies may also work in agronomically important crops. To achieve this, various computational approaches will be applied to a time series of gene expression during *B. cinerea* infection. Then inferred regulators of the defence response will be tested in a reverse genetics screen to validate their role in the defence response.

2.1 Introduction

2.1.1 The forward and reverse genetic approaches

In 1903 Wilhelm Johannsen introduced the terms genotype and phenotype, to distinguish between the material an organism inherits from parents and their observable

characteristics. He explained these concepts in a review in an American journal a few years later (Johannsen, 1911). The role of chromosomes in the genotypic inheritance of a phenotype was first demonstrated by William T H Morgan, who showed that a rare white-eye phenotype in the fruit-fly *Drosophila melanogaster* was inherited in a gender dependant way. This meant that the white-eye phenotype was recessive and was controlled by the X chromosome (Morgan, 1910).

One of Morgan's students, Hermann Muller, demonstrated that mutations induced by radiation could be stably inherited and showed mendelian patterns of inheritance (Muller, 1927). This approach, the identification of altered phenotypes in a population with induced mutations and the subsequent identification of the chromosomal location of the gene or genes responsible, became known as the 'forward genetics' approach. Through forward genetics Beadle and Tatum (1941) showed that some mutations changed both a specific gene and a specific protein, leading them to suggest that each gene encoded one protein.

Greater knowledge of genomic sequence has allowed researchers to apply an alternative to 'forward genetics' called 'reverse genetics' (for example Krysan et al., 1999). In this approach, lines of known genotype are tested for altered phenotypes, allowing researchers to determine the phenotypic effect of a given mutation. This allows researchers to prioritise the study of the effect of mutations in certain regions, such as the coding sequence of genes differentially expressed in certain conditions or with a similar coding sequence to genes already known to have an effect on phenotype. In principle 'reverse genetics' can be used to increase the proportion of altered phenotypes discovered by screening only lines whose known function is likely to be required for a wildtype phenotype. In practice this can be complicated by redundancy, where combinatorial mutants are required to observe an altered phenotype (reviewed in Kafri et al., 2009).

2.1.2 The effect of gene regulation on phenotype

It was originally hypothesised by Jacob and Monod (1961) that the regulation of gene expression was an important factor in phenotypic effects; later this would be demonstrated experimentally. The interruption of transcriptional regulation can lead to dramatic phenotypes; this is clearly demonstrated in the phenotypic effects of mutations of cis-regulatory sequences (reviewed in Wray, 2007), i.e. sequences that lie upstream of a gene and are important for control of its transcription. Mutations in cis-regulatory sequences can cause phenotypic differences by altering the expression of a gene in a given context, with a downstream effect on development, primary metabolism or environmental adaptation. Gene expression can be altered

by mutations in cis-regulatory sequences, which can modify the binding affinities of interacting TFs to the location of the mutation and therefore alter the transcription rate.

When organisms face environmental challenges they can use transcriptional regulation to modify their transcriptome to increase their chances of surviving and reproducing. Plants must adapt to the stresses that present themselves, and this is enabled in part by transcriptional reprogramming. The importance of transcriptional regulation in plant adaptation to stress has been demonstrated by genetic studies showing that many TFs are required (Singh et al., 2002; AbuQamar et al., 2006; Birkenbihl and Somssich, 2011), i.e. knockout or over-expressors of these TFs are unable to adapt normally to environmental stress.

2.1.3 The analysis of microarray time series

To study the changes in gene expression during an environmental stress, such as during *B. cinerea* infection, microarrays can be used. Gene expression analysis by microarrays is now a well-established approach in high-throughput molecular biology. Microarray experiments with a single time-point can be used to show which genes are differentially expressed in certain conditions. Time series microarray experiments are able to extend this to reveal the timing and dynamics of gene regulation (for a summary see Table 1 of Sima et al. (2009); for example primary literature see Spellman et al. (1998); Gasch et al. (2000); Arbeitman et al. (2002); Baugh et al. (2003); Orlando et al. (2008); Breeze et al. (2011)). These datasets allow the role of transcriptional regulation in an organism's response to developmental and environmental cues to be studied.

To gain an understanding of these time series researchers must deduce the relation of changes in gene expression to other changes occurring within the cell. This can be achieved by identifying regulators of expression in specific conditions and the effect of changes in expression on downstream processes. However, identification of transcriptional regulation is challenging given the context dependence, complexity and scale of genome-wide transcriptional dynamics. To help tackle these challenges many different computational approaches have been developed and applied.

Microarrays are a high-throughput experimental method for measuring the transcriptome, i.e. quantifying the relative abundance of all mRNA molecules in a sample. They contain DNA probes arranged in a grid. Each probe hybridises to cDNA (complementary DNA) generated from one specific mRNA sequence, as well as occasionally cross-hybridising weakly to other cDNA sequences. Fluorescent dyes are

incorporated into the cDNA and the intensity of fluorescence at each grid position is recorded to semi-quantitatively measure the concentration of the corresponding cDNA sequences. In this way the concentration of specific transcripts in the original sample can be quantified.

Microarray data is subject to both biological and technical variation. Biological variation can be intrinsic or due to poorly controlled experimental conditions. Technical variation is error caused by technical problems relating to sample dye incorporation, hybridisation and microarray printing defects. Technical variation can be decreased by applying quality control and normalisation methods. Quality control can be used to monitor technical variation and highlight when it gets beyond reasonable bounds. Microarrays that fail quality control can be repeated. Normalisation approaches can then be applied to further decrease technical variation by adjusting values to take into account dye biases, spatial hybridisation variation and print tip variations (for example, Wu et al., 2003).

The first stage of the analysis of normalised gene expression is to identify genes that are differentially expressed between different samples. These are genes which have been transcribed at a different rate, or whose mRNA has been degraded at a different rate between the different samples. Differentially expressed genes can be identified using hypothesis tests such as t-tests or ANOVA. Microarray experiments performed in a time series allow for temporal differential expression analysis by methods that explicitly take time into account, for example Timecourse (Tai and Speed, 2006) or Gaussian Process two-sample (Stegle et al., 2009). The timing of differential expression can be identified from two-sample time series, for example control and treated samples, by analysing time points separately or by probabilistic modelling of the time series (Stegle et al., 2009).

Notation

After identification of differentially expressed genes, they can be studied further by comparing their expression. To introduce this properly, some basic notation will be introduced.

A time series microarray dataset, $\mathbf{X}$, is a 3-dimensional table containing the expression values of genes $\mathcal{G} = \{1, 2, \cdots, G\}$, for time points $\mathcal{T} = \{1, 2, \cdots, T\}$, in replicates $\mathcal{R} = \{1, 2, \cdots, R\}$. Let X_{gtr} be the mRNA expression value of gene $g \in \mathcal{G}$ at time $t \in \mathcal{T}$ in replicate $r \in \mathcal{R}$. For simplicity of notation X_{gt} is taken to be the mean mRNA expression value of gene g at time t averaged over replicates. The average expression profile for gene g is denoted $X_{g\cdot} = [X_{g1}, X_{g2}, \cdots, X_{gT}]$.

A gene expression profile, $X_{g\cdot}$, details the expression of a gene over different experimental conditions. In the case of microarray time series the conditions are discrete time points. It can be informative to cluster gene expression profiles to find groups of genes that are co-expressed, as they may be under similar regulatory control (for examples Tavazoie et al. (1999); Yona et al. (2006)). Co-expression can be measured by comparing the expression profiles of genes, either by applying a similarity measure or by fitting statistical distributions to the data. Clustering algorithms are widely used for the purpose of inferring sets of co-expressed genes from gene expression data (e.g. Eisen et al. (1998); Tavazoie et al. (1999); Ghosh and Chinnaiyan (2002); Heard et al. (2005); Thalamuthu et al. (2006)). Such algorithms seek to partition the set of genes into subsets whose average expression profiles are more similar within subsets than between them. Some commonly used clustering methods for gene expression data are introduced in the next two sections.

Clustering gene expression based upon similarity measures

It is common for gene expression profiles to be clustered based upon their similarity (or negative distance), where similarity is defined as a function, S, mapping a pair of gene expression profiles $X_{i\cdot}$ and $X_{j\cdot}$ to a similarity score. Formally, a similarity measure is a function $S : \mathcal{G} \times \mathcal{G} \mapsto \mathbb{R}$.

Similarity measures The most widely used measures of similarity used for co-expression analysis include the (negative) Euclidean distance and PCC. Euclidean distance, d, is the standard measure of distance; it is based on the Euclidean norm. See Equation (2.1) for a definition of the Euclidean norm and Equation (2.2) for a definition of Euclidean distance.

$$Given \quad \underline{x} = [x_1, x_2, \ldots, x_n], \quad \|x\| := \sqrt{\sum_{i=1}^{n} x_i^2} \tag{2.1}$$

$$d(X_{i\cdot}, X_{j\cdot}) := \|X_{i\cdot} - X_{j\cdot}\| \tag{2.2}$$

Equation (2.2) treats the expression profiles as points in a T-dimensional space and defines the distance between pairs of expression profiles by the length of a straight line connecting the points that represent them. PCC is a similarity measure that scores pairs of expression profiles highly if they are linearly proportional, see Equation (2.3) for a definition.

$$PCC(X_{i\cdot}, X_{j\cdot}) := \frac{\sum_{t=1}^{T}(X_{it} - \bar{X}_{i\cdot})(X_{jt} - \bar{X}_{j\cdot})}{\sqrt{\sum_{t=1}^{T}(X_{it} - \bar{X}_{i\cdot})^2}\sqrt{\sum_{t=1}^{T}(X_{jt} - \bar{X}_{j\cdot})^2}} \tag{2.3}$$

Where $\bar{X}_{g\cdot} := \frac{1}{T}\sum_{t=1}^{T} X_{gt}$ is the mean expression value of gene $g \in \mathcal{G}$.

PCC and (negative) Euclidean distance are equivalent, in rank terms, when applied to standardised data, i.e. data that has been transformed to have mean equal to zero and standard deviation equal to one. Equivalent in rank terms means that there exists a strictly increasing function, f, mapping one to the other. The definition of strictly increasing for function f is as follows, $\forall a > b \Rightarrow f(a) > f(b)$.

$$\textit{Given} \quad \bar{X}_{k\cdot} = 0 \quad \textit{and} \quad \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(X_{kt} - \bar{X}_{k\cdot})^2} = 1 \quad \forall k \in \mathcal{G} \tag{2.4}$$

$$PCC(X_{i\cdot}, X_{j\cdot}) := \frac{\sum_{t=1}^{T}(X_{it} - \bar{X}_{i\cdot})(X_{jt} - \bar{X}_{j\cdot})}{(T-1)^{-1}\sqrt{\sum_{t=1}^{T}(X_{it} - \bar{X}_{i\cdot})^2}\sqrt{\sum_{t=1}^{T}(X_{jt} - \bar{X}_{j\cdot})^2}} \tag{2.5}$$

$$:= \frac{(T-1)\sum_{t=1}^{T}(X_{it} - \bar{X}_{i\cdot})(X_{jt} - \bar{X}_{j\cdot})}{\sqrt{\sum_{t=1}^{T}(X_{it} - \bar{X}_{i\cdot})^2}\sqrt{\sum_{t=1}^{T}(X_{jt} - \bar{X}_{j\cdot})^2}} \tag{2.6}$$

$$= (T-1)\sum_{t=1}^{T} X_{it}X_{jt} \tag{2.7}$$

Which is rank equivalent to $A(T-1)\sum_{t=1}^{T}(X_{it}X_{jt}) - C$, where A and C are positive constants. So we can choose any positive values for A and C, for example $A := \frac{2}{T-1}$ and $C := \sum_{t=1}^{T}(X_{it}^2 + X_{jt}^2)$. Then we have that:

$$A(T-1)\sum_{t=1}^{T} X_{it}X_{jt} - C := \sum_{t=1}^{T} 2X_{it}X_{jt} - \sum_{t=1}^{T}(X_{it}^2 + X_{jt}^2) \tag{2.8}$$

$$= \sum_{t=1}^{T}(2X_{it}X_{jt} - X_{it}^2 - X_{jt}^2) \tag{2.9}$$

$$= -\sum_{t=1}^{T}(X_{it} - X_{jt})^2 \tag{2.10}$$

$$= -\|X_{i\cdot} - X_{j\cdot}\|^2 \tag{2.11}$$

$$= -d(X_{i\cdot}, X_{j\cdot})^2 \tag{2.12}$$

Because Euclidean distance is strictly non-negative and squaring is a strictly in-

creasing function for non-negative numbers, $d(X_{i\cdot}, X_{j\cdot})^2$ is rank equivalent to the Euclidean distance $d(X_{i\cdot}, X_{j\cdot})$.

Qian et al. (2001) have introduced a similarity measure that is based on PCC, but allows the detection of time delayed correlation in time series. This has been shown to predict known and infer novel transcriptional regulation. It has also been shown to score functionally related gene pairs highly, as effectively as PCC and Euclidean distance in most of the test cases (Yona et al., 2006). Here the maximal match score introduced in Qian et al. (2001) is referred to as $\psi(i,j)$, which is defined in the following paragraph. The next three pages roughly reproduce the introduction of Kiddle et al. (2010).

Given a time series of gene expression measurements X_{gt} , Algorithm 1 returns a matrix $\psi(i,j)$ of similarity scores for all gene pairs $(i,j) \in \mathcal{G} \times \mathcal{G}$. Data $X_{g\cdot}$ for each gene profile are assumed to be normalized to mean zero and standard deviation one. For a given pair (i,j) the algorithm uses dynamic programming to build up a matrix Ω^+, which compares and scores each alignment between profiles $X_{i\cdot}$ and $X_{j\cdot}$. Time-delayed anti-correlation is captured in a second matrix Ω^-, whose entries are obtained in a similar manner. Finally, transient correlations are captured by explicitly forcing each entry of Ω^+ and Ω^- to be non-negative. Then, similarity score ψ is simply the highest entry in Ω^+ or Ω^-. The alignment matrices Ω^+ or Ω^- further yield a "match type", which may be positive/negative and simultaneous/delayed and describes the characteristics of the highest scoring alignment.

Algorithm 1 Computation of similarity measure ψ, following Qian et al. (2001), with the notation introduced in Kiddle et al. (2010).

(1) Initialise Ω^+_{t0}, Ω^+_{0t}, Ω^-_{t0} and Ω^-_{0t} equal to zero $\forall t \in \mathcal{T} \cup 0$.
(2) Initialise $t_1 = t_2 = 1$.
(3) Calculate $\Omega^+_{t_1t_2}$ and $\Omega^-_{t_1t_2}$:

$$\Omega^+_{t_1t_2} = \max(\Omega^+_{t_1-1t_2-1} + X_{it_1}X_{jt_2}, 0) \tag{2.13}$$

$$\Omega^-_{t_1t_2} = \max(\Omega^-_{t_1-1t_2-1} - X_{it_1}X_{jt_2}, 0) \tag{2.14}$$

(4) If $t_1 < T$ and $t_2 \leq T$ then set $t_1 = t_1 + 1$ and go to step 3.
(5) If $t_1 = T$ and $t_2 < T$ then set $t_1 = 1$ and $t_2 = t_2 + 1$ and go to step 3.
(6) Let $\omega^+ = \max_{t_1t_2}\{\Omega^+_{t_1t_2}\}$ and $\omega^- = \max_{t_1t_2}\{\Omega^-_{t_1t_2}\}$. Set:
$\psi(i,j) = \max\{\omega^+, \omega^-\}$.

Specifically, if $\omega^+ = \psi$ the profiles have a positive local correlation, whereas if $\omega^- = \psi$ the profiles have a negative local correlation. Likewise, if ψ is achieved at

$\Omega^{+}_{t_1 t_2}$ or $\Omega^{-}_{t_1 t_2}$ with $t_1 = t_2$ then the local correlation is simultaneous and equal to the absolute value of $PCC(X_{i\cdot}, X_{j\cdot})$ as long as the maximum of zero is never used in Equations (2.13) or (2.14) respectively. However, if ψ is achieved at $\Omega^{+}_{t_1 t_2}$ or $\Omega^{-}_{t_1 t_2}$ with $t_1 \neq t_2$ then the local correlation is time delayed and will be approximately equal to a time delayed PCC. This is an approximate relation because zeros may be used in Equations (2.13) or (2.14) respectively, and because normalisation of expression profiles to mean zero and standard deviation occurred across all time-points, not the subsets that are then being compared. The value of $\psi(i,j)$ remains the same if the order of all time points are reversed, reflecting the fact that it is calculated by a dynamic programming approach that is looking for similar sequences of gene expression values by aligning subsequences of gene expression profiles.

Many methods exist for clustering gene expression profiles for a given similarity measure, such as k-means (Steinhaus, 1956), Partitioning Around Medoids (PAM, Kaufman and Roussseeuw (1990)), Affinity Propagation (AP, Frey and Dueck (2007)) and hierarchical clustering (Ward, 1963).

K-means Given a user-set number of clusters K, (Euclidean) K-means seeks to find cluster assignments $c(i), c : \mathcal{G} \mapsto \mathcal{K} = \{1, ..., K\}$ and corresponding cluster means $\{\mu_k\}_{k \in \mathcal{K}}$ which minimise the following cost function:

$$J(\{c(g)\}, \{\mu_k\}) = \sum_{k \in \mathcal{K}} \sum_{g : c(g) = k} \|X_{g\cdot} - \mu_k\|^2 \tag{2.15}$$

Where $\|\cdot\|$ denotes the Euclidean norm, as defined in Equation (2.1). $\{c(g)\}_{g \in \mathcal{G}}$ and $\{\mu_k\}_{k \in \mathcal{K}}$ are cluster assignments and cluster means, respectively. Clusters, $\mathcal{C}_k$, are defined as $\mathcal{C}_k := \{\ g \in \mathcal{G} \mid c(g) = k\ \}$, and the cluster means are defined by $\mu_k := \frac{1}{|\mathcal{C}_k|} \sum_{g \in \mathcal{C}_k} X_{g\cdot}$.

K-means attempts to minimise cost function (2.15) by means of an iterative procedure in which the computation of cluster means alternates with the assignment of genes to clusters (Steinhaus, 1956). K-means can be applied to many other similarity measures, as long as a sensible mean can be defined for a group of gene expression profiles.

K-centres (also called K-medoids) The K-means cost function, Equation (2.15), directly uses cluster means $\{\mu_k\}$. In contrast, a matrix of similarities $S(i,j)$, $i, j \in \mathcal{I}$ between gene expression profiles may not give an analogue to the cluster mean. In this setting, a standard approach is to characterise a cluster by an observation within

that cluster, referred to as the centre (also known as the medoid or exemplar) of the cluster. This leads to the following cost function:

$$J(\{e(g)\}) \;=\; -\sum_{g\in\mathcal{G}\,:\,g\neq e(g)} S(g,e(g)) \tag{2.16}$$

Where $e : \mathcal{G} \mapsto \mathcal{E} \subset \mathcal{G}$ is a cluster assignment function, which in this case maps genes to the (index of) the corresponding cluster centre/medoid/exemplar, which is itself a gene. Conversely, $\mathcal{E} = \{E_1, E_2, \cdots, E_K\}$ is defined to be the set of cluster centres, i.e. $\mathcal{E} := e(\mathcal{G})$. Here clusters, $\mathcal{C}_k$, are defined as $\mathcal{C}_k := \{\, g \in \mathcal{G} \mid e(g) = E_k \,\}$.

PAM is a K-means-like algorithm for optimising Equation (2.16) for a user-set number of clusters K (Kaufman and Roussseeuw, 1990). Instead of using means to characterise a cluster, PAM uses data points as cluster centres. This allows PAM to cluster objects under any similarity measure defined upon them. Like K-means, PAM begins with an initial clustering and seeks to improve it iteratively. This means K-means and PAM are prone to finding local minima of their respective cost functions. Multiple initialisations are typically used to identify different local minima, in the hope that the lowest minimum discovered is the global minimum.

Affinity Propagation More recently a novel algorithm, AP, has been introduced to cluster data under the K-centres cost function (Equation 2.16). Unlike PAM, which considers each potential cluster centre in turn, AP considers all potential cluster centres at once. This is achieved by a message-passing algorithm, whose application to gene expression profiles is briefly described in this section, for full details of AP see Frey and Dueck (2007). Two different kinds of messages are exchanged between expression profiles: *responsibility* $r(i,j)$, which reflects profile j's suitability as a centre for profile i; and *availability* $a(i,j)$, which reflects evidence in favour of i choosing j as its centre.

Update equations. Initially, availabilities $a(i,j)$ are set to zero; "self-similarities" $S(i,i)$ are given a user-set value s; this is discussed below. Then, responsibilities and availabilities are updated sequentially using the following update equations:

$$r(i,j) \leftarrow S(i,j) - \max_{j':j'\neq j}\{a(i,j') + S(i,j')\} \tag{2.17}$$

$$\forall i \neq j, \quad a(i,j) \leftarrow \min\left\{0, r(j,j) + \sum_{i':i'\notin\{i,j\}} \max\{0, r(i',j)\}\right\} \tag{2.18}$$

$$a(j,j) \leftarrow \sum_{i':i' \neq j} \max\{0, r(i',j)\} \tag{2.19}$$

A damping factor $\lambda \in [0,1]$ is used to prevent numerical oscillations: each message is set to a weighted combination of its value from the previous iteration and its updated value, weighted by λ and $1-\lambda$ respectively. Update equations are iterated until cluster centres remain unchanged for a user-set number of iterations. Then cluster centres $e(i)$ are calculated by maximising over the sum of responsibility and availability:

$$e(i) \quad = \quad \underset{j \in \mathcal{I}}{\operatorname{argmax}}\, a(i,j) + r(i,j) \tag{2.20}$$

If $e(i) = i$, then i is a cluster centre.

Algorithm parameters. The self-similarity value s influences the number of clusters discovered, higher values giving a greater number of clusters. However, in contrast to the parameter K in K-means and K-centres, this is not a hard specification; rather, the number of clusters found emerges from data, but is influenced by self-similarity s. In this sense, self-similarity is closer in spirit to a shrinkage/regularization strength or Bayesian hyperparameter than a pre-specified number of clusters. Importantly, this means that a default value for s can give good results for a wide range of problems; as a default s can be set to the median of the (off-diagonal entries of the) similarity matrix S. The damping factor has a default value of $\lambda = 0.9$. The maximum number of iterations is given a default value of 1,000. Finally, by default, convergence is declared if cluster centres remain unchanged for 100 iterations.

Clustering gene expression based upon mixture models

An alternative to clustering based on a similarity measure is to define similarity in terms of statistical distributions. This can be achieved by fitting mixture models to the expression data, with each mixture component representing a group of co-expressed genes. Mixture models can take into account the distribution of noise in the data and allow 'fuzzy' assignments of genes to co-expressed groups, i.e. each gene is assigned a probability of belonging to each group rather than being assigned to a specific group. Thresholds can then be applied to assign genes to specific clusters.

The mixture component each gene belongs to is treated as a hidden variable and is chosen to maximise the fit of the mixture model. Typically mixture models are

fitted to gene expression by optimising their fit, using EM or EM-like approaches. These iterate between calculating the expectation of the mixture model and choosing an estimate of the hidden variables that maximises this expectation (Dempster et al., 1977). These can be applied to gene expression to find groups of co-expressed genes. For example, Heard et al. (2005) have introduced a Bayesian EM-like method to cluster time series, using splines to represent the distribution of gene expression over time (Heard et al., 2006).

Analysis of co-regulation

Co-expression, and clustering methods to reveal groups of co-expressed genes, have been introduced in the preceding sections. Co-expressed genes can be functionally related (Yona et al., 2006), for example they may be under similar regulatory control (Tavazoie et al., 1999). Genes under similar regulatory control are said to be co-regulated. Gene regulation can be inferred by identifying the TF or TFs likely to be responsible for the co-regulation of a group of genes. Sets of co-expressed genes can be analysed for further evidence of co-regulation by the analysis of their promoter sequences, for example by methods such as Multiple EM for Motif Elicitation (MEME), Motif Alignment and Search Tool (MAST) or Analysis of Plant Promoter Linked Elements (APPLES) (Baxter et al, in preparation; Bailey et al., 2006, 2009). These methods scan promoter sequences for over-represented motifs, relative to a background model, searching either for *de novo* motifs or motifs which have been shown in the literature to be the binding sequence of a known TF or group of TFs.

Over-represented motifs in the promoters of co-expressed genes are likely to be important for their regulation and probably represent the binding site of a TF that is regulating gene expression in particular conditions. If the motif is known to be associated with specific TFs, then it can help the researcher to identify a regulator of the co-expressed genes, inferring gene regulation, which can be tested experimentally. As an example, consider the plant specific WRKY TF family, which is named after the conserved amino acid sequence WRKYGQK (Eulgem et al., 2000). The WRKY box, (C/T)TGAC(T/C), is a consensus motif representing the core binding site of many (possibly all) WRKY TFs (de Pater et al., 1996; Rushton et al., 1996; Wang et al., 1998; Chen and Chen, 2000; Cormack et al., 2002). If this motif is found to be over-represented in the promoter of co-expressed genes, a WRKY TF is likely to bind there. However, it is not possible from motif information alone to determine which WRKY TF is binding. More recently, differences in the binding affinities of different WRKY TFs to different sequences adjacent to the core binding site have been found, but more needs to be done to allow specific direct transcriptional regulation to be inferred from sequence alone (Ciolkowski et al., 2008). Other

examples of conserved DNA-binding conservation have been documented in both plants (reviewed in Meshi and Iwabuchi, 1995) and mammals (reviewed in Mitchell and Tjian, 1989).

In summary, groups of co-regulated genes can be found using current approaches, but genome-wide inference of specific regulators in a given condition is an open problem. This is because the binding specificities of TFs to DNA sequences are not well characterised, and because some DNA sequences can be bound by many different TFs.

Inference of gene regulation from gene expression time series using network inference approaches

Many different approaches exist to infer the structure of GRNs from gene expression time series, many based around dynamic BNs or SSMs (for a review see Sima et al., 2009). In Chapter 1 one of these approaches, VBSSM, was introduced.

However, to infer regulation from gene expression these approaches typically require at least as many microarray samples, n, as genes, g. Typically many more genes are differentially expressed in a condition than samples exist, i.e. $g >> n$. This general problem is referred to in the literature as the "curse of dimensionality" (Bellman, 1961) and has been extensively studied in the field of model selection (Cleaskens and Hjort, 2008). A second problem is that the space of potential regulatory networks grows super-exponentially with the number of genes modelled (Robinson, 1973). These can both be overcome by grouping similarly expressed genes and treating them as a single variable (Segal et al., 2003). However, if group A is inferred to regulate group B, then it is not clear which TFs in group A are being inferred to regulate which genes in group B. These groupings make the inferences non-specific except in the case where a regulatory group contains only one TF.

In summary, network inference methods can be applied to gene expression time series to infer the specific regulators of genes even when no knowledge of DNA-binding specificity exists. Unfortunately, these approaches do not scale well, limiting the application of network inference to: instances where a few genes are differentially expressed, to subsets of differentially expressed genes, or to profiles which each represent a group of genes. In experiments where the number of differentially expressed genes, g, is much higher than the number of samples, n, network inference models are typically either genome-wide or specific, but not both. This limits the ability of network inference models to provide experimentally hypotheses about transcriptional regulation in many practical settings.

Phenotype screens as an experimentally tractable first step in the validation of inferred transcriptional regulation

To provide experimental hypotheses Windram (2010) used VBSSM to infer transcriptional regulation from a time series of Arabidopsis gene expression during infection by *B. cinerea* (Denby et al., in preparation). Phenotype testing of mutants of inferred regulatory hubs was then used to confirm their importance in the defence response. This is a very indirect test of the inferred regulation, but it benefits from experimental tractability. To test each inferred regulatory connection would require a test of binding and the effect on expression, whereas to test for altered susceptibility to *B. cinerea* the researcher only needs to infect a knockout or overexpressor mutant of the regulator and observe the result.

Genetic studies can reveal transcriptional regulators having a large effect on phenotype. These important regulators can then be analysed in greater detail by more intensive experimental approaches. In this way it is hoped that researcher time can be used most productively, to validate inferences that are likely to be important for plant adaptation to environmental stress.

2.1.4 Genetic approaches to study the function of genes involved in the defence response

Genetic modification of Arabidopsis by Agrobacteria

Plants can be genetically engineered by *Agrobacterium tumefaciens*, which secretes a plasmid into the plant to induce crown gall disease. This tumour inducing (Ti) plasmid, is capable of transferring part of its DNA sequence into the plant chromosomal DNA (Chilton et al., 1980). This segment of the Ti plasmid, capable of being incorporated into the plant genome, is referred to as Transfer-DNA (T-DNA). Naturally occuring T-DNA contains approximately eight genes, which can be expressed in the plant (Satchel and Nester, 1986; Schrammeijer et al., 2000). These genes cause crown gall disease by encouraging cell division, they also increase production of opines, an important nutrient for *A. tumefaciens*.

This naturally occurring process can be harnessed by researchers to manipulate the plant genome (Chilton, 1983). The genes in the T-DNA responsible for crown gall disease can be removed from the Ti plasmid without impeding its ability to transform plants (Leemans et al., 1981). Antibiotic resistance genes can be added into the T-DNA and then incorporated into the genome of plant cells in a liquid suspension, so that transformed cells can be selected on solid media containing an antibiotic (Bevan et al., 1983). Transformed shoots from solid media can be transferred to

soil and used to grow a transformed plant (Barton et al., 1983). Agrobacterium mediated plant transformation has been adapted, notably with vacuum infiltration (Bechtold and Pelletier, 1998) or floral dip steps (Clough and Bent, 1998) that eliminate the need for plant tissue cultures.

Incorporation of T-DNA occurs at semi-random locations in the plant genome (Zhang et al., 2007). This means that the T-DNA may disrupt functional elements within the genome, or be incorporated in regions where the expression of its genes are repressed. This allows the disruption of native genes, but also means that T-DNA may be inserted at several different locations. It also means that transgenes incorporated into the plant genome by Agrobacteria are subject to a 'location effect' that can modify their expression level.

Arabidopsis gene knockout by T-DNA incorporation

The location of incorporated T-DNA can be revealed by methods such as Thermal Asymmetric Interlaced-PCR (TAIL-PCR; Liu et al., 1995) or sequencing. When T-DNA is incorporated into the coding sequence of a gene, aberrant transcripts will be produced. This is referred to as a gene knockout (Krysan et al., 1999). If both alleles of that gene contain the same insertion, then the plant is homozygous for that gene and can no longer produce transcripts of the correct sequence. This typically prevents the production or function of the protein originally encoded by that gene. This is referred to as a homozygous knockout and can be used in a 'reverse genetics' screen to reveal phenotypic effects of gene disruption (Krysan et al., 1999).

An important resource for Arabidopsis 'reverse genetics' screens is the Salk Institute's homozygous T-DNA collection, a collection of transformed Arabidopsis lines (Alonso et al., 2003). The genome of each line will include incorporated 'disarmed' T-DNA at a known position. The Salk Institute's collection contains different lines with T-DNA inserts in approximately 24,476 genes, roughly two thirds of the Arabidopsis genome, many of which will be knockouts.

A disadvantage of gene knockout by T-DNA incorporation is that multiple T-DNA insertions sometimes occur, each lying at different locations on the chromosome (Zhang et al., 2007). This makes it harder for researchers to attribute phenotypic effect to individual insertions. Multiple independent gene knockouts can be screened to control for the effect of these other insertions.

Constitutive over-expression by the 35S promoter

Novel genes can be introduced into the plant genome, by inserting them into the T-DNA of *A. tumefaciens* before using them to transform plants. This can be used to alter the expression of native genes by combining them with a strong promoter such as the Cauliflower Mosiac Virus 35S promoter (Sanders et al., 1987). A transgene fused to a 35S promoter is usually prefixed with 35S::, for example *35S::MYBL2* which will be introduced later. The gene and strong promoter can be incorporated into T-DNA, and then into the plant genome.

Forward and reverse genetics approaches have revealed key regulators of the defence response of Arabidopsis to infection by *B. cinerea*

As discussed in the previous chapter, the defence response of Arabidopsis to *B. cinerea* typically leads to quantitative resistance, i.e. a reduction in pathogen growth. Pathogen growth can be quantified by RT-PCR of *B. cinerea* mRNA encoding housekeeping genes, whose transcripts are produced at a constant rate in most conditions. Commonly used *B. cinerea* 'housekeeping' gene are *BcActA* and *BcTubulin*; the mRNA level of these genes seems to correlate with infection lesion area in studies using single droplet inoculation on single leaves (Mengiste et al., 2003; Zheng et al., 2006). Alternatively, whole plants and/or spray inoculation can be used (Thomma et al., 1999; Berrocal-Lobo et al., 2002). Other measures of *B. cinerea* growth on Arabidopsis include disease severity ratings (Berrocal-Lobo et al., 2002; Pré et al., 2008) and proportion of decayed plants (Veronese et al., 2006). Altered susceptibility phenotypes observed through these different measures seem to be broadly comparable, for example Denby et al. (2004) have been shown that susceptibility of lines correlate between whole plant and detached leaf droplet infection assays.

One 'forward genetics' approach is to take many T-DNA lines, for which the location of insertion is not known and to screen them for altered phenotypes. Lines displaying phenotypes can then be studied to reveal the location of T-DNA, and therefore the genomic sequence responsible for the altered phenotype. This has been applied to study the genetics of quantitative resistance of Arabidopsis to infection by *B. cinerea*, and has revealed the *bos1* line to have a strong susceptibility phenotype and to be disrupted in the promoter and 5' untranslated region (UTR) of the TF encoding gene *MYB108* (Mengiste et al., 2003). This strong susceptibility phenotype will be used as a positive control in the phenotype screens presented later. An alternative 'forward genetics' approach was taken by Bessire et al. (2007) who tested a collection of Ethylmethanesulfonate (EMS)-mutagenized Arabidopsis plants for altered susceptibility to *B. cinerea*. This revealed a line, *botrytis-resistant1*

(*bre1*), that had a strong resistance phenotype. This phenotype was found to be due to a mutation in the coding sequence of the gene *LACS2* (Bessire et al., 2007).

A study by AbuQamar et al. (2006) shows a 'reverse genetics' approach, where T-DNA knockout lines of 14 TFs up-regulated during *B. cinerea* infection were screened for altered susceptibility to *B. cinerea*. Two of the 14 mutants, *wrky70* and *zfar1*, were found to have altered susceptibility to *B. cinerea* (AbuQamar et al., 2006). There is now substantially more data available on Arabidopsis gene expression during infection by *B. cinerea* (Denby et al., manuscript in preparation; Ferrari et al., 2007; Mulema and Denby, 2012). This has led to a greater knowledge of differential expression, with an order of magnitude more differentially expressed genes than had been previously published. It is possible that the phenotype rate of 'reverse genetics' can be increased by using this gene expression data to identify key regulators of the defence response.

The aim of this chapter is to discover novel regulators of the defence response by extending the approach of Windram (2010) i.e. using gene expression time series to infer gene regulation. Windram (2010) inferred gene regulation by applying network inference to gene lists chosen manually based on the literature. Inferred regulators were studied in a 'reverse genetics' screen and an increase in phenotype rate was seen. It would be desirable to be able to infer regulation among larger gene groups to allow more novel regulators to be found and to remove the need for manual and therefore subjective selection of genes to model together.

2.2 Results

2.2.1 Analysis of a time series of gene expression in Arabidopsis leaves during *B. cinerea* infection

Experimental conditions

The experiment was performed by Oliver Windram, Priyadharshini Madhou, Cunjin Zhang and Alex Tabrett, as described in this section. This section gives a brief overview of the experimental details, full technical details can be found in Denby et al., (in preparation).

192 Col0 Arabidopsis plants were grown in a 16:8 hour light:dark cycle (lights on 04:00 to 20:00), with leaf 7 marked after 25 days of growth. After three additional days, leaf 7 was detached from the plants and placed on 0.8% agar that had been allowed to set in the base of a propagator tray. The leaves were then treated with

5–7 (depending on leaf size) 10 μl droplets of either a mock or *B. cinerea* spore containing solution at 10am and transferred to a growth room with approximately 90% humidity, but the same lighting and temperature settings. Mock solution consisted of sterile half strength grape juice. *B. cinerea* spore solution similarly consisted of half strength grape juice, but also contained *B. cinerea* spores from the Pepper strain at a concentration of 10^5 spores/ml.

Analysis of transcripts by microarray hybridisation

Microarray analysis was performed by Oliver Windram, Priyadharshini Madhou, Cunjin Zhang and Alex Tabrett, as described in this section. Technical details will be presented in Denby et al., (in preparation).

Leaves were collected, snap frozen in liquid nitrogen and transferred to a $-80\,^{\circ}C$ freezer every 2 hours for a total of 48 hours post infection (hpi) or mock treatment respectively. Four biological replicates of both mock and Botrytis-treated leaves were collected at each time-point. This gave two time series (mock/infected) of 24 time-points each, with 4 biological replicates for each treatment and time-point.

All the steps in this paragraph are described fully in Breeze et al. (2011). Total RNA was extracted from the leaves, purified and amplified. Amplified RNA was reverse transcribed, with cye dye incorporated into the cDNA. Labelled cDNA was then purified, concentrated and hybridised to Complete Arabidopsis Transcriptome Microarray (CATMA) slides (Allemeersch et al., 2005). An average of 3 technical replicates were used to control for technical variation caused by dye, printing and spatial biases.

9,838 genes are differentially expressed between infected and mock time series

Expression values for each CATMA probe at each time point, in each biological replicate, were extracted using a mixed model ANOVA. This was performed by Oliver Windram in a version of the R package MicroArray ANOVA (MAANOVA, Wu et al., 2003), that had been adapted by Stuart McHattie (McHattie and Mead, in preparation).

Katherine Denby analysed differential expression between the mock and infected time series; this paragraph describes her work. A list of 10,600 differentially expressed genes were returned by a Gaussian process two-sample test with a cutoff decided by manual inspection. This list was supplemented by 236 additional genes found to be differentially expressed by an F-test performed within MAANOVA and

confirmed by manual inspection. 371 probes were removed when re-annotation indicated that they did not hybridise to open reading frames. 629 probes were removed as they were found to duplicate the gene matched by another differentially expressed probe.

The resulting list of differentially expressed genes included 9,838 genes, 643 of which are labelled as TFs by the *Arabidopsis thaliana* Transcription Factor DataBase (AtTFDB – http://arabidopsis.med.ohio-state.edu/AtTFDB).

Analysis of co-expression and co-regulation

The expression profiles of differentially expressed genes were analysed by cluster analysis. Two clustering methods were used, SplineCluster (Heard et al., 2005) which was applied by myself, Claire Hill and Oliver Windram, and AP clustering based on PCC (Frey and Dueck, 2007) which was applied by myself. Each clustering method was performed several times with a range of parameter values, returning different numbers of clusters.

Promoters of co-clustered/co-expressed genes were analysed by Richard Hickman; this paragraph describes his work. Clusters of co-expressed genes were analysed for over-representation of known TF binding sequences in the 500 bp upstream of their transcriptional start site, using APPLES (Baxter et al., in preparation), as described in Breeze et al. (2011). Known binding motifs for plant TFs were collected in the form of Position Specific Scoring Matrices (PSSMs) from the TRANSFAC® database (Matys et al., 2006) and the Plant Cis-acting regulatory DNA Elements (PLACE) database (Higo et al., 1999). In addition, two NAM/ATAF/CUC (NAC) TF binding motifs were taken from the literature (Olsen et al., 2004). Motif over-representation was performed by comparing the number of motif hits in a set of promoter sequences to the occurrence of motif hits in a randomly generated sequence of one million bp, generated from a 3rd order Markov model with parameters learned from the whole Arabidopsis genome.

The highest number of over-represented known motifs were found in a clustering performed using SplineCluster with iterative reclassification (Heard et al., 2005; Heard, Ahead of print), with a prior precision of 0.001, which clustered the differentially expressed genes into 44 groups (Supplemental Digital Information Table 1). These clusters contained a median of 214 genes, with a lower quartile of 96 genes and an upper quartile of 296.5 genes. While this clustering produced the optimal grouping for motif analysis, other clusterings did allow different known motifs to be found.

Over-represented known motifs are likely to represent the binding sites of regulators of the co-expressed genes in Arabidopsis during infection by *B. cinerea*. A key challenge is to identify the specific TF or TFs acting through this binding site to regulate expression in this experimental condition.

2.2.2 Inference of gene regulation by integrated analysis of gene expression and promoter sequence – a novel integrative approach

Cluster analysis can identify co-expressed genes, which may be co-regulated. Known motif over-representation analysis can then infer regulators that may be responsible. However, the DNA-binding specificity of TFs (i.e. the sequence motifs that they are able to bind) are poorly characterised in most organisms, even in model organisms such as Arabidopsis. An especially important knowledge gap involves how the sequence binding specificity varies among highly conserved TFs. This usually limits motif analysis to inferring TF families responsible for co-regulation. Here, a novel integrative approach is introduced to infer specific regulators of co-regulated genes, and is applied to the gene expression time series introduced in Section 2.2.1.

The approach that will be presented here is conceptual, consisting of three stages that can in principle be performed by any appropriate algorithm. These stages consist of: clustering co-expressed genes; over-representation analysis of experimentally derived binding motifs; and network inference based on the expression of potentially co-regulated genes and TFs associated with over-represented motifs found in their promoters.

This novel integrative approach was applied to the time series of gene expression introduced in Section 2.2.1. In the application that follows, clusters and over-represented motifs are taken from the analysis presented in Section 2.2.1. I.e. co-expressed genes had been clustered by SplineCluster (Heard et al., 2005), and motif over-representation analysed using the APPLES software package (Baxter et al., in preparation). The final stage, network inference, is performed in VBSSM (Beal et al., 2005).

Application to the time series of Arabidopsis gene expression during Botrytis infection

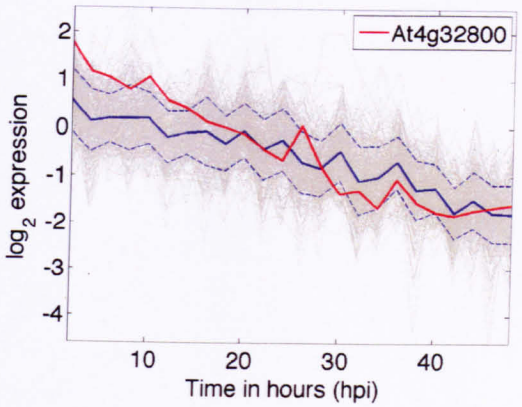
Groups of co-expressed genes in Arabidopsis during *B. cinerea* infection have been identified in Section 2.2.1, and these groups of genes have been analysed for over-representation of known TF binding motifs in their promoters. Here, clusters were studied that had the known DNA-binding motif of either the APETALA2-Ethylene

Responsive Element Binding Proteins (AP2-EREBPs; Riechmann and Meyerowitz, 1998), NAC (Olsen et al., 2004) or WRKY (Eulgem et al., 2000) defence related TF families over-represented in their promoters. Three additional clusters (16, 26, and 29, Supplementary Digital Information table 1) were over-represented for the WRKY binding motif; they were not analysed with the integrative approach because in each >100 genes had over-representation of the WRKY motif in their promoter sequence. The time series used contains a total of 96 microarray measurements, and so modelling >128 genes (genes in the clusters mentioned above and genes encoding differentially expressed WRKY TFs) using it is potentially problematic, as discussed in Section 2.1.3.

VBSSM was applied to data from the *B. cinerea* infection time series introduced in Section 2.2.1. VBSSM was chosen because it had performed comparably well against competitors in a benchmark study on a synthetic *in vivo* yeast GRN (Penfold and Wild, 2011). VBSSM was applied to $\{X_{gtr}\}$ for each of the four clusters, where $g \in$ the set of co-regulated genes or the set of potential regulators, $t \in \{1, 2, \ldots, 24\}$ and $r \in \{1, 2, 3, 4\}$. Figures 2.1–2.4 show the results for the four different clusters; in all cases 20 different VBSSM initialisations were used. The state space model was fitted with 1 to 20 hidden state vector dimensions, with the dimensionality of the hidden state vector that maximised the marginal likelihood (as determined by VBSSM) chosen as the final model. In all network diagrams, any edge whose Gaussian posterior probability distribution (as determined by VBSSM) had a mean over 3 standard deviations from zero in at least one initialisation is shown. This cut-off will be referred to as a z-score of 3. It is important to consider that inferred regulation may not in itself be significant, as randomly selected genes will be inferred to regulate each other, i.e. several randomly chosen groups of 50 genes were entered into VBSSM, and many inferred regulatory links were found with a threshold of z=3.[1]

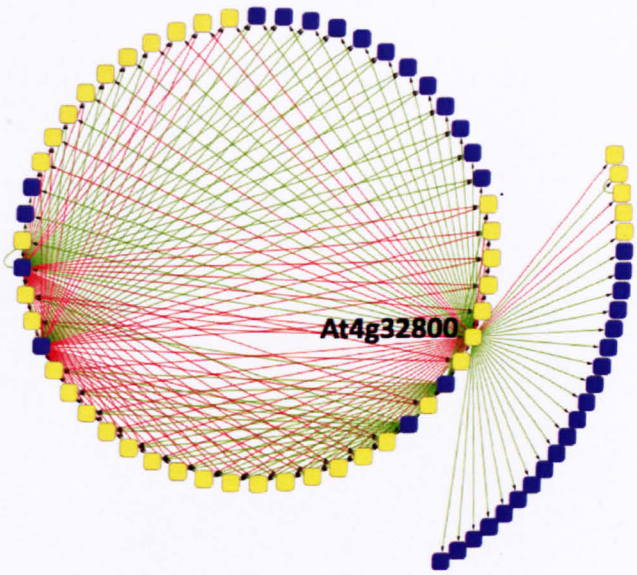
At4g32800 is inferred to co-regulate 38 genes Cluster 22 contained 307 genes (Figure 2.1(a) and Supplemental Digital Information Table 1), 56 of which had an Ethylene Responsive Element (ERE) (Figure 2.1(b) which matches the motif identified in Ohme-Takagi and Shinshi (1995), its TRANSFAC® identifier is M01057) within 500 bp of their transcriptional start site. The expression of these 56 potentially co-regulated genes were modelled together with that of the 53 differentially expressed AP2-EREBP TFs and the resulting inferred GRN is shown in Figure

[1]It may be that the marginal likelihood of the SSM, rather than that of specific edges, will allow SSMs inferred from random sets of genes to be distinguished from those inferred from the expression of genes which regulate each others expression.



(a) Average expression of genes in cluster and predicted regulators

(b) Ethylene Responsive Element



(c) Inferred gene regulation

Figure 2.1: Inferred regulation of potentially co-regulated genes. (a) The expression profiles of genes in cluster 22 and a TF (in red) which is inferred to regulate 38 of them. Gene expression profiles are averaged over biological and technical replicates, and then transformed to zero mean and standard deviation one. Expression of cluster members are shown in grey. The mean expression over all cluster members is plotted in blue. An interval of one standard deviation at each time point, for all cluster members, is indicated as a pair of dashed blue line. (b) shows the sequence logo of the ERE position specific scoring matrix (PSSM) which is over-represented in the promoters of these co-expressed genes. (c) shows gene regulation inferred by VBSSM, with 1 hidden state and a threshold z-score of 3. Blue nodes are co-expressed (co-clustered) genes and contain the known binding sequence of the AP2-ERE BP TF family. The yellow nodes indicate members of the AP2-ERE BP TF family. Green arrows indicate inferred positive regulation. Red arrows indicate predicated negative regulation.

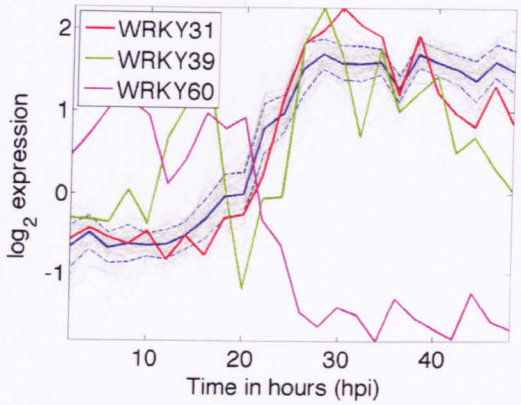
2.1(c). (For a list of the differentially expressed AP2-EREBP TFs, and potentially co-regulated targets, that were modelled together see Appendix tables A.1 and A.2).

At4g32800 is inferred to co-regulate the 38 genes: *At1g01960, At1g08720, At1g11960, At1g15240, At1g19580, At1g20693, At1g26900, At1g64520, At1g75010, At2g01820, At2g17670, At2g32400, At2g32410, At2g37195, At3g01340, At3g11250, At3g12280, At3g17300, At3g22320, At3g24200, At3g45890, At3g50860, At3g59990, At3g60830, At3g62370, At4g07410, At4g17910, At4g21710, At4g25550, At5g10910, At5g11030, At5g13850, At5g24740, At5g26830, At5g27990, At5g59980, At5g66880* and *At5g67530.* To assess the robustness of this inference, VBSSM was applied again to the same dataset without the first time-point, inferring *RAP2.6* to be the main regulator instead (Appendix figure B.1).

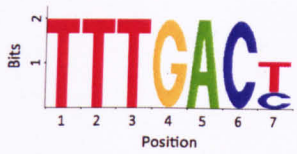
WRKY31 is inferred to co-regulate 9 genes Cluster 24 contained 56 genes (Figure 2.2(a) and Supplemental Digital Information Table 1), 12 of which had a WRKY binding motif (Figure 2.2(b) which matches the WRKY binding sequence identified in de Pater et al. (1996), its PLACE identifier is S000310) within 500 bp of their transcriptional start site. The expression of these 12 potentially co-regulated genes were modelled together with that of the 29 differentially expressed WRKY TFs and the resulting inferred GRN is shown in Figure 2.2(c). (For a list of the differentially expressed WRKY TFs, and inferred co-regulated targets, that were modelled together see Appendix table A.3).

WRKY31 is inferred to co-regulate the 9 genes; *At1g05575, At1g26380, At1g30700, At1g69930, At1g76600, At2g17500, At2g47190, At3g25250* and *At4g21830.* In addition, *WRKY60* is inferred to regulate *At2g17500* and *At2g21830*, and *WRKY39* is inferred to regulate *At1g75500.* To assess the robustness of this inference, VBSSM was applied again to the same dataset without the first time-point, inferring *WRKY31* to be the main regulator again (Appendix figure B.2).

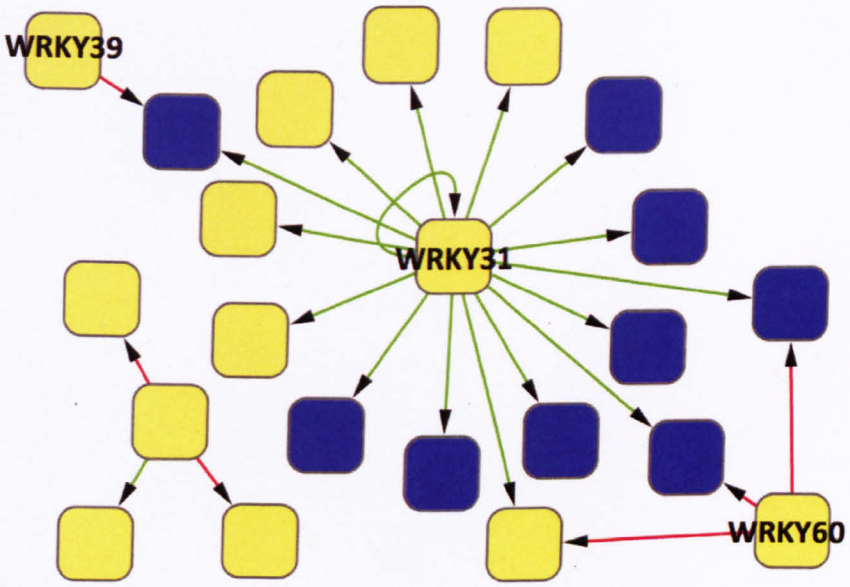
ANAC092 and ANAC019 are inferred to co-regulate 9 genes each, including two genes inferred to be jointly regulated Cluster 27 contained 195 genes (Figure 2.3(a) and Supplemental Digital Information Table 1), 34 of which had a NAC-like binding motif (PSSM M00040 from TRANSFAC® as shown in Figure 2.3(b) which fits to a subsequence of the consensus sequence presented in Olson et al. (2005)) within 500 bp of their transcriptional start site. The expression of these 56 potentially co-regulated genes were modelled together with that of the 34 differentially expressed NAC TFs and the resulting inferred GRN is shown in Figure



(a) Average expression of genes in cluster and predicted regulators

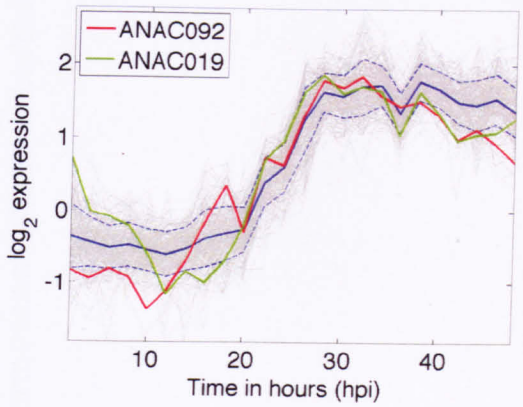


(b) WRKY binding motif



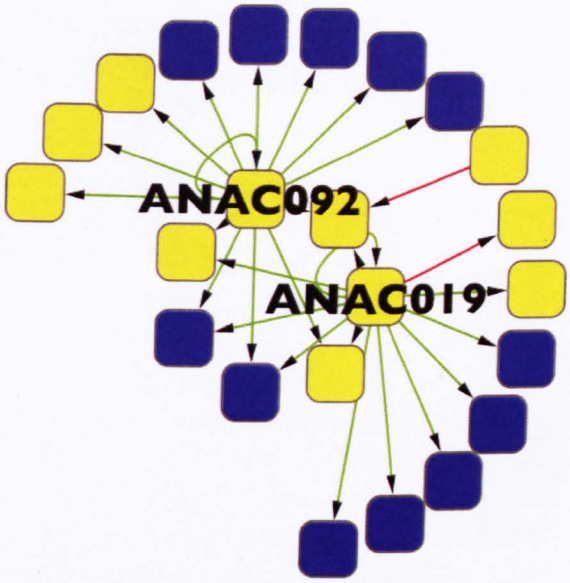
(c) Inferred gene regulation

Figure 2.2: Inferred regulation of potentially co-regulated genes. (a) shows the expression profile of genes in cluster 24 and 3 TFs (in red, green and mauve) which are inferred to regulate some of them. Gene expression profiles are averaged over biological and technical replicates, and then transformed to zero mean and standard deviation one. Expression of cluster members are shown in grey. The mean expression over all cluster members is plotted in blue. An interval of one standard deviation at each time point, for all cluster members, is indicated as a pair of dashed blue line. (b) shows the sequence logo of the WRKY PSSM which is over-represented in the promoters of these co-expressed genes. (c) shows gene regulation inferred by VBSSM, with 7 hidden states and a threshold z-score of 3. Blue nodes are co-expressed (co-clustered) genes and contain the known binding sequence of the WRKY TF family. The yellow nodes indicate members of the WRKY TF family.



(a) Average expression of genes in cluster and predicted regulators

(b) NAC-like binding motif



(c) Inferred gene regulation

Figure 2.3: Inferred regulation of potentially co-regulated genes. (a) shows the expression profile of genes in cluster 27 and two TFs (in red and green) which are inferred to regulate some of them. Gene expression profiles are averaged over biological and technical replicates, and then transformed to zero mean and standard deviation one. Expression of cluster members are shown in grey. The mean expression over all cluster members is plotted in blue. An interval of one standard deviation at each time point, for all cluster members, is indicated as a pair of dashed blue line. (b) shows the sequence logo of the NAC-like PSSM M00040, from TRANSFAC®, which is over-represented in the promoters of these co-expressed genes. (c) shows gene regulation inferred by VBSSM, with 7 hidden states and a threshold z-score of 3. Blue nodes are co-expressed (co-clustered) genes and contain the known binding sequence of the NAC TF family. The yellow nodes indicate members of the NAC TF family.

2.3(c). (For a list of the differentially expressed NAC TFs, and inferred co-regulated targets, that were modelled together see Appendix table A.4).

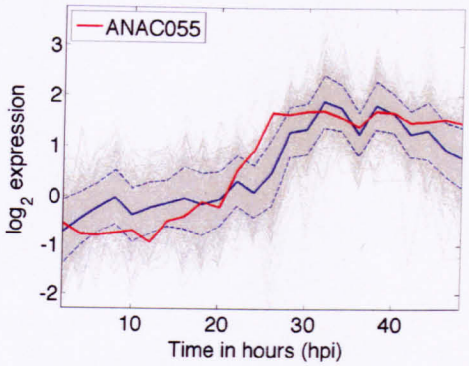
ANAC092 is inferred to co-regulate the 7 genes; *At1g32120, At3g48890, At3g52540, At3g53400, At3g60130, At4g14680* and *At5g27520. ANAC019* is inferred to co-regulate the 7 genes; *At1g21310, At1g28480, At1g71100, At3g48890, At4g18950, At5g13500* and *At5g27520.* Two of these, *At3g48890* and *At5g27520*, are inferred to be regulated by both *ANAC019* and *ANAC092.* To assess the robustness of this inference, VBSSM was applied again to the same dataset without the first time-point, inferring *ANAC055* to be the main regulator instead (Appendix figure B.3).

ANAC055 is inferred to co-regulate 43 genes Cluster 38 contained 326 genes (Figure 2.4(a) and Supplemental Digital Information Table 1), 47 of which had a NAC-like binding site (Figure 2.4(b) which fits to a subsequence of the consensus sequence presented in Olson et al. (2005)) within 500 bp of their transcriptional start site. The expression of these 47 potentially co-regulated genes were modelled together with that of the 34 differentially expressed NAC TFs and the resulting inferred GRN is shown in Figure 2.4(c). (For a list of the differentially expressed NAC TFs, and potentially co-regulated targets, that were modelled together see Appendix tables A.4(a) and A.5).

ANAC055 is inferred to co-regulate the 43 genes; *At1g09180, At1g09960, At1g12200, At1g12820, At1g23100, At1g27000, At1g27100, At1g52550, At1g71180, At1g75270, At2g01850, At2g26230, At2g29700, At2g39780, At2g40420, At2g43540, At3g11200, At3g12100, At3g18520, At3g48140, At3g51990, At3g55390, At3g57785, At3g60020, At3g61680, AT3G62830, At3g63260, At4g01410, At4g05590, At4g14010, At4g23530, At4g29580, At4g31300, At5g03290, At5g11090, At5g11960, At5g20120, At5g20650, At5g25050, At5g27710, At5g47200, At5g60580* and *At5g64250.* To assess the robustness of this inference, VBSSM was applied again to the same dataset without the first time-point, inferring *ANAC055* to be the main regulator again (Appendix figure B.4).

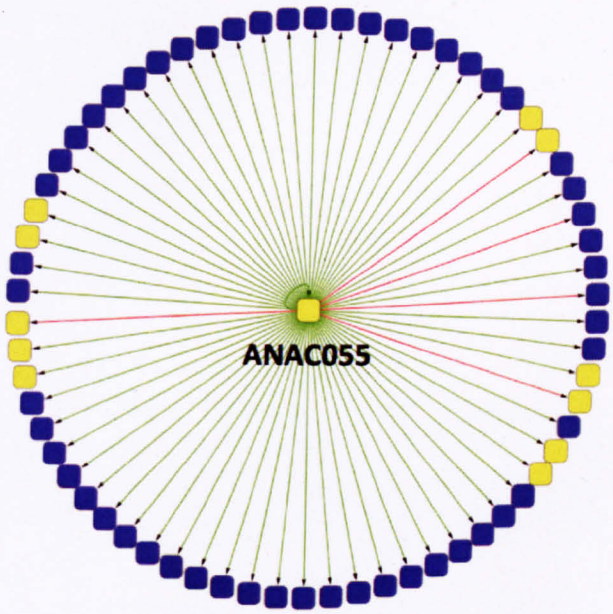
Comparison of results to existing literature

Experimental studies of the TFs that have been inferred to regulate *B. cinerea* responsive gene expression, can be used to assess the plausibility of the inferred regulation. Altered *B. cinerea* susceptibility phenotypes and experimental analysis of transcriptional regulation can be compared to what has been inferred.



(a) Average expression of genes in cluster and predicted regulators

(b) NAC-like binding motif



(c) Inferred gene regulation

Figure 2.4: Inferred regulation of potentially co-regulated genes. (a) shows the expression profile of genes in cluster 38 and a TF (in red) that is inferred to regulate 43 of them. Gene expression profiles are averaged over biological and technical replicates, and then transformed to zero mean and standard deviation one. Expression of cluster members are shown in grey. The mean expression over all cluster members is plotted in blue. An interval of one standard deviation at each time point, for all cluster members, is indicated as a pair of dashed blue line. (b) shows the sequence logo of the NAC-like PSSM M00040 (from TRANSFAC®) which is over-represented in the promoters of these co-expressed genes. (c) shows gene regulation inferred by VBSSM, with 9 hidden states and a threshold z-score of 3. Blue nodes are co-expressed (co-clustered) genes and contain the known binding sequence of the NAC TF family. The yellow nodes indicate members of the NAC TF family. Green arrows indicate inferred positive regulation. Red arrows indicate predicated negative regulation.

Phenotype *ANAC019*, *ANAC055*, *ANAC092*, *WRKY31* and the AP2-ERE BP TF *At4g32800* have been inferred in this chapter to regulate genes in response to infection. All three NAC TFs inferred to regulate co-regulated genes have mutants which are known to have altered susceptibility to *B. cinerea* (Bu et al., 2008; Windram, 2010). This confirms the importance of these NAC TFs in the defence response, but does not test the inferred regulation. Unfortunately mutants of *WRKY31* or *At4g32800* were not available and so it was not possible to test their effect on susceptibility to *B. cinerea*.

Regulation In the case of *ANAC092*, microarray data already exists for an estradiol inducible (i.e. transient) overexpressor compared to wildtype in liquid media grown seedlings (Balazadeh et al., 2010). This data has been analysed in Windram (2010), with 123 genes found to be differentially expressed. The differentially expressed gene list has no overlap with the seven genes inferred to be targets in Figure 2.3(c).

In the case of *ANAC019* and *ANAC055*, microarray data exists for knockout plants compared to wildtype during leaf senescence (Hickman et al., in preparation). None of the inferred targets of *ANAC019* were differentially expressed in the *anac019* knockout leaves. One of the 43 inferred targets of *ANAC055*, *At5g27710* is differentially expressed in the *anac055* knockout plant, but this overlap is likely to occur by chance. (Number of differentially expressed genes after *B. cinerea* infection = 9,838 ; number of these differentially expressed in senescing *anac055* plant = 230; number of inferred targets of *ANAC055* = 43 and an overlap of 1 giving a cumulative hypergeometric p-value of 0.6392 ascalculated in MATLAB®).

In this section specific transcriptional regulation has been inferred, based on both gene expression and known binding sequences of TF families. This methodology is less biased to the literature than the manual selection used in Windram (2010), but is limited to inferring regulation by TF families with well characterised DNA-binding sequences.

2.2.3 Temporal clustering by affinity propagation predicts transcriptional regulation genome-wide

All of the work presented in this section, except for the application of TCAP to the expression all 9,838 differentially expressed genes, was originally published in Kiddle et al. (2010).

Identifying functionally related genes is an important task in the exploratory analy-

sis of gene expression time series. One such relation is transcriptional regulation and can be studied by clustering gene expression, as shown in Section 2.2.1. However, most clustering methods are designed to find co-expressed genes rather than to infer regulation, this is why network inference was used in the previous section. The gene expression profiles of the targets of a transcriptional activator could be correlated to it, but if a time series has sufficient temporal resolution then it should be possible to observe a delay between the transcriptional activation of a TF and its targets (Qian et al., 2001). Similarly, the expression profile of a transcriptional repressor may be anti-correlated to its targets (Qian et al., 2001, 2003). The aim is predict genome-wide transcriptional regulation during *B. cinerea* by studying time-delayed correlation in the time series of gene expression introduced in Section 2.2.1.

TCAP algorithm

Here a novel method, Temporal Clustering by Affinity Propagation (TCAP), is introduced that infers gene regulation by taking into account these features in gene expression time series. The aim of TCAP is to group genes whose expression profiles show time-delayed correlation/anti-correlation. Later this will be shown to be relevant to predicting gene regulation. TCAP, which is implemented in MATLAB®, consists of three main stages: calculation of Qian similarity, clustering by AP and the creation of output files. It is available to download from http://www.wsbc.warwick.ac.uk/stevenkiddle/tcap.html.

Calculation of Qian similarity, ψ, is performed for every pair of genes following the definition in Section 2.1.3. As Qian similarity is symmetric (i.e. $\psi(i,j) = \psi(j,i)$), it is only calculated for $i > j$. These similarity scores are then clustered by AP with the self-similarity parameter set to the median value of ψ across all gene pairs. The resulting clusters are stored in a spreadsheet and several plots are produced. These plots show the expression profiles of co-clustered genes in various ways, for example colour coding or aligning profiles by their time-delay, as returned by the Qian similarity algorithm defined in Section 2.1.3.

Benchmarking results

Validation of similarity measure ψ Expression time series, of genes and their known regulators, were used to validate the ability of similarity measure ψ to correctly predict transcriptional regulation. To this end two biological examples were used, from *Saccharomyces cerevisiae* (yeast) and Arabidopsis, respectively, in which the underlying biology is relatively well understood.

The *S. cerevisiae* genome has been well studied and provides a number of vali-

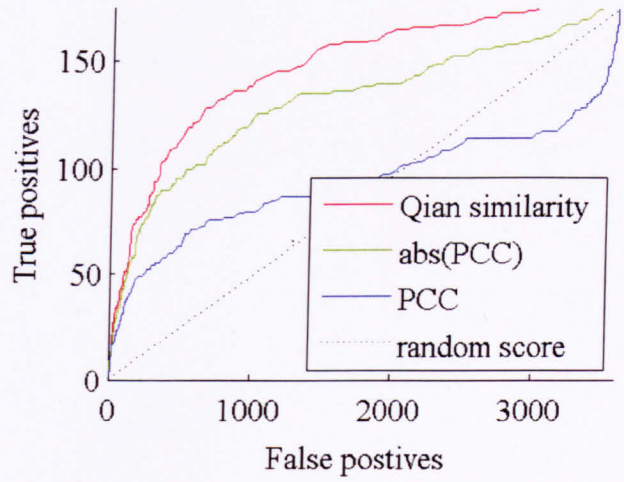
dated TF-target pairs. Published microarray time series (Gasch et al., 2000; Qian et al., 2003; Spellman et al., 1998) of such regulatory pairs, consisting of validated positive examples and randomly selected negative examples, were used. The TF-target pairs and the compiled expression data was obtained from the supplementary data of Qian et al. (2003), which is described in the next paragraph.

The positive examples had been chosen from TRANSFAC®(Matys et al., 2006) and the Saccharomyces cerevisiae Promoter Database (Zhu and Zhang, 1999); negative examples had been identified by finding genes without the known binding site of the TF or permuting the target gene's (but not the TF's) expression profile. The expression profiles cover a total of 79 time points, which gives a relatively high time resolution. Unfortunately, expression profiles from different experiments had been concatenated, with insufficient labelling to allow them to be separated. The concatenated time series also contained experiments with different sampling frequencies. Whilst not ideal for analysis using ψ, which assumes even sampling frequencies and separate experiments, this dataset has well characterised interactions and a relatively large number of observations.

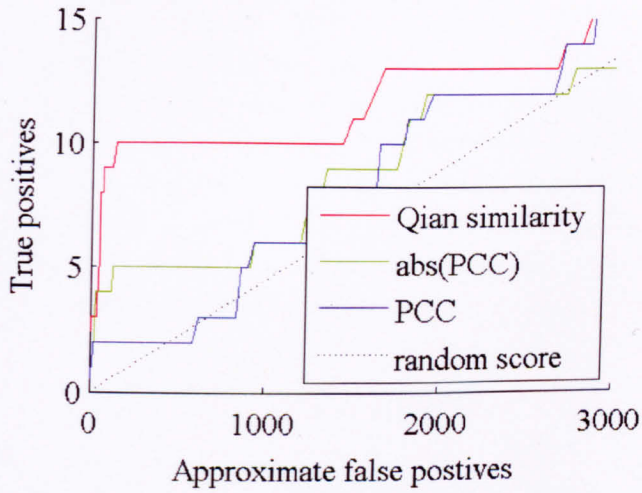
The ability of the similarity score ψ to correctly distinguish between positive and negative examples of transcriptional regulation was studied by means of a receiver operator characteristic (ROC) analysis. Similarity scores ψ(i,j) for each TF-target pair – positive and negative – were thresholded to yield inferred positive TF-target pairs. The inferred regulatory TF-target pairs were then compared with the list of known positive and negative pairs to yield true positive and false positive rates as a function of threshold level. Varying the threshold gives a curve which is referred to as a ROC; this shows the sensitivity and specificity of the inferences across all possible thresholds on a single plot, giving a comprehensive view of the ability of the score to distinguish between positive and negative examples.

Figure 2.5(a) shows ROC curves obtained from the *S. cerevisiae* dataset for the similarity score ψ, the widely used PCC and the absolute value of the PCC ($|PCC|$). The (expected) curve which would be obtained by chance is also shown for comparison. Similarity score ψ performs better than both PCC and the absolute value of PCC in this instance, suggesting that the score is indeed able to detect instances of direct regulation.

The results presented above pertain to direct regulatory relationships between TFs and validated targets. However, the complete set of pairwise relationships in a GRN naturally includes indirect as well as direct influences; e.g. if TF A directly regu-



(a) Yeast TF-target pairs



(b) Arabidopsis clock module

Figure 2.5: Analysis of the ability of ψ to predict transcriptional regulation. (a) ROC plots obtained from microarray data for validated examples of TF-target pairs in yeast (microarray data from Spellman et al. (1998); Gasch et al. (2000), TF-target pairs from supplementary data of Qian et al. (2003)). Similarity score ψ outperforms both Pearson's correlation coefficient (PCC) and its absolute value. The dotted line corresponds to random guesswork. (b) ROC plots obtained from microarray data, comparing the expression profiles of genes from the *A. thaliana* circadian clock with that of random genes. Similarity score ψ outperforms the other measures of similarity, performing roughly twice as well as measures neglecting time lags.

lates the expression of target gene B, which in turn directly regulates the expression of target gene C, the pair (A,C) is an example of an indirect transcriptional relationship. The ability of ψ to correctly distinguish indirect regulation, from no regulation, was tested by applying it to the gene expression of a well-studied GRN in Arabidopsis. A small network of six genes (*LHY*, *CCA1*, *TOC1*, *GI*, *PRR7* and *PRR9*), has been shown to form the core of the circadian clock GRN in Arabidopsis (Locke et al., 2006; McClung, 2008). The average expression profiles, of these six genes and 560 genes chosen at random from the Arabidopsis genome, were taken from the mock time series presented in Section 2.2.1. None of the 560 randomly chosen genes were annotated as belonging to the circadian clock (Ashburner et al., 2000; Swarbreck et al., 2008). In the resulting set of pairs, those including only members of the known circadian clock module were treated as positive examples, while those with only one member of the circadian clock were considered to be false positives. As the similarity measure, ψ, is symmetric there is $\binom{6}{2} \times \frac{1}{2} = 15$ (the number of different undirected pairs of the 6 circadian clock genes) positive examples and $6 \times 560 = 3,360$ negative examples.

ROC curves were constructed in a similar manner to the TF-target case above (Figure 2.5(b)). Similarity score ψ very clearly outperforms PCC and its absolute value in this instance. For example, 10 (out of 15) true positives are obtained at a cost of 141 false positives; in comparison, PCC requires 1,649 and absolute PCC requires 1,783 false positives. This suggests that ψ is indeed able to detect both direct and indirect regulations, even under highly sparse conditions, i.e. when true positives are scarce relative to false positives.

Comparison of AP and PAM as methods to cluster based on Qian similarity The similarity measure ψ captures a quite different notion of similarity than a straightforward vector distance. In Figure 2.5 ψ was demonstrated to be an appropriate similarity measure for predicting transcriptional regulation from gene expression time series. In other studies, such as those by Qian et al. (2001) and Yona et al. (2006), ψ is shown to perform comparably with other similarity measures at predicting other types of functional relations.

Because ψ is not a simple (negative) vector distance, clustering under ψ represents a fundamentally different formulation of the clustering problem than many widely used methods (for examples Hastie et al., 2001; Ghosh and Chinnaiyan, 2002; Heard et al., 2005; Thalamuthu et al., 2006). In this sense, TCAP and these widely used

methods address different questions, which makes them difficult to compare directly. However, PAM (Kaufman and Roussseeuw, 1990) represents a natural choice for clustering under the similarity measure ψ; indeed, it has been suggested for this purpose in previous work[2] (Qian et al., 2001).

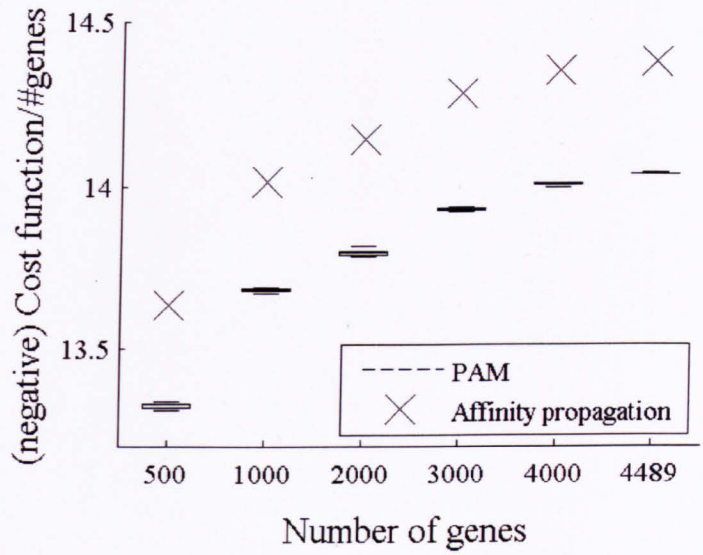
Here PAM and AP are applied to Qian similarity matrices generated from gene expression data. Their effectiveness is assessed by comparing the cost, as defined by Equation (2.16), of the resulting clusterings (i.e. partitions of all genes into non-overlapping gene groups). These clustering methods were applied to two microarray time series: 4,489 genes over 18 time points from a published *S. cerevisiae* experiment (Spellman et al., 1998) and 6,000 genes over 24 time points from the *B. cinerea* infection experiment presented in Section 2.2.1.

For each dataset both methods were applied to the full set of genes and also to smaller, randomly selected subsets, to investigate dependence on dimensionality. For each regime of dimensionality 10 runs of PAM and one run of AP (which is deterministic) was applied to the data. Since the same similarity measure was used in both cases, the underlying cost function (2.16) is identical. AP was applied using default parameters; AP is able to automatically learn a good number of clusters by setting the self-similarity to the median similarity value (Frey and Dueck, 2007). To ensure a fair comparison, the number of clusters returned by PAM was set to equal the number of clusters discovered by AP in each case.

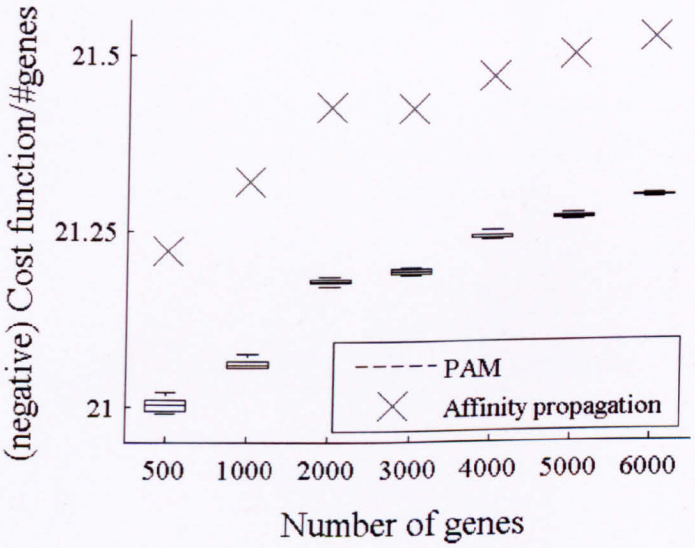
Figure 2.6(a) shows results obtained using the *S. cerevisiae* dataset of Spellman et al. (1998), which is a time course of expression profiles of genes from cells synchronized by the addition of alpha pheromone. Genes were selected because they did not having any missing values in their expression profiles.

The Arabidopsis dataset contains the expression profiles of 6,000 genes in the *B. cinerea* infection time series introduced in Section 2.2.1. The 6,000 most differentially expressed genes, between infected and mock samples, were used for computational tractability (because each method was applied to the data multiple times). Figure 2.6(b) shows results on the Arabidopsis data. In each case, box plots show values of the objective function obtained using PAM; AP is deterministic and gives a single result in each case. Figure 2.7 shows an analysis in which 400 PAM runs were used on the full Qian similarity matrix of the Arabidopsis dataset, with each run allowed the same compute time as a single run of AP. AP is completely deter-

[2]K-means rather than PAM is actually suggested, but because of the inappropriateness of k-means for clustering based on ψ it can be assumed that authors actually meant the mediod equivalent to k-means, i.e. PAM.



(a) AP vs. PAM in yeast data



(b) AP vs. PAM in Arabidopsis data

Figure 2.6: Here the method proposed in Qian et al. (2001) is compared to TCAP. (a) They were both applied to data from Spellman et al. (1998), a time series consisting of 4,489 genes over 18 time points. Various subsets of this were clustered and the cost function, as given in Equation (2.16) and then divided by the number of genes in the subset, is reported. 10 runs of PAM, each allowed to take as long as a single run of AP, were applied to the data. (b) Both methods were applied to the time series introduced in Section 2.2.1. Various subsets of this were clustered and the cost function, as given in Equation (2.16) and then divided by the number of genes in the subset, is reported. 10 runs of PAM, each allowed to take as long as a single run of AP, were applied to the data.

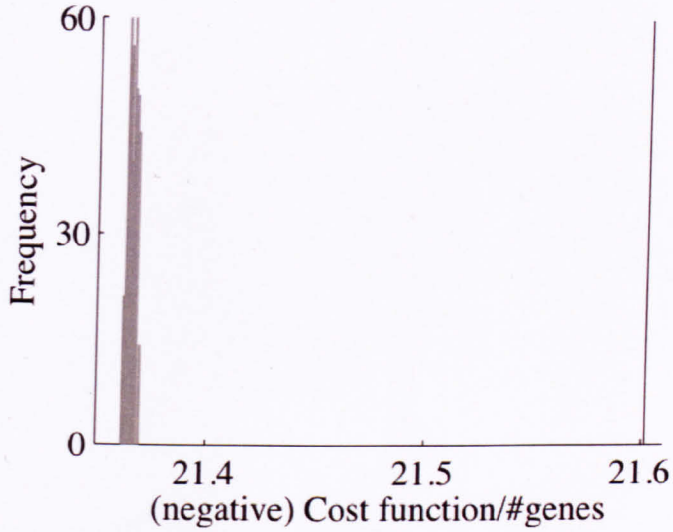
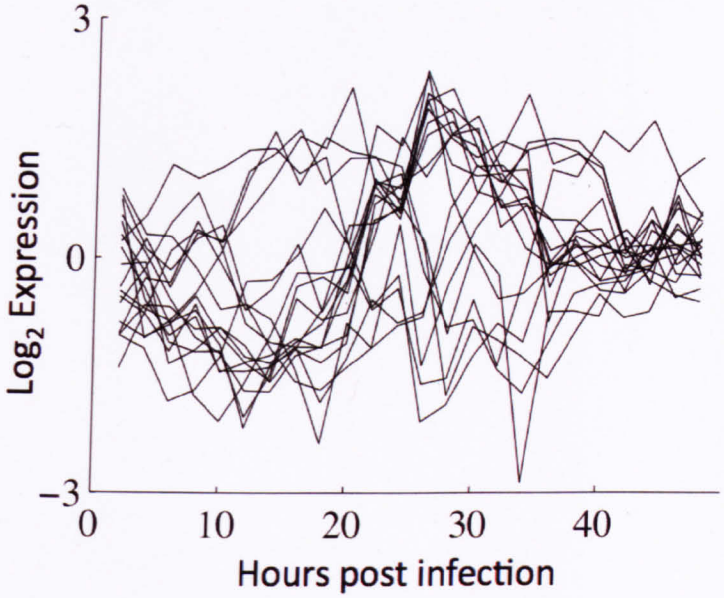


Figure 2.7: Here the method proposed in Qian et al. (2001) is compared to TCAP. Here the *A. thaliana* data was clustered again by both methods, but with 400 runs of PAM (shown in the grey histogram) each allowed to take as long as a single run of AP (black line, representing the result of a single run of AP).

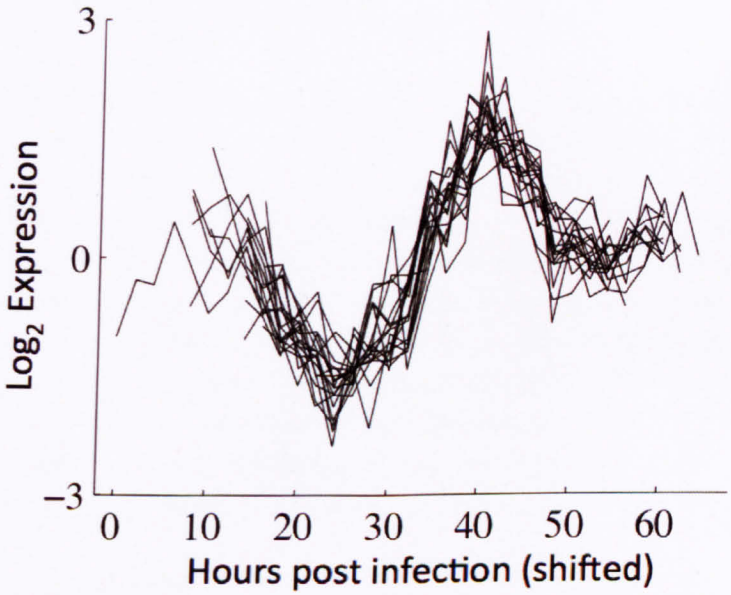
ministic, and therefore not subject to variation due to initial conditions or stochastic steps. It is clear that PAM is performing significantly worse than AP at producing clusters to minimize cost function Equation (2.16).

Application to the time series of Arabidopsis gene expression during infection by *B. cinerea*

TCAP was applied to the time series expression profiles of the 6,000 most differently expressed genes in Arabidopsis leaves during infection by *B. cinerea*, from the dataset introduced in Section 2.2.1. A default self-similarity of 0.762 (median of the off-diagonal entries of the similarity matrix) was used which resulted in 492 clusters, 153 of these were singleton clusters and so were ignored (Supplemental Digital Information Table 2). The remaining 339 clusters contain a median value of 13 genes (lower quartile = 6 and upper quartile = 24.5). The VirtualPlant software platform was used for Gene Ontology (GO) term overrepresentation analysis, with P-values calculated using the hypergeometric distribution (Ashburner et al., 2000; Gútierrez et al., 2005; Katari et al., 2010). First clusters that demonstrate the method will be highlighted; the results obtained by applying TCAP to the expression of all 9,838 differentially expressed genes will be shown later.



(a) Complex temporal cluster



(b) Complex temporal cluster, adjusted for delay and anti-correlation

Figure 2.8: TCAP finds time-delayed correlation in gene expression time series. Gene expression profiles are averaged over biological and technical replicates, and then transformed to zero mean and standard deviation one. (a) A cluster returned by TCAP (cluster 208, Supplemental Digital Information Table 2). (b) The same cluster as in the previous figure, adjusted for time delays and anti-correlation. Some profiles in this plot have been shifted in time and/or vertically inverted according to their original match type, as determined by the algorithm to calculate ψ that was introduced in Section 2.1.3.

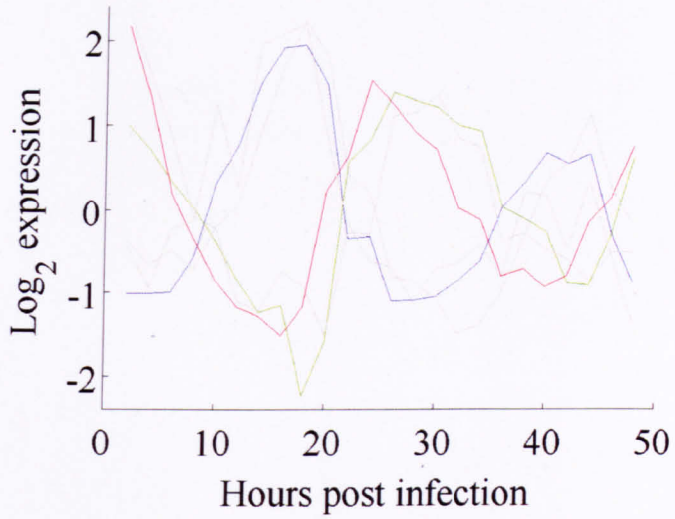
TCAP can detect time delayed correlation First, the results of TCAP were studied to see if time delayed correlation had been detected. Here an example is given to demonstrate that TCAP was indeed able to uncover time-delayed correlation, and therefore to produce clusters that would not be found by methods based on simple vector distances. In Figure 2.8(a) a TCAP cluster is shown whose underlying temporal patterns are sufficiently complex as to make the cluster appear, at first glance, devoid of any coherent pattern (cluster 208, Supplemental Digital Information Table 2). Figure 2.8(b) shows the same cluster, adjusted for time delayed correlations/anti-correlations: this is now highly coherent.

TCAP clusters recover published regulation Every TCAP cluster which contains a TF can be interpreted as a inferred transcriptional module, with the TFs inferred as regulators, although it is biologically intuitive that delays between TFs and their targets will be non-negative, unless other regulators or feedbacks are involved in their regulation. To test the ability of TCAP to accurately infer regulation, its inferences were compared to regulation known in the literature. In Figure 2.9 two clusters are shown which recapitulate known regulatory interactions.

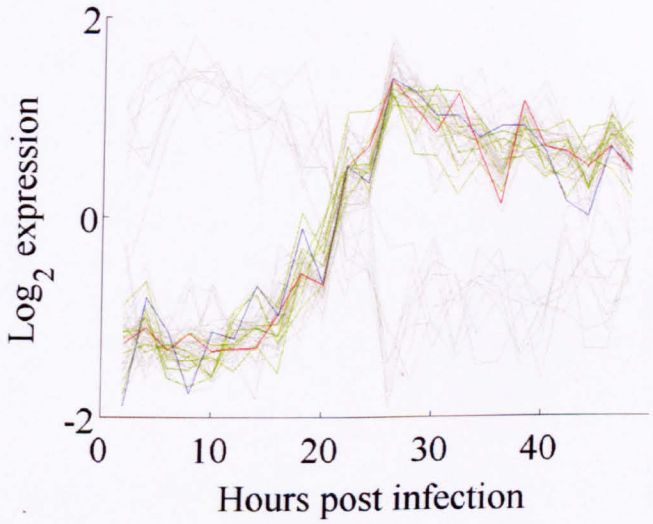
Figure 2.9(a) shows a cluster of 6 genes (cluster 258, Supplemental Digital Information Table 2) whose expression profiles seem to approximately repeat every 24 hours. The cluster contains two genes, *LHY* and *GI*, encoding known components of the core circadian clock GRN, as well as 4 other genes (*At1g56300, At3g47380, At3g54500* and *At4g15430*). Gene *GI* was found to score highly with *LHY* with a time-delayed and anti-correlated match. The time-delayed and anti-correlated relationship between the two expression profiles fits extremely well with the known role of LHY as a transcriptional repressor of GI (Locke et al., 2006; McClung, 2008). In addition, another member of the cluster, *At1g56300*, belongs to a class of genes known as rapid wounding response (RWR) genes, which are also known to be regulated by the circadian clock (Walley et al., 2007).

The de novo discovery of a small cluster containing these genes is striking in light of the fact that the relationship between these genes took many years and much research effort to uncover. The remaining cluster members appear to have no known link to the circadian clock; however, given the highly validated nature of other cluster members, these further genes provide intriguing hypotheses for additional downstream targets.

Figure 2.9(b) shows a second cluster (cluster 12, which has 57 members, Supplemental Digital Information Table 2) whose members form a striking and biologically



(a) Circadian cluster



(b) Cluster containing *ORA59* and *ERF1*

Figure 2.9: Clusters found by TCAP that recover known transcriptional regulation. Gene expression profiles are averaged over biological and technical replicates, and then transformed to zero mean and standard deviation one. (a) A circadian module. *LHY* (in blue) is known to be a transcriptional repressor of *GI* (in red). At1g56300 (in green) is a Rapid Wounding Response gene, which are known to be regulated by the circadian clock. Here grey lines represent the expression levels of three additional cluster members. (b) A cluster containing 6 genes co-regulated by *ORA59* (in green), *ORA59* (in blue) and *ERF1* (in red) that is believed to jointly regulate *PDF1.2* with *ORA59* (Pré et al., 2008). The expression of other cluster members are shown in grey.

coherent group. It is noteworthy that this cluster contains a regulator and known target genes of this regulator. The TF *ORA59* is in this cluster, along with six genes (*At1g59950, At2g43580, At3g23550, At3g56710, At4g11280* and *At4g24350*) that have been previously found to be up-regulated in an inducible overexpressor line of *ORA59* (Pré et al., 2008). This overlap is unlikely to occur by chance. (Number of differentially expressed genes after *B. cinerea* infection excluding *ORA59* = 9,837 , number of these differentially expressed in an inducible overexpressor of *ORA59* = 46 (Pré et al., 2008), number of inferred targets of *ORA59* = 56 and an overlap of 6 giving a cumulative hypergeometric p-value of 2.03×10^{-7} calculated using MATLAB®). The genes in this overlap are also up-regulated in the *B. cinerea* time series. Moreover, *ORA59* and another TF, *ERF1*, are believed to jointly regulate *PDF1.2* (Pré et al., 2008) and *ERF1* is also found in this cluster. *PDF1.2* itself is not in the dataset as there is no probe for it on the microarrays used. Both *ORA59* and *ERF1* are known to respond to the plant hormone ethylene; the cluster also has an over-representation, significant at 1%, of the GO term "response to ethylene stimulus".

Little is known in Arabidopsis about the relative timing of expression of TFs and their direct targets, i.e. how long a delay there is in general between differential expression of a TF and a noticeable change in the expression of its targets. However, in the case of 2.9(b), the temporal resolution of the dataset is apparently not sufficient to pick up a delayed correlation between the expression of the regulator *ORA59* and its targets. As the targets of *ORA59* were not originally studied in the context of the defence response to *B. cinerea* (Pré et al., 2008), it is also possible that these genes are not targets of *ORA59* in this condition, and are instead being co-regulated with it.

TCAP can group functionally related genes TCAP was applied to the time series expression profiles of the full list of 9,838 differently expressed genes introduced in Section 2.2.1, with a default self-similarity value of 0.731 (median of the off-diagonal entries of the similarity matrix) and a maximum of 3,000 iterations to ensure convergence which produced 579 clusters; 111 of the clusters were singleton clusters and so were ignored (Supplemental Digital Information Table 3). The remaining 468 clusters contain a median value of 16 genes (lower quartile = 8 and upper quartile = 27).

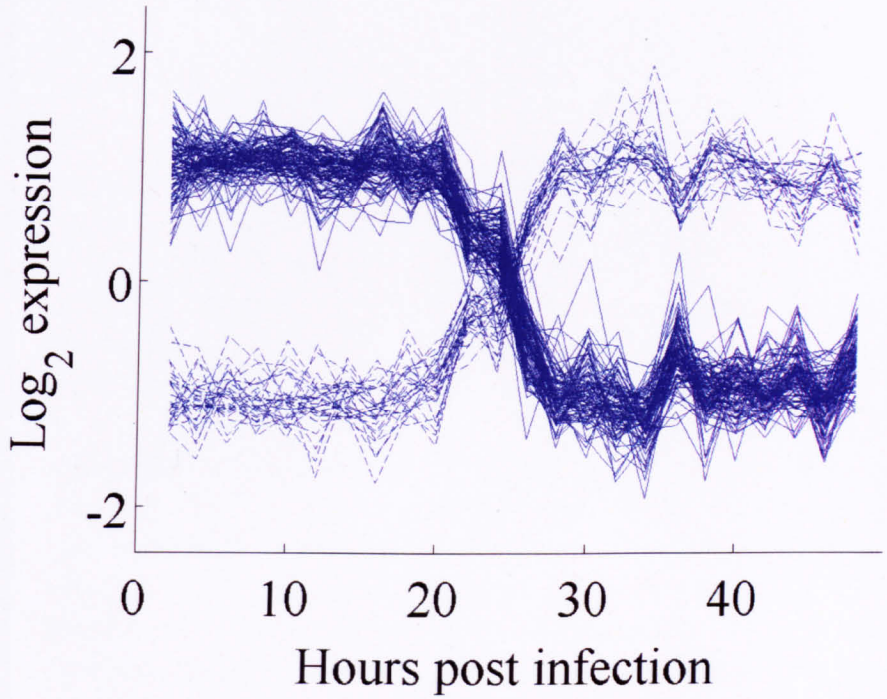
Cluster 334 (Supplemental Digital Information Table 3) has an overlap with the circadian cluster presented in Figure 2.9(a), i.e. it also contains *LHY, GI* and *At1g56300*, but all the other members are different. This shows that some results

can be robust against input gene lists/clustering parameters. The cluster presented in Figure 2.9(b) appears to be less robust as there is no cluster containing *ORA59, ERF1* and all of the 6 *ORA59* targets presented previously. However, cluster 8 which has 56 members (Supplemental Digital Information Table 3) appears to contain 17 of the 57 genes originally found in the cluster presented in Figure 2.9(b). *ERF1* and two of the original *ORA59* targets (*At1g59950* and *At2g43580*) are in cluster 8, as well as two other *ORA59* targets; *At5g22300* and *At5g27420* (Pré et al., 2008). *ORA59* itself now appears in cluster 153 (Supplemental Digital Information Table 3) along with another *ORA59* target, *At3g49630*, which was not present in the cluster shown in 2.9(b) (Pré et al., 2008). Although the exact *ORA59* cluster is not robust in this case, there does appear to be robustness towards the grouping of *ORA59* targets. Moreover, robustness should increase if larger clusters are used.

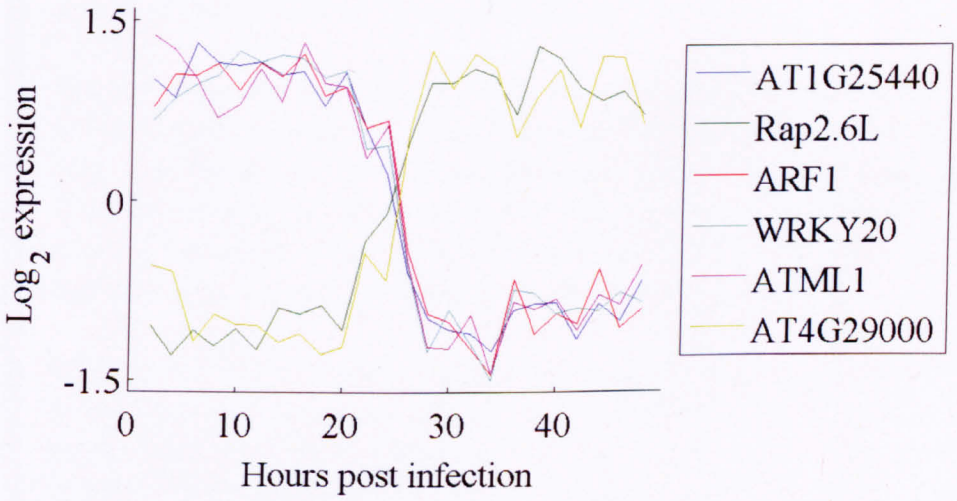
To see if TCAP could group functionally related genes, these clusters were analysed for over-represented GO terms against a background of all the genes differentially expressed during *B. cinerea* infection. 45/468 of the clusters were found by VirtualPlant to have over-represented GO terms at a significance of 1% (FDR corrected p-value against GO terms but not against the number of clusters compared). For comparison, the cluster assignment function was shuffled to produce clusters of the same size with random members (using randsample.m in MATLAB®) and GO term overrepresentation analysis was repeated, this gave 5/468 clusters with over-represented GO terms (assessed as before). This suggests that TCAP is finding functionally related genes at a greater than random rate. Many more GO terms will be found over-represented against the background of all genes, but many of these will be found without clustering, this was why only the differentially expressed gene list was used for the previous analysis. GO terms over-represented against the background of all genes will be discussed as relevant to specific clusters in later sections.

TCAP infers novel regulators of the defence response As well as identifying gene regulation that is known in the literature and grouping functionally related genes, TCAP grouped many genes with TFs that are inferred to control their expression. In this section novel transcriptional regulation inferred by TCAP is shown and compared to the literature.

Module 1 - TCAP cluster with the highest average Qian similarity. TCAP cluster 1 (Supplemental Digital Information Table 3) contains the TFs: *ARF1, WRKY20, RAP2.6L* (AP2-EREBP TF), *ATML1, At1g25440* and *At1g29000*. Several gene expression plots are shown in Figure 2.10 (a) and (b). This module has the highest average Qian similarity of all the TCAP clusters. In Figure 2.5, higher values of



(a) Module 1



(b) TFs in module 1

Figure 2.10: Transcriptional module inferred by TCAP. Here, module 1 was plotted in various ways to show time delayed correlations between gene expression profiles. Gene expression profiles are averaged over biological and technical replicates, and then transformed to zero mean and standard deviation one. (a) The average expression profiles of all genes in module 1, anti-correlated genes are drawn as dashed lines. (b) The average expression profiles of all TFs in module 1.

ψ have been shown to be more likely to correspond to correctly inferred TF-target pairs, and so this cluster is inferred to be linked by some transcriptional regulation, i.e. one or more of the TFs is inferred to regulate the other genes. Nothing appears to be known about regulatory targets of the TFs in this cluster.

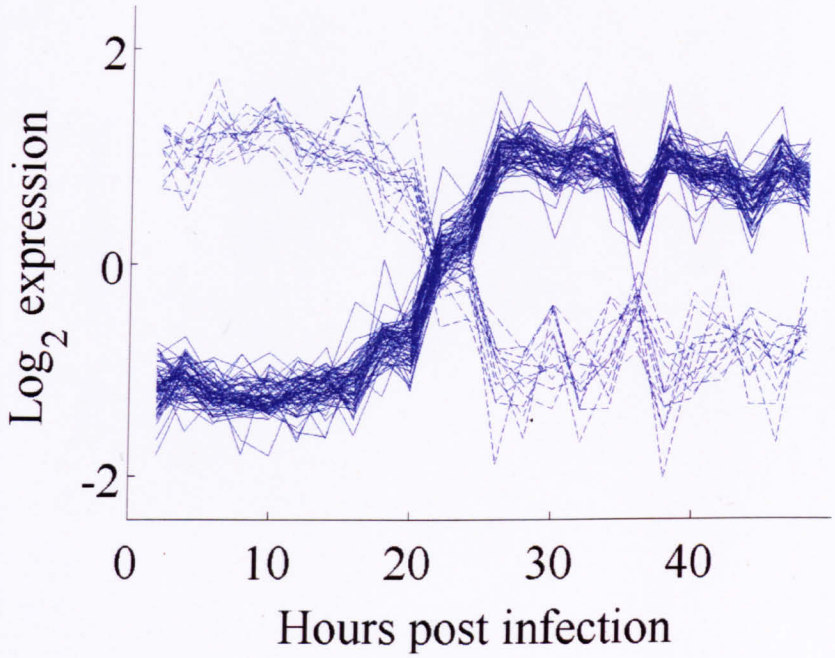
Module 2 - TCAP cluster with the second highest average Qian similarity. TCAP cluster 2 (Supplemental Digital Information Table 3) contains the TFs: *NUB, ATERF11, LBD41, ANAC055, AtERF1, At1g71520 (AP2-EREBP TF), At2g33710, At3g53600, At4g28811, At5g14280* and *At5g56960.* Several gene expression plots are shown in Figure 2.11 (a)–(b). This module has the second highest average Qian similarity of all the TCAP clusters.

ANAC055 is already known to be important in the defence response, an overexpressor of *ANAC055* is more susceptible to infection by *B. cinerea* (Bu et al., 2008). One of the 73 inferred targets of *ANAC055* from module 5, *AT2G28860*, is differentially expressed in a *anac055* knockout during leaf senescence (Hickman et al., in preparation). This overlap is likely to occur by chance. (Number of differentially expressed genes after *B. cinerea* infection excluding *ANAC055* = 9,838 , number of these differentially expressed in senescing *anac055* plant = 230 , number of inferred targets of *ANAC055* = 73 and an overlap of 1 giving a cumulative hypergeometric p-value of 0.8233 as calculated in MATLAB®).

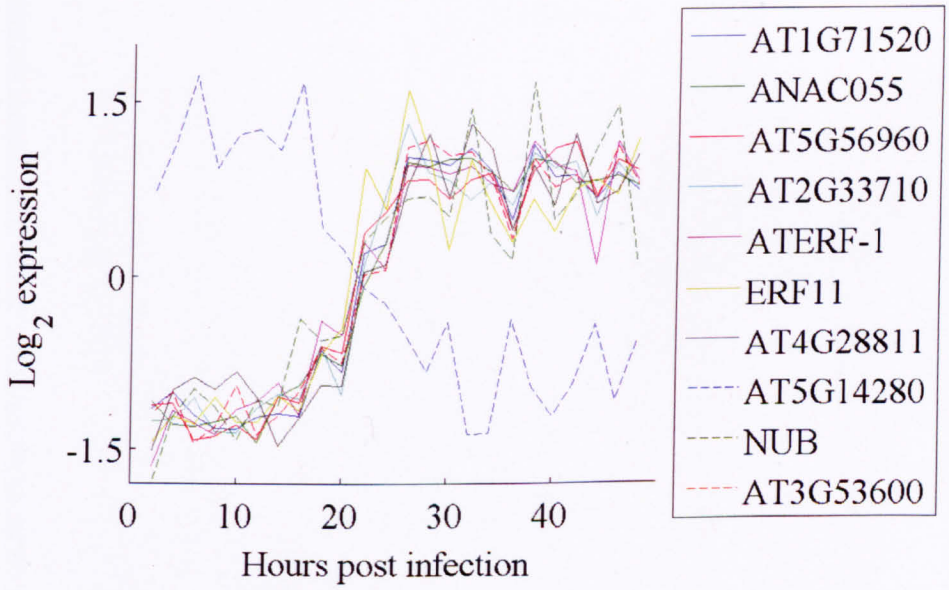
The regulation may also be by another TF in this cluster, unfortunately mutant versus wildtype microarray studies of the other TFs in this cluster were not available. Even though modules 1–2 have the highest average Qian similarities, because they both contain several TFs and none of their members show a time-delayed correlation, the inferences are not specific. As with the *ORA59* cluster, time delays may have been observed if a higher temporal resolution was used.

The next four TCAP clusters have been selected because they contained a TF and genes with a time-delayed correlation/anti-correlation to it; therefore these clusters provide inferences that are more specific, i.e. while a time delay between the transcription of a TF and its target is not always observed in gene expression time series, presumably because of complex regulation or low temporal resolution, when present it can suggest the direction of gene regulation. If the gene expression of two TFs correlates with a time delay, the earlier TF is presumably more likely to regulate the latter.

Module 3 - time-delayed correlation to a TF. TCAP cluster 71 (Supplemental Digital

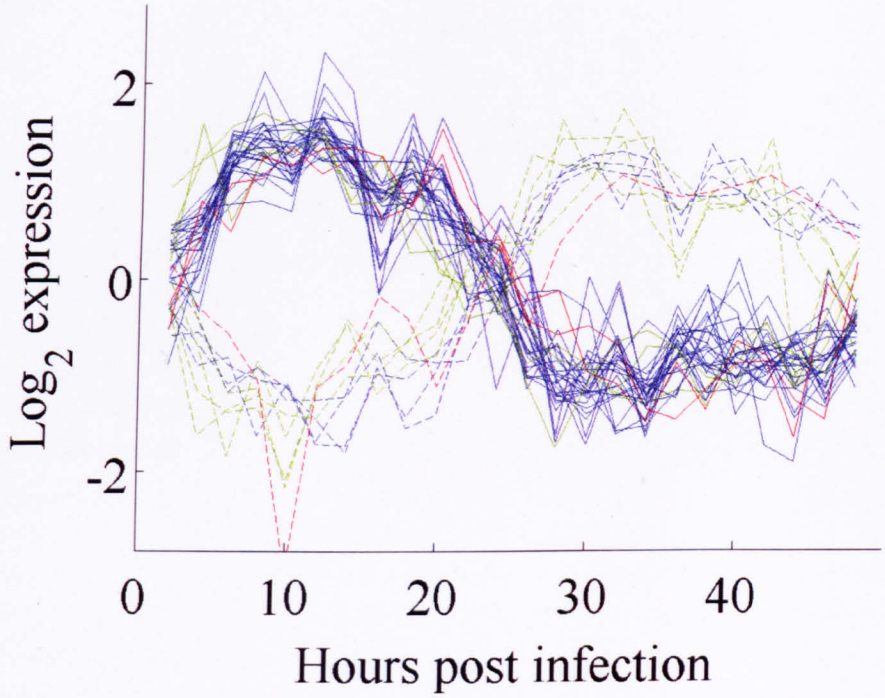


(a) Module 2

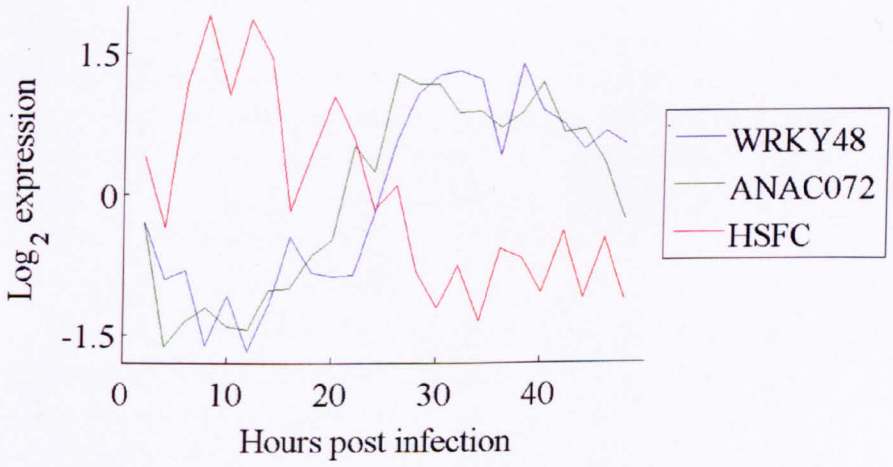


(b) TFs in module 2

Figure 2.11: Transcriptional module inferred by TCAP. Here, module 2 was plotted in various ways to show time delayed correlations between gene expression profiles. Gene expression profiles are averaged over biological and technical replicates, and then transformed to zero mean and standard deviation one. (a) The expression of all genes in module 2, anti-correlated genes are drawn as dashed lines. (b) The expression of all TFs in module 2.



(a) Module 3



(b) TFs in module 3

Figure 2.12: Transcriptional module inferred by TCAP. Here, module 3 was plotted in various ways to show time delayed correlations between gene expression profiles. Gene expression profiles are averaged over biological and technical replicates, and then transformed to zero mean and standard deviation one. (a) The expression of all genes in module 3, coloured according to which time delayed component they belong to. The genes are coloured as follows: anti-correlated genes are drawn as dashed lines, red lines have a positive delay of 1, blue lines are not delayed, green lines have a negative delay of 1. (b) The expression of all TFs in module 3.

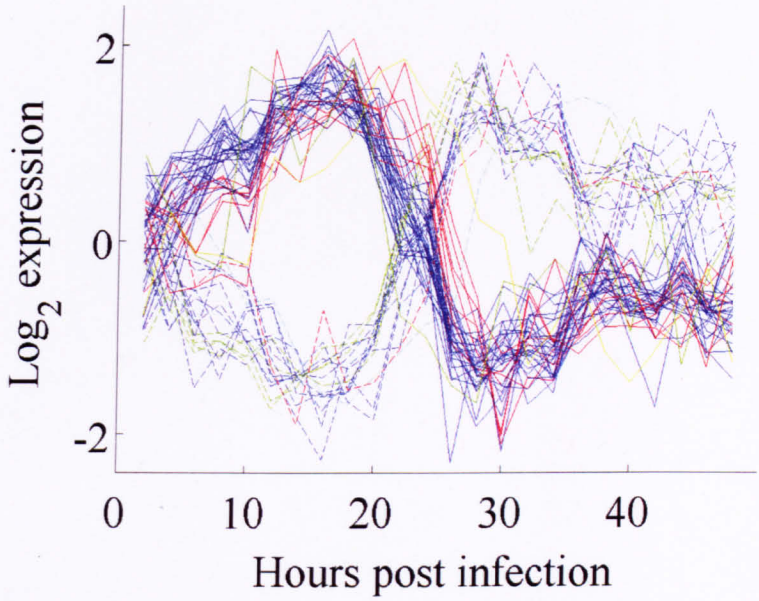
Information Table 3) contains the TFs: *HSFC, ANAC072* and *WRKY48*. Several gene expression plots are shown in Figure 2.12 (a)–(b). Cluster 71 is highlighted because an early negative delay component contained the TF *ANAC072*.

Several microarray experiments have been performed to identify possible targets of *ANAC072* (Fujita et al., 2004; Tran et al., 2004). Fujita et al. (2004) used a *35S::ANAC072* line as well as a constitutively expressed chimeric *ANAC072* with a fused repression domain. Inferred *ANAC072* targets from cluster 71 were compared with differentially expressed genes in these *ANAC072* mutants; the only inferred target differentially expressed in this mutant was *At4g37990*. Considering only the *ANAC072* targets that are differentially expressed after *B. cinerea* infection, and excluding *ANAC072* itself from the analysis as its presence in its own cluster is a given, then the overlap of one gene is unlikely to occur by chance (the number of differentially expressed genes excluding *ANAC072* = 9,837 , the number of differentially expressed *ANAC072* targets = 13, the number of genes in *ANAC072* cluster excluding *ANAC072* = 38 and an overlap of 1, giving a cumulative hypergeometric p-value of 0.0491 as calculated using MATLAB®). Tran et al. (2004) identified *ANAC072* targets using a 35S::ANAC072 mutant, 21 of which are differentially expressed during infection by *B. cinerea*.

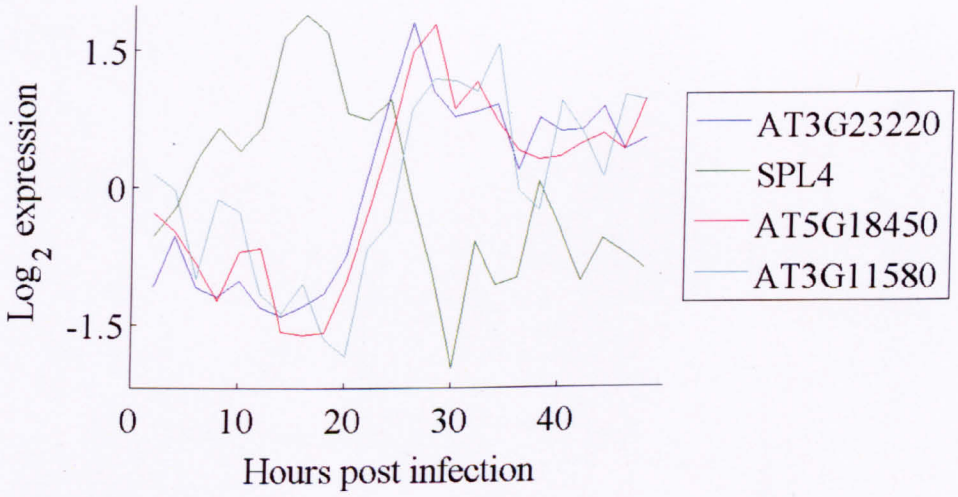
The mutant versus wildtype microarray experiments of Fujita et al. (2004) and Tran et al. (2004) have not been conducted during *B. cinerea* infection, and so these findings do not rule out the inferences made in cluster 71. Given that more overlap is found with the differentially expressed genes found in the study of Fujita et al. (2004), than in the study by Tran et al. (2004), it is also possible that study by Fujita et al. (2004) was performed with experimental conditions closer to that used in the *B. cinerea* infection experiment.

Module 4 - Time-delayed correlation to a TF. TCAP cluster 60 (Supplemental Digital Information Table 3) contains the TFs: *SPL4, At3g11580, At3g23220* (AP2-ERE BP TF) and *At5g18450*. Several gene expression plots are shown in Figure 2.13 (a)–(b). Cluster 60 was chosen because it has TFs in three different delay components, representing two layers of inferred directed gene regulation. Nothing appears to be known about regulatory targets of the four TFs in this cluster.

Module 5 - Time-delayed correlation to a TF. TCAP cluster 166 (Supplemental Digital Information Table 3) contains the TFs: *RGL1, DREB2A, CDF* and *NF-YB5*. Several gene expression plots are shown in Figure 2.14 (a)–(b). Cluster 166 was chosen because it has TFs in several different delay components.

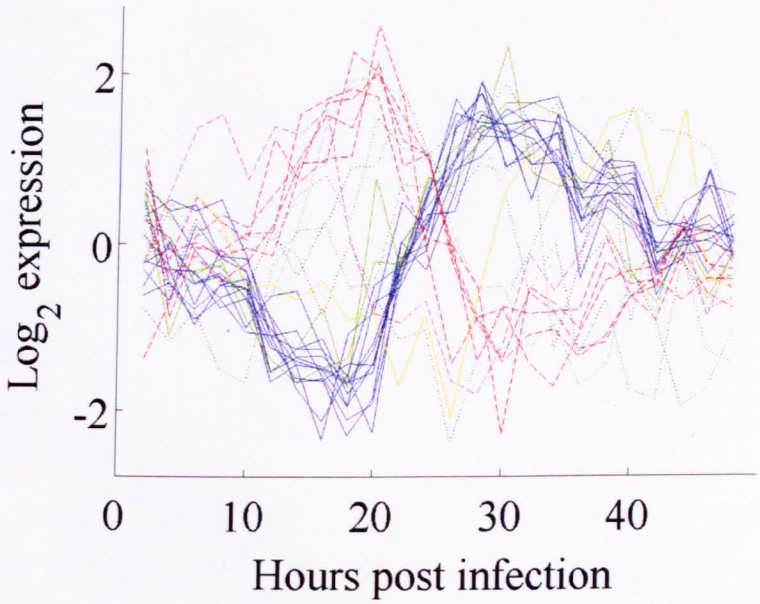


(a) Module 4

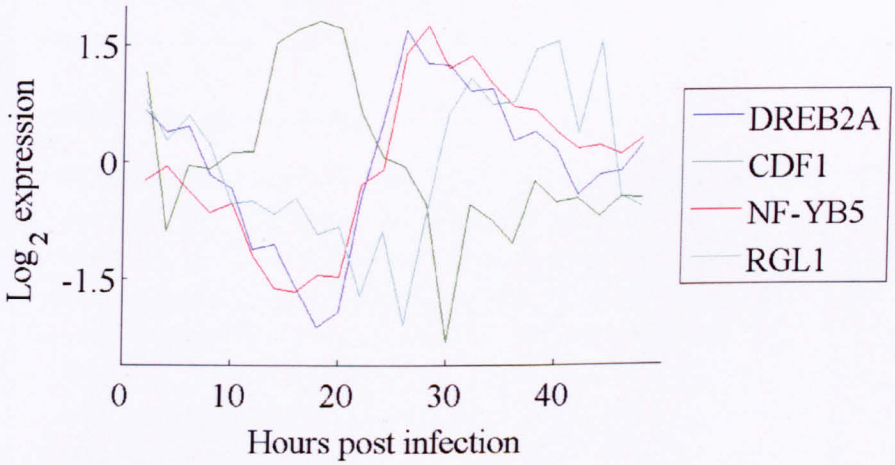


(b) TFs in module 4

Figure 2.13: Transcriptional module inferred by TCAP. Here, module 4 was plotted in various ways to show time delayed correlations between gene expression profiles. Gene expression profiles are averaged over biological and technical replicates, and then transformed to zero mean and standard deviation one. (a) The expression of all genes in module 4, coloured according to which time delayed component they belong to. The genes are coloured as follows: anti-correlated genes are drawn as dashed lines, red lines have a positive delay of 1, blue lines are not delayed, green lines have a negative delay of 1, cyan lines have a negative delay of 2 and yellow lines have a positive delay of 3. (b) The expression of all TFs in module 4.



(a) Module 5



(b) TFs in module 5

Figure 2.14: Transcriptional module inferred by TCAP. Here, module 5 was plotted in various ways to show time delayed correlations between gene expression profiles. Gene expression profiles are averaged over biological and technical replicates, and then transformed to zero mean and standard deviation one. (a) The expression of all genes in module 5, coloured according to which time delayed component they belong to. The genes are coloured as follows: anti-correlated genes are drawn as dashed lines, black and dotted lines have a positive or negative delay of more than 3, black lines have a positive delay of 3, purple lines have a negative delay of 2, red lines have a positive delay of 1, blue lines are not delayed, green lines have a negative delay of 1, cyan lines have a negative delay of 2 and yellow lines have a positive delay of 3. (b) The expression of all TFs in module 5.

Module 6 - Time-delayed correlation to a TF. TCAP cluster 262 (Supplemental Digital Information Table 3) contains the TFs: *MYBL2* and *AtMYB15*. Several gene expression plots are shown in Figure 2.15 (a)–(b). A TF in this cluster, *MYBL2*, scores highly for a match with the other genes with a time delay of 6 hours. This cluster has an over-representation, significant at 1%, of the GO term "response to abscisic acid (ABA)". ABA has been shown to play a role in the interaction between *B. cinerea* and plant hosts (AbuQamar et al., 2006; Audenaert et al., 2002), hence this cluster may represent a transcriptional module involved in the response to this hormone.

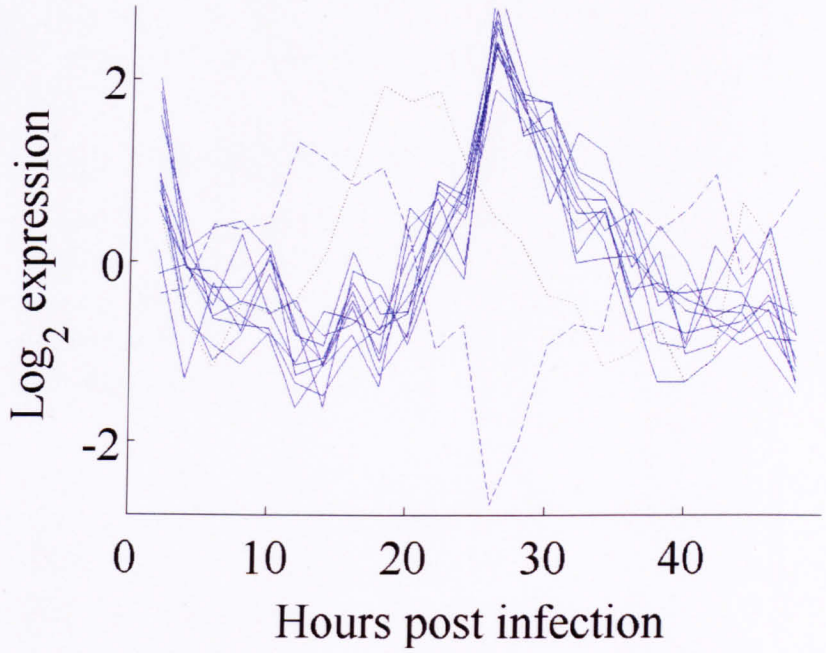
2.2.4 Reverse genetics screen of inferred regulators of the defence response

In the previous two sections two novel approaches to infer specific transcriptional regulation genome-wide have been introduced. The first approach inferred regulation either by TFs for which there were no mutants available, or TFs with mutants that had already been shown to have altered susceptibility to *B. cinerea*. TCAP modules 1–6 show inferred regulation by TFs. Mutants of some of these TFs were available and had not been screened for altered susceptibility to *B. cinerea*.

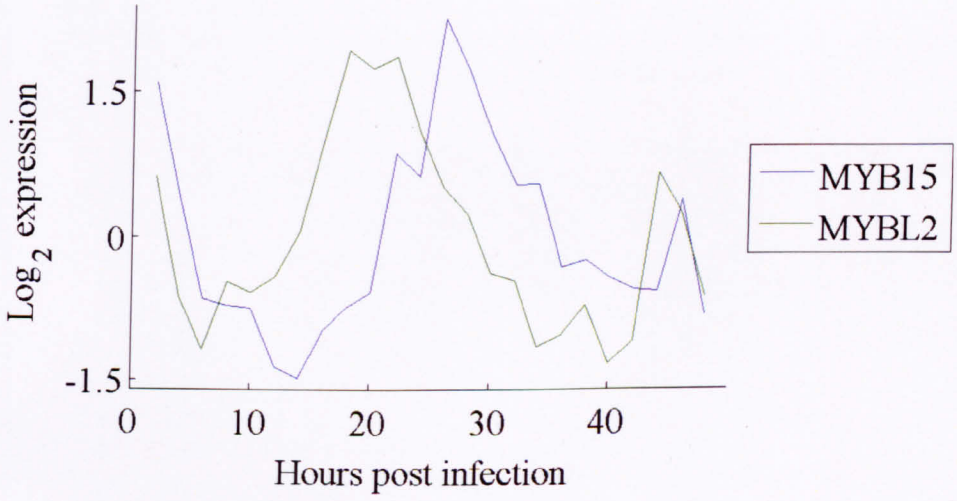
Materials and methods

Reverse genetics screen for altered susceptibility to Botrytis Arabidopsis seeds - for Col4 (wildtype), *bos1* (positive control, see section 2.1.4 or Mengiste et al. (2003)) and lines from Table 2.1 - were stratified in 0.1% agar at 4 °C for 3 days and then transferred to Arabidopsis soil mix (Scotts Levingtons F2s compost:sand:fine grade vermiculite in a ratio of 6:1:1). Plants were grown in a controlled environment with a 16:8 hour light:dark cycle at 20 °C, with 60% humidity and a light intensity of 100 μmol photons.m^{-2}.s^{-1}. These plants were allowed to grow for 4–5 weeks before mature leaves were detached and tested for susceptibility to *B. cinerea*.

Detached leaves were placed onto 0.8% agar that had been allowed to set in the base of propagator trays. Spores were collected from *B. cinerea* cultures from the Pepper isolate, which had been grown on apricot halves (Tesco) incubated at 20 °C for two weeks, by scraping *B. cinerea* into sterile water and filtering with glass wool to remove hyphae. Spores were diluted to 10^5 spores/ml in 1:1 (v/v) water and red pressed grape juice (Tesco). Each detached leaf was infected with a single 10 μl droplet of spore solution and incubated in a controlled environment at 90% humidity for 3 days, with a 16:8 hour light:dark cycle at 20 °C and a light intensity of 100 μmol photons.m^{-2}.s^{-1}. A time course of photographs were taken of the leaves



(a) Module 6



(b) TFs in 6

Figure 2.15: Transcriptional modules inferred by TCAP. Here, module 6 was plotted in various ways to show time delayed correlations between gene expression profiles. Gene expression profiles are averaged over biological and technical replicates, and then transformed to zero mean and standard deviation one. (a) The expression of all genes in module 6, coloured according to which time delayed component they belong to. The genes are coloured as follows: anti-correlated genes are drawn as dashed lines, black and dotted lines have a negative delay of 4. (b) The expression of all TFs in module 6.

at several time-points between 36 and 82 hpi, and the area of infection recorded from each photograph using ImageJ. A scale bar was used in the agar trays to provide a scale reference for the ImageJ analysis.

Reverse genetics screening of inferred regulators of the defence response

Available mutants of inferred regulators of defence responsive gene expression were acquired to allow their susceptibility to infection by *B. cinerea* to be assessed. This reverse genetics approach is similar to that used in the studies by AbuQamar et al. (2006) and Windram (2010), but in this study TCAP was used to prioritise differentially expressed TFs to screen.

Altered expression mutants Knockout and overexpressor mutants of inferred regulators that were available from the literature, Salk Institute homozygous T-DNA collection (Alonso et al., 2003) or generated at Warwick were collected for use in reverse genetics screens (listed in Table 2.1).

Table 2.1: Lines used in reverse genetics screen of inferred regulators of the defence response. All lines are T-DNA knockouts, except for *35S::MYBL2* which is a constitutive overexpressor. All lines except *mybl2_1* and *35S::MYBL2* were obtained from Salk Institutes homozygous T-DNA collection. *mybl2_1* was kindly provided by Dubos et al. (2008). *35S::MYBL2* was generated at Warwick by the PRESTA consortium.

Line Name	SALK name	Gene name	AGI	Module number
nub_1	SALK_004964c	*NUB*	At1g13400	2
nub_2	SALK_100548c	*NUB*	At1g13400	2
lbd41_1	SALK_078678c	*LBD41*	At3g02550	2
at3g53600_1	SALK_132289c		At3g53600	2
at3g53600_2	SALK_027144c		At3g53600	2
at5g14280_1	SALK_011661c		At5g14280	2
at5g14280_2	SALK_054183c		At5g14280	2
anac072_1	SALK_072276	*ANAC072*	At4g27410	3
anac072_2	SALK_072286	*ANAC072*	At4g27410	3
anac072_3	SALK_063576c	*ANAC072*	At4g27410	3
wrky48_1	SALK_066438c	*WRKY48*	At5g49520	3
wrky48_2	SALK_144719c	*WRKY48*	At5g49520	3
at3g23220_1	SALK_128736c		At3g23220	4
rgl_1	SALK_041897c	*RGL1*	At1g66350	5
rgl_2	SALK_136162c	*RGL1*	At1g66350	5
mybl2_1	SALK_107780	*MYBL2*	At1g71030	6
35S::MYBL2		*MYBL2*	At1g71030	6

Phenotype screen results The resulting lesion size data for each line at each time point is compared to those for wildtype samples to determine altered susceptibility. Comparisons are performed by hypothesis testing within MATLAB®, with the null hypothesis being that both mutant and wildtype lines have the same mean (or median) lesion area when measured at the same time. This means that the alternative hypothesis is that mutant and wildtype lines have a different mean (or median) lesion area when measured at the same time, which would mean that the mutation has resulted in altered susceptibility of the plant to *B. cinerea*. Hypothesis testing was performed using a t-test (two tailed and not assuming equal variances) (Student, 1908) and a non-parametric equivalent known as the Mann-Whitney-Wilcoxon (MWW) test (Wilcoxon, 1945). The normality of the data for each line was tested using the Kolmogorov-Smirnov test with the Lilliefors table (Lillefors, 1967), which can indicate whether a normal or non-parametric test is most appropriate (results re shown in Appendix C). A significance threshold of 5% was used in the hypothesis tests, i.e. when the p-value is less than 0.05 the null hypothesis was rejected. Multiple testing corrections have not been applied, but any mutant displaying altered susceptibility was retested to control for spurious phenotypes. The results are shown in Appendix C, and are summarised in Table 2.2.

A novel altered susceptibility phenotype was observed in mutants of *ANAC072*, as found in three independent T-DNA knockout lines (Table 2.2 and example screen photos in Figure 2.16). The phenotype of these *ANAC072* independent T-DNA knockout lines (*anac072_1*, *anac072_1* and *anac072_3*) is a slightly decreased susceptibility to infection by *B. cinerea*. Screen photos are shown in Figure 2.16 to demonstrate that even the strongest novel altered susceptibility phenotype found in this study is weak in comparison with the difference between the wildtype (Col4) and positive control (*bos1*). It also demonstrates the importance of biological replication, because of the variability in lesion size within lines.

Another novel phenotype is found in two independent *NUB* T-DNA knockout lines, *nub_1* and *nub_2*, which also show a slightly decreased susceptibility to infection by *B. cinerea*. A contradictory phenotype is observed for *nub_1* in one screen, but that screen had fewer biological replicates than the other screens. Given that decreased susceptibility was shown for *nub_1* in 4 other more highly replicated screens and in a second independent knockout of *NUB* (*nub_2*), the contradictory result is likely to be spurious.

An altered phenotype is observed in a *LBD41* T-DNA knockout line (*lbd41_1*). This too shows a slightly decreased susceptibility to *B. cinerea*. Unfortunately an inde-

Table 2.2: Summary of results of reverse genetic screen for altered susceptibility to *B. cinerea*. Phenotype is observed if either the t-test or MWW test reject the null hypothesis at the 5% significance level.

Line Name	Proportion of screens with altered phenotype observed	Results in each screen. (M=more susceptible, L=Less susceptible. N=neither)	Proportion of time-points with altered phenotype observed	Susceptibility screen figure / table number (See Appendix C)
anac072_1	3/3	LLL	3/7	C.4, C.6 and C.8
anac072_2	2/2	LL	3/4	C.4 and C.8
anac072_3	5/5	LLLLL	10/13	C.4, C.5, C.6, C.7 and C.8
wrky48_1	3/5	MLMNN	4/12	C.1, C.3, C.5, C.7 and C.8
wrky48_2	2/2	LL	2/6	C.3 and C.5
at3g23220_1	0/1	N	0/2	C.1
rgl_1	0/1	N	0/2	C.1
rgl_2	2/2	MM	2/5	C.1 and C.5
mybl2_1	0/1	N	0/3	C.2
35S::MYBL2	1/1	M	3/3	C.2
nub_1	5/5	MLLLL	6/12	C.1, C.3, C.5, C.7 and C.8
nub_2	2/2	LL	4/6	C.3 and C.5
lbd41_1	4/4	LLLL	8/9	C.1, C.4, C.3 and C.8
at3g53600_1	0/1	N	0/2	C.1
at3g53600_2	2/2	ML	3/4	C.1 and C.8
at5g14280_1	3/4	LLNL	6/12	C.1, C.4, C.7, and C.8
at5g14280_2	2/2	LL	3/4	C.1 and C.4

(a) From top to bottom: *anac072_1* (here labelled 72_1), Col4 and *anac072_2* (here labelled 72_2) at 50hpi

(b) *anac072_3* (here labelled im297) at 50hpi

(c) *bos1* at 50hpi

Figure 2.16: Photos from the screen summarised in C.8, showing the weak resistance phenotype of mutants of *ANAC072*. Scale bars are shown in each photo. (a) shows the lesion size of lines Col4 (which is the wildtype control), *anac072_1* and *anac072_2* at 50 hpi. (b) shows the lesion size of line *anac072_3* at 50 hpi. (c) shows the lesion size of line *bos1*, which is used as a positive control, at 50 hpi.

pendent T-DNA knockout was not available. An increased susceptibility phenotype seen in *rgl_2* is not seen in an independent knockout (*rgl_1*). The T-DNA location in the *rgl_1* line is very near to the 3' UTR, meaning that it is possible that some functional truncated transcript is still produced. An alternative possibility is that the *rgl_1* line is heterozygous for the T-DNA insertion, which has not been tested in this study. A promising result for *35S::MYBL2* proved hard to repeat due to the extreme dwarfing observed in this line.

In this chapter two novel methods to infer transcriptional regulation from gene expression time series have been introduced. They were used to infer regulators of the defence response, some of which were then screened in a 'reverse genetics' screen. Mutants of some inferred regulators showed altered susceptibility to *B. cinerea*, although the altered phenotypes observed were weak in comparison to that of a previously published mutant.

2.3 Discussion

2.3.1 Computational and statistical discussion

Multiple testing problems in bioinformatic analysis

In bioinformatic analysis of high-throughput data, it is common to apply statistical hypothesis tests to many thousands of objects at once. A good example is the analysis of differential expression in microarray datasets, where hypothesis tests are applied to tens of thousands of DNA probes. If this multiple testing is not taken into account then there is a danger of obtaining many false positives by chance. In practice biologists typically use multiple testing corrections such as the Bonferroni (Bonferroni, 1936; Miller, 1981), Benjamini-Hochberg (Benjamini and Hochberg, 1995), or SAM corrections (Tusher et al., 2001), or resort to manual threshold selection. The appropriate balance between minimising false positives and minimising false negatives will always depend on the tolerance of downstream analysis to these errors.

In Section 2.2.1 approximately a third of the Arabidopsis genome was found to be differentially expressed between the mock and *B. cinerea* infected samples. This is a comparatively high proportion of differentially expressed genes compared to other microarray studies, and has a number of potential explanations: the additional statistical power provided by the relatively large number of biological samples used (192 biological samples in total); the dramatic nature of the biological treatment (a pathogen that is trying to kill Arabidopsis); and the desire, conscious or otherwise,

of the researchers to minimise false negatives. The final point is especially relevant in Systems Biology, where the aim is to study the relation between all important components of a process. In general it has not been shown that the most differentially expressed genes during a biological process are the most important for the progression of that process, and so a stringent list of differentially expressed genes may miss key components. However, the disadvantage of a less stringent list is that downstream analyses or experiments may be sensitive either to its false positive rate or to its size.

Other multiple testing problems occur in the bioinformatic analyses that typically follow, for example in over-representation analysis of: differentially expressed genes from other microarray experiments, GO terms and promoter sequence motifs. A good example of this is the cumulative hypergeometric p-value between inferred targets of a TF and genes differentially expressed between a mutant of that TF and wildtype. While an overlap of a single gene between inferred and experimentally suggested targets of *ANAC072* was found to be unlikely to happen by chance, this was not the only hypothesis test of this sort performed, and so the overlap is not so unlikely to occur by chance.

'Optimal' clustering of similarly expressed genes

When clustering gene expression a common challenge is to choose an appropriate number of groups to partition the genes into. This choice can be direct, as in choosing K for K-means, or indirect, as in choosing s for AP. Statistical methods can be used to determine the 'optimal' number of clusters from the data, for example through the use of the silhouette plot (Kaufman and Roussseeuw, 1990).

Both SplineCluster and AP have default values used to determine the number of clusters to return, both based on an optimisation process where the number of clusters emerges from both this default value and the data (Heard et al., 2005; Frey and Dueck, 2007). Therefore it is surprising that the number of clusters returned, using default parameter values, vary so much between SplineCluster and TCAP (when applied to the 9,838 differentially expressed genes with default values SplineCluster returned 38 clusters, whereas TCAP returned 468 non-singular clusters). This difference appears to depend on both the similarity measure and the clustering method used, as clustering based on PCC rather than ψ in AP produced fewer clusters (when applied to the 9,838 differentially expressed genes with default values, PCC-AP returned 164 clusters, none of which were singular). This may reflect the fact that ψ allows profiles to be deemed 'similar' in more ways than PCC, reducing the variance of all pairwise similarities, i.e. if two profiles are similar according to PCC then they

will be similar according to ψ, which does not necessarily hold in reverse.

Ultimately a biologist would consider the 'optimal' number of clusters the one with the best balance between specificity and sensitivity with regards to grouping together functionally related genes. Typically a biologist assesses this using GO terms, which represent known functional relations (Ashburner et al., 2000). An example of a way to analyse this is the biological homogeneity index (Datta and Datta, 2006), which assesses the consistency of functional annotation within clusters for a given clustering. It is not clear whether the statistical and biological 'optimal' clustering number will be equivalent.

In this study the biologically 'optimal' partition of differentially expressed genes was considered to be the one that showed the greatest diversity in over-represented known binding motifs in the promoters of the groups of genes. This choice reflected the interest of the group in discovering groups of co-regulated genes. This was achieved by SplineCluster with a prior precision of 0.001, rather than with the default prior precision of 0.0001. However, it was noted that other partitions allowed the discovery of additional motifs not over-represented in the final choice. This demonstrates both that a single partition of genes into groups may lose important biological information and the benefit of selecting groups of genes based on independent biological information. However, if over-representation analysis is applied to many different gene partitions, multiple testing corrections may be required. Methods exist for clustering jointly based on both expression data and information on functional relations (for example Reiss et al., 2006; Meng et al., 2009), which would avoid the need for multiple testing. However, if independent data is used to inform clusters there is the danger that the groupings will be biased towards known, rather than unknown, functional relations.

Network inference

Network inference, based on Markov process assumptions, is a sensible way to attempt to model transcriptional regulation from gene expression time series. However network inference applied to purely wildtype experiments attempts to model causal relationships based purely on the co-variation between gene expression profiles over time. This is a modelling limitation imposed by the data, which did not include perturbations/interventions at the level of individual genes. This is likely to be the reason that VBSSM infers regulatory connections among randomly selected gene expression profiles. It may be that the posterior probability of a SSM can be used to distinguish between useful and spurious inferences, but I am not aware that this has been demonstrated with VBSSM or any alternative Markov process based net-

work inference approach.

Because Markov process modelling approaches, like VBSSM, model co-variation of gene expression over time, they will not be able to distinguish between heavily correlated variables. This is perhaps the reason that all network diagrams in Section 2.2.2 appear to contain regulatory 'hubs', i.e. genes inferred to regulate a larger than average number of targets. Because co-expressed genes have been chosen because they correlate highly, it is not surprising that VBSSM infers that they are regulated in a similar way. In the case of Figure 2.3 where two TFs, *ANAC019* and *ANAC092*, are inferred to have the same number of targets among SplineCluster cluster 27, this suggests that either: the list of co-expressed/potentially co-regulated genes are not sufficiently well correlated to be modelled in the same way (suggesting that the cluster is larger than optimal); the replicates, which were averaged over for clustering, contain additional useful information; or that VBSSM is highly sensitive to noise and is therefore spuriously modelling the co-expressed genes. The fact that correlated variables may not be sufficiently different to be modelled separately in network inference suggests that network inference applied to members of the same TCAP cluster is likely to be relatively uninformative.

Removal of the first time point of expression profiles in Section 2.2.2 and the application of VBSSM, revealed that some regulatory inferences are more sensitive than others to relatively small changes in the data. Results that are more robust may be more accurate than those that aren't.

It may be useful to combine motif-informed network inference with the module network approach, where co-expressed genes are treated as a single variable for network inference. Motif information could then be used to design hard constraints or informative priors, possibly with a mixture of nodes representing either clusters of genes or single TFs with associated known binding motifs.

Temporal clustering by affinity propagation

Except for the earlier methods of Qian et al. (2001) and Balasubramaniyan et al. (2004), the ability to use time-delays within a cluster to infer regulators is a relatively unique feature of TCAP as a clustering method. AP is more effective at clustering under similarity measure ψ than its alternative, PAM, as has been shown in Figure 2.6. TCAP is of comparable runtime to clustering methods such as SplineCluster (Heard et al., 2005), which makes it more suited to exploratory data analysis than more computationally intensive methods such as the one presented by Balasubramaniyan et al. (2004).

TCAP, like VBSSM, infers the most likely regulation among a group of genes, rather than discriminating between the expression profiles of genes that regulate each other and those that do not. As with VBSSM this is probably a limitation imposed by the use of purely wildtype experiments. However, in the case of the circadian clock, which may or may not be representative of Arabidopsis GRNs in general, Figure 2.5(b) demonstrates that ψ can be used to discriminate between members of a GRN and random genes.

The similarity measure ψ detects time-delayed correlation by aligning gene expression during contiguous subsets of time-points. More specifically, given T time-points, ψ will align $T - |d|$ time-points from each expression profile, where d is the number of time-points delay detected, i.e. if ψ is achieved at $\Omega^{+}_{t_1 t_2}$ or $\Omega^{-}_{t_1 t_2}$ then $d = t_1 - t_2$. A time-delayed correlation with a delay of d is based on the expression of each gene across $(T - |d|)$ time points, it is clear that the higher the number of timepoints under which a gene correlates with another, the less likely that the correlation arose by chance alone. For example, a PCC of 0.95 across 3 time-points is less convincing than a PCC of 0.95 across 10.

The similarity measure ψ penalises longer delays by definition (Algorithm 1 in section 2.1.3). This can be demonstrated by observing that ψ is optimised, for a fixed d, when the expression profile subsets compared are perfectly correlated or anti-correlated, in which case $\text{argmax}_{X_{i\cdot}, X_{j\cdot}|d} \psi(X_{i\cdot}, X_{j\cdot}) = T - |d| - 1$. ψ is used 'raw' in TCAP, whereas in Qian et al. (2001) various methods to reduce this delay penalisation are discussed. These were not implemented in TCAP because of their additional computational cost. This is perhaps why in the TCAP clustering of 9,828 genes, which can be seen in Supplemental Digital Information Table 3, only 558 genes were observed to have a time-delayed match to their cluster centre.

In this chapter attempts have been made to benchmark the similarity measure ψ, as a predictor of transcriptional regulation from gene expression time series. Benchmarking requires known positive and negative cases of transcriptional regulation, as well as expression profiles of the genes involved. First this was performed on a yeast gene expression time series (Figure 2.5(a)), with the positive and negative examples, as well as expression data taken from Qian et al. (2003). The way this dataset was originally generated may positively or negatively bias the inferences. Of all the potential problems, the permutation of negative example target profiles is probably the most problematic, because it may positively bias the ROC curve. Permuting the time-points in an expression profile reduces its auto-correlation, which may reduce

the ψ score of anything compared to it. This was originally performed by Qian et al. (2003) to ensure that all negative examples were reliably negative, but this could lead to a misleading analysis of predictive ability. Other problems with this dataset, such as concatenation of separate experiments and variable time spacing, are less worrying because they would be unlikely to positively bias the analysis.

In the second ROC analysis Arabidopsis gene expression time series data was used to benchmark the ability of ψ to predict transcriptional regulation (Figure 2.5(b)). In this application many of the methodological flaws of the yeast ROC have been avoided: approximate negative examples were chosen at random and target expression profiles were not permuted; data from a single experiment is used; and all time points are evenly spaced. However the circadian clock network is still a topic of ongoing research and so the 'true positives' reflect literature knowledge at the time, rather than perfectly reflecting the underlying biology. Another caveat with this analysis is that the circadian clock network is responsible for generating an endogenous oscillator of gene expression, as it controls the time-of-day dependent (diurnal) patterns of gene expression in Arabidopsis. This biological function may make the circadian clock less representative of Arabidopsis GRNs in general, as the time delayed correlation of gene expression is directly linked to the GRNs function. These caveats also apply to the TCAP cluster that grouped together *LHY* and *GI* (Figure 2.9(a)). While this result may not be representative, it is still striking, grouping together a transcriptional repressor and a target based on temporal features that are not usually analysed genome-wide. Currently, few good datasets for benchmarking of predictors of transcriptional regulation exist, although this may change as synthetic biology matures (Cantone et al., 2009).

TCAP builds on the temporal clustering method introduced in Qian et al. (2001), producing clusters with a higher average similarity to cluster centres (Figure 2.6). It is faster than comparable methods, for example that of Balasubramaniyan et al. (2004), but achieves this at the cost of being approximate and linear. Non-linear time delays, such as in the formulation of Balasubramaniyan et al. (2004), have not yet been benchmarked as a predictor of transcriptional regulation, but if they are found to outperform ψ then they could replace ψ in the TCAP algorithm. This is possible due to the flexibility of AP which is able to cluster under any similarity measure. A key point then would be to develop efficient algorithms for calculating or approximating this measure, to make the modified TCAP fast enough for exploratory analysis of gene expression time series. This is important as the similarity of all pairs of gene expression profiles must be calculated before AP is applied.

AP was shown to substantially outperform PAM at clustering under ψ in Figure 2.6. This is similar to the results of Frey and Dueck (2007), where AP is shown to outperform PAM when applied to clustering: images, text, putative exons and American airports. This shows that AP outperforms PAM at clustering under a wide range of similarity measures. One of the main benefits of AP or PAM, in comparison to K-means, is their ability to cluster under unconventional similarity measures. This has been shown to be useful in this chapter, in Frey and Dueck (2007) as discussed and in a range of other bioinformatic applications such as in: sequence analysis, structural biology and biological network analysis (reviewed in Bodenhofer et al., 2011).

The flexibility of AP with regard to similarity measure will allow the future improvement of TCAP through measures that are, for example: more sensitive to time-delayed correlation; more sensitive to transient correlations; that take into account uneven time-point spacing; or that can detect non-linear correlations. It may be possible to develop a probabilistic equivalent to TCAP based on hidden Markov models, maybe taking inspiration from the literature on probabilistic dynamic time warping (for example Oates et al., 1999). Applications of TCAP, and variants thereof, to other types of biological time series is a topic of ongoing collaborations.

Differences and similarities between VBSSM and TCAP

Both VBSSM and TCAP are methods that allow inference of transcriptional regulation from gene expression. Some of their similarities and differences are discussed in this section. Both VBSSM and TCAP infer transcriptional regulation by detecting time-delayed covariation of expression profiles, in the case of VBSSM this is a result of the Markov process assumption, whereas in TCAP this is a result of approximately calculating the correlation corresponding to all possible time delays. A key difference between VBSSM and TCAP is that the graphical model approach of VBSSM allows it to model combinatorial regulation, whereas TCAP relies on a pair-wise measure, ψ, to infer transcriptional regulation.

Another difference relates to the trade-off between genome-wide and specific inference, which was introduced at the start of this chapter. TCAP inferences are unspecific in the cases where multiple TFs are found to correlate positively and simultaneously (i.e. d=0, for examples see Figures 2.10 and 2.11), whereas VBSSM typically infers specific regulation (see for examples figures 2.1 - 2.4(c)). However, this decrease in specificity of inferences allows TCAP to be applied robustly to the expression of longer lists of genes. This is likely to be due to the relatively simple model of gene regulation (time-delayed correlation) used in TCAP, relative to the

complexity of the model (a SSM) used in VBSSM; the SSM has many parameters that must be fitted to the data, requiring sufficient data to constrain the model. Finally, VBSSM infers the value of unobserved variables. This may make it more robust by taking into account unobserved external influences.

The results of VBSSM and TCAP are different, but related. For example, both methods sometimes infers that TFs and their targets are co-expressed (for example the expression of inferred regulator and targets in: VBSSM inferences presented in Figures 2.1(a) and TCAP Figures 2.10 and 2.11). Figure 2.3 shows an example of a inference of combinatorial regulation by VBSSM, two genes (*At3g48890* and *At5g27520*) are inferred to be regulated by both *ANAC019* and *ANAC092*. Figure 2.4(a) shows the expression of a cluster of genes and its regulator as inferred by VBSSM; visually at least the TF and its targets show time-delayed correlation.

2.3.2 Biological discussion

In this chapter, a time series of (almost) genome-wide gene expression, in Arabidopsis leaves during infection by *B. cinerea*, has been studied. Cluster analysis has been used to reveal co-expressed genes, some of which were found to have known TF binding motifs over-represented in their promoters. Transcriptional regulators of the defence response were inferred from their expression, either by finding time-delayed correlation or by applying network inference to TFs and groups of genes with the TFs known binding motif over-represented in their promoters. Finally, inferred regulators of the defence response were studied in a 'reverse genetics' screen, which revealed several novel but weak altered susceptibility to *B. cinerea* phenotypes.

Botrytis infection time series

The time series of Arabidopsis gene expression during *B. cinerea* infection analysed in this Chapter represents a substantially richer dataset, with 24 timepoint instead of the three or less presented in previous studies (AbuQamar et al., 2006; Ferrari et al., 2007). One feature that has not been explored here is the relation between circadian and defence response related gene regulation, a natural topic given the span of the experiment over 48 hours. Recent work has shown that cold treatment responsive changes in gene expression are partially controlled by circadian/diurnal rhythms, and that changes to the expression of circadian clock components propagate to have a wide effect on the cold responsive transcriptome (Bieniawska et al., 2008). The relation of B. cinerea responsive and circadian regulated gene expression could be investigated by meta analysis of *B. cinerea* responsive changes in gene expression, or by investigation of the expression of circadian clock components during the time series. Indeed, it has recently been shown that the circadian clock is

dampened during *B. cinerea* infection (Windram et al., manuscript in preparation), and that pathogen growth is affected by time of infection (R Smith and K Denby, unpublished). This suggests that *B. cinerea* responsive gene expression may depend on time of infection, and time of day, in non-trivial ways.

Validation of regulation inferred by VBSSM

VBSSM was applied to the expression time series of potentially co-regulated genes and associated TFs to infer regulators of the defence response. Three of the seven inferred regulators, *ANAC019/055/092*, have already been shown to be important for the defence response of Arabidopsis to infection by *B. cinerea* (Bu et al., 2008; Windram, 2010).

Inferred targets of these TFs were compared with targets suggested by the biological literature, to see if known regulation had been recovered. One of the most common methods of finding targets of a TF is see which genes are differentially expressed in a mutant of that TF, for example knockout or overexpressor versus wildtype microarray experiments. No over-representation was seen in the overlap between these lists, produced from published microarray experiments, and the genes inferred to be regulated by them.

Validation of regulation inferred by TCAP

TCAP was used to infer transcriptional regulation from gene expression time series, it was shown to predict known and infer novel transcriptional regulation. One problem with using TCAP to infer regulators of the defence response is that technically it infers regulation for all TFs that fall into non-singleton clusters. Practically, inferred regulation was highlighted if the average Qian similarity was very high (modules 1–2) or if a TF was found to have a positive time delayed correlation to other genes (modules 3–6).

Some of the highlighted inferred regulators belong to TF families already associated with the defence response of Arabidopsis to *B. cinerea*, for example the MYB (Martin and Paz-Ares, 1997; Mengiste et al., 2003), NAC (Bu et al., 2008), AP2-ERE BP (Berrocal-Lobo et al., 2002) and WRKY (Zheng et al., 2006) TF families. Of all the inferred regulators only *ANAC055* had previously been shown to be important for the defence response of Arabidopsis to infection by *B. cinerea* (Bu et al., 2008). VBSSM also inferred that *ANAC055* was a regulator of the defence response (Figure 2.4).

Inferred targets of these TFs were compared with targets suggested by the bio-

logical literature, to see if known regulation had been recovered. Relevant data existed for: a knockout of *ANAC055* during senescence; a constitutive overexpressor of *ANAC072*; a constitutively expressed chimeric *ANAC072* with a fused repression domain; and an overexpressor of a constitutively active version of DREB2A. Of these, only one had an over-represented overlap, and this was only by one gene, excluding the TF in question (At4g37990 was seen in both module 3 and in Fujita et al. (2004) as differentially expressed in a *35S::ANAC072* line).

The vast majority of inferred gene regulation made in this chapter, by both TCAP and VBSSM, currently have no experimental backing. This can be partially attributed to the lack of appropriate experimental data in the literature, even mutant versus wildtype microarray experiments could not be found for most inferred regulators. For those that did exist, very little if any over-representation of known targets in lists of inferred targets was observed. This suggests either that they are false positives, that they regulate redundantly or that the validation data used was not appropriate. None of the literature experiments had been performed during *B. cinerea* infection and so it is possible the literature datasets were not similar enough in biological context to be appropriate for validation of these inferences. Ideally Chromatin Immuno-Precipitation (ChIP) performed with an antibody that precipitates the TF of interest, from leaf samples infected with *B. cinerea*, would be used for validation, but this sort of data is almost non-existent in the literature. Another approach to validate the predicted regulation would be to use Yeast-1-hybrid to test the ability of the inferred regulatory TF to bind the promoter of inferred targets.

Reverse genetics screen of inferred regulators

Most of the inferred regulators of the defence response had not been tested in reverse genetics screens in the literature, and so altered expression mutants were screened for altered susceptibility to *B. cinerea*. TFs whose mutants showed altered susceptibility to *B. cinerea* were likely to be important regulators of the defence response. Altered susceptibility to *B. cinerea* was found to be variable across screens, and not all data was distributed normally (Table 2.2 and Appendix C). Nevertheless, some mutants of inferred regulators were found to give repeatable but weak quantitative resistance against *B. cinerea*.

This study was performed to see which inferred regulators had mutants which caused altered susceptibility to *B. cinerea*. Because the study was not designed to test whether the methodology used was effective at predicting regulators that would have mutants with an altered phenotype, no appropriate control group for that was used. This means that there is no internal control, such as randomly chosen TFs or

differentially expressed TFs, against which to test for over-representation of mutants giving altered susceptibility to *B. cinerea*. The closest substitute for an internal control is the study by AbuQamar et al. (2006) who showed differential susceptibility in mutants of 2 out of 14 up-regulated TFs. Unfortunately, AbuQamar et al. (2006) used spray-infected whole plants to qualitatively show altered susceptibility. This makes it tricky to compare to the screen presented in this chapter, especially as they did not present biological replicates. The screens presented in this chapter revealed that at least one (*ANAC072*), and possibly a second (*NUB*), TF/s out of 9 had two independent T-DNA knockouts that displayed altered susceptibility to *B. cinerea*. This is comparable to the altered phenotype rate in the reverse genetic screen presented in AbuQamar et al. (2006). However, a visual comparison suggests that the novel altered phenotypes found in AbuQamar et al. (2006) are considerably more dramatic.

Most of the gene knockout lines that show *B. cinerea* susceptibility phenotypes in the literature have shown increased susceptibility (for examples Mengiste et al., 2003; AbuQamar et al., 2006; Zheng et al., 2006; Pré et al., 2008), whereas decreased susceptibility was observed in all repeatable phenotypes of mutants of TCAP inferred regulators. Although all novel phenotypes presented here are subtle, it is known that resistance to necrotrophic pathogens is quantitative and polygenic, i.e. relies on many different genes with each contributing only slightly to measurable resistance (reviewed in Poland et al., 2008).

This chapter demonstrates the variability of the *B. cinerea* susceptibility phenotype and shows how this can be taken into account by using biological replicates, measuring susceptibility quantitatively, applying hypothesis testing and by performing multiple independent screens.

Redundantly acting TFs

It is also obvious from the literature that it is not possible to use a lack of phenotype in a mutant of a single gene to show non-involvement of that gene in a given process, because that gene may act redundantly. This has been shown with the *TGA2/5/6*, *WRKY18/40* and *WRKY18/60* combinations; knockouts of these genes show no altered susceptibility to *B. cinerea* whereas combinatorial knockouts do (Zander et al., 2010; Xu et al., 2006b). In fact *ANAC072* is suspected to act at least partially redundantly. This was shown in Fujita et al. (2004) where a knockout did not show an altered ABA sensitivity phenotype, but a constitutively expressed chimeric *ANAC072* with a fused repression domain did. This may also explain the weakness of the altered susceptibility to *B. cinerea* observed with the *ANAC072* knockout.

Transcriptional regulation can also act redundantly, which partly explains disparities between expression and binding experiments (reviewed in Gitter et al., 2009).

If a TF acts redundantly it will be necessary to use other mutants, such as over-expressors or combinatorial knockouts, in reverse genetics screens to observe altered phenotypes. These mutants are frequently not available and can be time consuming to produce. If a researcher is interested in a set of n genes, and they wish to screen double knockouts of all possible combinations, then they must screen $\frac{n}{2}(n-1)$ lines which grows rapidly in n. The situation is worse if a researcher wishes to use combinatorial knockouts of larger numbers of genes. Therefore, combinatorial mutants are more appropriate if a researcher already has a suspicion that the relevant genes may be acting redundantly together. Alternatively, constitutive over-expressors can be used to bypass functional redundancy, altering phenotypes. Unfortunately, over-expressors are also rare and time consuming to produce, and ectopic constitutive over-expression can cause large-scale downstream effects which can confound the phenotype. Ideally, phenotype screens of inducible over-expressors of genes of interest could be used to bypass both redundancy and some of the confounding effects of ectopic expression. Unfortunately, inducible over-expressors are even rarer than constitutive over-expressors.

2.3.3 Conclusions

While it is hard to make genome-wide specific inferences about transcriptional regulation from gene expression time series, it can be achieved either by incorporating additional biological information, such as promoter sequences, or by searching for transcriptional regulation that causes a time-delayed correlation between the expression of a TF and its target. While it is encouraging that both approaches have recovered some known regulators of the defence response, most specific inferred regulation remain to be validated. Two exceptions are the prediction by TCAP that *LHY* regulates *GI* and that *ORA59* regulates a number of downstream genes. Inferred novel regulators of the defence response were investigated in a reverse genetics screen, and novel altered susceptibility phenotypes were observed. However, none of the phenotypes had a susceptibility phenotype as far from wildtype as the positive control, *bos1*, possibly suggesting that forward genetics screens are better suited to identifying genes with a strong non-redundant influence on susceptibility.

Chapter 3

Development and validation of a qualitative model of gene regulation during the defence response

In the previous chapter predicted transcriptional regulation was compared to transcriptional regulation demonstrated in literature studies. It was clear that although targets of some TFs have been studied in some contexts, few had been tested during *B. cinerea* infection. This made it hard to assess the accuracy of the predictions, which relate to transcriptional regulation during *B. cinerea* infection. The context-specificity of transcriptional regulation is not known in general, and so it is hard to know the degree to which regulation observed in one context can be extrapolated to another. This is an important question because although knowledge of transcriptional regulation during infection by *B. cinerea* is sparse, knowledge of transcriptional regulation in other contexts is more substantial. If transcriptional regulation does not depend too heavily on biological context, then knowledge of transcriptional regulation in other contexts can be extrapolated to predict regulation during infection by *B. cinerea*. This could then be used as a qualitative model of the dynamics of pathogen responsive transcriptional regulation, which could be experimentally validated and/or used to guide quantitative modelling approaches.

The aim of this chapter is to develop and validate a qualitative model of gene regulation during the defence response of Arabidopsis to infection by *B. cinerea*. To achieve this, literature knowledge of transcriptional regulation between genes known to be involved in the defence response, derived from experiments performed in any biological context, will be compiled. A qualitative model of the defence response

GRN will then be generated by extrapolating this regulation to the context of *B. cinerea* infection. This qualitative model will then be experimentally validated in relatively 'context free' conditions. Finally, context dependence will be investigated by comparative transcriptomics, to link this regulation to the context of *B. cinerea* infection.

3.1 Introduction

3.1.1 Biological contexts

If a TF can regulate the expression of its target and can bind close to its TSS, it is likely that it affects its target's expression directly by transcriptional regulation. However, this regulation may only occur in certain biological contexts, such as the contexts under which binding and regulation of expression has been observed. In this chapter, the aim is to build a qualitative model of transcriptional regulation in one context, *B. cinerea* infection, from transcriptional regulation observed in other contexts.

Experimental evidence of *in planta* transcriptional regulation, such as those introduced in the previous section, is always derived from a specific biological context. Such contexts are usually physiological – such as different developmental stages, cell types or during specific stresses – or non-physiological – such as during treatment with chemicals at physiologically extreme levels or in mutants with altered expression of certain genes. Transcriptional regulation observed in one biological context can be hypothesised, but not assumed, to occur in a different context. This will be referred to in this thesis as 'out of context' evidence of transcriptional regulation. For example, a knockout of *ARF2* has been shown to be less susceptible to infection by *B. cinerea* (Youn-Sung Kim et al., in preparation), and targets have been identified in the context of seedlings (Vert et al., 2008). The biological context in seedlings, or during *B. cinerea* infection, may affect the genes *ARF2* is able to transcriptionally regulate. Therefore, with respect to transcriptional regulation during *B. cinerea* infection, this evidence is 'out of context'. By comparison, experimental evidence of Arabidopsis transcriptional regulation that has been obtained *in vitro*, in yeast or in bacteria, is 'context free', i.e. it is not known if, and in what conditions, this transcriptional regulation occurs *in planta*.

3.1.2 Yeast one-hybrid

One source of 'context free' evidence of transcriptional regulation is Y1H. Y1H is an experimental approach to identify TFs that can bind to a gene's promoter in yeast. In Y1H the coding sequence of a TF is fused to that for the GAL4 activation

domain (GAL4 AD), which can activate transcription in yeast. The transcription of any gene whose promoter can bind to that TF will then be increased. A promoter-reporter fusion, of the promoter of a given gene, can then be used to observe the binding of the TF-GAL4 AD protein fusion (Li and Herskowitz, 1993). This is a similar technique to yeast two-hybrid (Fields and Song, 1989), but used to study protein-DNA rather than protein-protein interactions.

Y1H can be performed against a library of TFs to reveal which of the TFs can interact with the promoter fragment in yeast. Traditionally such libraries have been constructed by generating cDNA from total RNA in a sample (Li and Herskowitz, 1993), however this library will be biased towards highly expressed genes. More recently Y1H has been made Gateway compatible (Deplancke et al., 2004) and normalised clone libraries have been developed for *Caenorhabditis elegans* and Arabidopsis TFs (Deplancke et al., 2004; Ou et al., 2011), reducing the bias towards highly expressed TFs. The cloned library Y1H method is illustrated in Figure 3.1. Cloned libraries can be combinatorially pooled to make screening high throughput.

Limitations of the Y1H approach

While Y1H can identify potential regulators of a promoter fragment, there are some experimental limitations that can lead to false positives or false negatives. In the case of false positives it is not obvious how a given interacting TF-promoter pair identified by Y1H could be conclusively shown not to occur, in any context relevant to the organism from which the TF and promoter originate. This makes it hard to estimate a false positive rate for Y1H. However most Y1H screens will be performed to identify potential regulators that will then be tested in a specific condition. In this case a false positive is a TF-promoter pair seen to bind in yeast that fails to either bind to the promoter or regulate the expression of the corresponding target gene in that context. In Deplancke et al. (2004) 2 of 6 Y1H interactors tested had a significant effect on the expression of the target gene, giving a false positive rate of ~67% in this setting. This suggests that some Y1H interactors are prevented from affecting transcriptional regulation in certain, or possibly all, experimental conditions. This reinforces the idea that biological contexts need to be considered when considering evidence of transcriptional regulation.

The *pDEST22* plasmid expresses the chimeric TF in relatively high levels that may not correspond to the levels of that TF in physiological conditions *in planta*. This is important because a TF must be present in sufficient levels to regulate its targets. One way a false positive may occur is if the TF is not expressed at a sufficient level in the experimental condition tested. Y1H is performed with activation do-

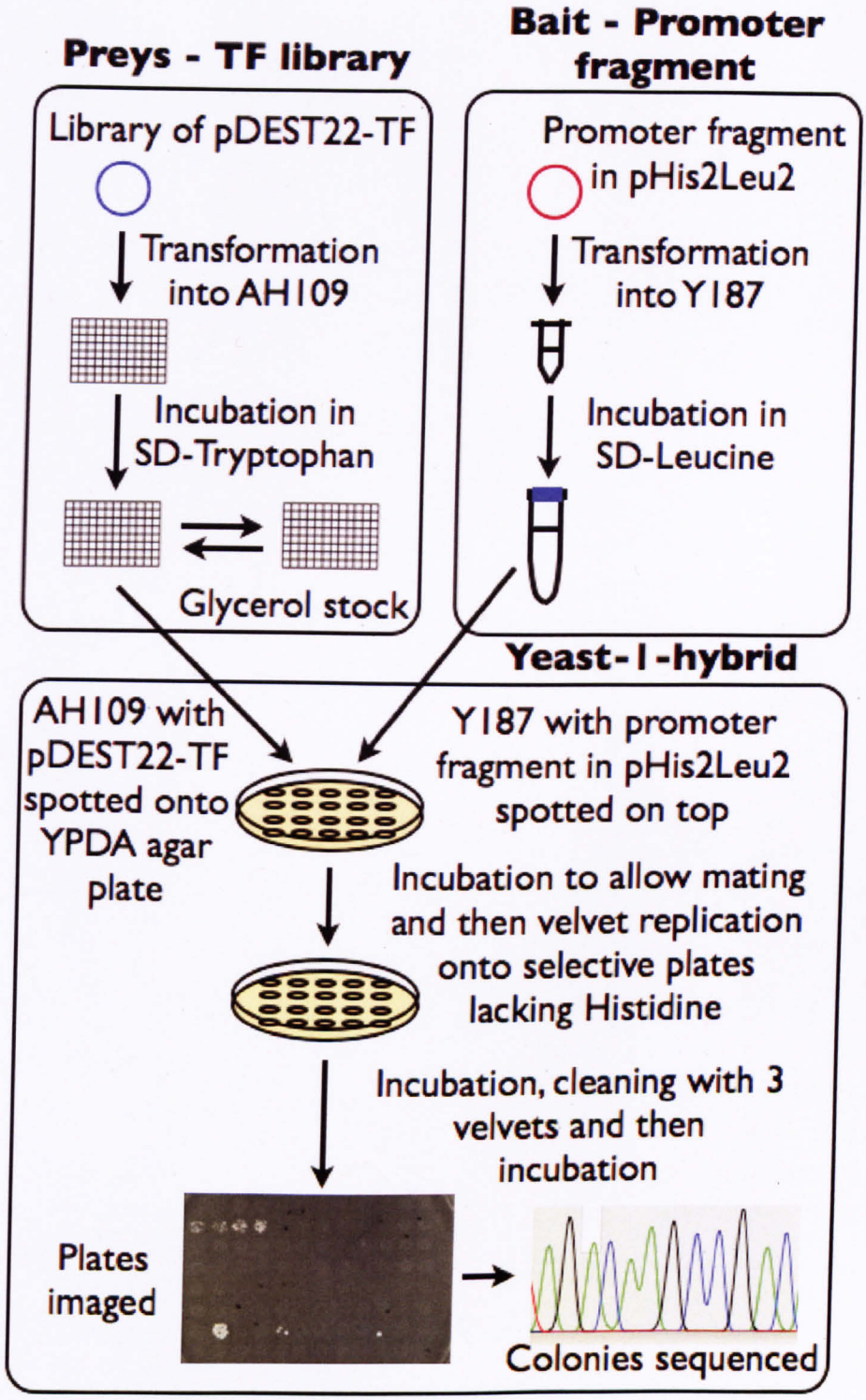


Figure 3.1: Diagram of cloned library Y1H, redrawn and substantially adapted from Ou et al. (2011).

mains fused to TFs, and so an interaction could fail to regulate expression if the native activation/repression domain of the TF was not effective in that context. Another possibility is that the activity of the TF is deactivated in some manner in that experimental context. Because of this, a positive interaction observed in Y1H is 'context-free', in the sense that a researcher does not know which 'natural' conditions the interaction is relevant to, if any. Validation approaches have their own associated false negative rates, and so a true positive identified by Y1H can be spuriously found to be negative in a validation experiment. A final possibility is that a positive interaction is observed in Y1H spuriously due to the variability of auto-activation of the reporter in the promoter-reporter construct.

False negatives are easier to analyse, because positives have been identified by other experimental methods in a range of conditions. For example Y1H was able to identify 3 out of 7 (~40%) known promoter interactors in a study of *C. elegans* promoters (Deplancke et al., 2004). One of the ways a false negative can occur is that the promoter-reporter plasmid is not usually incorporated into the yeast genome, and so it will not be in a proper chromatin context. Additionally false negatives may arise with TFs that require activation or interacting proteins to bind to DNA. False negatives can also arise due to technical issues such as pooling and colony picking, or PCR/sequencing failures.

Estimates of the false positive and false negative rate of Y1H are currently based on one small scale study in *C. elegans* (Deplancke et al., 2004). More accurate estimates will require larger scale studies, ideally with TFs and promoters from a range of organisms.

Arabidopsis transcriptional regulators correctly identified by cloned TF library yeast one-hybrid

Because of the limitations of Y1H it is important to validate positive interactions in the native organism. Here novel interactors of Arabidopsis promoters revealed in cloned TF library Y1H screens that have been subsequently validated *in planta*, from the literature and from personal communication with Richard Hickman, are introduced.

Pruneda-Paz et al. (2009) observed that the TCP TF, CHE, could interact with a fragment of the *CCA1* promoter in a cloned TF library Y1H screen. This binding was validated *in vitro* by EMSA and then *in planta* by ChIP-PCR (Pruneda-Paz et al., 2009). The function of the binding site was demonstrated by mutating these sites in a promoter-reporter fusion. This demonstrated that Y1H interactions, iden-

tified using cloned TF libraries, can occur *in planta* and can impact a given biological process, in this case the circadian clock GRN.

In a cloned TF library Y1H screen for interactors of the *ANAC019* and *ANAC055* promoters, Richard Hickman et al., (in preparation) identified a novel interactor of their promoter fragments in yeast. They subsequently showed that this interactor could regulate their expression *in planta* by performing knockout versus wildtype microarray experiments. In addition, BES1 was found to interact with *ANAC072* promoter fragments in a cloned TF library Y1H screen (Richard Hickman et al., in preparation), and this interaction had already been found by Yu et al. (2011) to occur *in planta* as revealed by ChIP-chip (ChIP with microarray identification of immunoprecipitated DNA).

Validation of Y1H interactions with transient transactivation assays

Another way that transcriptional regulation can be validated is with transient transactivation assays, where cells (of the organism from which the TF and promoter originate) are transiently transformed with both a TF over-expressing plasmid and a promoter-reporter plasmid. The effect of this TF on the promoter is demonstrated by monitoring the effect of over-expression of the TF on the level of the reporter. Transient transactivation assays have not yet been applied in the literature to validate interactions between Arabidopsis TFs and promoters that had been first identified in cloned TF library Y1H screens. However, they have been successfully applied to validate interacting *C. elegans* TF and promoters identified in this way (Deplancke et al., 2004). The advantage of transient transactivation assays, as a way to validate novel Y1H interactors, is that they can be performed rapidly in comparison to other experimental approaches, e.g. EMSA, mutant versus wildtype microarray and ChIP. One disadvantage, in comparison to mutant versus wildtype microarray or ChIP, is that transient transformation has not been adapted to applications during stresses such as *B. cinerea* infection. This means that evidence from transient transactivation assays is 'out of context' with respect to *B. cinerea* infection. Another disadvantage is that transient transactivation assays are not always amenable to high-throughput applications.

In summary, positive Y1H interactors have been validated *in planta* either by ChIP or mutant versus wildtype gene expression measurements. This shows the ability of cloned TF library Y1H to reveal novel regulators of a given Arabidopsis promoter, despite false positives and negatives. Additionally, transient transactivation assays offer a rapid validation of Y1H interactors, but have not yet been applied to validate Arabidopsis cloned library Y1H interactors in the literature.

The aim of this chapter is to develop a qualitative model of the defence response GRN and to experimentally validate it. A qualitative model will be developed from 'in context', 'out of context' and 'context free' evidence from the literature. This will be experimentally validated by relatively 'context free' approaches. Finally, context-specificity of transcriptional regulation will be investigated.

3.2 Materials and Methods

3.2.1 Yeast one-Hybrid

Cloning of promoter fragments with restriction enzymes

Oligonucleotides were designed to produce promoter fragments of approximately 400 base pairs (bp), such that all the fragments for a promoter covered at least the first 1,000 bp upstream of the transcriptional start site (TSS). Oligonucleotides were designed in this way to generate 3 fragments each for the *PGIP1* and *LACS2* promoters. Additionally, oligonucleotides for four fragments of the *WRKY33* promoter were designed in the same way by Laura Butler, see Appendix table D.1(a). Oligonucleotides were extended to add SacI and SpeI restriction sites to the fragments, these additions are shown by lower case letters in the oligonucleotides which are detailed in Tables D.1 and D.2 in Appendix D.

Promoter regions for *LACS2, PGIP1* and *WRKY33* were amplified from genomic DNA (Col4) using these oligonucleotides and KOD polymerase (Roche, Welwyn) according to manufacturer's instructions. PCR products were loaded onto agarose gels, cut out of the gels and cleaned with the QIAquick gel extraction kit (Qiagen, West Sussex) according to the manufacturer's instructions.

Cleaned fragments and the *pHis2Leu2* vector (kindly provided by Claire Hill) were digested with restriction enzymes SacI and SpeI (New England Biolabs, Hertfordshire) according to the manufacturer's instructions. Digested fragments were cleaned using a Qiaquick PCR Cleanup kit (Qiagen, West Sussex) according to the manufacturer's instructions. Digested *pHis2Leu2* was run on an agarose gel and extracted using QIAquick gel extraction kit (Qiagen, West Sussex) according to the manufacturer's instructions. Digested *pHis2Leu2* and promoter fragments were ligated together using T4 DNA Ligase (Invitrogen, Paisley) according to the manufacturer's instructions.

Alpha select Gold efficiency competent cells (Bioline, London) were defrosted for

ten mins on ice. Ligated fragments in *pHis2Leu2* vectors were added to 10 μl of competent cells, mixed gently and incubated on ice for 30 mins. Cells were heat shocked at 42 °C for 30 seconds and then incubated on ice for 2 mins. 500 μl of SOC media (Table 3.1) was added and the cells incubated on a vigorous shaker at 37 °C for 1–1.5 hours. Cells were centrifuged at 1,700 g for 3 mins and 300 μl of supernatant was removed. Cells were resuspended in the remaining media and were transferred to LB agar (both from Sigma-Aldritch, Gillingham) plates containing Kanamycin (50 μg/mL). Cells were grown overnight at 37 °C. Colony PCR was performed with Taq polymerase according to manufacturer's instructions (with oligonucleotides: forward – 5'-CTATCAGGGCGATGGCCCACTA-3', and reverse – 5'-AATGCACTCAACGATTAGCG-3', and Taq polymerase from Invitrogen, Paisley), to check for the presence of the insert. PCR positive colonies were grown overnight in LB containing Kanamycin (50 μg/mL) at 37 °C on a vigorous shaker. Then plasmids were extracted with a QIAprep spin miniprep kit (Qiagen, West Sussex) and sequenced using the primers above (one at a time) and a BigDye®Terminator v3.1 cycle sequencing kit (Applied Biosystems, Warrington) according to the manufacturer's instructions.

Table 3.1: SOC media

Reagents (Sigma-Aldritch, Gillingham)
2% (w/v) bacto-tryptone (20 g)
0.5% (w/v) bacto-yeast extract (5 g)
8.56 mM NaCl (0.5 g)
2.5 mM KCl (0.186 g)
10 mM MgCl2 (0.952 g)
20 mM glucose (3.603 g)
ddH2O to 1000 mL

In addition to the 10 promoter fragments cloned (4 for *WRKY33* and 3 each for *LACS2* and *PGIP1*), an additional 4 were obtained from colleagues. A fragment of the *ARF2* promoter cloned into *pHis2Leu2* was kindly provided by Laura Butler. Three fragments of the *ORA59* promoter cloned into *pHis2Leu2* were kindly provided by Peijun Zhang. The primers that had been used to generate these clones are shown in Appendix tables D.2(b) and D.3.

Transformation of yeast strain Y187 with pHis2Leu2 plasmids

An α strain of *Saccharomyces cerevisiae*, Y187, was grown overnight in 10 ml of YPDA (Clontech, Saint-Germain-en-Laye) at 30 °C on a vigorous shaker. 1 ml of the culture was centrifuged at 400 g for 5 mins for each ten transformations. Cells

were resuspended in 1 ml of 0.1 M LiAc, centrifuged again and then resuspended in 1 ml of 0.1 M LiAc. Cells were incubated at 30 °C in a water bath for 1 hour.

0.5–1 μg of the promoter fragment in *pHis2Leu2* was combined with 40 μg of denatured salmon sperm carrier DNA (Clontech, Saint-Germain-en-Laye) and mixed with 290 μl 50% (v/v) polyethylene glycol (PEG) 3350. The DNA/PEG mix was heated to 30 °C.

100 μl of cell suspension was added to the DNA/PEG mix and mixed gently. Cell/ DNA/ PEG mix was incubated for 50 mins at 30 °C in the water bath. Cells were heat shocked by incubating at 42 °C for 15 mins and then centrifuged at 1,000 g for 5 mins. Supernatant was removed, resuspended in sterile water and then spread on SD minus Leucine (SD-L; minimal SD and amino acid dropout supplements from Clontech, Saint-Germain-en-Laye) agar plates. Plates were incubated at 30 °C for 1–2 days, restreaked onto SD-L (minimal SD and amino acid dropout supplements from Clontech, Saint-Germain-en-Laye) agar plates and incubated again at 30 °C for 1–2 days.

Cloned transcription factor library

The TF library was constructed by Claire Hill and Alexandra Tabrett, as previously described in Windram (2010). The TF library contains TF clones with the N-terminal fused to the GAL4 activation domain in the yeast expression vector *pDEST22* (Invitrogen, Paisley). The library contains 1037 TF clones, pooled 12 clones to a well in two 96-well plates, in two alternative arrangements giving a total of four 96-well plates.

Transformation of yeast with cloned TF library pDEST22s

Claire Hill kindly provided the TF library pre-transformed into an *a* strain of *S. cerevisiae*, AH109 (Clontech, Saint-Germain-en-Laye).

Additionally, for individual (i.e. non-pooled/pair-wise) Y1H screens, *pDEST22* plasmids, kindly provided by Alison Jackson, for *ABI4, ORA59, ERF1, ERF15, MYC2, ARF2, WRKY25, WRKY33* and *At2g23230* were individually transformed into an *a* strain of yeast, AH109, as described for Y187 except using AH109, *pDEST22-TF* clones and SD-T (minimal SD and amino acid dropout supplements from Clontech, Saint-Germain-en-Laye) agar plates for selection. Additionally, a *pDONR::GFP* entry clone kindly provided by Volkan Cevik was used to generate a *pDEST22::GFP* plasmid by the Gateway LR reaction (Invitrogen, Paisley) following the manufacturer's instructions. This *pDEST22::GFP* was also used to transform AH109, as

described above.

Transcription factor library subculture

For each 96-well glycerol stock library plate 500 μl SD-T (minimal SD and amino acid dropout supplements from Clontech, Saint-Germain-en-Laye) was added to each well in a 2.2 ml deep 96-well plate. Transcription factor library glycerol stocks were taken from $-80\,^{\circ}$C storage and placed on ice. Library plates were subcultured using a 96-deep well replicator (V and P Scientific Inc, San Diego) into the 96-well plates containing SD-T (minimal SD and amino acid dropout supplements from Clontech, Saint-Germain-en-Laye) media. Plates were then closed using a gas-permeable seal and incubated at $30\,^{\circ}$C on a vigorous shaker for 4 days.

Pooled library yeast one-hybrid by mating and auxotrophic selection

S. cerevisiae cultures, of Y187 that had been transformed with the promoter fragment containing *pHis2Leu2* plasmid, were made in 10 ml of SD-L (minimal SD and amino acid dropout supplements from Clontech, Saint-Germain-en-Laye) and incubated overnight on a vigorous shaker at $30\,^{\circ}$C. 3 μl of the overnight culture was spotted onto each gridspot of a 96-well arrangement on a YPDA (Clontech, Saint-Germain-en-Laye) agar plate. 3 μl of each well of the transcription factor library subculture was spotted on top of the Y187 spots, at the corresponding library grid position. Yeast were allowed to mate overnight by incubation at $30\,^{\circ}$C.

YPDA (Clontech, Saint-Germain-en-Laye) agar plates were replicated using velvets onto agar plates containing the following growth media (minimal SD and amino acid dropout supplements from Clontech, Saint-Germain-en-Laye):

- SD minus Leucine and Tryptophan (SD-LT).
- SD minus Leucine, Tryptophan and Histidine (SD-LTH).
- SD-LTH with various concentrations of 3-Amino-1,2,4-triazole (3AT).

Plates were incubated at $30\,^{\circ}$C overnight. Then the plates were cleaned with 3 velvets before being incubated at $30\,^{\circ}$C for 3–4 days. Finally, the plates were imaged with upper white light in a G:BOX (SynGene, Cambridge). Growing colonies on SD-LTH and SD-LTH 3AT agar plates were picked into 10 μl of 20 mM NaOH. Then the plate was shaken, sealed and then incubated at $99\,^{\circ}$C for 10 mins. Then colony PCR was performed on 1.2 μl of the boiled yeast extract using Taq polymerase (oligonucleotides: forward – 5'-CTAACGTTCATGATAACTTCATG-3', reverse – 5'-GAAGTGTCAACAACGTATCTACC-3'; and Taq polymerase from Invitrogen,

Paisley), according to the manufacturer's instructions. PCR products were cleaned using a MultiScreen HTS PCR 96-well plate (Millipore, Watford) according to the manufacturer's instructions. Cleaned PCR products were sequenced to identify interacting TFs, using the forward oligonucleotide above and a BigDye®Terminator v3.1 cycle sequencing kit (Applied Biosystems, Warrington) according to the manufacturer's instructions.

Individual yeast one-hybrid by mating and auxotrophic selection

To test individual TF-promoter pairs, Y1H by mating and auxotrophic selection can be performed without pooling. *S. cerevisiae* cultures of Y187 which have already been transformed with the promoter fragment containing *pHis2Leu2* plasmid were made in 10 ml of SD-L (minimal SD and amino acid dropout supplements from Clontech, Saint-Germain-en-Laye). The cultures were incubated overnight on a vigorous shaker at 30 °C. 3 μl of the overnight culture was spotted onto as many grid-spots as required of a 96-well arrangement on a YPDA (Clontech, Saint-Germain-en-Laye) agar plate. Then 3 μl individual transcription factor *pDEST22* transformed AH109 yeast, that had been incubated overnight in 10 ml of SD-T (minimal SD and amino acid dropout supplements from Clontech, Saint-Germain-en-Laye) at 30 °C on a vigorous shaker, were spotted on top of the *pHis2Leu2* transformed Y187 spots. Yeast were allowed to mate overnight by incubation at 30 °C.

YPDA agar plates replicated using velvets onto agar plates with the following growth media (minimal SD and amino acid dropout supplements from Clontech, Saint-Germain-en-Laye):

- SD-LT.
- SD-LTH.
- SD-LTH with various concentrations of 3AT.

Plates were incubated at 30 °C overnight. Then the plates were cleaned with 3 velvets before being incubated at 30 °C for 3–4 days. Finally, the plates were imaged with upper white light in a G:BOX (SynGene, Cambridge).

Individual yeast one-hybrid by co-transformation and auxotrophic selection

Individual Y1H can also be performed by co-transformation. First, individual TFs in *pDEST22* plasmids were co-transformed into yeast strain Y187, that had already been transformed with a *pHis2Leu2* plasmid containing a promoter fragment. This

was performed by the yeast transformation protocol presented earlier, but with SD-L (minimal SD and amino acid dropout supplements from Clontech, Saint-Germain-en-Laye) media used to grow the initial overnight culture and SD-LT (minimal SD and amino acid dropout supplements from Clontech, Saint-Germain-en-Laye) agar plates used to select the transformed yeast.

S. cerevisiae cultures of strain Y187 containing a promoter fragment in a *pHis2Leu2* plasmid, as well as the coding sequence of a TF in a *pDEST22* plasmid, were made in 10 ml of SD-LT (minimal SD and amino acid dropout supplements from Clontech, Saint-Germain-en-Laye) and were incubated overnight on a vigorous shaker at 30 °C. All cultures were concentrated by centrifugation at 300 g for 10 mins, and removal of supernatant to give a concentration of 10^8 cells per ml as determined by optical density and an OD600 table. A serial dilution of each co-transformed culture was performed in 96-well plates. 200 μl of the overnight culture was pipetted into the first well. Then 20 μl of this was pipetted into the second well, followed by 180 μl of H_2O and mixed by pipetting. The previous step/sentence was repeated for the third, fourth and fifth wells. For the screen of interactors of *ARF2* promoter fragment 1, an additional 180 μl of H_2O was added to the first well. Then 3 μl of each serial dilution, of each co-transformed culture, was spotted onto agar plates with the following growth media (minimal SD and amino acid dropout supplements from Clontech, Saint-Germain-en-Laye):

- SD-L.
- SD-T.
- SD-LT.
- SD-LTH.
- SD-LTH with various concentrations of 3AT.

Plates were incubated at 30 °C for 2–3 days. Plates were imaged with upper white light in a G:BOX (SynGene, Cambridge) and photographed.

3.2.2 Biolistic transactivation experiments

Plant Growth

Arabidopsis Col4 (wildtype) seeds were stratified in 0.1% (w/v) agar at 4 °C for 3 days and then transferred to Arabidopsis soil mix (Scotts Levingtons F2s compost:sand:fine grade vermiculite in a ratio of 6:1:1). Plants were grown in a controlled environment with a 10:14 hour light:dark cycle at 19.5 °C, with 60% humidity and

a light intensity of 100 μmol photons.m^{-2}.s^{-1}. These plants were allowed to grow for 6–8 weeks before mature leaves were transformed by biolistic infiltration.

Plasmids

Promoter-reporter fusion plasmids for *WRKY33* (*P1::GUS*, *P4::GUS* and *P4m1-4::GUS*) were kindly donated by Imre Sommisch, their construction is detailed in Lippok et al. (2007). The coding sequences of *ARF2*, *WRKY25* and *WRKY33*, in a Gateway *pDONR* entry vector (Invitrogen, Paisley) were kindly provided by Youn-Sung Kim (*ARF2*) and Alison Jackson (*WRKY25* and *WRKY33*). The coding sequences were then transferred by an LR reaction (Invitrogen, Paisley), performed according manufacturer's instructions, to a destination vector containing a constitutive promoter (*p35S::gateway*) kindly donated by Volkan Cevik. *P1::GUS*, *P4::GUS*, *P4m1-4::GUS*, *p35S::WRKY25*, *p35S::WRKY33*, *p35S::ARF2*, as well as *p35S::MYC2*, *p35S::GAL4DB* and *p35S::LUC* (over-expressor of luciferase) kindly donated by Volkan Cevik, were transformed into Alpha select Gold efficiency competent cells (Bioline, London) as detailed in section 3.2.1. Cells were spread onto LB Agar plates containing Carbenicillin (100 μg/mL) and incubated overnight at 37 °C. Plasmids were then extracted using a QIAprep Midiprep kit (Qiagen, West Sussex) according to manufacturer's instructions. Extracted plasmids were then sequence verified (oligonucleotides used in separate reactions: forward – 5'-CTAACGTTCATGATAACTTCATG-3', reverse – 5'-GAAGTGTCAACAACGTATCTACC-3', from Invitrogen, Paisley) using a BigDye®Terminator v3.1 cycle sequencing kit (Applied Biosystems, Warrington) according to manufacturer's instructions.

Biolistic transformation

DNA mixes for Biolistic transformation were prepared as follows, with volumes equalised with ddH_2O:

- For the first assay : 2.5 μg of *p35S::LUC*, 2 μg of *p35S::[TF of interest]* and 4 μg of *P1::GUS*.
- For the second and third assays : 2.5 μg of *p35S::LUC*, 3 μg of *p35S::[TF of interest]* and 4 μg of *[promoter fragment of interest]::GUS*.

p35S::GAL4 DB, an over-expressor plasmid for the GAL4 DNA-binding domain, was used as a control plasmid. Macrocarriers (Bio-Rad, Hemel Hempstead) were attached to Macrocarrier holders (Bio-Rad, Hemel Hempstead). The rest of this paragraph was performed by Volkan Cevik. DNA mixes were mixed with 50 μl of 2.5 M CaCl and 20 μl of 0.1 M Spermidine trihydrochloride (Sigma-Aldritch,

Gillingham). This was mixed with 50 μl of Tungsten M-17 Microcarriers (Bio-Rad, Hemel Hempstead) that had been suspended in 100% EtOH at 60 mg/ml, vortexed vigourously for 5 mins and then spread onto the Microcarriers.

Col4 leaves were detached by scalpel and placed in the centre of a petri dish containing 1/4 strength Murashige and Skoog medium (Duchefa Biochemie, Haarlem, The Netherlands) in 1.8% (w/v) agar. Transformation was performed with a PDS-1,000/He Biolistic® particle delivery system (Bio-Rad, Hemel Hempstead), using 1,100 psi Rupture discs (Bio-Rad, Hemel Hempstead) and stopping screens (Bio-Rad, Hemel Hempstead). Transformed leaves were then placed in 30 ml tubes (Griener Bio-One, Stonehouse) which had 5 ml of 1/4 strength MS in 1.8% (w/v) agar set along one side of the tube, and approximately 0.5 ml of 1/4 strength MS (Duchefa Biochemie, Haarlem, The Netherlands) at the bottom of the tube.

Three separate biolistic transformations were performed for each DNA mix, with between 2 and 3 leaves transformed in each transformation. Tubes of transformed leaves were transferred to a controlled environment with a 16:8 hour light:dark cycle at 23.5 °C, with 60% humidity and a light intensity of 100 μmol photons.m^{-2}.s^{-1}. They were left in the controlled environment for 24 hours.

Protein extraction

All leaves transformed with the same DNA mix were pooled, frozen in liquid N_2 and ground to a fine powder with a pestle and mortar. The powder was transferred to three separate 2 ml tubes and refrozen in liquid N_2. Approximately 300 μl of Passive Lysis Buffer (Promega, Mannheim, Germany), depending on sample volume, was added to each tube. Tubes were transferred to ice, and then tubes where vortexed briefly twice and then returned to ice, as often as possible until all samples were thoroughly thawed. All tubes were then centrifuged at 12,000 g for 20 mins at 4 °C, after which the supernatant was removed and placed in a deep 96-well plate.

Reporter measurements

For the first two screens samples were loaded sequentially into a 96-well plate. For the larger third assay, samples were loaded in wells assigned by the MATLAB® function randsample.m. This randomisation was used to control for technical variability due to the plate reader. For the Luciferase assays 20 μl of each protein extract (in Passive Lysis Buffer) was mixed with 100 μl of Luciferase Assay Reagent (Promega, Mannheim, Germany). For the GUS assays 80 μl of protein extracts (in Passive Lysis Buffer) was added to 200 μl GUS assay buffer which contained 20 mg of methylumbelliferyl β-D-galactopyranoside (MUG; dissolved in 1% (v/v)

dimethylformamide, both from Sigma-Aldritch, Gillingham) per 25 ml of GUS extraction buffer, which is detailed in Table 3.2.

Table 3.2: GUS extraction buffer.

Reagents (Sigma-Aldritch, Gillingham)	Stock	Volume
50 mM Sodium phosphate pH 7.0	1 M	500 μl
1 mM Ethylenediaminetetraacetic acid (EDTA)	0.5 M	20 μl
10 mM Dithiothreitol (DTT)	0.1 M	1,000 μl
0.1% Triton X-100	10% (v/v)	100 μl
0.1% Sarkosyl	30% (v/v)	334 μl
H_2O up to 10 ml		8,046 μl

20 μl of GUS assay was taken immediately and mixed with 180 μl of GUS stopping buffer (200 mM sodium carbonate), this was used for the zero hour GUS reading. The remaining GUS assay solution was incubated at 37 °C for 12 hours, after which the GUS assay was stopped as in the previous step, this was used for the 12 hour GUS reading.

All reporter measurements were made with a GENios Microplate Reader (Tecan, Männedorf, Switzerland). GUS fluorescence measurements were taken with an excitation wavelength of 360 nm, an emission wavelength of 465 nm, a gain of 60, 3 flashes, 40 μs integration time and a shake and settle time of 3 seconds each. Luciferase luminescence measurements were taken with a 595 nm filter, 5,000 ms integration time, a gain of 150, and a shake and settle time of 3 seconds each.

GUS readings were taken to be the difference between the readings from the 0 and 12 hour GUS reaction, minus the difference between untransformed leaves at 0 and 12 hours. This normalised the readings against the background of fluorescence. These values were divided by Luciferase readings, i.e. the Luciferase readings were used as a transformation control.

T-tests to compare transient transactivation results

All t-test p-values were calculated in MATLAB®, using a two-sample two tail t-test without assuming equal variances (using script ttest2.m).

3.3 Results

Knowledge of the structure of the GRN underlying the defence response of Arabidopsis to infection by *B. cinerea* is currently very sparse. In this section literature

evidence, experimental analysis and bioinformatic approaches will be used to develop hypotheses about the structure of this GRN.

3.3.1 A qualitative model of the defence response GRN

To develop an initial hypothesis about the structure of the GRN, 'in context', 'out of context' and 'context free' evidence of transcriptional regulation was used. As well as regulation of TFs by other TFs, regulation of genes in response to elictors of the defence response and regulation of physiological outputs affecting resistance will be considered. This will be used to generate a model that spans from pathogen perception to activation of resistance mechanisms.

Genes that have been shown to affect the susceptibility of Arabidopsis to infection by *B. cinerea* were selected, with an initial focus on genes encoding TFs or potential physiological outputs, i.e. TFs – *ARF2, EIN3, OCP3, MYC2, ERF1, ORA59, ANAC019, ANAC055, ATAF1, WRKY33, CAMTA3, TGA2, TGA3, TGA5, TGA6* and *WRKY70* (introduced in Section 1.4) – and physiological outputs – *BAP1, CHIB, LACS2, PAD3* and *PGIP1* (introduced in Section 1.4.2).

Evidence from the literature suggesting transcriptional regulation between these genes was compiled. *SDG8, COI1, EIN2, MPK3, MPK6, NPR1* and *JAZs* were added to the list of genes because they affect *B. cinerea* susceptibility, and evidence in the literature suggested they could alter the expression of at least one of the genes in the original list (Berr et al., 2010; Thomma et al., 1998, 1999; Galletti et al., 2011; Chini et al., 2007). Additionally *ATG18a*, a gene encoding a protein required for the autophagy pathway involved in nutrient recycling during programmed cell death (Lai et al., 2011b), was added because it was found to be regulated by *ANAC055* during senescence (Hickman et al., in preparation) and *WRKY33* during *B. cinerea* infection (Lai et al., 2011b). A knockout mutant of *ATG18a* also showed an increased susceptibility to infection by *B. cinerea* (Lai et al., 2011b). Because of this it is possible that *ATG18a* acts as a physiological output of the defence response GRN. Finally, evidence of transcriptional regulation of these genes in response to chitin or OGs, elictors of the defence response, was compiled and is summarised in Table 3.3. Evidence of transcriptional regulation between the list of genes is summarised in Tables 3.4–3.5 and visualised in Figure 3.2.

All genes in Figure 3.2 except *PAD3, PDF1.2, COI1, MYC2, NPR1, SDG8, TGA6* and some of the JAZs were found to be differentially expressed during *B. cinerea* infection (as determined in Section 2.2.1). By visual inspection of the expression profiles of genes in Botrytis-infected versus mock-infected leaves (in the dataset in-

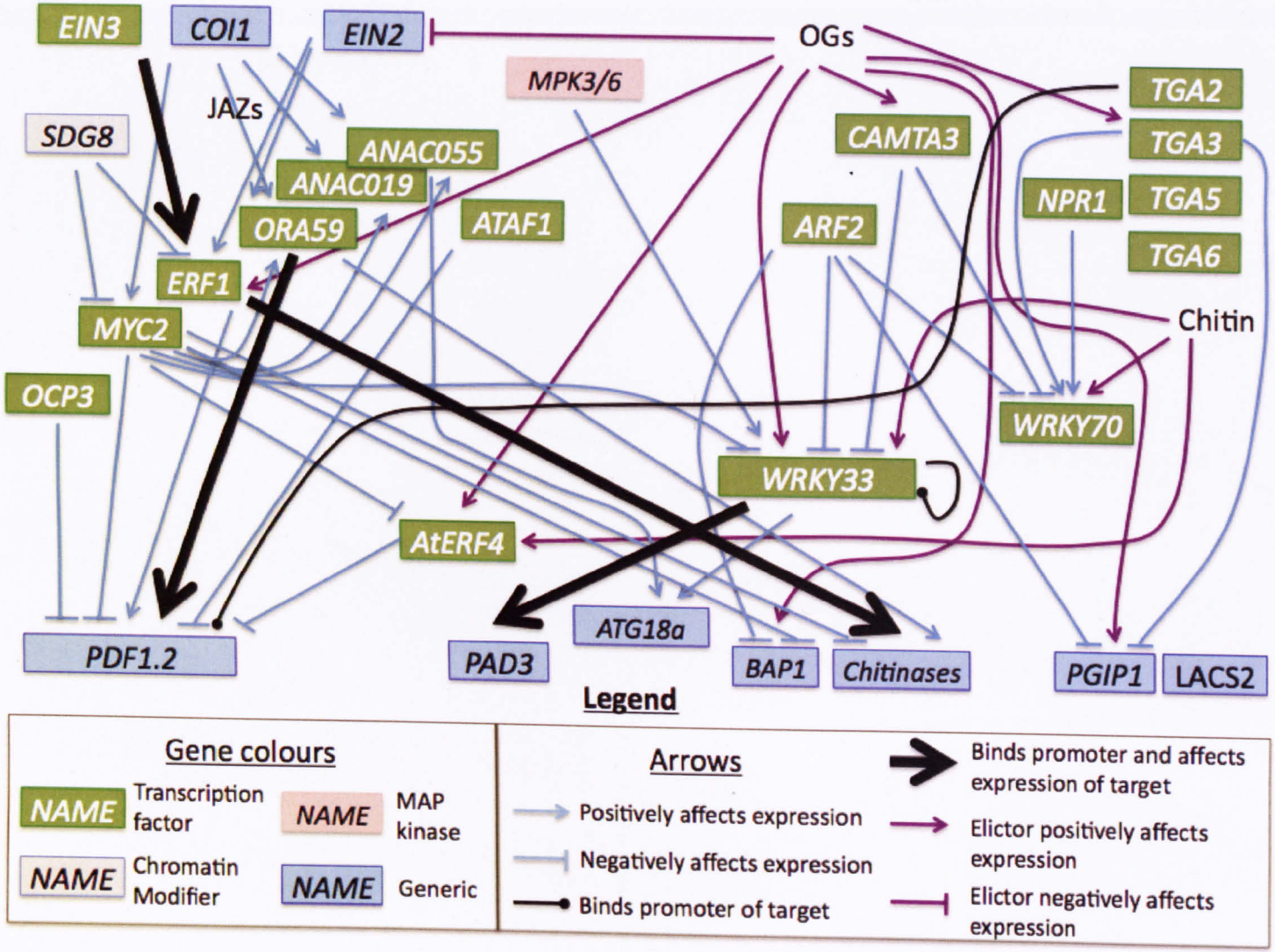


Figure 3.2: Summary of literature on transcriptional regulation of Arabidopsis defence response genes. For references see Tables 3.3–3.5.

Table 3.3: A table summarising genes known to be responsive to elictors of the defence response.

Elicitor	Target	Evidence	Source
Chitin	*ERF4*	qRT-PCR	Libault et al. (2007)
Chitin	*WRKY33*	qRT-PCR	Libault et al. (2007)
Chitin	*WRKY70*	qRT-PCR	Libault et al. (2007)
OGs	*EIN2*	Microarray	Ferrari et al. (2007)
OGs	*ERF1*	Microarray	Ferrari et al. (2007)
OGs	*ERF4*	Microarray	Ferrari et al. (2007)
OGs	*WRKY33*	Microarray	Ferrari et al. (2007)
OGs	*CAMTA3*	Microarray	Ferrari et al. (2007)
OGs	*BAP1*	Microarray	Ferrari et al. (2007)
OGs	*PGIP1*	Microarray	Ferrari et al. (2007)
OGs	*TGA3*	Microarray	Ferrari et al. (2007)

troduced in Section 2.2.1) it was determined that *JAZ1, JAZ6, JAZ7, JAZ8, JAZ9* and *JAZ10* are up-regulated, and *JAZ12* down-regulated during infection by *B. cinerea*. In all of these cases the up- or down-regulation was consistent after differential expression, i.e. genes upregulated at a timepoint in the Botrytis-infected time series, relative to the mock-infected, stayed up-regulated at later time points and *visa versa*. PAD3 and PDF1.2 were not represented by microarray probes on the CATMA cDNA arrays used in Section 2.2.1, but they have been shown previously to be up-regulated during *B. cinerea* infection (Manners et al., 1998; Lai et al., 2011a). While more than half of the genes featured in Figure 3.2 are up-regulated, *ARF2, CAMTA3, EIN2, EIN3, LACS2, MPK3, OCP3, TGA2, TGA3, TGA5* and *WRKY70* are down-regulated during infection. This suggests that downregulation of TFs is part of the defence response, suggesting that focusing on up-regulated TFs as in AbuQamar et al. (2006) may miss important regulators of the defence response. (Differential expression of genes during *B. cinerea* infection determined according to Section 2.2.1).

The vast majority of relevant literature evidence is 'out of context' with respect to *B. cinerea* infection and relates to regulation of expression, with relatively few cases of TF-promoter binding known. However, if 'out of context' regulation can be extrapolated to the context of *B. cinerea* infection then this evidence provides a good basis for an initial model of the structure of the defence response GRN. Additionally, nothing is known about the transcriptional regulation of *LACS2*, which may be a physiological output controlled by the defence response (Section 1.4.2).

Table 3.4: Literature and unpublished evidence of transcriptional regulation relating to the defence response.

Regulator	Target	Context	Evidence	Source
EIN3	*ERF1*	*35S::EIN3*	RNA blot	Solano et al. (1998)
EIN3	*ERF1*	*in vitro*	EMSA	Solano et al. (1998)
ERF1	*CHIB*	*35S::ERF1*	RNA blot	Solano et al. (1998)
ERF1	*CHIB*	*in vitro*	EMSA	Solano et al. (1998)
ORA59	*CHIB*	Inducible over-expressor of *ORA59*	Microarray	Pré et al. (2008)
ANAC055	*ATG18a*	Senescing *ANAC055* knockout	Microarray	Hickman et al., (in preparation)
WRKY33	*ATG18a*	*B. cinerea* infected *WRKY33* knockout	RNA blot	Lai et al. (2011b)
COI1	*MYC2*	MeJA treated *COI1* knockout	RNA blot	Bu et al. (2008)
COI1	*ORA59*	Two week old liquid media grown	RNA blot	Pré et al. (2008)
COI1	*ANA019*	MeJA treated *COI1* knockout	RNA blot	Bu et al. (2008)
COI1	*ANA055*	MeJA treated *COI1* knockout	RNA blot	Bu et al. (2008)
EIN2	*ERF1*	JA and/or ET treatment	RNA blot	Lorenzo et al. (2003)
NPR1	*WRKY70*	SA treated plants	Microarray	Wang et al. (2006)
MAPK3	*WRKY33*	*MAPK3/6* knockout seedlings	qRT-PCR	Mao et al. (2011)
MAPK6	*WRKY33*	*MAPK3/6* knockout seedlings	qRT-PCR	Mao et al. (2011)
OCP3	*PDF1.2*	Leaf	RT-PCR	Coego et al. (2005)
ERF1	*PDF1.2*	*35S::EIN3*	RNA blot	Solano et al. (1998)
ORA59	*PDF1.2*	Two week old liquid media grown	RNA blot	Pré et al. (2008)
ORA59	*PDF1.2*	*in vitro*	EMSA	Zarei et al. (2011)
ATAF1	*PDF1.2*	*B. cinerea* infected plants	RT-PCR	Wang et al. (2009)
TGA2	*PDF1.2*	*in vitro*	EMSA	Spoel et al. (2003)
ERF4	*PDF1.2*	*35S::ERF4*	RT-PCR	McGrath et al. (2005)

Table 3.5: Literature and unpublished evidence of transcriptional regulation relating to the defence response continued.

Regulator	Targets	Context	Evidence	Source
MYC2	*ERF4*, *WRKY33*, *ORA59*, *BAP1* and *CHIB*	*MYC2* knockout, treated or untreated with JA	Microarray	Dombrecht et al. (2007)
ARF2	*WRKY33*, *BAP1*, *PGIP1* and *WRKY70*	*ARF2* knockout seedlings	Microarray	Vert et al. (2008)
CAMTA3	*WRKY33* and *WRKY70*	*CAMTA3* knockout	Microarray	Galon et al. (2008)
SDG8	*ERF1* and *MYC2*	methyl-JA treated *SDG8* knockout	Microarray	Berr et al. (2010)
TGA3	*PGIP1* and *WRKY70*	*B. cinerea* infected *TGA3* knockout	Microarray	Windram (2010)

3.3.2 Context free validation and extension of the qualitative model by yeast one-hybrid

The relative lack of literature on direct regulators of the defence response motivates the application of cloned library Y1H to identify TFs capable of binding to the promoters of genes highlighted in Figure 3.2. In the Y1H screens that will be presented, three levels of 3AT were used to reduce auto-activation by inhibiting Histidine biosythesis; in the cloned library Y1H screen of *WRKY33* promoter fragments 25, 50 and 100 mM 3AT were used, in all other Y1H screens these levels were lower (5, 25 and 50 mM) as 5 mM was found sufficient to prevent auto-activation in many cases.

WRKY33

Only one TF, WRKY33 itself, is known to be able to bind to the *WRKY33* promoter (Mao et al., 2011).

Pooled library Y1H To see if any of the TFs which are known to be able to transcriptionally regulate *WRKY33* can also bind to the *WRKY33* promoter, as well as identifing other TFs that are able to interact with fragments of the *WRKY33* promoter, cloned library Y1H was performed. Four promoter fragments were used, each approximately 400 bp long, covering the 1,000 bp upstream of the TSS of *WRKY33*. Y1H was performed on all four fragments, against all four library plates. The TFs found to interact with the promoter fragments are summarised in Table 3.6. These are TFs that caused growth beyond auto-activation on at least one selective plate.

Table 3.6: Pooled TF library Y1H screen by mating and auxotrophic selection. Interactors of *WRKY33* promoter fragments 1–4 are shown. TFs highlighted in red were found in screens with both arrangements of the library.

Fragment number	Interactors
1	ERF10, AT3G12910 (NAC)
2	None
3	WRKY33, WRKY25, TCP3, TCP8, TCP14, TCP15, TCP16, TCP20, AT1G35560 (TCP)
4	TCP1, TCP3, TCP15, TCP16

The known WRKY33-*WRKY33* promoter interaction was recovered by Y1H, as

well as a WRKY25-*WRKY33* promoter interaction that is novel. *WRKY25* and *WRKY33* have high sequence similarity (Eulgem et al., 2000), and both have been found to interact with the MAPK substrate MKS1 (Andreasson et al., 2005), suggesting a degree of redundancy. Many TCP TFs were found to interact with the *WRKY33* promoter in two fragments. Little is known of the role of TCP TFs in the defence response. It should be noted that fragments with no identified interactors might have been found to interact with more TFs if different levels of 3AT were used.

A degree of technical variability is observed between the alternative pooling replicates; this is consistent with the known limitations of cloned library Y1H screens and Y1H in general (introduced in Section 3.1.2).

Retested Y1H by individual mating Given the technical variability observed, some of these interactions were retested in an individual (i.e. pair-wise/non-pooled) Y1H screen to see if the results were repeatable. Additionally, a negative control was used to allow better characterisation of auto-activation. This negative control was transformed with *pDEST22::GFP*, which expresses the green fluorescent protein (GFP) which has no reported DNA-binding interactions. For experimental tractability only a small number of interactors were re-tested. WRKY33 was re-screened because the WRKY33-*WRKY33* promoter interaction has already been shown to occur *in planta* (Mao et al., 2011) and therefore this interaction can act as a positive control. WRKY25 was rescreened because it has a high sequence similarity to WRKY33 (Eulgem et al., 2000), this means that it is plausible that this interaction also occurs *in planta*. Additionally, it was noted that *WRKY33* expression was increased in knockouts of either *ARF2* (Vert et al., 2008) or *MYC2* (Dombrecht et al., 2007) and that binding motifs for each, TGTCTC (Wang et al., 2011) and CACATG (Abe et al., 1997; Badis et al., 2009) respectively, were found to be present in the 1 kb upstream of the TSS of *WRKY33*. This suggests the potential for direct transcriptional regulation of *WRKY33* by these TFs, and so they were re-screened even though they were not seen to interact in the cloned library Y1H screen.

Individual Y1H screens were performed by mating to retest the interaction of WRKY25, WRKY33, ARF2 and MYC2 with *WRKY33* promoter fragments 1–4. The results of the individual Y1H by mating for *WRKY33* fragments 1–4 are shown in Figure 3.3. WRKY33 was again shown to be able to bind to a fragment of its own promoter, confirming the result of the pooled library Y1H screen (Table 3.6). Only screens using SD-LTH and SD-LTH 5 mM 3AT are shown, WRKY33 (prey) + *WRKY33* promoter fragment 3 (bait) was able to grow at 25 and 50 mM 3AT, but no other growth beyond control was seen on these selections. Auto-activation, as demon-

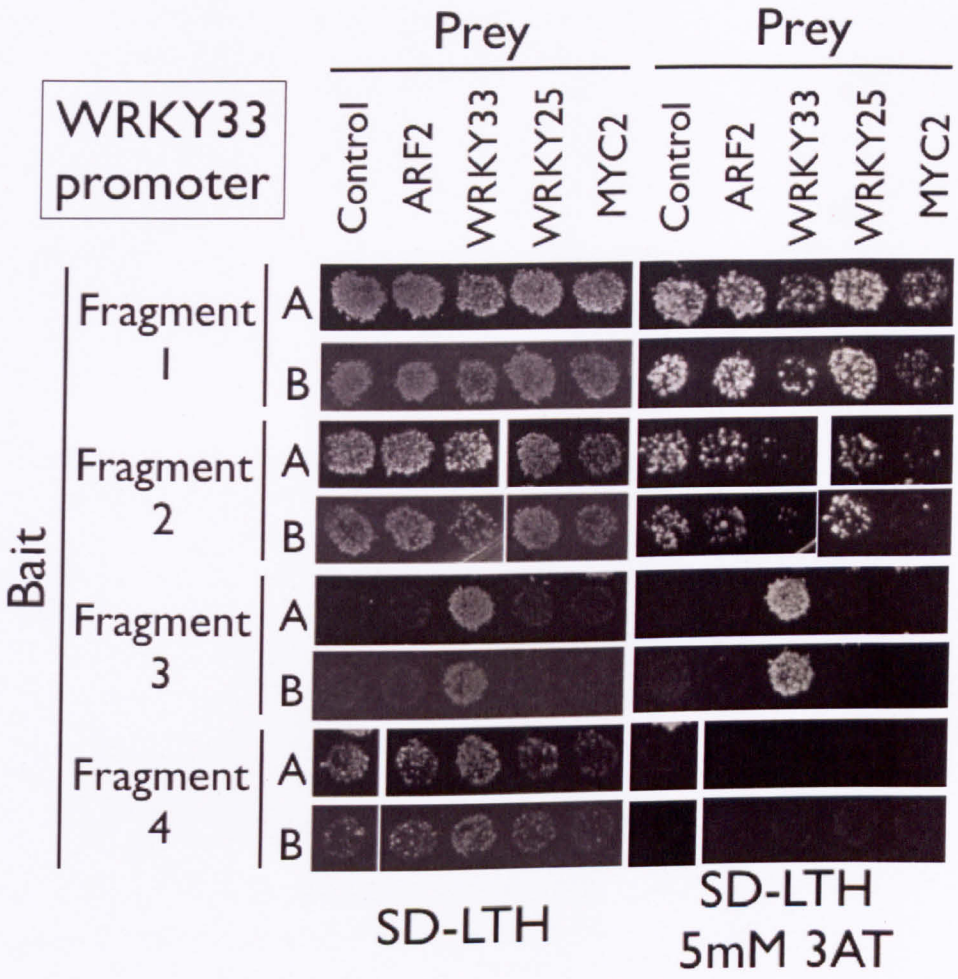


Figure 3.3: WRKY33 can interact with the promoter of *WRKY33* in yeast. Photos are of individual Y1H screens, of *WRKY33* promoter fragments, by mating and auxotrophic selection. Baits, preys and selections were as shown. Two replicates, A and B, were used to control for experimental variability. WRKY33 was shown to interact with *WRKY33* promoter fragment 3.

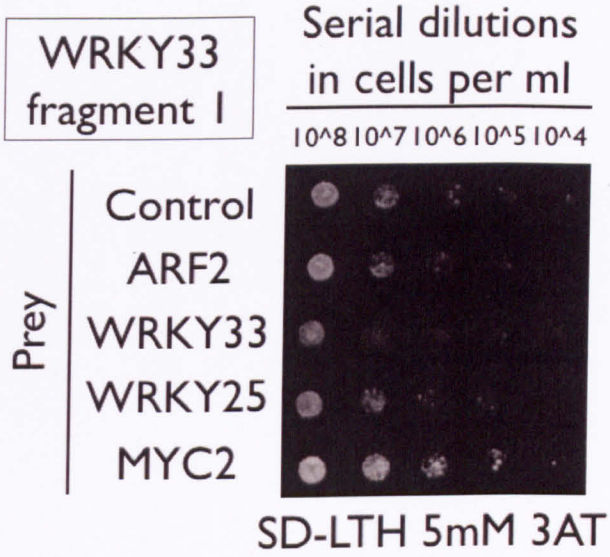
strated by growth of the control, was observed with all other promoter fragments. WRKY25 was not observed to interact with *WRKY33* promoter fragment 3 in either of the replicates, raising the possibility that WRKY25 was a false positive in the cloned library Y1H screen.

Retested Y1H by co-transformation Because of the variability observed between the two Y1H screens by mating, that is the lack of a WRKY25-*WRKY33* promoter fragment interaction in the second screen, it was desirable to retest these interactions with a less variable Y1H screen. This was achieved by performing co-transformation instead of mating and by controlling the concentration of cells in each spot. Additionally, because of the high levels of auto-activation observed in the previous two screens, serial dilutions were used to give a higher resolution of the differences in growth rate. Individual Y1H screens were performed by co-transformation, to retest the potential interactions of WRKY25, WRKY33, ARF2 and MYC2 with *WRKY33* promoter fragments 1–4.

The results of the individual Y1H screens, by co-transformation, for *WRKY33* promoter fragments 1 and 3 are shown in Figure 3.4. The other two fragments had no growth beyond the control and so are not shown. In Figure 3.4(a) MYC2 was shown to interact with *WRKY33* promoter fragment 1 in yeast. In Figure 3.4(b) WRKY25 and WRKY33 were shown to interact with *WRKY33* promoter fragment 3 in yeast, in a fragment containing the 38 bp stretch that has been shown to be required for pathogen response in Lippok et al. (2007) and overlapping with the sequence that WRKY33 was found to bind *in planta* (Mao et al., 2011). (All other fragments and selections had no growth beyond the control, except for WRKY25 on *WRKY33* fragment 3 with 25 mM and 50 mM selection and so are not shown.)

Validation of Y1H results by motif analysis The Y1H screens presented above have been performed against promoter fragments approximately 400 bp long, whereas TFs have been shown to bind to specific short DNA sequences about 5-8 base pairs (bp) long (reviewed in Wray et al., 2003). It is not possible to tell which short sequences were bound by interacting TFs in the screens presented here, but known binding motifs of these TFs can be used to identify potential binding sites, this is shown in Figure 3.5. Additionally, two promoter fragments (P1 and P4) used in a study by Lippok et al. (2007) are shown which will be used to validate some of these interactors *in planta* in Section 3.3.3. For every Y1H interactor, a match of at least four bp with a known binding motif associated with that TF was found at positions within the relevant promoter fragments.

While WRKY33 has already been shown to bind to its own promoter *in planta*



(a) MYC2 can interact with *WRKY33* promoter fragment 1 in yeast.

(b) WRKY25 and WRKY33 can interact with *WRKY33* promoter fragment 3 in yeast.

Figure 3.4: MYC2, WRKY25 and WRKY33 can interact with the promoter of *WRKY33* in yeast. Photos are of individual Y1H screens, of *WRKY33* promoter fragments, by co-transformation and auxotrophic selection. (a) Y1H of *WRKY33* promoter fragment 1 showed that MYC2 can interact. (b) Y1H of *WRKY33* promoter fragment 3 showed that both WRKY33 and WRKY25 can interact.

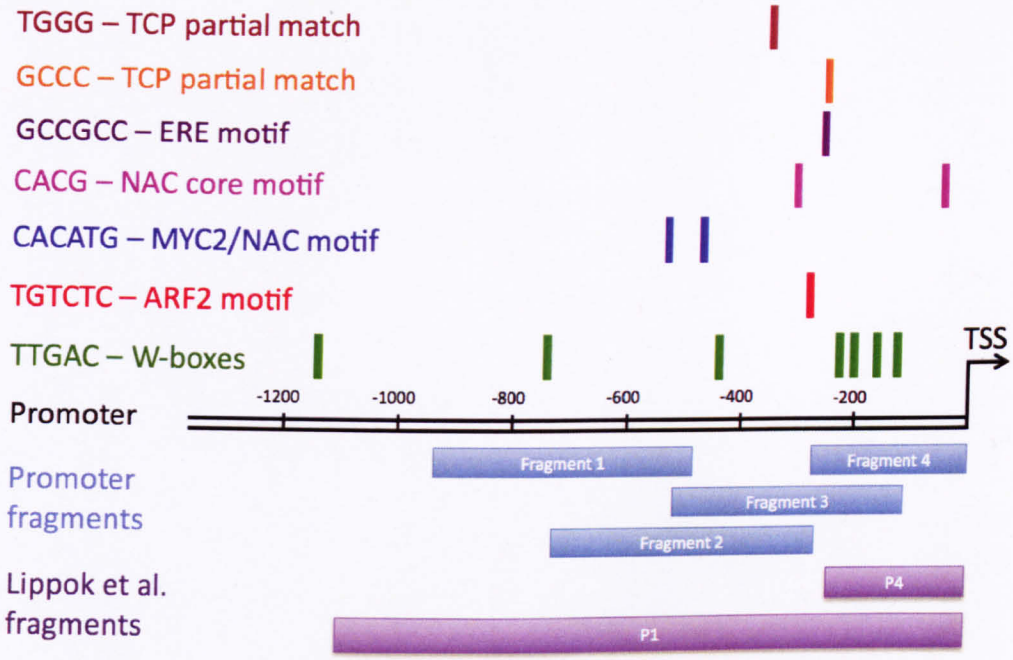


Figure 3.5: The *WRKY33* promoter, with TF binding motifs and promoter fragments displayed. Promoter fragments are those screened by Y1H in this chapter. Lippok et al. fragments correspond to the fragments constructed by Lippok et al. (2007). WRKY-box (de Pater et al., 1996; Ciolkowski et al., 2008), ARF2 (Wang et al., 2011), MYC2/NAC (Abe et al., 1997; Badis et al., 2009; Tran et al., 2004), NAC core (Tran et al., 2004) and ERE (Ohme-Takagi and Shinshi, 1995; Badis et al., 2009) motif matches are shown. Additionally, partial matches of the TCP binding motif TGGGC[C/T] (Giraud et al., 2010), are shown.

(Mao et al., 2011), the other interactors, including MYC2 and WRKY25, represent possible novel interactors *in planta*. This adds to the 'context free' evidence of transcriptional regulators of *WRKY33*, which can be used to extend the qualitative model presented in Figure 3.2. MYC2 was only shown to interact with a fragment of the *WRKY33* promoter in the Y1H by co-transformation. This is likely to be because the serial dilutions provided a better resolution of the difference in growth of this colony in comparison to the control. The interaction of WRKY25 with the promoter of *WRKY33* was seen in the original library Y1H screen and in the co-transformation Y1H screen, and so the negative result in the individual mating Y1H screen is likely to be a false negative. Motifs corresponding to these interactors present in the *WRKY33* promoter adds plausibility and provides potential binding sites for the interactors identified by cloned library Y1H.

ORA59

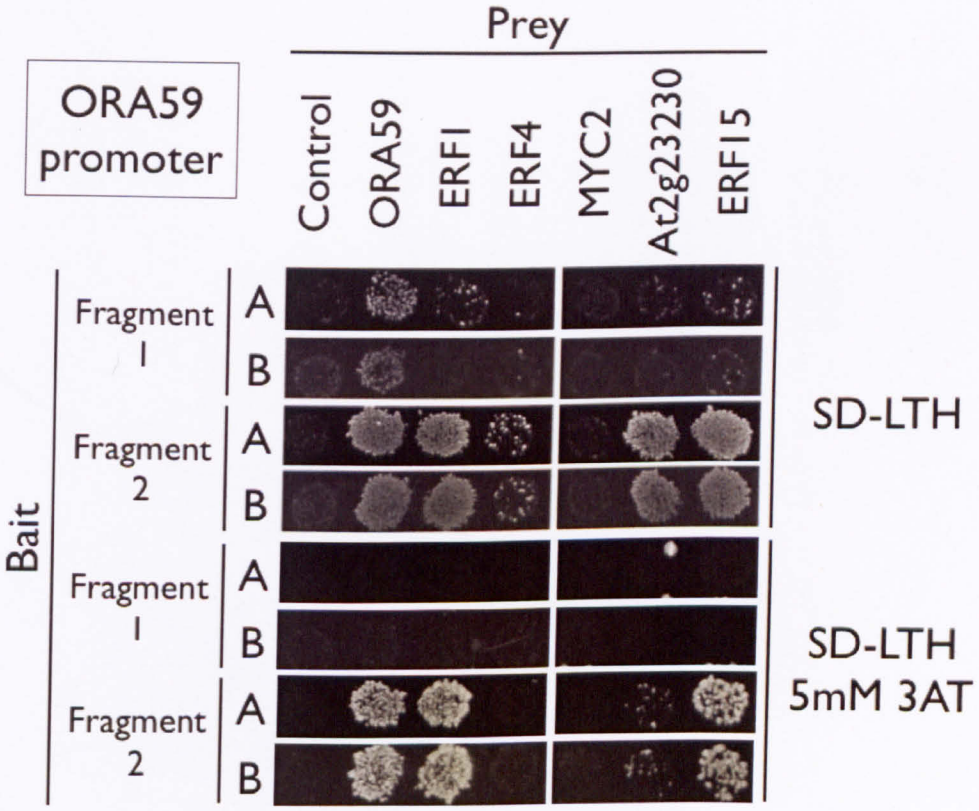
ORA59 is an AP2-ERE BP TF that is up-regulated in response to *B. cinerea* infection, and has been shown to be up-regulated synergistically in response to JA and ET treatment (Pré et al., 2008). However, nothing is known about the direct transcriptional regulators responsible for the upregulation of *ORA59* expression in these conditions.

Pooled library Y1H To identify potential transcriptional regulators of *ORA59* three promoter fragments, each approximately 400 bp long, covering the 1,000 bp upstream of the TSS of *ORA59* were screened with a pooled library Y1H screen. Y1H was performed on all three fragments, against all four library plates. The TFs that were found to interact with fragments of the *ORA59* promoter are summarised in Table 3.7.

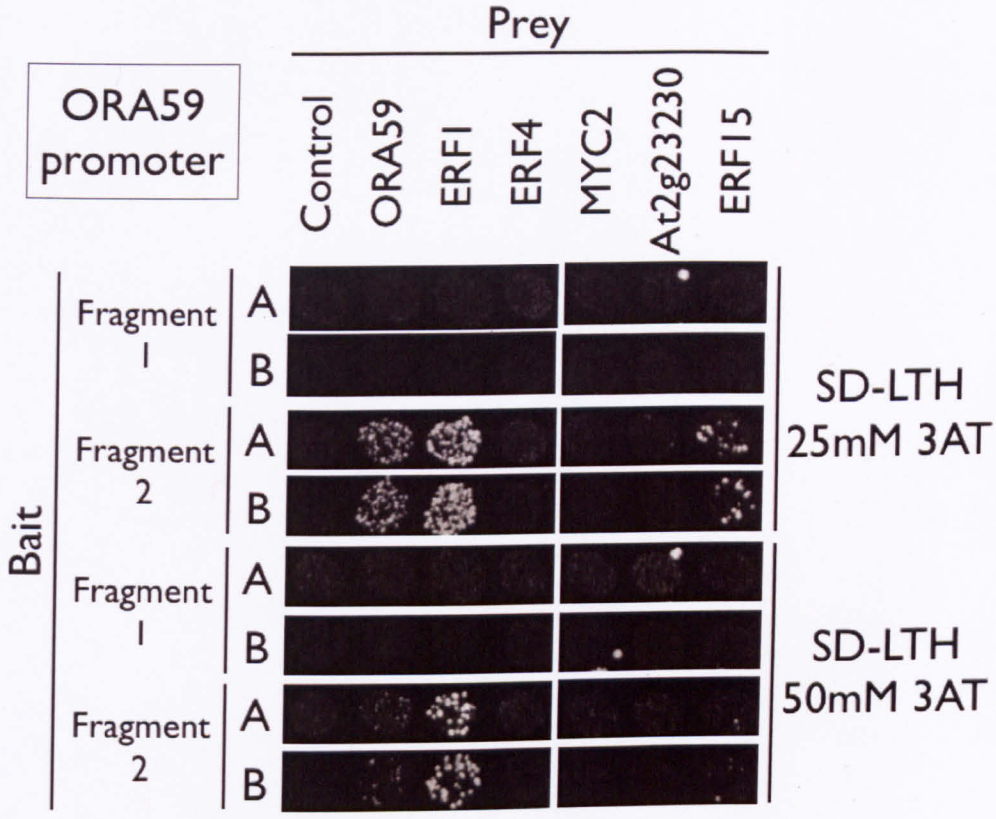
TCPs and AP2-ERE BPs, as well as the NAC TF At3g12910 that has also been found to bind to the *WRKY33* promoter (Table 3.6), were found to interact with fragments of the *ORA59* promoter. Similarly to the WRKY33-*WRKY33* promoter interaction, ORA59 was found to interact with fragments from its own promoter (fragments 1 and 2 shown in Table 3.7). Other than the TCP and NAC TFs, no TF was found to interact with both the *WRKY33* and *ORA59* promoter fragments, suggesting that cloned library Y1H is able to identify novel interactors specific to each promoter. Additionally, the fact that WRKY TFs were found to interact with the promoter of the WRKY TF *WRKY33* and AP2-ERE BPs were found to interact with the promoter of the AP2-ERE BP *ORA59* suggests intra-family transcriptional regulation.

Table 3.7: Pooled TF library Y1H screen by mating and auxotrophic selection. Interactors of *ORA59* promoter fragments 1–3 are shown. TFs highlighted in red were found in screens of both arrangements of the library.

Fragment number	Interactors
1	ERF1, ERF14, ERF15, At3g23220 (ERF), ORA59, At2g42920, PIF7, TCP1, TCP3, TCP4, TCP14, TCP15, TCP16, TCP20, At1g35560 (TCP)
2	ERF1, ERF2, ERF5, ERF6, ERF10, ERF13, ERF14, ERF15, ERF71, At3g23220 (ERF), AT3G23230 (ERF), At5g43410 (ERF), ORA59, ABI4, TCP1, TCP3, TCP16
3	At3g12910 (NAC), TCP14



(a) SD-LTH and SD-LTH 5 mM 3AT selections



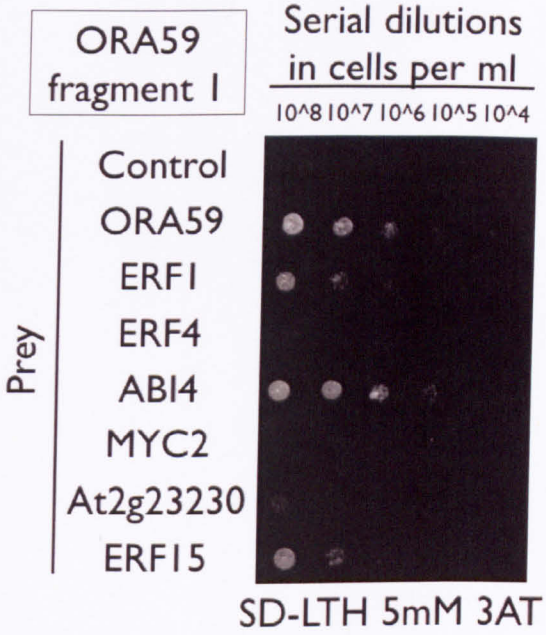
(b) SD-LTH 25 mM 3AT and SD-LTH 50 mM 3AT selections

Figure 3.6: ORA59, ERF1, ERF4, AT2G23230 and ERF15 can interact with the promoter of *ORA59* in yeast. Photos are of individual Y1H screens, of *ORA59* promoter fragments, by mating and auxotrophic selection. Baits, preys and selections were as shown. Two replicates, A and B, were used to control for experimental variability. ORA59, ERF1 and ERF15 were shown to interact with *ORA59* promoter fragments 1 and 2. ERF4 and AT2G23230 were shown to interact with *ORA59* promoter fragment 2. (a) The results from selections on SD-LTH and SD-LTH 5 mM 3AT media. (b) The results from selections on SD-LTH 25 mM 3AT media and SD-LTH 50 mM 3AT media.

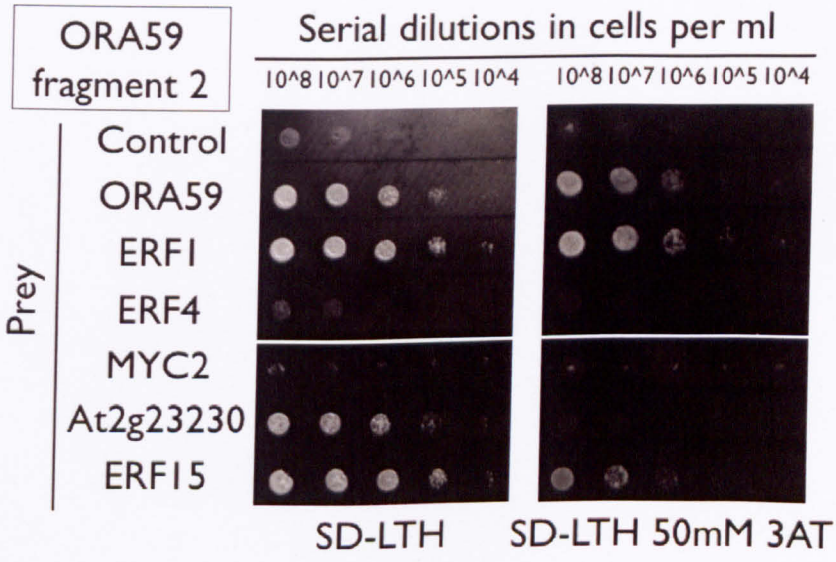
Retested Y1H by individual mating Lots of AP2-ERE BPs interacted with the *ORA59*, but not the *WRKY33*, promoter fragments in Y1H and so it was assumed that they were more promoter-specific than the TCP TFs. For experimental tractability a small subset of the AP2-ERE BPs were rescreened individually to validate the results of the cloned library Y1H screen. Although *MYC2* was not seen to bind to the *ORA59* promoter fragments in the pooled library Y1H screen, *ORA59* has been shown to be up-regulated in a knockout of MYC2 (Dombrecht et al., 2007) showing that it is able to regulate *ORA59* expression and so it was rescreened to see if this regulation was likely to occur directly. Individual Y1H screens were performed by mating to retest the interaction of ORA59, ERF1, ERF4, MYC2, AT2G23230 and ERF15 with *ORA59* promoter fragments 1–2. The results of the individual Y1H by mating for *ORA59* fragments 1–2 are shown in Figure 3.6. ORA59, ERF1, ERF4, AT2G23230 and ERF15 were again shown to be able to bind to fragments of the *ORA59* promoter, confirming the results of the pooled library Y1H screen (Table 3.7).

Retested Y1H by co-transformation Individual Y1H screens were performed by co-transformation to retest the interaction of ORA59, ERF1, ERF4, ABI4, AT2G23230 and ERF15, as well as MYC2, with *ORA59* promoter fragments 1–2. Unfortunately, the plasmid containing ABI4 did not successfully co-transform with the plasmid containing *ORA59* fragment 2 and so this interaction could not be tested in this screen. The results of the individual Y1H by co-transformation for *ORA59* promoter fragments 1–2 are shown in Figures 3.7(a)–(b). ORA59, ERF1, ABI4, AT2G23230 and ERF15 were shown to interact with *ORA59* promoter fragment 2. In addition ORA59, ERF1, ABI4, ERF15 and possibly AT2G23230, were shown to interact with *ORA59* promoter fragment 1. For *ORA59* fragment 1 most selections other than those which are shown were not very informative; most interactors except ABI4 are not seen at higher levels of 3AT. For *ORA59* fragment 2 most other selections were also not very informative; most interactors have similar growth on intermediate levels of 3AT, except for AT2G23230 which shows intermediate growth between the selections shown.

Validation of Y1H results by motif analysis Motif analysis was used to identify potential binding sites of the interacting TFs. An ERE, the binding motif associated with the AP2-ERE BP TFs, is present in the *ORA59* promoter. Its position, at the overlap of promoter fragments 1 and 2, fits well with the finding that AP2-ERE BPs can interact with these fragments. It is worth noting that the ERE is approximately 300 bp closer to the start of fragment 2 then it is to the start of fragment 1, which may account for the increased growth of the colonies corresponding to AP2 ERE-BPs in the co-transformation Y1H screen and the greater amount



(a) *ORA59* promoter fragment 1



(b) *ORA59* promoter fragment 2

Figure 3.7: ORA59, ERF1, ABI4, AT2G23230 and ERF15 can interact with the promoter of *ORA59* in yeast. Photos are of individual Y1H screens, of *ORA59* promoter fragments, by co-transformation and auxotrophic selection. (a) Y1H of *ORA59* promoter fragment 1 showed that ORA59, ERF1, ABI4 and ERF15 can interact. (b) Y1H of *ORA59* promoter fragment 2 showed that ORA59, ERF1, AT2G23230 and ERF15 can interact.

of AP2-ERE BP interactors with fragment 2 in the pooled library Y1H screen. The only full TCP motif match is in fragment 1, which was only found to have one interacting TCP TF (Table 3.7), however TCP TFs were also found to interact with fragments 1 and 2. To identify potential binding sites, partial matches with the TCP motif were identified and are shown in Figure 3.8. It is not known whether TCP TFs are able to bind to these partial matches, but given that many different TCPs were identified as interacting with fragments of the *ORA59* promoter, it is unlikely that these interactions occurred spuriously.

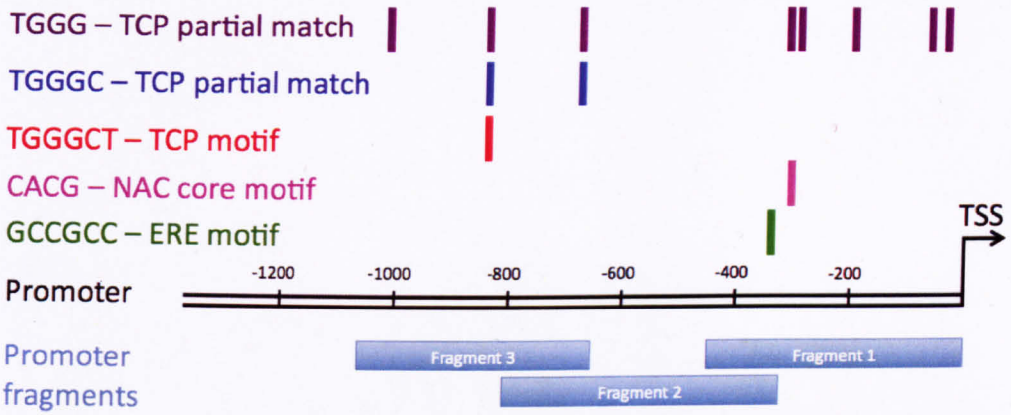


Figure 3.8: The *ORA59* promoter, with TF binding motifs and promoter fragments displayed. Promoter fragments are those screened by Y1H in this chapter. TCP (Giraud et al., 2010), NAC core (Tran et al., 2004) and ERE (Ohme-Takagi and Shinshi, 1995; Badis et al., 2009) motif matches are shown. Additionally, partial matches of the TCP binding motif TGGGC[C/T] (Giraud et al., 2010), are shown.

In these Y1H screens many AP2-ERE BPs, including ORA59 itself, have been seen interacting with fragments of the *ORA59* promoter in yeast. The TFs ABI4, ORA59, ERF1, ERF4 , ERF15 and AT2G23230 were seen to interact with fragments of the *ORA59* promoter in at least two independent screens, many in all three. This, and the fact that many of the Y1H interactors belong to the same TF families suggest that these interactors did not arise spuriously. The presence of only one ERE suggests the likely location of the binding site for these AP2-ERE BPs. TCP motifs were not found in all relevant promoter fragments, but the degree of specificity of TCP TFs is not known.

ARF2

ARF2 has been shown to be down-regulated in response to *B. cinerea* infection (Section 2.2.1). However, no direct transcriptional regulators of *ARF2* expression are

known in the literature. Previously, Laura Butler had cloned a promoter fragment of *ARF2*, and Phillip Law had shown that ORA59, ERF15 and ERF1 could interact with this fragment in yeast, by performing a pooled library Y1H screen.

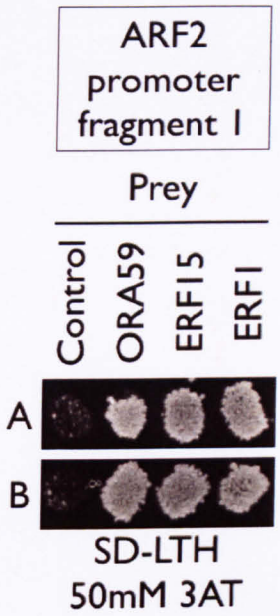


Figure 3.9: ORA59, ERF15 and ERF1 can interact with the promoter of *ARF2* in yeast. The screen photo shows an individual Y1H screen, of *ARF2* promoter fragment 1, by mating and auxotrophic selection. Preys and selections were as shown. Two replicates, A and B, were used to control for experimental variability. ORA59, ERF15 and ERF1 were shown to interact with *ARF2* promoter fragment 1.

Retested Y1H by individual mating Individual Y1H screens were performed by mating to retest the interaction of ORA59, ERF15 and ERF1 with *ARF2* promoter fragment 1. The results of the individual Y1H by mating for *ARF2* fragment 1 are shown in Figure 3.9. ORA59, ERF15 and ERF1 were again shown to be able to bind to *ARF2* promoter fragment 1. This was demonstrated with a selection of SD-LTH with 50 mM 3AT. Autoactivation of the control transformed yeast meant that no additional growth was seen beyond the control with SD-LTH and SD-LTH 5 mM 3AT selections. Intermediate growth of the control transformed yeast was seen with a selection of 25 mM 3AT. This is indicative of strong autoactivation of this promoter-reporter, but the difference between the control and other colonies at 50 mM 3AT are clear.

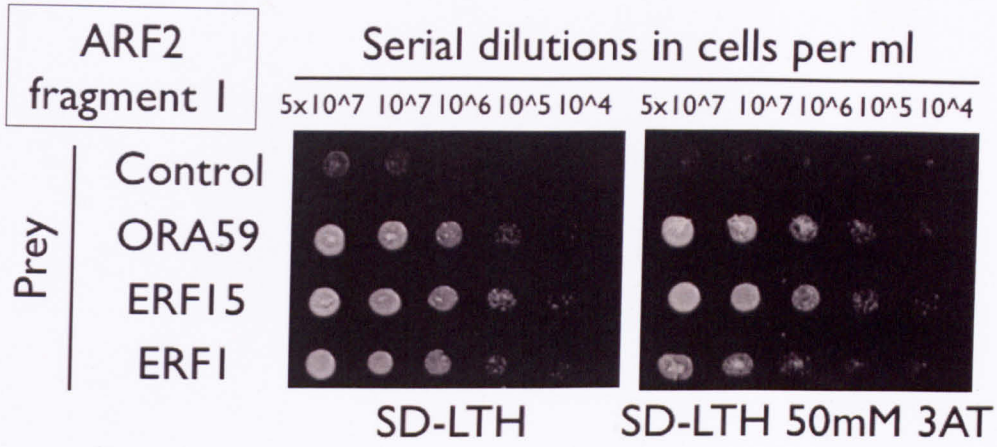


Figure 3.10: ORA59, ERF15 and ERF1 can interact with the promoter of *ARF2* in yeast. Photos are of individual Y1H screens, of *ARF2* promoter fragment 1, by co-transformation and auxotrophic selection. ORA59, ERF15 and ERF1 were shown to interact with *ARF2* promoter fragment 1, even with a selection that included 50mM 3AT.

Retested Y1H by co-transformation Individual Y1H screens were performed by co-transformation to retest the interaction of ORA59, ERF15 and ERF1 with *ARF2* promoter fragment 1. The results of the individual Y1H screens by co-transformation for *ARF2* promoter fragment 1 are shown in Figure 3.10. ORA59, ERF15 and ERF1 were shown to interact with *ARF2* promoter fragment 1. Co-transformed yeast on intermediate selections, SD-LTH 5 mM 3AT and SD-LTH 25 mM 3AT, showed intermediate growth.

The Y1H results presented in this section confirms the ability of ORA59, ERF15 and ERF1 to bind to a promoter fragment of *ARF2* in yeast. These interactions have been observed in the original pooled library Y1H screen by Phillip Law, as well as both the individual mating and co-transformation screens presented here. Because these results have been shown in three independent Y1H screens, they are unlikely to be spurious. However, no ERE motif was found in the *ARF2* promoter 3.11. This is odd considering that the ERE has been fairly well characterised by a recent protein binding microarray study of the ERF1 DNA-binding specificity (Godoy et al., 2011).

PGIP1

As well as the transcriptional regulation of TFs it would be desirable to be able to model the regulators of physiological outputs controlling susceptibility of Arabidopsis to infection by *B. cinerea*. This would allow gene regulation to be linked

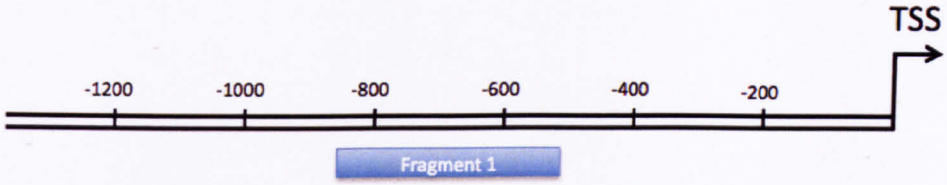


Figure 3.11: The *ARF2* promoter, with promoter fragment 1 displayed. Promoter fragment 1 was used in the Y1H screens.

to resistance mechanisms. For example, over-expression of *PGIP1* leads to reduced susceptibility of Arabidopsis to infection by *B. cinerea* (Ferrari et al., 2003b). This is believed to be due to the ability of the protein it encodes to inhibit fungal poly-galacturonases which would otherwise harm plant cells. *PGIP1* is up-regulated in response to *B. cinerea* infection and treatment with the OGs, which act as DAMPs (Ferrari et al., 2003b). Additionally, *PGIP1* is down-regulated by *ARF2* (Vert et al., 2008) and *TGA3* (Windram, 2010), which suggests they have some role in the regulation of *PGIP1* expression.

Pooled library Y1H To see if these roles are direct and to identify other potential transcriptional regulators of *PGIP1*, cloned library Y1H was performed against all four library plates. Three promoter fragments were used, each approximately 400 bp long, covering the 1,000 bp upstream of the TSS of *PGIP1*. The TFs that were found to interact with fragments of the *PGIP1* promoter are summarised in Table 3.8.

Table 3.8: Pooled TF library Y1H screen by mating and auxotrophic selection. Interactors of *PGIP1* promoter fragments 1–4 are shown. TFs highlighted in red were found in screens of both arrangements of the library.

Fragment number	Interactors
1	None
2	TCP3, TCP14, TCP15, TCP16, At1g35560 (TCP)
3	BHLH100, TCP3, TCP14, TCP15, TCP16, TCP20, At1g35560 (TCP)

Again, many TCPs were found to interact with the promoter fragments in yeast. Also again, an interactor was identified that had not been seen to interact with any other promoters in the Y1H screens presented in this chapter. This interactor, the

only non-TCP interactor of promoter fragments of *PGIP1* identified, was BHLH100, which has no known role in the defence response.

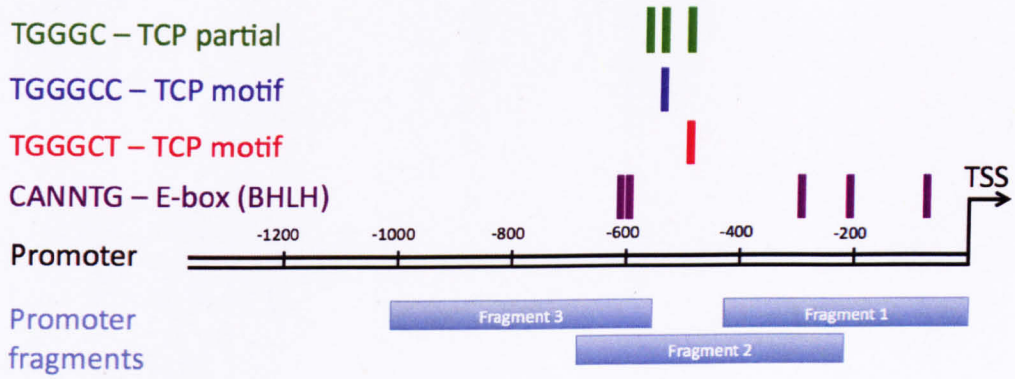


Figure 3.12: The *PGIP1* promoter, with TF binding motifs and promoter fragments displayed. Promoter fragments are those screened by Y1H in this chapter. Matches to the BHLH motif CANNTG, where N's can be any nucleotide, are shown (Toledo-Ortiz et al., 2003). Additionally, full and partial matches of the TCP binding motif TGGGC[C/T] (Giraud et al., 2010), are shown.

Validation of Y1H results by motif analysis In Figure 3.12 TCP motifs are found in promoter fragment 2 which was found to interact with TCP TFs in the Y1H screen (Table 3.8). Fragment 3 was also found to interact with TCP TFs in the Y1H screen, but only partial TCP motif matches are found in it. BHLH TFs are known to bind to the E-box, CANNTG (Toledo-Ortiz et al., 2003), and this motif is found to occur in all fragments, including fragment 3 which is the fragment that was found to interact with BHLH100 in the pooled library Y1H screen.

In summary, TCP TFs and BHLH100 were found to interact with the promoter of *PGIP1* in a Y1H screen. Matches to the full TCP binding motif were found in 2/3 of the fragments found to interact with TCP TFs in Y1H. Two E-box motif matches were found in the fragment that interacted with BHLH100 in the Y1H screen. This adds plausibility and provides potential binding sites for the interactors identified by cloned library Y1H.

LACS2

Another potential physiological output is *LACS2*, whose expression affects cuticle permeability and susceptibility to *B. cinerea* (Bessire et al., 2007). It was down-regulated during infection by *B. cinerea* (Section 2.2.1), but nothing seems to be known about its indirect or direct transcriptional regulation in any context.

Pooled library Y1H Y1H was used to identify potential regulators of *LACS2* expression. Y1H was performed on all three fragments against all four library plates. The TFs that were found to interact with fragments of the *LACS2* promoter are summarised in Table 3.9. Some TCPs, two AP2-ERE BPs and MYB49 were found to bind to the promoter of *LACS2*. MYB49 is known to be up-regulated by ABA (Yanhui et al., 2006), which is involved in resistance to *B. cinerea* (Audenaert et al., 2002) and regulation of cuticle permeability (Curvers et al., 2010).

Table 3.9: Pooled library Y1H by mating and auxotrophic selection. Interactors of *LACS2* promoter fragments 1–3 are shown. No interactors were found with both arrangements of the library.

Fragment number	Interactors
1	MYB49
2	ABI4, At5g21960 (ERF), TCP1, TCP3, TCP14 TCP15, TCP16, At1g35560 (TCP)
3	TCP16

Validation of Y1H results by motif analysis In Figure 3.13 it can be seen that three MYB binding motifs were found in fragment 1, the fragment found to interact with MYB49 in a Y1H screen. Additionally a TCP binding motif is found in fragment 3, which has been found to interact with TCP16 in a Y1H screen. Although no TCP binding motifs are found in fragment 2, which also interacted with TCP TFs in a Y1H screen, partial motif matches are found. No EREs, the known binding motifs of AP2-ERE BPs are found, and therefore a possible location for the binding site corresponding to the two AP2-ERE BPs interacting with fragment 2 (ABI4 and At5g21960) has not been identified.

In summary, TCP TFs, two AP2-ERE BPs and *MYB49* were found to interact with the promoter of *LACS2* in Y1H screens. Again, plausible binding sites are found for most of the TFs found to interact with the promoter fragments in Y1H. TCP TFs were found to interact with fragments of the promoters of all 4 genes tested by cloned library Y1H in this chapter, suggesting that their binding is fairly ubiquitous.

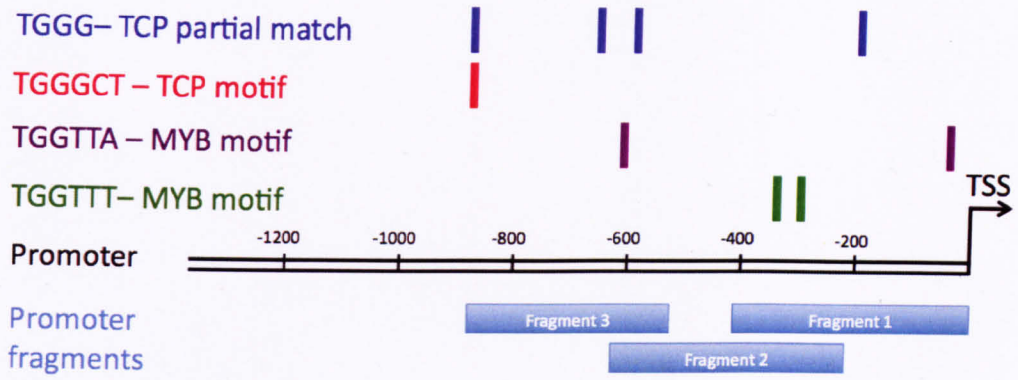
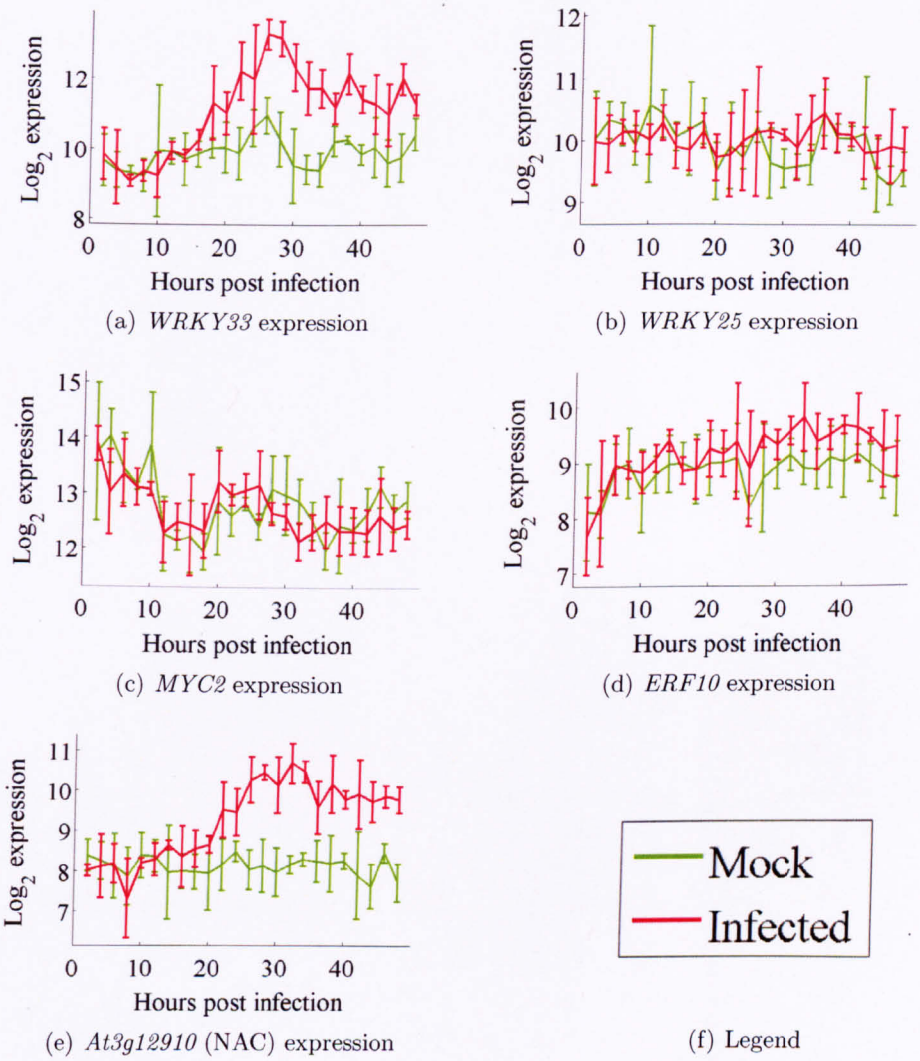


Figure 3.13: The *LACS2* promoter, with TF binding motifs and promoter fragments displayed. Promoter fragments are those screened by Y1H in this chapter. Matches of the MYB binding motif [A/T]AACCA (Abe et al., 2003), are shown. Additionally, full and partial matches of the TCP binding motif TGGGC[C/T] (Giraud et al., 2010), are shown.

Comparison of expression profiles of TF-target pairs

Y1H has generated lists of potential regulators of *ARF2, LACS2, ORA59, PGIP1* and *WRKY33*, but for the reasons discussed in Section 3.1.2 these interactions are 'context free' with respect to Arabidopsis. Given the expression profiles of potential TF-target pairs during *B. cinerea* infection, it may be possible to identify which targets are most likely to be regulated in that context. Here, the expression profiles of TF-target pairs are plotted and compared.

The expression profiles during *B. cinerea* infection of *WRKY33* and TFs that have been found to interact with its promoter in the Y1H screen are shown in Figures 3.14–3.15. Because *WRKY33* is differentially expressed earlier than its potential transcriptional regulators, it is hard to infer anything from their expression. It is possible, given the interaction of WRKY33 with it's own promoter, that the early differential expression of *WRKY33* is caused by WRKY33 itself, after activation by a signalling pathway. A candidate would be a MAPK pathway as WRKY33 is known to bind the MAPK substrate MKS1 and can be phosphorylated by MPK3, 4, or 6 (Andreasson et al., 2005; Qiu et al., 2008; Mao et al., 2011). Additionally, because *MYC2* is not differentially expressed, if it regulates *WRKY33* in this context then it must itself be regulated post-transcriptionally, which is plausible given its known interaction with the JAZ co-factors (Chini et al., 2007), many of which are differentially expressed during *B. cinerea* infection (Section 2.2.1). *WRKY25* is not found to be differentially expressed in this dataset; however it has been shown to be differentially expressed during *B. cinerea* infection in other experiments (AbuQamar



(a) *WRKY33* expression

(b) *WRKY25* expression

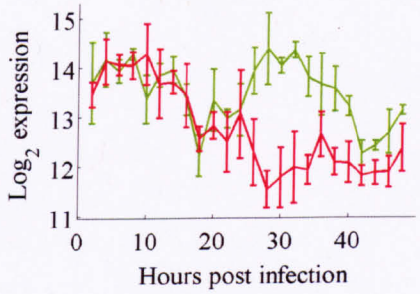
(c) *MYC2* expression

(d) *ERF10* expression

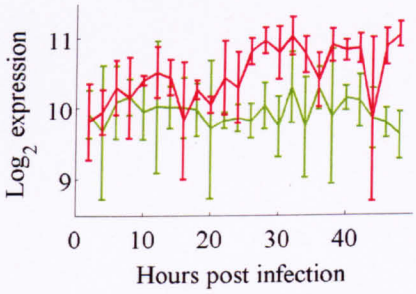
(e) *At3g12910* (NAC) expression

(f) Legend

Figure 3.14: Expression of Y1H interactors of *WRKY33* in Arabidopsis leaves during infection by *B. cinerea*. (Expression data from the experiment introduced in Section 2.2.1. Experiment will be published in Denby et al., manuscript in preparation, and is also presented in Windram, 2010). Lines show the mean expression profile, while bars represent standard deviations. Of these genes, only *WRKY33* and *At3g12910* were found to be differentially expressed in this time series (Section 2.2.1).

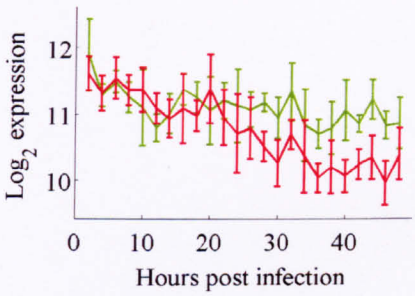


(a) *TCP3* expression

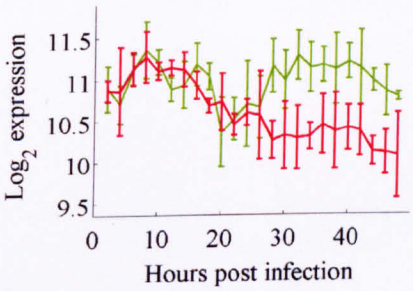


(b) *TCP8* expression

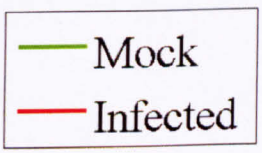
t



(c) *TCP15* expression



(d) *TCP20* expression



(e) Legend

Figure 3.15: Profiles of differentially expressed common Y1H interactors in Arabidopsis leaves during infection by *B. cinerea*. (Expression data from the experiment introduced in Section 2.2.1. Experiment will be published in Denby et al., manuscript in preparation, and is also presented in Windram, 2010). Lines show the mean expression profile, while bars represent standard deviations.

et al., 2006). In those experiments it is possible that WRKY25 may be responsible for regulating the expression of *WRKY33*.

The expression profiles during *B. cinerea* infection of *ORA59* and TFs that have been found to interact with its promoter in the Y1H screen are shown in Figures 3.15–3.17. Many of the AP2-ERE BPs are differentially expressed at the same time as *ORA59*, around 18-20 hpi, but none before it. This makes it hard to use the expression profiles to suggest which of the promoter interactors are the most plausible regulators of *ORA59* expression during *B. cinerea* infection.

The expression of *ARF2* and the three TFs that have been shown here to bind to a fragment of the *ARF2* promoter in yeast, during infection by *B. cinerea* are shown in Figures 3.16(a)-(d). All the potential interactors are up-regulated at around 18-20 hpi, whereas *ARF2* is down-regulated at around 30 hpi. This means that all potential interactors are equally plausible regulators of *ARF2* expression.

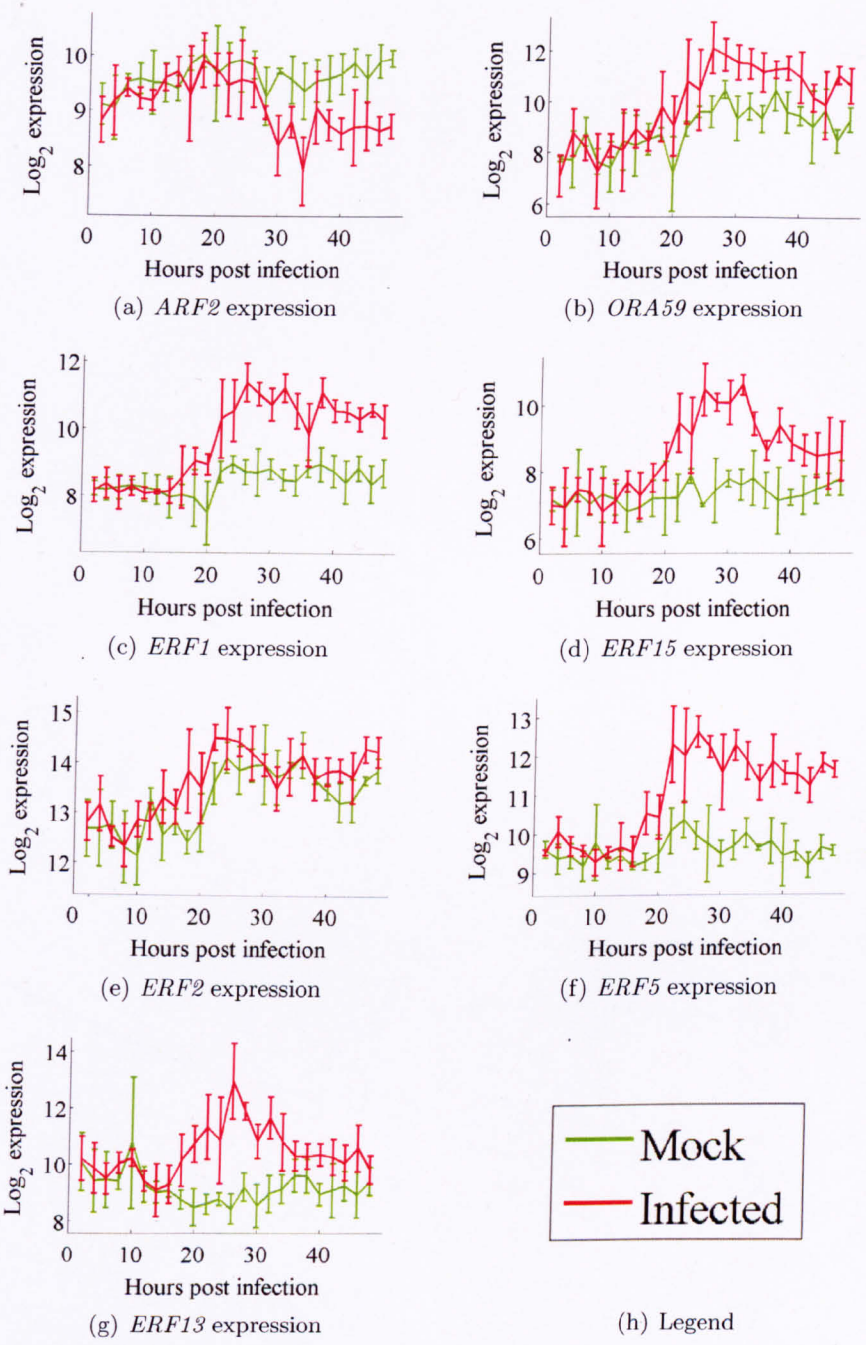
The expression profiles during *B. cinerea* infection of *PGIP1* and TFs that have been found to interact with its promoter in the Y1H screen are shown in Figures 3.18 and 3.15. The differentially expressed interacting TCPs appear to be differentially expressed slightly earlier than *PGIP1*. This makes the differentially expressed TCP TFs plausible transcriptional regulators of *PGIP1*. While *BHLH100* is not differentially expressed, post-transcriptional activation can't be ruled out.

The expression profiles during *B. cinerea* infection of *LACS2* and TFs that have been found to interact with its promoter in the Y1H screen are shown in Figures 3.19 and 3.15. *At5g21960* is not differentially expressed, but as with *BHLH100*, post-transcriptional activation can't be ruled out. *MYB49* is up-regulated slightly earlier than *LACS2* is down-regulated, which means that it is possible that MYB49 represses the transcription of *LACS2*.

Overall it seems hard to assign context to these potential regulators based on their gene expression profiles during infection by *B. cinerea*, although they can be used to suggest whether transcriptional regulation may be regulated post-transcriptionally if it is known to regulate the target in that specific context.

3.3.3 In planta validation of the qualitative model by transient transactivation assays

While Y1H can identify TFs that are able to bind to a given promoter, it cannot reveal the effect these TFs have on the expression of the gene linked to that promoter



(a) *ARF2* expression

(b) *ORA59* expression

(c) *ERF1* expression

(d) *ERF15* expression

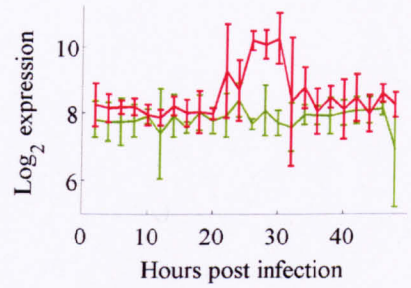
(e) *ERF2* expression

(f) *ERF5* expression

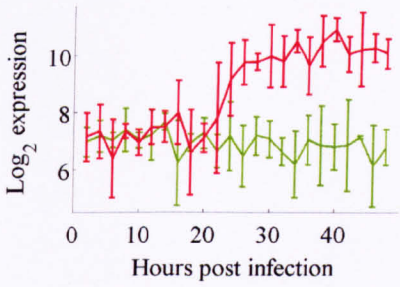
(g) *ERF13* expression

(h) Legend

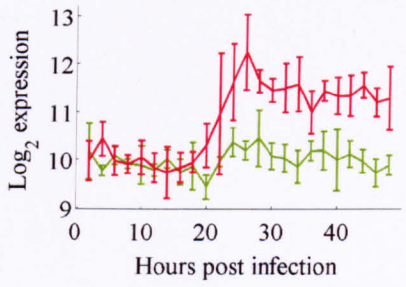
Figure 3.16: Expression of genes in mock and B. cinerea infected leaves. Expression profiles of: (a) *ARF2*, (b) *ORA59*; (b)-(d) Y1H interactors of *ARF2*; (b)-(g) Y1H interactors of *ORA59* in Arabidopsis leaves during infection by *B. cinerea*. (Expression data from the experiment introduced in Section 2.2.1. Experiment will be published in Denby et al., manuscript in preparation, and is also presented in Windram, 2010). Lines show the mean expression profile, while bars represent standard deviations.



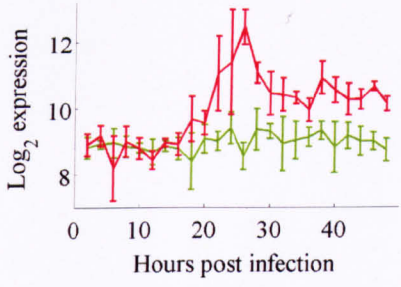
(a) *ERF14* expression



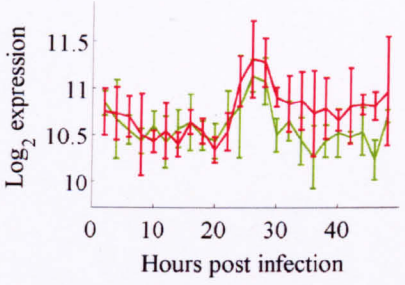
(b) *ERF71* expression



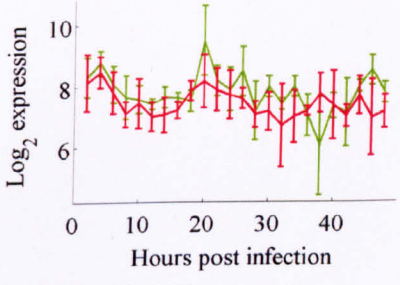
(c) *At3g23220* (ERF) expression



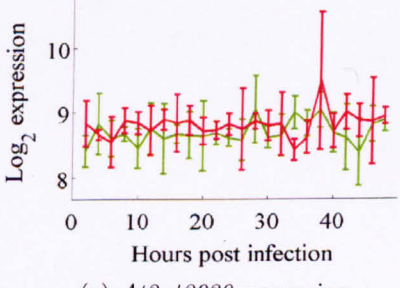
(d) *AT3G23230* (ERF) expression



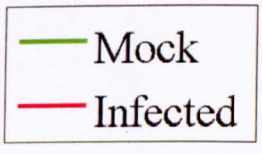
(e) *ABI4* expression



(f) *PIF7* expression

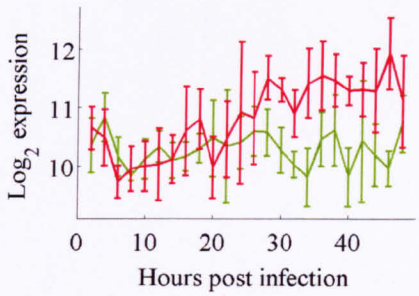


(g) *At2g42920* expression

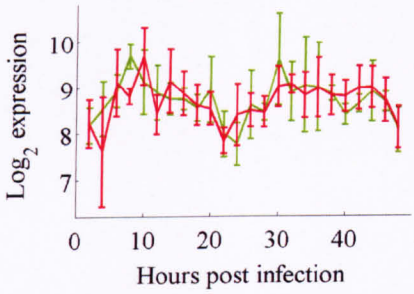


(h) Legend

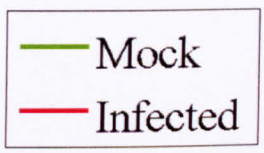
Figure 3.17: Expression of Y1H interactors of *ORA59* in Arabidopsis leaves during infection by *B. cinerea*. (Expression data from the experiment introduced in Section 2.2.1. Experiment will be published in Denby et al., manuscript in preparation, and is also presented in Windram, 2010). Lines show the mean expression profile, while bars represent standard deviations.



(a) *PGIP1* expression

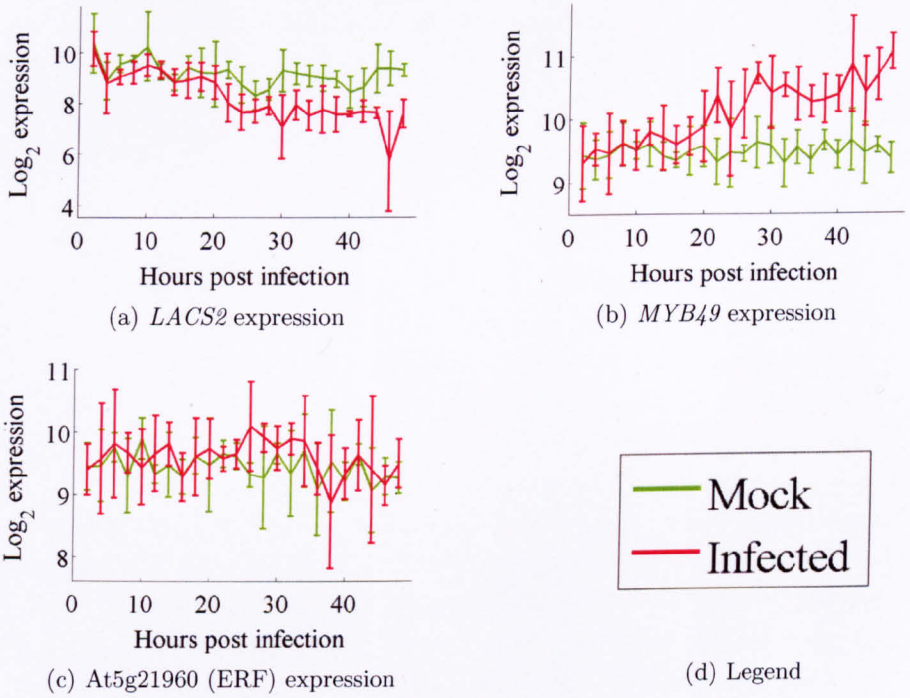


(b) *BHLH100* expression



(c) Legend

Figure 3.18: Expression of *PGIP1* and Y1H interactors of *PGIP1* in Arabidopsis leaves during infection by *B. cinerea*. (Expression data from the experiment introduced in Section 2.2.1. Experiment will be published in Denby et al., manuscript in preparation, and is also presented in Windram, 2010). Lines show the mean expression profile, while bars represent standard deviations.



(a) *LACS2* expression

(b) *MYB49* expression

(c) At5g21960 (ERF) expression

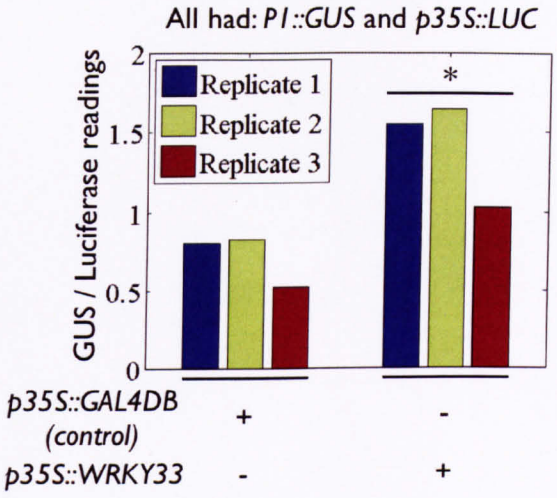
(d) Legend

Figure 3.19: Expression of *LACS2* and Y1H interactors of *LACS2* in Arabidopsis leaves during infection by *B. cinerea*. (Expression data from the experiment introduced in Section 2.2.1. Experiment will be published in Denby et al., manuscript in preparation, and is also presented in Windram, 2010). Lines show the mean expression profile, while bars represent standard deviations.

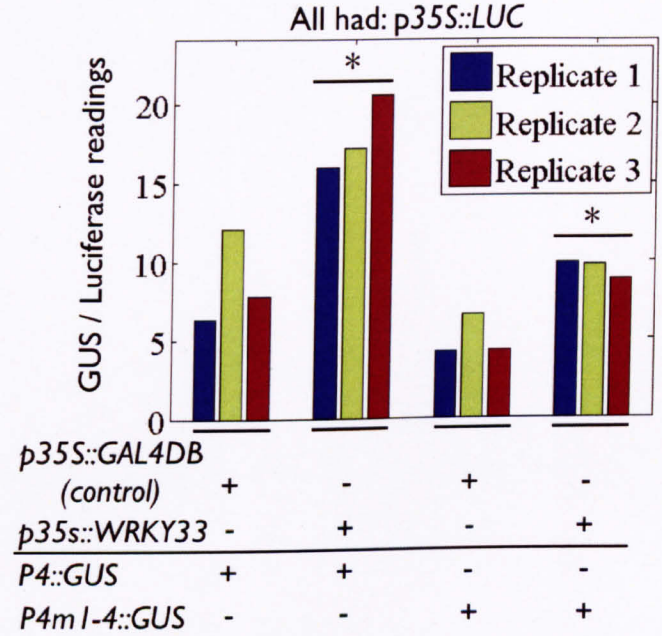
or whether the TF-promoter pair can interact *in planta*. For example, TFs that can regulate *WRKY33* expression have been identified from the literature (*ARF2* and *MYC2*), additionally Y1H showed that WRKY25 and WRKY33 can bind to the promoter *WRKY33* in yeast. To test the ability of these TFs to both bind to the promoter and regulate the expression of *WRKY33 in planta*, a transactivation assay was performed with promoter-reporter constructs that either contained or did not contain the predicted binding sites of these TFs. The location of the promoter fragments, P1 and P4, used in this screen are shown in Figure 3.5. An additional fragment, P4m1-4, was used, which was identical to P4 except that all WRKY boxes had been mutated. The reporter GUS was used in the promoter-reporter constructs. The construction of the *P1::GUS*, *P4::GUS* and *P4m1-4::GUS* constructs is detailed in Lippok et al. (2007). Leaves were transiently transformed by biolistic delivery of plasmids, to over-express a TF of interest and to introduce one of these promoter-reporter fusions.

WRKY33 was found to be capable of activating its own expression approximately two-fold *in planta*, e.g. Figures 3.20(a)–(b). This fits with the recent finding that WRKY33 is a transcriptional activator (Lai et al., 2011a). In the experiment shown in Figure 3.20(a) the difference, between the measurements made on samples transformed with the control plasmid versus *p35S::WRKY33* (*WRKY33* over-expressor) plasmid, was found to be significant at the 5% level in a t-test (p=0.0495). In the experiment shown in Figure 3.20(b) the difference, between the measurements made on samples transformed with the control plasmid versus *p35S::WRKY33* (*WRKY33* over-expressor) plasmid, was found to be significant at the 5% level in a t-test (p=0.0156). Strangely, a similar result was seen when a promoter-reporter plasmid was used which only differed at the sites of the first four WRKY-boxes, which are shown in Figure 3.5. WRKY TFs are believed to bind to sites whose sequence matches the WRKY-binding motif (de Pater et al., 1996) and the four WRKY-boxes in the first 250 bp have been shown to be important for *WRKY33* expression (Lippok et al., 2007). With mutated WRKY-boxes (*P4m1-4::GUS*) the difference, between the measurements made on samples transformed with the control plasmid versus *p35S::WRKY33* (WRKY33 over-expressor), was found to be significant at the 5% level in a t-test (p=0.0175). This is counterintuitive, as WRKY TFs are believed to act through WRKY-boxes, so it would be expected that the ability of WRKY TFs to regulate expression would depend on the presence of WRKY-boxes.

To validate this counterintuitive result, as well as observing the effect of other potential regulators, a further transactivation assay was performed. The results for this assay are shown in Figure 3.21, comparing the ability of ARF2, MYC2, WRKY25



(a) Effect of *WRKY33* over-expression on P1::GUS



(b) Effect of *WRKY33* over-expression on P4::GUS and P4m1-4::GUS

Figure 3.20: WRKY33 was shown to increase expression of a reporter, GUS, fused to fragments of the *WRKY33* promoter. Leaves were transformed with three plasmids; *35S::LUC*, an over-expressor of a given TF (or GAL4 DB which acts as a negative control) and a promoter-reporter (promoter::GUS) plasmid. The construction of the promoter-reporter plasmids, *P1::GUS*, *P4::GUS* and *P4m1-4::GUS*, is detailed in Lippok et al. (2007). Luciferase readings were used as a transformation control. Results were compared to controls with the same promoter-reporter, by a two-tailed unequal variance two-sample t-test, results that were significant at the 5% level are indicated by a star (*). (a) WRKY33 was shown to increase expression of GUS from *P1::GUS*. (b) WRKY33 was shown to increase expression of GUS from both *P4::GUS* and *P4m1-4::GUS*.

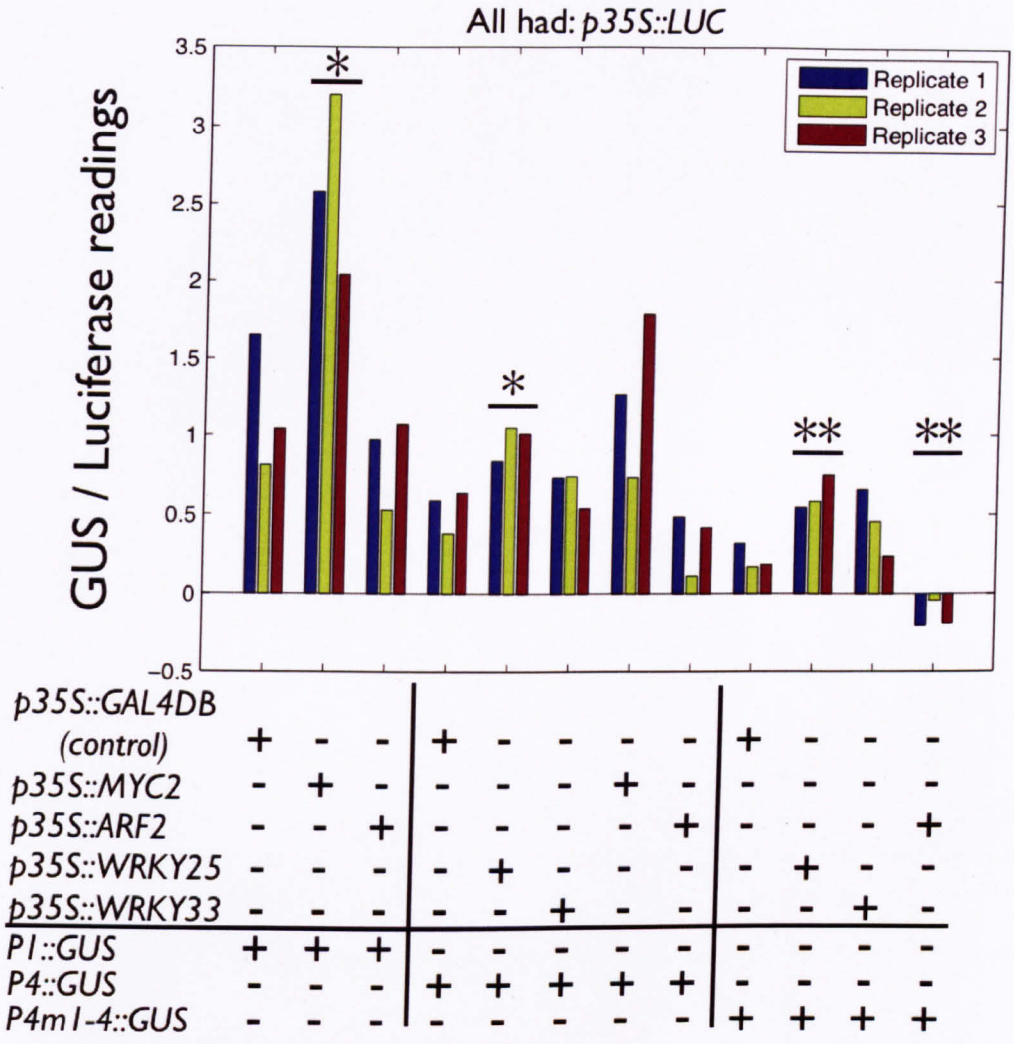


Figure 3.21: WRKY25 and MYC2 were shown to increase, and ARF2 was shown to decrease, expression of a reporter, GUS, fused to *WRKY33* promoter. Leaves were transformed with three plasmids; *35S::LUC*, an over-expressor of a given TF (or GAL4 DB which acts as a negative control) and a promoter-reporter (promoter::GUS) plasmid. The construction of the promoter-reporter plasmids, *P1::GUS*, *P4::GUS* and *P4m1-4::GUS*, is detailed in Lippok et al. (2007). Luciferase readings were used as a transformation control. Results were compared to controls with the same promoter-reporter, by a two-tailed unequal variance two-sample t-test, results that were significant at the 5% or 1% level are indicated by a star (*) or two stars (**) respectively.

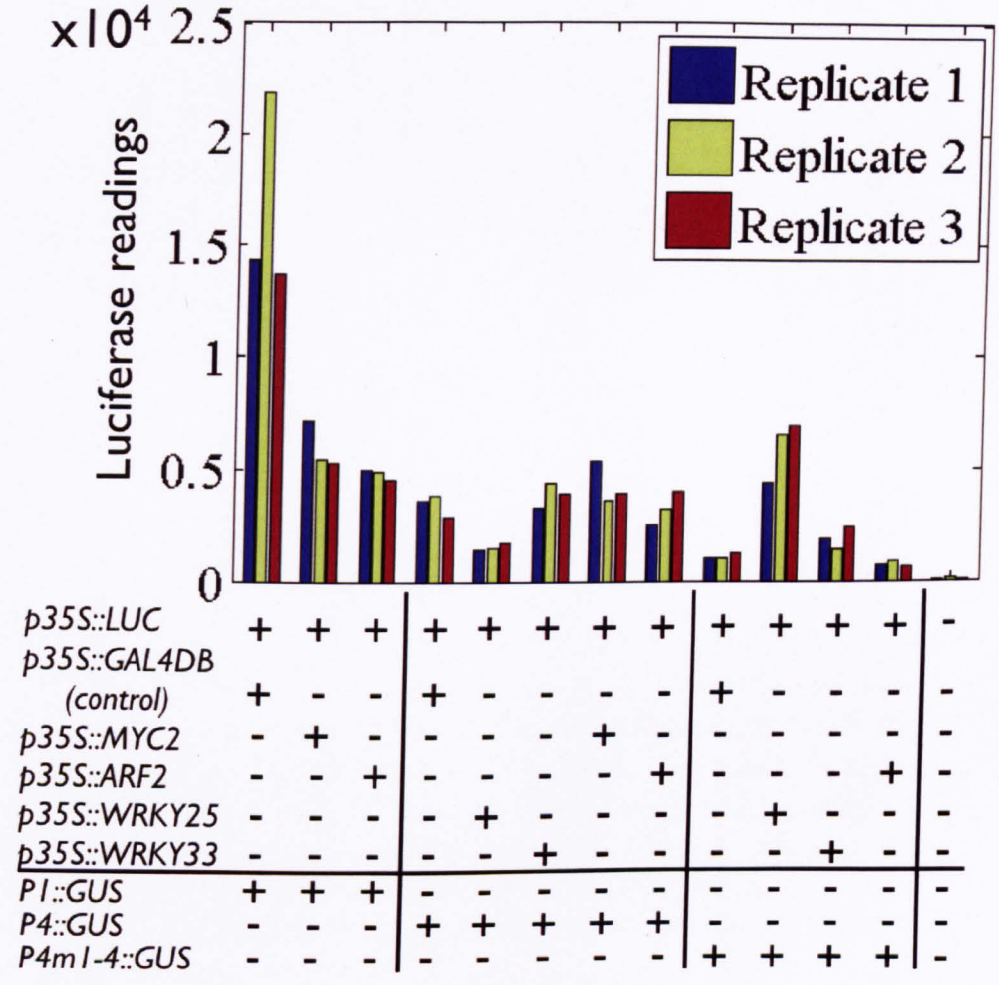


Figure 3.22: Technical variability in the transactivation experiment shown in Figure 3.21. Leaves were transformed with three plasmids; *35S::LUC*, an over-expressor of a given TF (or GAL4 DB which acts as a negative control) and a promoter-reporter (promoter::GUS) plasmid. The construction of the promoter-reporter plasmids, *P1::GUS*, *P4::GUS* and *P4m1-4::GUS*, is detailed in Lippok et al. (2007). Here the transformation control, Luciferase, measurement is shown to demonstrate technical variability.

and WRKY33 to alter expression of a reporter fused to segments of the *WRKY33* promoter. In this assay WRKY33 was not shown to activate expression of the reporter. Technical variability of the assay shown in Figure 3.21 is plotted in Figure 3.22. This demonstrates that over-expression of TFs is not causing systematic reductions in production of reporters.

MYC2 was shown to increase the expression of GUS approximately two-fold, from a *P1::GUS* promoter-reporter fusion. This fits with the known role of MYC2 as a transcriptional activator, as MYC2 is able to activate the expression of a reporter fused to the promoter of *RD22* in protoplasts (Abe et al., 2003). The difference between the measurements made on samples transformed with the control plasmid versus *p35S::MYC2* (MYC2 over-expressor) plasmid was found to be significant at the 5% level in a t-test (p=0.0296). While a significant difference was not seen between control and *p35S::MYC2* transformed samples, co-transformed with the *P4::GUS* promoter-reporter plasmid, the median expression is raised approximately two-fold in the *p35S::MYC2* transformed samples.

WRKY25 was shown to increase the expression of GUS from *P4::GUS* and *P4m1-4::GUS* promoter-reporter fusions. This counterintuitive result is similar to that seen for *WRKY33* in Figure 3.20(b), i.e. mutation of the suspected binding site of WRKY25 or WRKY33 does not appear to abolish their ability to activate the expression of GUS. The difference between the measurements made on samples transformed with the control plasmid versus the *p35S::WRKY25* plasmid (WRKY25 over-expressor), each co-transformed with the promoter-reporter plasmid *P4::GUS*, was found to be significant at the 5% level in a t-test (p=0.0142). The difference between the measurements made on samples transformed with the control plasmid versus the *p35S::WRKY25* plasmid (WRKY25 over-expressor), each co-transformed with the promoter-reporter plasmid *P4m1-4::GUS*, was found to be significant at the 1% level in a t-test (p=0.0083).

ARF2 over-expression (using *p35S::ARF2*) was shown to decrease expression of GUS from the *P4m1-4::GUS* plasmid, this difference was found to be significant at the 1% level in a t-test (p=0.0066). A negative GUS/Luciferase reading resulted from GUS readings of less than background, presumably because of technical variability, i.e. less fluorescence corresponding to GUS was detected in wells loaded with protein extracts from samples transformed with *p35S::ARF2* and *P4m1-4::GUS* then was detected in a well loaded with protein extracts from untransformed leaves, this indicates that next to zero GUS was produced in the sample transformed with *p35S::ARF2*. This suggests that ARF2 is repressing the expression of the GUS re-

porter.

In summary WRKY25, WRKY33 and MYC2 have been found to positively, whereas ARF2 has been found to negatively, regulate the expression of a reporter fused to the promoter of *WRKY33*. WRKY33 was found to up-regulate the reporter in two of the three assays, which suggests it can regulate its own expression. This is consistent with the finding that it can activate expression (Lai et al., 2011a) and binds to the *WRKY33* promoter *in planta* (Mao et al., 2011). The repression of the reporter by ARF2 was affected by the presence of the four WRKY-boxes mutated in *P4m1-4::GUS*, suggesting some link between regulation by ARF2 and WRKY TFs.

Transient transactivation assays have proved useful in testing the ability of TFs to regulate expression *in planta*. This is complementary to Y1H, which provides candidate regulators but does not by itself reveal the regulatory effect of the TFs interacting with promoter fragments.

3.3.4 Analysis of context-dependence by comparative transcriptomics

Y1H has revealed that some regulation known in the literature is likely to be direct, for example MYC2 regulating *WRKY33*. Transient transactivation assays have been used to show the effect of TFs on the expression of a reporter fused to the promoter of *WRKY33*. This suggests that the general approach of this chapter is allowing direct transcriptional regulators to be revealed. However, none of this regulation has been shown to occur during *B. cinerea* infection. In this section context-dependence of transcriptional regulation is studied by comparative transcriptomics. The aim is to reveal the elictors, signalling pathways and regulators responsible for regulation of the defence response of Arabidopsis to infection by *B. cinerea*.

Overlaps between genes differentially expressed during *B. cinerea* infection and genes differentially expressed in conditions relating to elictors, hormones and regulators, were analysed using the cumulative hypergeometric distribution against the null hypothesis that the overlaps had occurred by chance (in MATLAB®). The results are summarised in Table 3.10. (Hypergeometric parameters: 23,802 the number of unique genes which have a CATMA probe and 9,838 the number of genes differentially expressed during *B. cinerea* infection).

Reassuringly, the overlap with the lowest p-value was from a list of genes found to be differentially expressed at 48 hpi by *B. cinerea* (Ferrari et al., 2007) In fact, overlaps with p-values less than 10^{-10} were found with all lists of genes differentially

Table 3.10: Comparative transcriptomics, comparing the overlaps between lists of genes differentially expressed in different conditions and genes differentially expressed during *B. cinerea* infection (Section 2.2.1). The line shows a Bonferroni corrected 5% significance threshold. All lists originating from TAIR are lists of genes corresponding to the GO term given (Ashburner et al., 2000; Swarbreck et al., 2008). Lists in black correspond to *B. cinerea* infection experiments. Lists in blue correspond to TF mutant experiments. Lists in red correspond to elictor (PAMP/DAMP) treatment experiments. Lists in magenta correspond to hormone treatment experiments or GO lists. Cumulative hypergeometric p-values test the null hypothesis that the overlap size occurred by chance.

Gene list	Source of list	Number of genes in list	Number also differentially expressed in Botrytis time series	Cumulative hypergeometric p-value
Botrytis 48 hpi	Ferrari et al. (2007)	4,881	3,031	1.37×10^{-236}
TGA3 knockout	Windram (2010)	1,855	1,232	5.86×10^{-114}
Botrytis 24 hpi	Mulema and Denby (2012)	2,437	1,363	4.40×10^{-53}
MeJA treatment	Dombrecht et al. (2007)	3,497	1,809	2.75×10^{-41}
OG 1 hour	Ferrari et al. (2007)	1,827	1,019	2.09×10^{-38}
Botrytis 12 hpi	Mulema and Denby (2012)	984	564	4.92×10^{-25}
Botrytis 0–6 mm	Mulema and Denby (2012)	1,770	927	1.49×10^{-22}
Botrytis 18 hpi	Ferrari et al. (2007)	252	171	1.21×10^{-17}
Botrytis Induced Genes	AbuQamar et al. (2006)	457	270	9.54×10^{-15}
Botrytis 6–12 mm	Mulema and Denby (2012)	1,561	767	7.36×10^{-11}
ARF2 knockout	Vert et al. (2008)	895	456	2.10×10^{-9}
35S::ERF1	Lorenzo et al. (2003)	192	120	2.43×10^{-9}
OG 3 hour	Ferrari et al. (2007)	440	242	3.87×10^{-9}
ET and JA	Lorenzo et al. (2003)	146	84	5.51×10^{-5}
35S::ANAC072	Tran et al. (2004)	29	22	1.69×10^{-4}
Response to JA stimulus	TAIR	140	77	7.36×10^{-4}
CAMTA3 knockout	Galon et al. (2008)	103	56	5.06×10^{-3}
Response to ET stimulus	TAIR	169	86	7.41×10^{-3}
ORA59 inducible overexpressor	Pré et al. (2008)	86	46	1.51×10^{-2}
Response to SA stimulus	TAIR	127	64	2.40×10^{-2}
Chitin Induced TFs	Libault et al. (2007)	118	59	3.49×10^{-2}
ANAC072 targets	Fujita et al. (2004)	21	13	4.62×10^{-2}
ANAC055 knockout	Hickman et al., (in preparation)	514	226	0.119
MYC2 knockout	Dombrecht et al. (2007)	778	334	0.188
ANAC019 knockout	Hickman et al., (in preparation)	287	114	0.731
Response to auxin stimulus	TAIR	272	101	0.931
AUX/IAA, ARFs, or TIR/ABP	Llorente et al. (2008)	56	18	0.939
MYC2 knockout and MeJA responsive	Dombrecht et al. (2007)	375	140	0.950
35S::ANAC055	Tran et al. (2004)	14	1	0.999
35S::ANAC019	Tran et al. (2004)	17	1	0.999

expressed during *B. cinerea* infection experiments (AbuQamar et al., 2006; Ferrari et al., 2007; Mulema and Denby, 2012). This demonstrates that some genes respond similarly to *B. cinerea* infection under varying experimental conditions, which suggests that the results may be able to be extrapolated. An overlap with a low p-value, 2.09×10^{-38}, was found between a list of genes differentially expressed 1 hour after treatment with OGs (Ferrari et al., 2007) and the list of genes differentially expressed during *B. cinerea* infection. This suggests the importance of this elictor for activating *B. cinerea* responsive changes in gene expression.

The list of genes differentially expressed in a TF mutant, whose overlap with the list of genes differentially expressed during *B. cinerea* infection gave the lowest p-value, was from an experiment with a *TGA3* knockout. The *TGA3* knockout was performed during *B. cinerea* infection, see Windram (2010), which may at least partially account for its low p-value. The high significance of the overlap of genes differentially expressed in an *ARF2* knockout and during *B. cinerea* infection, is consistent with *ARF2* being a regulator of *B. cinerea* responsive gene expression. Similarly, the high significance overlap of the genes differentially expressed in an over-expressor of *ERF1* and during *B. cinerea* infection, is consistent with *ERF1* being a regulator of *B. cinerea* responsive gene expression. Although much less significant, the overlap of genes differentially expressed in an over-expressor of *ANAC072* and during *B. cinerea* infection, suggests that *ANAC072* is also a regulator of *B. cinerea* responsive gene expression, fitting with the novel altered susceptibility phenotype observed in a knockout of *ANAC072* in Chapter 2. The list of genes differentially expressed in mutants of *ARF2, ERF1* and *ANAC072* were not derived from experiments in which the samples were infected with *B. cinerea.* This is consistent with some of the regulation observed in these experiments also occurring during *B. cinerea* infection. This fits with the altered susceptibility of mutants of *ARF2, ERF1* and *ANAC072* to infection by *B. cinerea.*

Genes differentially expressed in response to treatment with MeJA (Methyl Jasmonate), JA, and both JA and ET have significant overlaps with the list of genes differentially expressed during *B. cinerea* infection. This fits with the known importance of JA and ET signalling in regulation gene expression in response to necrotrophic pathogens (Glazebrook, 2005; Pieterse et al., 2009).

In this section elicitors (OGs), hormones (JA and ET) and TFs (TGA3, ARF2, ERF1 and ANAC072) regulating gene expression during infection by *B. cinerea* have been predicted. The significant overlaps demonstrate that some transcriptional regulation can be extrapolated from one context to another, and suggests an

important but currently unknown role for *ARF2* in regulating the defence response. This fits with the recent finding that a knockout of *ARF2* has a substantial reduced susceptibility to infection by *B. cinerea* (Youn-Sung Kim et al., in preparation).

In this chapter a qualitative model of the defence response GRN has been developed, and some of the transcriptional regulation has subsequently been validated experimentally. For example, MYC2 was shown to bind to the promoter of, and to regulate the expression of a reporter fused to the promoter of *WRKY33*. Cloned TF library Y1H also revealed many other TFs that are capable of binding to the promoters of *ARF2, LACS2, ORA59, PGIP1* and *WRKY33*, many of which are known to be able to bind to sequence motifs that are present in the relevant promoter fragments. These novel interactors extend the qualitative model, providing more hypotheses about the structure of the defence response GRN. The effect of some of these regulators on the expression of a reporter fused to the promoter *WRKY33* were shown by transient transactivation assays. Finally, comparative transcriptomics showed that genes known to be downstream of the TFs TGA3, ARF2, ERF1 and ANAC072 are over-represented in the list of genes differentially expressed during *B. cinerea* infection, implicating them in the GRN controlling defence response gene expression during *B. cinerea* infection.

3.4 Discussion

3.4.1 Y1H predicts novel direct transcriptional regulators

In Section 3.3.2 Y1H was applied to test the ability of TFs to bind to the promoters of *ARF2, LACS2, ORA59, PGIP1* and *WRKY33*. This revealed many different TFs are able to interact with these promoters and many of these interactors have known binding motifs which were present in the promoter fragments they were able to bind to. In addition many related TFs were found to bind to similar fragments, suggesting that the results are not spurious. While some TFs, mostly belonging to the TCP TF family, were found to interact with many different promoter fragments, the majority were highly specific as can be seen in Figure 3.23. This provides rich 'context free' information about the local structure of the defence response GRN, confirming that known regulation can be direct and providing novel candidates for direct transcriptional regulators of *ARF2, LACS2, ORA59, PGIP1* and *WRKY33*.

For example, a knockout of *MYC2* has been shown to have higher expression of *WRKY33* (Dombrecht et al., 2007). Y1H revealed that MYC2 can bind to the promoter of *WRKY33*, suggesting that this regulation is direct. This novel finding could be validated *in planta* by ChIP, and could be linked to *B. cinerea* infection by

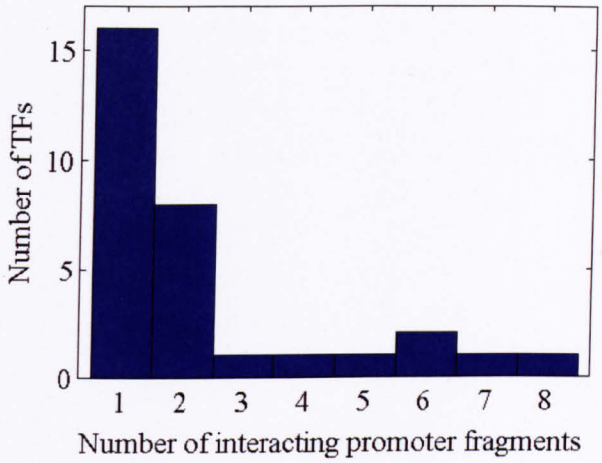


Figure 3.23: The specificity of the TF-promoter fragment interactions found by cloned TF library Y1H is shown in a histogram. The interactions can be seen in Tables 3.6–3.9.

ChIP on infected leaves.

Novel TF-promoter interactors revealed by Y1H provide candidate direct regulators, some of which make sense given the literature. For example, ABI4 and MYB49 were found to interact with the promoter of *LACS2* in a Y1H screen. These novel interactors are known to be linked with ABA signalling (Finkelstein et al., 1998; Yanhui et al., 2006), and ABA has been shown to affect the susceptibility of tomato to infection by *B. cinerea* (Audenaert et al., 2002). Additionally, both ABA treatment and mutations in *LACS2* have been shown to affect cuticle permeability, suggesting a role for ABA in the downregulation of *LACS2* during infection (Curvers et al., 2010; Bessire et al., 2007).

Because Y1H screens have been performed with approximately 400 bp promoter fragments, it is not possible to directly relate results to specific short DNA sequences. To relate Y1H interactors to specific short sequences in the promoter fragment, site directed mutagenesis can be used to remove that sequence from the *pHis2Leu2* plasmid.

In Section 3.3.2, the gene expression profiles of *ARF2, LACS2, PGIP1, ORA59* and *WRKY33*, and TFs capable of binding to their promoters, were visually compared. A different approach would be to model the expression of both the potential regulators and the potential target, in both mock and Botrytis-infected conditions, using modelling approaches such as the non-linear multiple time series approach re-

cently introduced by Penfold et al. (2012). This approach takes into account the fact that the underlying regulation in each time series may be independent or related, and also allows prior knowledge, such as Y1H interactors, to be taken incorporated into the model. This would have the added benefit that it would allow a more objective and non-trivial prediction of which potential regulators may be important for the regulation of *B. cinerea* responsive changes in the Arabidopsis transcriptome.

While the cloned TF library is not comprehensive, it does contain all the TFs in the qualitative model presented here except EIN3. In the future the usefulness of this library can improved by expanding the number of TFs contained within it.

3.4.2 Transient transactivation assays validate transcriptional regulation *in planta*

Cloned TF library Y1H provides many candidate transcriptional regulators for a gene, but the ability of these candidates to regulate its expression need to tested separately. Transient transactivation assays provide a relatively rapid method by which to test this, which has confirmed the role of some known regulators and characterised the effect of novel candidates on the expression of a reporter fused to the target genes promoter. For example a transient transactivation assay showed the ability of ARF2 to repress the expression of a reporter fused to the *WRKY33* promoter, confirming the ability of ARF2 to repress expression of *WRKY33* which had been shown previously shown by Vert et al. (2008) using a knockout versus wildtype microarray experiment.

In a knockout mutant of *MYC2*, *WRKY33* expression was found to be higher than in wildtype (Dombrecht et al., 2007). This seems to suggest that MYC2 represses *WRKY33* expression, and that the higher expression resulted from de-repression of *WRKY33* expression. However MYC2 appeared to activate expression of *WRKY33* in the transactivation assay presented in this chapter. Similarly, in another transient transactivation assay Abe et al. (2003) showed that MYC2 could activate expression of a reporter fused to the *RD22* promoter. These seemingly contradictory results, repression/activation of genes by one TF, make sense when it is considered that the literature suggests that repression by MYC2 is mediated by JAZ proteins that interact directly with it (Chini et al., 2007; Memelink, 2009; Pauwels et al., 2010). This leaves room for MYC2 to function as a transcriptional activator when mRNA levels of MYC2 are suitably high relative to that of the JAZ proteins, as would be expected in transient transformation assays where *MYC2* is expressed from a plasmid resulting in high levels. MYC2 could then act to repress gene expression when sufficient levels of JAZ proteins exist, by helping to anchor the direct and indirect interactors

of JAZ proteins, such as the co-repressor TOPLESS, close to the TSS. Because of this both *ARF2* and *MYC2* are candidates for the strong repressor of *WRKY33* expression suggested by Lippok et al. (2007), who showed that *WRKY33* expression increased rapidly after cycloheximide treatment. As *MYC2* has been found to bind to the *WRKY33* promoter it is currently the stronger candidate. This could be tested by repeating the cycloheximide experiment in a knockout of *MYC2*.

As well as confirming known transcriptional regulation, transient transactivation assays characterised the currently unknown effect of WRKY25 and WRKY33 on the expression of a reporter fused to the *WRKY33* promoter. Both WRKY25 and WRKY33 were found to activate expression of the reporter, suggesting that the binding of WRKY33 to the *WRKY33* promoter, observed in Mao et al. (2011) and the Y1H screen, mediates transcriptional activation. This is consistent with the requirement of WRKY-boxes in the promoter of *WRKY33* for strong pathogen induced expression of a fused reporter (Lippok et al., 2007). It is also consistent with the recent finding that WRKY33 is a transcriptional activator (Lai et al., 2011a). The promoter-reporter constructs (e.g. *P1::GUS, P4::GUS* and *P4m1-4::GUS*) could be stably transformed into the genome of knockouts of *MYC2, ARF2, WRKY25* and *WRKY33* to test their role in regulation of *WRKY33* expression during *B. cinerea* infection.

It was not possible to demonstrate binding was direct with the transient transactivation assays in this chapter. This could result from indirect regulation, either with or without the direct transcriptional regulation. Also it is possible that overexpression of the TFs is resulting in ectopic binding (binding at irrelevant genomic locations due to physiologically high protein levels) that is confounding the results. This suggests that ChIP may be required to demonstrate that these TFs are binding to the given promoters *in planta*.

While the biolistic transactivation assay used in this study proved effective, the methodology could be improved by: higher technical/biological replication, non-parametric hypothesis testing (possibly using the Kruskal-Wallis χ^2 test), constitutive promoters that are less likely to be affected by TF overexpression and a dual reporter system that had more similar protein stability/kinetics.

In summary, transient transactivation assays appear to be a complementary approach to cloned library Y1H, allowing the effect of binding to be studied rapidly. ChIP is likely to be needed to demonstrate direct TF-promoter interactions *in planta*.

3.4.3 A role for ARF2 in the defence response

The only literature on the role of *ARF2* in the defence response is by Stotz et al. (2011) who have shown that *ARF2* affects the susceptibility of Arabidopsis to the necrotrophic pathogen *Sclerotinia sclerotiorum*. Additionally *ARF2* has recently been found to affect the susceptibility of Arabidopsis to infection by *B. cinerea* (Youn-Sung Kim et al., in preparation). 'Out of context' evidence gave a list of 895 genes differentially expressed in a knockout of *ARF2* versus wildtype experiment (Vert et al., 2008). The over-representation of these genes in the list of genes differentially expressed during *B. cinerea* infection suggests that the role of *ARF2* during the defence response is to regulate the expression of the genes in this overlap. One of these targets, *WRKY33*, was confirmed by a transient transactivation assay. Two other targets are the likely physiological outputs *BAP1* and *PGIP1* (Vert et al., 2008), suggesting that *ARF2* inhibits the defence response against *B. cinerea* which could explain the decreased susceptibility of the *ARF2* knockout to infection by *B. cinerea* (Youn-Sung Kim et al., in preparation). This could be tested by screening double and single knockouts for epistatic *B. cinerea* susceptibility phenotypes.

3.4.4 TOPLESS may play a role in the defence response of Arabidopsis to infection by *B. cinerea*

TFs commonly act by recruiting co-factors that encourage or discourage transcription. In the case of *MYC2*, a seemingly contradictory result can be explained by the interaction of *MYC2* with repressive co-factors such as JAZ proteins, NINJA and TOPLESS. This has been discussed in more detail in section 3.4.2. The latter co-factor, TOPLESS, has also been shown to mediate repression by ARF5, by binding its co-factor BDL (Szemenyei et al., 2008). ARF TFs share conserved domains such as the C-terminal dimerization domain, which facilitates interactions with the IAAs co-factors such as BDL (Ulmasov et al., 1999; Reed, 2001). This raises the possibility that the role of both MYC2 and ARF2 is to recruit the repressive co-factor TOPLESS to a genes promoter, allowing a rapid activation by de-repression following detection of pathogen attack and the subsequent activation of signalling pathways. This could be studied by testing whether ARF2 can interact with BDL and TOPLESS.

3.4.5 A role for ANAC072 in the defence response

In the previous chapter *ANAC072* was inferred to regulate the expression of 38 genes during *B. cinerea* infection. Consequently, it was screened in a reverse genetics screen which showed that knockout mutants of *ANAC072* had a weak but repeatable reduction in susceptibility to *B. cinerea*. In this chapter known downstream targets,

found in an 'out of context' experiment, were found to be over-represented in the list of genes differentially expressed during *B. cinerea* infection. The altered phenotype and potential regulatory targets during *B. cinerea* infection suggest a role for ANAC072 in regulation of the defence response. This could be tested in a mutant versus wildtype microarray experiment performed during *B. cinerea* infection.

3.4.6 TCPs

TCP TFs were found binding to 9/13 of the promoter fragments screened using cloned library Y1H in this chapter. Some fragments still had auto-activation with the highest level of 3AT used, allowing for the possibility that TCPs would have been found binding to more of the fragments if a higher 3AT level was used. Overall this suggests that the ability of TCPs to interact with the promoter-reporter plasmids in yeast is fairly ubiquitous. This could be due to a technical problem, such as a TCP binding motif in the promoter-reporter plasmid. This is considered unlikely due to other results within the group, but could be tested by performing a cloned library Y1H with a *pHis2Leu2* vector that contains no promoter fragment.

Assuming that a technical problem is not behind the binding of TCP TFs to the promoter-reporter plasmids, then the Y1H results presented in this chapter suggests that TCP TF-promoter binding is fairly ubiquitous, at least in the promoters of genes associated with the defence response. To see if TCP binding is unique to the defence response, cloned library Y1H could be performed on promoter fragments of genes that are not believed to be linked to the defence response. It is worth noting that the TF, CHE, revealed by Pruneda-Paz et al. (2009) to interact with the promoter of *CCA1* in Y1H and *in planta* was itself a TCP TF, showing that TCP binding occurs in the promoter of at least one gene that has not yet been linked to the defence response. The ability of interacting TCPs to regulate the expression of the promoters they can bind to could be tested with transient transactivation assays or mutant versus wildtype microarrays, a good candidate for a TCP target is *PGIP1* which is differentially expressed after some of its interacting TCP TFs during *B. cinerea* infection.

3.4.7 Extended qualitative model of the defence response gene regulatory network

The experimental results presented in Sections 3.3.2–3.3.3 can be used to extend the qualitative model introduced in Section 3.3.1. Before this screen the qualitative model included four edges where evidence existed that the regulator could both bind to the promoter and regulate the expression of the target (EIN3 $\rightarrow$*ERF1*, ERF1 $\rightarrow$ *CHIB*, ORA59 $\rightarrow$ *PDF1.2* and WRKY33 $\rightarrow$ *PAD3*). The Y1H screens presented

in this chapter showed that both MYC2 and WRKY25 could bind to the promoter of *WRKY33*. Together with literature evidence showing that they could regulate the expression of *WRKY33* (Dombrecht et al., 2007; Li et al., 2011), this adds two more binding/regulation confirmed edges to the qualitative model (i.e. MYC2 $-|$ *WRKY33* and WRKY25 $\rightarrow$ *WRKY33*). These two edges were supported by the transient transactivation assay, which demonstrated that both WRKY25 and MYC2 can affect the expression of a reporter fused to the *WRKY33* promoter. Additionally, transient transactivation assays also revealed that WRKY33 could regulate the expression of a reporter fused to its own promoter, which together with the WRKY33-*WRKY33* promoter binding evidence from Mao et al. (2011) adds another binding/regulation confirmed edge to the qualitative model (i.e. WRKY33$\rightarrow$ *WRKY33*). Taken together these three additional binding and regulation confirmed edges almost doubles the number of direct transcriptional edges in the qualitative model, this is shown in Figure 3.24 together with the additional 'context free' regulation evidence provided by the Y1H screens.

This demonstrates the utility of both the cloned TF library Y1H screens and transient transactivation assays in elucidating the local structure of a GRN. Further Y1H and transient transaction assays could be used to validate and extend the network structure around other network components, leading to a GRN structure that could then be modelled and/or tested during *B. cinerea* infection. The relevance of the regulatory edges in the qualitative model to the defence response of Arabidopsis to *B. cinerea* infection could be tested by mutant versus wildtype microarray and ChIP experiments on *B. cinerea* infected leaves.

Comparative transcriptomics has shown that targets of TGA3, ARF2, ERF1 and ARF2 are over-represented in the overlap with the list of genes differentially expressed during *B. cinerea* infection. This is consistent with the regulation of *B. cinerea* responsive gene expression by these TFs, this could be tested by Y1H or ChIP to extend the qualitative model presented here.

3.4.8 Comparative transcriptomics could be handled within a GO analysis package to identify regulators of different contexts

The comparative transcriptomic analysis presented in Table 3.10 was performed in the same way that GO terms are typically analysed; by ranking them by cumulative hypergeometric p-values. Currently genes differentially expressed in mutant versus wildtype experiments are not annotated as such using GO terms, meaning that comparative transcriptomics can only be performed on manually selected lists. The downside of manual selection of these lists is that it will be biased by the expectations

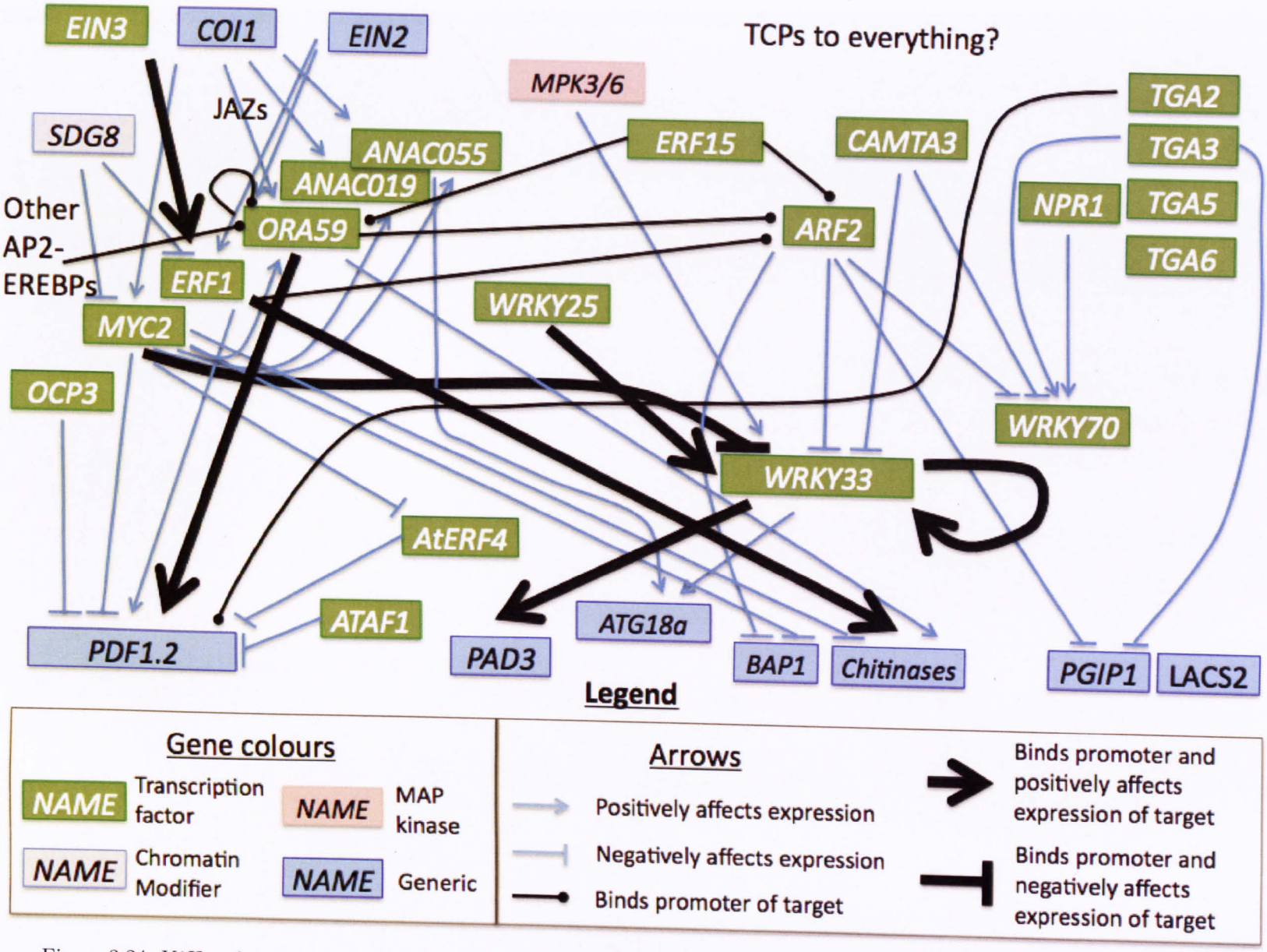


Figure 3.24: Y1H and transactivation results extending the qualitative model of Figure 3.2 (for references see Tables 3.4–3.5).

of the researcher, i.e. there may be mutant versus wildtype experiments whose differentially expressed gene list overlaps are more significant than those observed in Table 3.10, which have not been observed as these lists have not been selected for analysis. An extension of this manual approach would be to annotate genes as being differentially expressed in response to mutations of specific TFs, either within the GO framework or within a custom database. An example of a custom database is Genevestigator which can be used for meta-analysis of transcriptomics, but which currently only stores data from Affymetrix microarrays (Hruz et al., 2008). If this information could be incorporated within the GO framework then the research community would be more likely to re-use published mutant versus wildtype data. Re-use of data is good, both for groups using and producing mutant versus wildtype datasets, and may lead to novel findings that would otherwise have been missed.

3.4.9 Conclusion

Cloned library Y1H and transient transactivation assays are complementary approaches allowing the identification and characterisation of direct transcriptional regulators. This approach has validated and extended an initial qualitative model based upon 'out of context' and 'context free' evidence of transcriptional regulation. For example, Y1H demonstrated that MYC2 could bind to the promoter of *WRKY33*, showing that transcriptional regulation known in the literature was probably direct. Additionally Y1H identified novel promoter interactors, such as WRKY25, which interacted with the *WRKY33* promoter in Y1H. The ability of WRKY25 to activate the expression of a reporter fused to the promoter of *WRKY33 in planta* was then shown by a transient transactivation assay. Transactivation assays also showed that the binding of WRKY33 to its own promoter mediates transcriptional activation. The relevance of some of the regulators in the qualitative model to the defence response to *B. cinerea* infection was demonstrated by comparative transcriptomics. For example, this highlighted two TFs, ARF2 and ANAC072, that have recently been found to have mutants with altered susceptibility to *B. cinerea* infection. This suggests roles for these TFs in regulating part of the defence response to infection by *B. cinerea*.

Chapter 4

Dynamic modelling of the gene regulatory network mediating plant defence

In the previous chapter a qualitative model of the GRN mediating plant defence was built from existing qualitative and quantitative data, before being validated and extended experimentally. This qualitative model is an important first step towards a predictive quantitative model of the GRN mediating plant defence. In this chapter the aim is to develop a quantitative model of the defence response. To achieve this network inference will be applied to the expression of these genes over time to infer the structure of the GRN underpinning the plant defence response. Additionally, the qualitative model generated in the previous chapter will be used as informative priors to guide inference.

4.1 Introduction

4.1.1 Using Bayesian priors to take current knowledge into account during network inference

The Bayesian approach allows prior information to be used during statistical inference; prior information is used to define prior distributions, often referred to simply as priors, over parameters. Priors can be either vague or specific, known as uninformative or informative priors respectively. Priors are taken into account by the application of Bayes theorem which is defined in Equation 1.1. In this equation it can be seen that the posterior probability distribution of a Bayesian variable is affected by the choice of a prior distribution for it. For example, our *a priori* belief of the probability of a coin landing on the queen side ('heads') is 0.5, because we assume that the two possible outcomes ('heads' or 'tails', ignoring the chance of the

coin landing on its edge) are equally likely. This is an uninformative prior because it does not suggest that any one discrete outcome is more likely to happen than the other. If we were to consider the outcome that a coin lands on its edge, then we would probably presume *a priori* that it was unlikely relative to 'heads' or 'tails'. This is an informative prior, whose value/weight would be based, at least initially, on a subjective judgement. More formally, an uninformative prior distribution over a set of discrete outcomes is one that assigns the same probability to each.

In the case of GRN structure inference, priors can be defined for specific edges (directed regulatory pairs, i.e. TF $\rightarrow$ its target). A prior relating to the probability of a specific edge is called a 'prior edge', and is an informative prior because it suggests certain edges are more likely than others. The use of informative priors in biological network structure inference has been explored in Mukherjee and Speed (2008), and will be applied in this chapter in an effort to increase the accuracy of inferences of the structure of the defence response GRN. Informative priors can help to constrain network inference, which is especially useful for the relatively small and noisy datasets encountered in molecular biology applications (Mukherjee and Speed, 2008). The Bayesian approach allows the flexible integration of current knowledge into network inference. This flexibility allows inference to disagree with the prior given strong enough evidence; for example in a paper by Mukherjee and Speed (2008) even an incorrect informative prior was found to improve predictive accuracy.

4.1.2 Prior edges in VBSSM

The matrix D in the SSM shown in Equation (1.7) summarises the effect of each gene on the expression of the other genes at the next time point. For the default (uninformative) prior, every entry of D is assumed to be normally distributed with a mean of zero and with a unit variance (Beal et al., 2005). Within the VBSSM GUI (developed by Paul Brown and David Wild) there is an option to specify prior edges, i.e. cases of transcriptional regulation that are known *a priori*. These prior edges shift the mean of the prior distribution of the relevant entry of D; a positive mean represents positive regulation, while a negative mean represents repression. These prior edges will affect the transcriptional regulation that will be inferred by affecting the posterior probability over parameters given observed data by Bayes theorem (Equation 1.1). The value of the offset of the mean of the distribution of a given parameter, in standard deviations, is termed its z-score. The z-score can be used to specify a shift in the prior or can be calculated from the posterior distribution of the parameter, given the data and prior, to infer that parameter from the data.

The aim of this chapter is to develop a quantitative model of gene regulation during *B. cinerea* infection, which would allow predictions to be made about the effect of genetic perturbations to the network. It may also allow central regulators of the defence response to be identified. Time series of gene expression during infection and prior literature knowledge about trancsriptional regulation can be used to develop a quantitative model of gene regulation.

4.2 Results

In the previous chapter a qualitative model of the defence response GRN was built, based on experimental evidence from the literature and the unpublished work of colleagues, this was shown in Figure 3.2. Here a quantitative modelling approach, VBSSM, is used to infer the structure of the GRN underpinning the defence response based on time series of gene expression during infection by *B. cinerea*.

4.2.1 Application of VBSSM to the gene regulatory network underpinning the defence response

VBSSM was used because it has been shown to perform relatively well at predicting the network structure of a synthetic yeast GRN (Penfold and Wild, 2011). Genes were selected for modelling based on current knowledge of the genetics, gene regulation and physiological outputs of this defence response, as summarised in Figure 3.2. This was discussed in greater detail in Section 3.3.1. The expression profiles of these genes in the mock and Botrytis-infected time series from Section 2.2.1 are shown in Appendix E.

Without prior knowledge

VBSSM was applied to the time series expression data of these genes, with the default uninformative prior used (data described in Section 2.2.1). The resulting GRN structure inferred by VBSSM is shown in Figure 4.1. The structure contains two edges that agree with the literature on downstream targets of ANAC055 and TGA3, namely that ANAC055 positively regulates the expression of *ATG18a* and that TGA3 positively regulates the expression of *WRKY70*. *ATG18a* has been found to be differentially expressed in a knockout of *ANAC055* during senescence (Hickman et al., in preparation). *WRKY70* has been shown to be differentially expressed in a knockout of *TGA3* during infection by *B. cinerea* (Windram, 2010). If self-regulation is excluded then the p-value of obtaining at least two edges that are known in the literature, by chance, is 0.106 (as calculated in MATLAB ®using the hypergeometric distribution to test the null hypothesis that this overlap occurred

by chance), which is not significant at the 5% level. ($32 \times 31 = 992$ possible extra edges, 16 'true' positives, see Table 4.1, and 35 inferred edges).

Another inference which is plausible given current knowledge is the regulation of *WRKY33* expression by MPK3, which is known to occur by post-translational modifications (Mao et al., 2011). However, this regulation is not known to be affected by the level of *MPK3* expression and so it is not obvious that this inference makes sense. If true this inference suggests that this post-transcriptional regulation is affected by the expression level of *MPK3*, which is plausible in activating conditions. If this is interpreted as a correct inference then the cumulative hypergeometric p-value decreases to 0.016, which is significant at the 5% level. (Overlap of three, all other parameters as in previous paragraph). To assess the sensitivity of this inference to changes in the data, inference was repeated on the same dataset leaving off the first timepoint, resulting in a broadly comparable inferred network structure with a few minor differences (Figure B.5).

With priors based on literature knowledge of direct regulation

In this section a prior based on literature knowledge of direct regulation was used to inform network structure inference. The prior related to two regulatory connections: EIN3 binding to the promoter of *ERF1* and positively regulating its expression; and ERF1 binding to the promoter of *CHIB* and positively regulating its expression (Solano et al., 1998). A prior weight of 0.5 standard deviations was used for each prior edge.

The resulting GRN structure inferred by VBSSM is shown in Figure 4.2. It contains the two edges specified by the prior. In addition, it also infers the two known edges, TGA3 $\rightarrow$ *WRKY70* and ANAC055 $\rightarrow$ *ATG18a*, which were also inferred in Figure 4.1. Again, to test sensitivity inference was repeated on the same dataset leaving off the first timepoint, with the informative prior described in the previous paragraph used, resulting in a broadly comparable inferred network structure with a few minor differences to that produced with both the informative prior and the full dataset. The most noticeable difference was that MPK3 was no longer inferred to regulate any of the other genes (Figure B.6).

With priors based on literature knowledge of regulation that is possibly indirect

In this section the prior used in the last section is extended to take literature on indirect regulation into account. This prior is summarised in Table 4.1. A prior weight of 0.5 standard deviations was used for each prior edge. The resulting

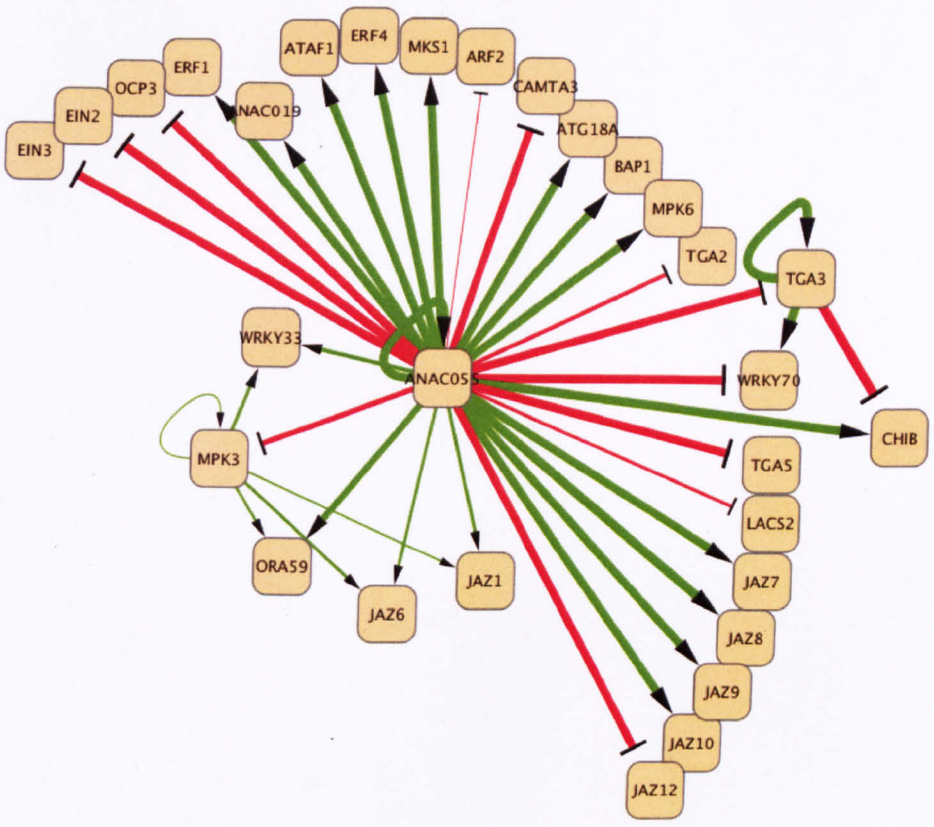


Figure 4.1: Inferred structure of the defence response GRN, using an uninformative prior. Network structure inferred by VBSSM from the expression of the genes shown in Figure 3.2 and the differentially expressed JAZs, using 20 initialisations and 4 hidden states. Green arrows indicate positive regulation and red arrows indicate negative regulation. The thickness of the arrows correspond to the number of initialisations that led to that inferred edge.

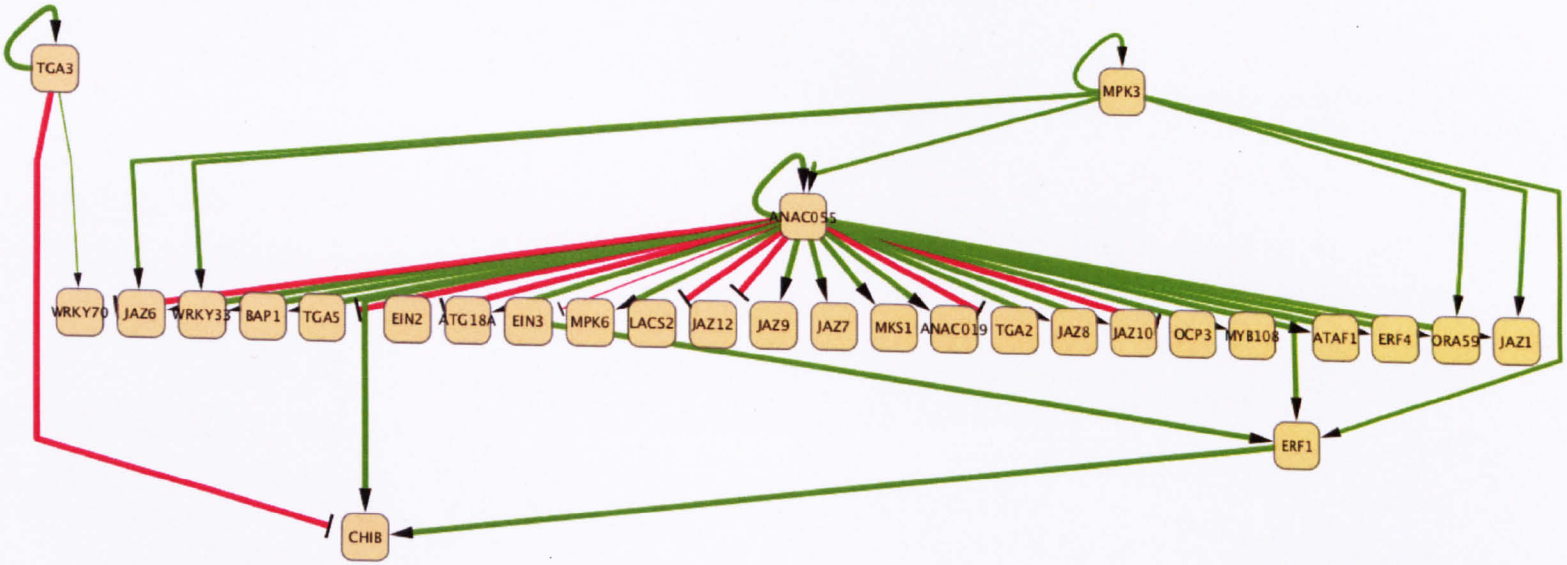


Figure 4.2: Inferred structure of the defence response GRN, made using a prior representing literature on direct regulation. Network structure inferred by VBSSM from the expression of the genes shown in Figure 3.2 and the differentially expressed JAZs, using 20 initialisations and 3 hidden states. Green arrows indicate positive regulation and red arrows indicate negative regulation. The thickness of the arrows correspond to the number of initialisations that led to that inferred edge. A prior was used to reflect knowledge of transcriptional regulation, each with a prior weight of 0.5 standard deviations.

GRN structure inferred by VBSSM is shown in Figure 4.3. Half of the prior edges were recovered in the inferred network structure (ANAC055 → *ATG18a*, WRKY33 → *ATG18a*, CAMTA3 -| *WRKY33*, CAMTA3 -| *WRKY70*, TGA3 → *WRKY70*, TGA3 -| *PGIP1* and ORA59 → *CHIB*), and some novel regulation was also inferred. One prior edge was inferred with the opposite sign to that used in the prior, i.e. ARF2 is inferred to *positively* regulate the expression of *WRKY33*. The inference was repeated with the same prior on the same dataset, leaving off the first timepoint, resulting in a broadly comparable inferred structure with a few minor differences. The most obvious differences were JAZ12 → *WRKY33* and WRKY33 → *MPK3* which were additionally inferred (Figure B.7).

Table 4.1: A table summarising the prior edges used in VBSSM, to generate Figure 4.3, based on the literature summarised in Figure 3.2.

Regulator	Target	Nature of regulation	Source
EIN3	*ERF1*	Positive	Solano et al. (1998)
ERF1	*CHIB*	Positive	Solano et al. (1998)
TGA3	*PGIP1*	Negative	Windram (2010)
TGA3	*JAZ9*	Negative	Windram (2010)
TGA3	*WRKY70*	Positive	Windram (2010)
ORA59	*CHIB*	Positive	Pré et al. (2008)
ANAC055	*ATG18a*	Positive	Hickman et al., (in preparation)
WRKY33	*ATG18a*	Positive	Lai et al. (2011b)
CAMTA3	*WRKY33*	Negative	Galon et al. (2008)
CAMTA3	*WRKY70*	Negative	Galon et al. (2008)
ARF2	*WRKY33*	Negative	Vert et al. (2008)
ARF2	*BAP1*	Negative	Vert et al. (2008)
ARF2	*PGIP1*	Negative	Vert et al. (2008)
ARF2	*WRKY70*	Negative	Vert et al. (2008)
ARF2	*JAZ1*	Negative	Vert et al. (2008)
ARF2	*JAZ6*	Negative	Vert et al. (2008)

With priors derived from the previous chapter and the literature

In this section the prior that was used in the last section is extended to take the Y1H results of the previous chapter into account. The additional prior edges are summarised in Table 4.2. A prior weight of 0.5 standard deviations was used for each prior edge. Because these priors are just based on Y1H results there is no knowledge of whether the regulator activates or represses the expression of its targets. *In lieu* of this PCC was used to infer whether the regulation is positive or negative in nature, i.e. regulation was assumed to be positive if the expression profiles of the TF and its target were positively correlated over the time series. The resulting GRN

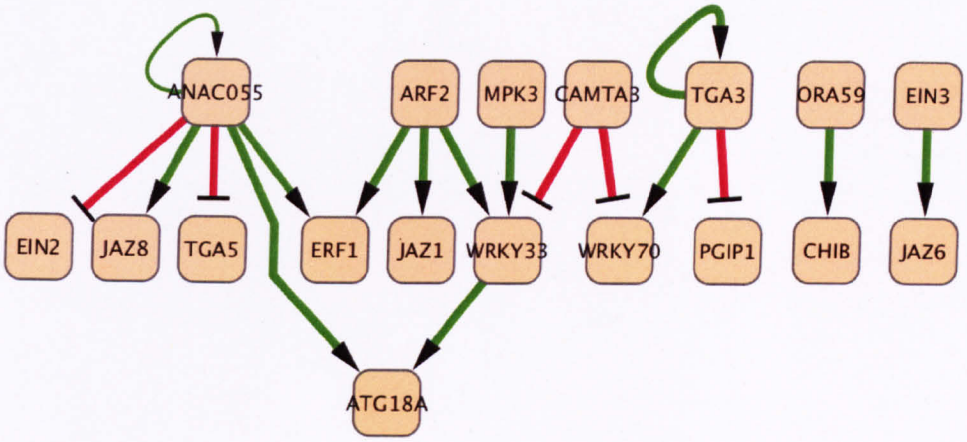


Figure 4.3: Inferred structure of the defence response GRN, made using a prior representing literature on both direct and indirect regulation. Network structure inferred by VBSSM from the expression of the genes shown in Figure 3.2 and the differentially expressed JAZs, using 20 initialisations and 4 hidden states. Green arrows indicate positive regulation and red arrows indicate negative regulation. The thickness of the arrows correspond to the number of initialisations that led to that inferred edge.

structure inferred by VBSSM is shown in Figure 4.4. None of the additional prior edges have been inferred. In fact the inferred structure is very similar to Figure 4.3, identical except for the loss of the inference that CAMTA3 negatively regulates the expression of *WRKY70*. Inference with the same prior was repeated on the same dataset, leaving off the first timepoint, resulting in a broadly comparable inferred network structure with a few minor differences. The most obvious of which were JAZ12 $\rightarrow$ *WRKY33* and WRKY33 $\rightarrow$ *MPK3* which were again additionally inferred (Figure B.8).

Table 4.2: A table summarising the additional prior edges, used in VBSSM to generate Figure 4.4, based on the previous chapter.

Regulator	Target	PCC	Inferred nature of regulation
ORA59	*ARF2*	-0.373	Negative
ERF1	*ARF2*	-0.465	Negative
ERF1	*ORA59*	0.956	Positive

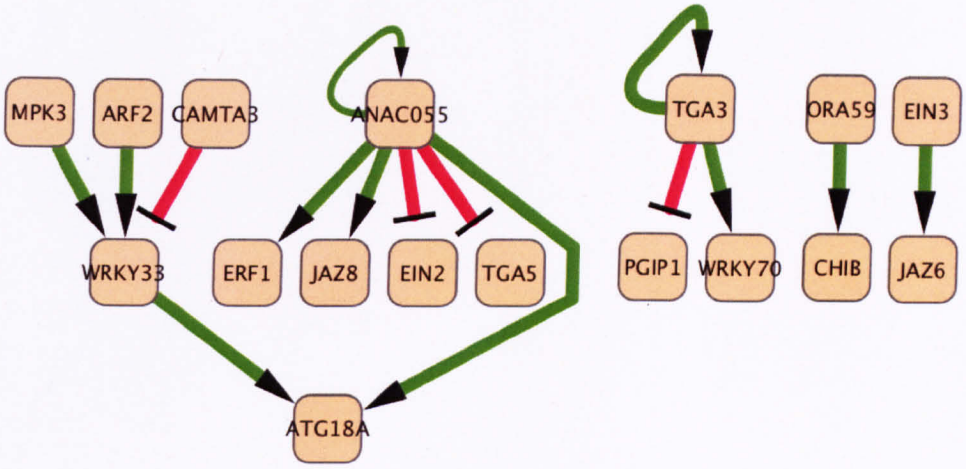


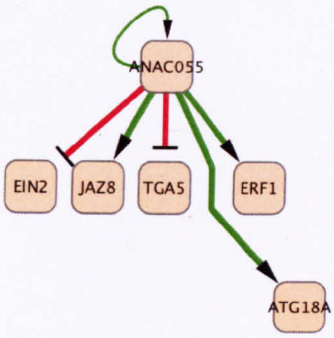
Figure 4.4: Inferred structure of the defence response GRN, made using a prior representing the Y1H results of the previous chapter, as well as literature on both direct and indirect regulation. Network structure inferred by VBSSM from the expression of the genes shown in Figure 3.2 and the differentially expressed JAZs, using 20 initialisations and 4 hidden states. Green arrows indicate positive regulation and red arrows indicate negative regulation. The thickness of the arrows correspond to the number of initialisations that led to that inferred edge.

Comparison of expression profiles of inferred TF-target pairs

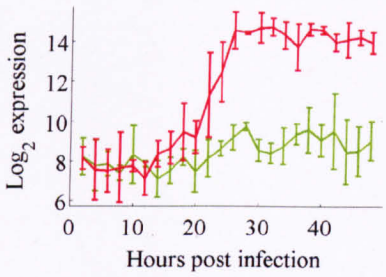
VBSSM has been used to infer the structure of the GRN underlying the defence response of Arabidopsis to infection by *B. cinerea*. Inferences were made based on the expression of genes over time during infection and also based on prior knowledge of transcriptional regulation. In this section these inferences are analysed visually to highlight biologically plausible inferences. For example, data showing the expression of these genes in the mock infection time series from Section 2.2.1 can be used to see if changes in the expression of regulators precede changes in the expression of their inferred targets. Because the mock data has not been used in network inference it provides independent information on the inferred GRN structure[1]. This was applied to three demonstrative examples: consistently inferred targets of ANAC055; inferred regulators and targets of *WRKY33*; and regulators of *CHIB*.

ANAC055 was inferred to regulate many downstream genes in Figure 4.1. While fewer targets were inferred when informative priors were used in inference, in all inferred network structures ANAC055 was inferred to be a central regulator (Figures

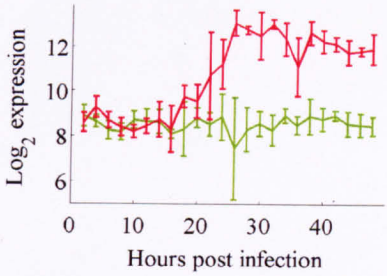
[1]Mock data was used to determine differential expression, as explained in Section 2.2.1, and only differentially expressed genes were modelled together in VBSSM. Other than this mild dependence, the mock data can be considered to represent independent data that was not used in network inference.



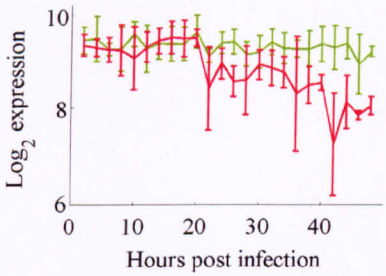
(a) Inferred *ANAC055* local network



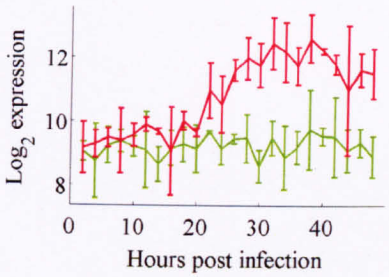
(b) *ANAC055* expression



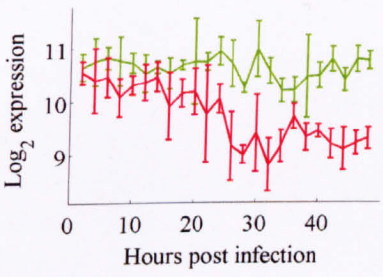
(c) *JAZ8* expression



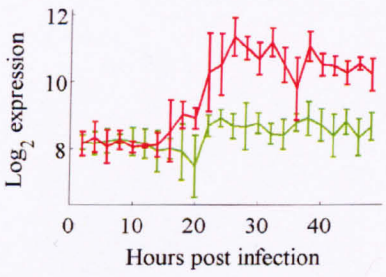
(d) *TGA5* expression



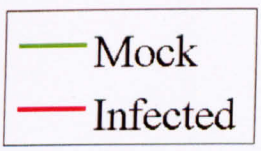
(e) *ATG18a* expression



(f) *EIN2* expression



(g) *ERF1* expression



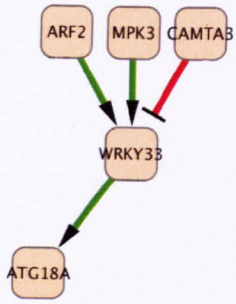
(h) Legend for (b)–(g)

Figure 4.5: Inferred targets of *ANAC055* from Figures 4.1–4.3. (a) Diagram showing genes inferred to be transcriptionally regulated by *ANAC055* in all of Figures 4.1–4.3. (b)–(g) Expression profiles of the genes shown in (a). (h) Legend for expression plots. (Expression data from the experiment introduced in Section 2.2.1. Experiment will be published in Denby et al., manuscript in preparation, and is also presented in Windram, 2010).

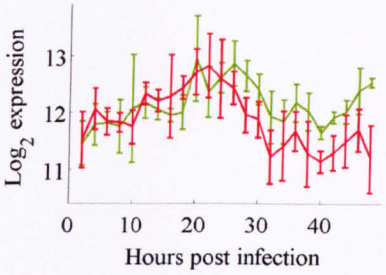
4.1–4.4). The expression of consistently inferred targets of ANAC055 in Figures 4.1–4.3 were compared to that of *ANAC055*, this is shown in Figure 4.5. Both *JAZ8* and *TGA5* appear visually to be first differentially expressed at a later timepoint than ANAC055, which appears visually to be first differentially expressed at around 22 hpi. Additionally, the expression of *JAZ8* and *TGA5* visually appear to be up or down regulated versus the mock respectively, both of which are consistent with the inferred positive or negative regulation by *ANAC055*, which is itself up-regulated. This means that *JAZ8* and *TGA5* are plausible downstream targets of ANAC055. The expression of *EIN2* and *ERF1* visually appears to be first differentially expressed at around 20 hpi, slightly earlier than *ANAC055*, suggesting that their regulation is also regulated by other factors or by post-transcriptional regulation of *ANAC055*. *ATG18a* visually appears to be first differentially expressed at roughly the same time as *ANAC055* and so it is possible that it is regulated by *ANAC055*, which fits with the differential expression of *ATG18a* in a knockout of *ANAC055* (Hickman et al., in preparation).

In the GRN structure inferred without an informative prior in Figure 4.1, ANAC055 and MPK3 were inferred to regulate the expression of *WRKY33*. The same inference was made in Figure 4.2 when a few prior edges were used. When more prior edges were added and inference was applied again some prior edges were recovered in the inferred structure, for example the subnetwork shown in Figure 4.6(a) was inferred in Figures 4.3–4.4. All of the inferred regulators (MPK3, ARF2 and CAMTA3) of *WRKY33* expression in Figure 4.6(a) visually appear to be differentially expressed after *WRKY33*, this suggests that either an additional regulator controls *WRKY33* expression or that at least one of its inferred regulators is post-transcriptionally activated. *CAMTA3* expression visually appears to be down-regulated at roughly the same time as *WRKY33*, making the inference that CAMTA3 negatively regulates *WRKY33* expression less plausible. *ATG18a* visually appears to be first differentially expressed at a similar time to *WRKY33*, making the positive regulation of *ATG18a* expression by WRKY33 plausible.

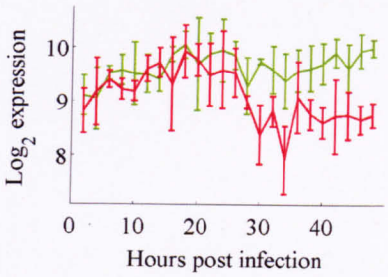
In Figures 4.3–4.4 *CHIB* expression is inferred to be regulated by ORA59, whereas in Figure 4.2 *CHIB* expression is inferred to be regulated by TGA3, ANAC055 and ERF1. In Figures 4.7–4.8 these possibilities are compared to the expression data. In Figure 4.7 it visually appears that first differential expression of *ORA59* and *CHIB* occurs at similar times, making the regulation of *CHIB* expression by ORA59 plausible. However, in Figure 4.8 two inferred regulators, *ANAC055* and *ERF1*, visually appear to be differentially expressed prior to *CHIB*. This makes them both good candidates to regulate the expression of *CHIB*. *CHIB* has already been shown



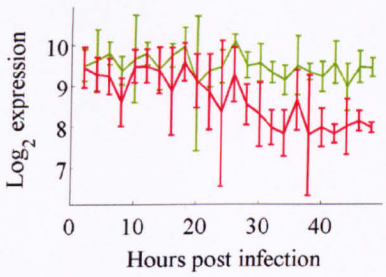
(a) *WRKY33* inferred local network



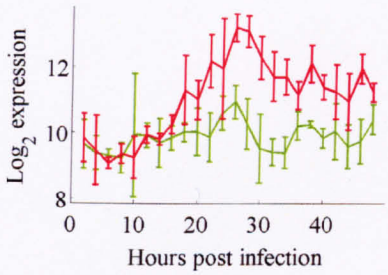
(b) *MPK3* expression



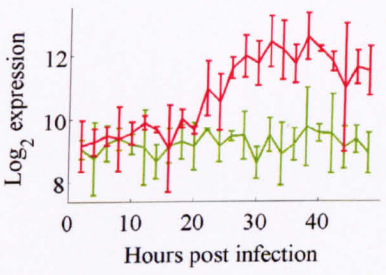
(c) *ARF2* expression



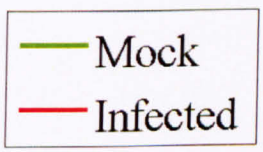
(d) *CAMTA3* expression



(e) *WRKY33* expression



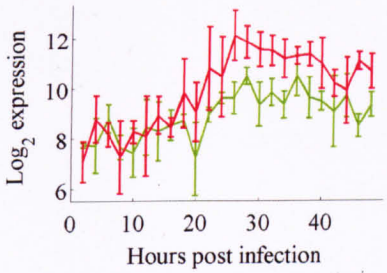
(f) *ATG18a* expression



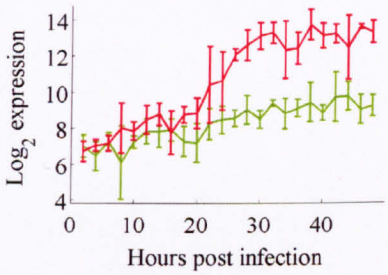
(g) Legend for (b)–(f)

Figure 4.6: Inferred regulators of *WRKY33* from Figures 4.3–4.4. (a) Diagram showing genes inferred to transcriptionally regulate *WRKY33* in Figures 4.3–4.4. (b)–(f) Expression profiles of the genes shown in (a). (g) Legend for expression plots. (Expression data from the experiment introduced in Section 2.2.1. Experiment will be published in Denby et al., manuscript in preparation, and is also presented in Windram, 2010). Expression profiles show average gene expression, bars represent standard deviation.

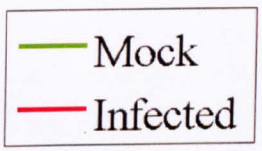
(a) *CHIB* inferred local network, priors edges from Table 4.1



(b) *ORA59* expression



(c) *CHIB* expression



(d) Legend for (b)–(c)

Figure 4.7: *CHIB* and the inferred regulator *ORA59* from Figures 4.3–4.4. (a) Diagram showing that *ORA59* is inferred to transcriptionally regulate *CHIB*. For details see Figures 4.3–4.4. (b)–(c) Expression profiles of the genes shown in (a). (d) Legend for expression plots. (Expression data from the experiment introduced in Section 2.2.1. Experiment will be published in Denby et al., manuscript in preparation, and is also presented in Windram, 2010). Expression profiles show average gene expression, bars represent standard deviation.

to be up-regulated in over-expressors of either *ERF1* or *ORA59* (Solano et al., 1998; Pré et al., 2008), and ERF1 has already been shown to capable of binding to the promoter of CHIB *in vitro* (Solano et al., 1998).

In summary, time series of gene expression and literature knowledge has allowed known and novel regulation to be inferred. Mock expression profiles allowed inferred transcriptional regulation to be compared to the temporal precedence of differential expression, allowing good candidates for future validation work to be selected.

4.3 Discussion

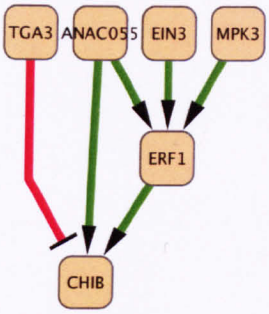
4.3.1 Quantitative models of the GRN underpinning the defence response

The use of prior edges

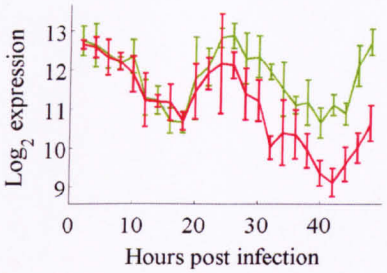
In this chapter informative Bayesian priors have been used to try to improve the accuracy of inference of the structure of the GRN underpinning the defence response. The flexibility of Bayesian priors has been demonstrated in Figures 4.3–4.4 where some prior edges have not been inferred. This shows that informative priors can be overruled by data. Bayesian priors also allow data to overrule the nature of the prior, for example in Figures 4.3–4.4 where ARF2 has been inferred to *positively* regulate *WRKY33*, in opposition to the literature (Vert et al., 2008) and the results of the previous chapter.

One challenge with the use of Bayesian priors is how to set the prior weight, i.e. how much to bias the inference towards certain prior expectations. In this chapter the prior weight was set arbitrarily, this could be improved upon either by: sensitivity analysis like that used in Mukherjee and Speed (2008), where the stability of inference relative to prior weight is verified; an empirical Bayesian approach (Robbins, 1956; Werhli and Husmeier, 2007), where the prior is informed by the data; or by analysis of the effect of various prior weights on accuracy of the inference, in a benchmark study similar to that performed in Mukherjee and Speed (2008) or Marbach et al. (2010).

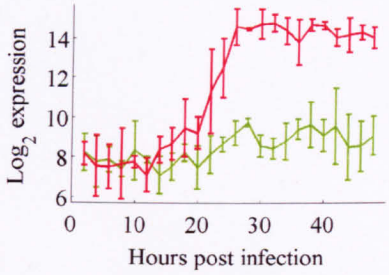
Without these, the accuracy of the inferences made in this chapter must be analysed by comparison to current knowledge or by validation work. This will be discussed further in the next two sections. However, the priors used for Figure 4.2, which relate to the most well validated knowledge of direct transcriptional regulation in the defence response, were both recovered in the inferred GRN structure. This suggests



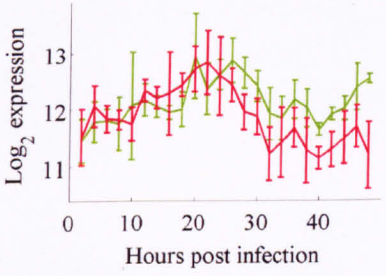
(a) *CHIB* inferred local network, using top two priors edges from Table 4.1



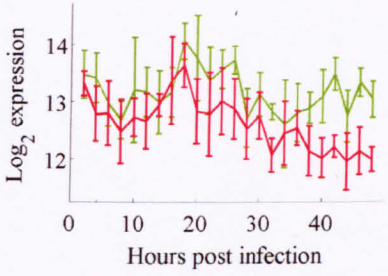
(b) *TGA3* expression



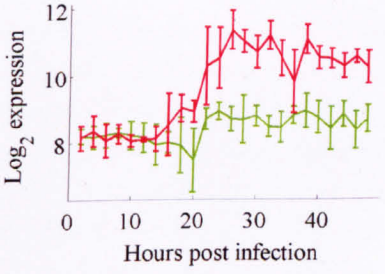
(c) *ANAC055* expression



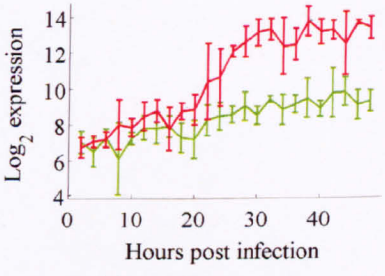
(d) *MPK3* expression



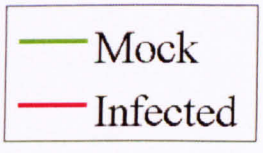
(e) *EIN3* expression



(f) *ERF1* expression



(g) *CHIB* expression



(h) Legend for (b)–(g)

Figure 4.8: Inferred regulators of *CHIB* and *ERF1* from Figure 4.2. (a) Diagram showing genes inferred to transcriptionally regulate *CHIB* or *ERF1* in Figure 4.2. (b)–(g) Expression profiles of the genes shown in (a). (h) Legend for expression plots. (Expression data from the experiment introduced in Section 2.2.1. Experiment will be published in Denby et al., manuscript in preparation, and is also presented in Windram, 2010).

that they are plausible with respect to the data and the modelling approach used. In the future the novel transcriptional regulation inferred in this chapter should be able to be rigorously tested; with current knowledge it is hard to assess the accuracy of the inferred regulation.

Where priors edges are not inferred, it is hard to know whether this is because of the model assumptions, the data, the prior weight, or the validity/relevance of the literature knowledge that the prior was based on.

As well as these experimentally derived priors, priors inferred from large scale gene expression databases such as the Nottingham Arabidopsis Stock Centre arrays database could be used. Network structures have been inferred from this database and can be found in Needham et al. (2009) and Carrera et al. (2009). These structures could be used in conjunction with the time series data used here to infer the network structure of the GRN underpinning the plant defence response to infection by *B. cinerea*.

Interpretation of inferred transcriptional regulation

One of the strongest inferences made in this chapter is the central role of ANAC055 in the regulation of the defence response. This is inferred in Figures 4.1 and 4.2, and to a lesser extent in Figures 4.3 and 4.4. Additionally, the comparatively early differential expression of *ANAC055* makes it a plausible early regulator of the defence response. The altered susceptibility of an overexpressor of *ANAC055* to infection by *B. cinerea* suggests that *ANAC055* plays an important role in the defence response (Bu et al., 2008). As the expression of *ANAC055* is itself controlled by MYC2 these inferences suggest a possible second wave of MYC2 mediated transcriptional regulation (Bu et al., 2008), i.e. MYC2 $\rightarrow$ ANAC055 $\rightarrow$ inferred targets. Given that *MYC2* is not differentially expressed during *B. cinerea* infection it is possible that this regulation depends on post-transcriptional regulation.

It is encouraging that in Figure 4.1 VBSSM inferred the plausible regulation of the expression of *ATG18a* by ANAC055. Regulation of *ATG18a* by ANAC055 has been demonstrated during leaf senescence (Hickman et al., in preparation), rather than during *B. cinerea* infection. This suggests that this regulation may also occur during *B. cinerea* infection, which is both plausible and testable. It is plausible because although *ATG18a* expression is reduced during *B. cinerea* in a knockout of *WRKY33*, it is still induced upon infection (Lai et al., 2011a). This suggests a TF other than WRKY33 can up-regulate *ATG18a* during *B. cinerea* infection. The regulation of *ATG18a* expression, by *WRKY33*, is taken into account by the use of

a prior in Figures 4.3–4.4.

Although MPK3 is known to regulate *WRKY33* expression, this regulation has not yet been related to the expression of *MPK3*, as it is known to be mediated by a kinase signalling pathway (Mao et al., 2011). The inference that *MPK3* expression increases the expression of *WRKY33* during *B. cinerea* infection could be tested with an overexpressor of *MPK3*.

In summary, some regulation was inferred that was known, and some that was plausible. Without further validation it is not known whether the number of correct inferences in Figures 4.1–4.4 is better than what could be achieved by random guesswork. Good candidates for validation were highlighted by visual examination of the timing and nature of differential expression of inferred regulatory pairs.

Network validation

The systems biology ideal is to experimentally validate model predictions and for this to guide future modelling in a virtuous cycle. In reality this is hard to achieve because of the difficulty of modelling and experimenting on complex biological phenomena. For example, both modelling and experimental validation of the structure of a GRN is extremely challenging. It is not clear which model assumptions about the dynamics of transcriptional regulation are realistic enough, especially with regard to hard to measure variables like the combinatorial logic of promoters as well as the level and activity of all relevant proteins.

On the experimental side, the strongest validation requires both *in planta* binding assays and mutant versus wildtype expression comparisons. The latter being comparatively easy to perform, with RT-PCR or microarrays for example. Unfortunately, *in planta* binding assays, such as ChIP, are considerably harder to perform, usually requiring either the generation of a tagged TF or the production of a specific antibody. Both of these approaches are time consuming. This means that experiments to strongly validate inferences that may have a high false positive rate are unattractive to a molecular biologist with other options.

This can be dealt with by starting with weaker validations, which can convince the researcher of the utility of further validation work. Following this, stronger validations can be made. For example, a prediction can be initially tested with a mutant versus wildtype expression experiment, with either a transient or stably transformed background. An alternative approach is to test TF-promoter binding outside of its natural context, for example in Yeast. These two approaches have been applied in

the previous chapter, to find regulators of TFs or physiological outputs.

Even weaker validations can be time consuming, and so there is a clear need for inferred transcriptional regulation to be as accurate as possible, and for the validation approaches to be streamlined and made higher-throughput. It is hoped that this will lead to a more effective application of the 'systems biology cycle' of modelling and experimental analysis.

An additional problem is that while it is theoretically possible to validate regulation using currently available methods, it is not practically possible to invalidate a prediction, so long as it is not overly specific. For example, a TF can regulate a target without binding near to its TSS, for example by indirect regulation. This means that an indirect regulatory prediction cannot be invalidated with assays measuring the binding of the inferred regulator to the targets promoter, such as ChIP, for example if the inferred regulator regulates an inferred target through transcriptional regulation of an intermediate TF it would not be detected binding to the targets promoter. Similarly the prediction that a TF regulates a target cannot be invalidated by a single knockout versus wildtype experiment because functional redundancy cannot be ruled out.

Because of these challenges, and more, high-throughput elucidation of gene regulatory networks remains an open problem.

4.3.2 Future directions

Relating genotype to phenotype via transcriptional regulation

In the longer term, if a predictive model of gene regulation in the defence response can be made it may be possible to use this to predict the effect of genetic perturbations on susceptibility to infection by *B. cinerea*. For example, some genes are known to encode enzymes which are mechanistically involved in resistance, such as the physiological outputs described in Section 1.4.2. The expression of these physiological outputs may act as markers of resistance. If the effect of the expression of physiological outputs on resistance to infection by *B. cinerea* can be modelled, then this could be combined with a model of the regulation of their expression by TFs. This model should then predict the impact of genetic perturbations on susceptibility, by first predicting the change in expression of physiological outputs. From a systems biology perspective this would be desirable, as susceptibility is considerably easier to test than network structure. This would allow model validation to be performed relatively quickly, facilitating a Systems Biology cycle of modelling and experimental validation.

This would require substantial research on physiological outputs, and also on the possible network structure regulating them. In this chapter and the previous, the second of these steps was pursued, but much additional work is required. It is unfortunate that the expression of two possible physiological outputs, *PAD3* and *PDF1.2*, were not measured in the gene expression time series experiment introduced in Section 2.2.1, as these could have been modelled in this chapter. Especially as relatively good literature knowledge exists on their regulation (for example in the papers of: Mao et al., 2011; Lai et al., 2011a; Lorenzo et al., 2004; McGrath et al., 2005; Coego et al., 2005; Pré et al., 2008; Wang et al., 2009).

ODE model of the defence response GRN

In this chapter VBSSM was used to generate quantitative models of transcriptional regulation during the defence response. It would be desirable to develop an ODE model when the network structure is better known, as this would allow the model to explicitly incorporate biochemical kinetics such as non-linear effects. An example of a simple ODE model of transcriptional regulation was given in Equation 1.12, this includes a transcriptional activation and a mRNA degradation term, α and δ respectively, which could be experimentally derived. Methods to determine genome-wide degradation rates already exits, and could be applied during *B. cinerea* infection to see if degradation is differentially regulated (for example Narsai et al. (2007)). Transcriptional activation could be studied independently of gene specific degradation rates through promoter-reporter assays. Additionally, it may be possible to get estimates of the sensitivity of transcriptional activation to the level of regulating TFs by adapting the transient transactivation assay used in the previous chapter. For example, the effect of different levels of overexpression of a TF on the levels of a reporter fused to the promoter of interest could be determined. Mock infection expression profiles could be used to determine a background model of diurnal expression, something that isn't taken into account in the current model. This could be important to determine the difference between diurnal and pathogen response transcriptional regulation.

4.3.3 Conclusions

In this chapter the expression profiles of genes with a known role in the defence response were used to predict the structure of the GRN underpinning the defence response. Several predictions were found to fit with the literature, i.e. the inferred target has been found to be differentially expressed in a knockout of the inferred regulator. These are the regulation of *WRKY70* expression by TGA3 (Windram, 2010); the regulation of *ATG18a* expression by ANAC055, which has been observed during

leaf senescence (Hickman et al., in preparation); and the regulation of *WRKY33* expression by MPK3 (Mao et al., 2011), which is inferred to depend on the expression of *MPK3*. In addition, literature knowledge was incorporated into the models through the use of Bayesian priors. For example: the regulation of *ERF1* expression by EIN3; and the regulation of *CHIB* expression by ERF1 (Solano et al., 1998). Overall the models predict the central role of ANAC055 in the regulation of the defence response.

Chapter 5

General conclusions

5.1 Inferring gene regulation from gene expression time series

An overall aim of this thesis was to develop a predictive model of the GRN underpinning the defence response. To do this unsupervised learning and graphical model inference were used. In each case improvements on existing analyses were made: TCAP was developed and it was shown cluster together some TFs with their known targets; VBSSM was applied to sets of genes that were likely to be involved together in the GRN; and literature knowledge was used to generate informative priors that could guide network inference. Because knowledge of transcriptional regulation in plants is sparse, it is hard to assess the predictive accuracy of these approaches. However, known regulation was recovered by each approach.

5.1.1 TCAP

TCAP was developed to infer transcriptional regulation between large numbers of genes. To achieve this a simple model of the effect of transcriptional regulation on the expression of TF-target pairs, time-delayed correlation, was used (Qian et al., 2001). This model was shown to have some predictive accuracy in a benchmarking exercise performed on experimentally derived datasets from the literature. In addition, groups of genes with a strong time-delayed correlation were discovered by clustering the genes. This was achieved with a recently introduced clustering approach, AP (Frey and Dueck, 2007), that was shown to be effective in this setting.

TCAP was applied to cluster the time series expression profiles of genes differentially expressed during *B. cinerea* infection. Although the sensitivity of TCAP to time-delayed correlation could be improved, many cases of strong time-delayed correlation between groups of genes were found. The most convincing recovery of

known transcriptional regulation was the grouping of *LHY* and *GI* in a cluster of 7 genes. *GI* was found to have a time-delayed anti-correlation with *LHY*, consistent with the known role of *LHY* as a transcriptional repressor of *GI*. Other clusters also grouped genes with a known regulatory connection, such as: *ORA59*, *ERF1* and six known targets of *ORA59* (*At1g59950*, *At2g43580*, *At3g23550*, *At3g56710*, *At4g11280* and *At4g24350*, from Pré et al. (2008)); and *ANAC072* with one of its known targets (*At4g37990*, from Fujita et al. (2004)). In the case of the cluster containing *ORA59*, no time-delay was observed in the correlation of *ORA59* and its targets. This could because of the temporal resolution of the dataset, or because of the context difference between the two studies. In Chapter 3 ERF1 was found to be capable of binding to the promoter of *ORA59* in Yeast, this suggests that it may regulate the expression of *ORA59*. If this can be shown to occur during *B. cinerea* infection, then it further validates the predictions of the *ORA59* cluster. In the case of the cluster containing *ANAC072*, the overlap with known targets is not statistically significant. Further experiments such as *ANAC072* mutant versus wildtype gene expression experiments or ChIP performed on leaf samples infected with *B. cinerea* are needed to test the validity of this inference.

TCAP can be extended in many ways: the sensitivity to time-delayed correlation can be improved, possibly by calculating an approximate p-value for each score in a manner similar to that introduced in Qian et al. (2001); time-delayed correlation can be analysed with a probabilistic model such as cubic splines with a delay parameter inferred for each gene expression profile; time series with uneven spacing could be handled with an interpolation system such as cubic splines, or by conventional dynamic time-warping if the number of time points is suitably high, as in Oates et al. (1999); the sensitivity to transient correlation could be increased, possibly using a biclustering approach that allowed time delayed comparisons and could use only contiguous subsets of timepoints (for a review of biclustering approaches see Madeira and Oliveira (2004)); the application of TCAP to multiple time series could be implemented, for example by adding the Qian score for each time series together; non-linear time-delayed correlation, such as aligned Spearman's rank correlation (Balasubramaniyan et al., 2004), could be benchmarked similarly to Figures 2.5, and then used instead of the Qian score in TCAP; the clustering of large datasets by AP could be performed in parallel to reduce runtime, as discussed in Blasberg and Gobbert (2008); and TCAP could be made available as a web tool to increase usage by biologists. All extensions would have to be optimised with respect to runtime, as a good feature of the current implementation of TCAP is that its runtime is not considerably higher than standard clustering approaches. This is important for methods that aim to allow exploratory analysis of data.

At the time of writing, few existing methods other than TCAP allow specific transcriptional regulation to be inferred from the expression of thousands of genes. Because of this TCAP is a valuable contribution to the GRN modelling field, as recognised by its recent publication (Kiddle et al., 2010).

5.1.2 VBSSM

In this thesis VBSSM was used to infer the structure of the GRN underpinning the defence response. Application of VBSSM, and all comparable methods that could be found in the literature, are limited to a small number of genes (or groups of genes), given the gene expression dataset available. Therefore in this thesis the focus was on ways to select genes to model together, the accuracy of these models and the effect of literature based informative priors.

In Chapter 2 analysis of co-regulation and literature knowledge of TF binding preferences were used to select genes to model in VBSSM. This approach, while dependent upon literature knowledge of TF binding preferences, is unbiased with respect to the literature in terms of the targets it can infer. In this way targets of the AP2-EREBPs, NAC and WRKY TFs during *B. cinerea* infection were inferred. The NAC TFs *ANAC019, ANAC055* and *ANAC092*, which were inferred to regulate co-expressed genes, were already known to affect the susceptibility of Arabidopsis to infection by *B. cinerea* (Bu et al., 2008; Windram, 2010). The overlap of inferred targets, with known targets from literature experiments, was not statistically significant. However, none of the literature experiments were performed during *B. cinerea* infection, and regulation can be context-specific as shown in Chapter 3. To properly test these predictions *in planta* and during *B. cinerea*, mutant versus wildtype gene expression or ChIP experiments, performed on leaf samples infected with *B. cinerea* are required.

In Chapter 4 genes whose mutants have altered susceptibility to *B. cinerea*, genes that can regulate the expression of genes that do, and genes that are a potential physiological outputs were modelled together. The rationale was twofold: TFs whose mutants have altered susceptibility to infection by *B. cinerea* are likely to regulate the pathogen-responsive expression of genes; and some evidence already exists, in the literature and in Chapter 3, that these genes are regulated by each other. Application of VBSSM to their expression correctly inferred that TGA3 can positively regulate the expression of *WRKY70*, i.e. *WRKY70* is differentially expressed in infected leaves in a knockout of *TGA3* during *B. cinerea* infection versus infected wildtype samples (Windram, 2010). VBSSM also inferred that ANAC055

positively regulates the expression of *ATG18a*, this is known to occur during senescence, i.e. *ATG18a* is differentially expressed during senescence in a knockout of *ANAC055* versus wildtype (Hickman et al., in preparation). Although the regulation of *ATG18a* expression by ANAC055 has not yet been shown during *B. cinerea* infection, the comparative transcriptomics analysis in Chapter 3 shows that not all transcriptional regulation is highly context specific. Additionally, ATG18a has been shown to be regulated by, but still induced in a knockout of, *WRKY33* (Lai et al., 2011b). This suggests that another TF is also important for the *B. cinerea* related up-regulation of *ATG18a* expression, and this inference suggests that ANAC055 is a possible candidate. This could be tested by mutant versus wildtype gene expression or ChIP experiments, performed on leaf samples infected with *B. cinerea*. A final prediction that is biologically plausible is that MPK3 expression regulates the expression of *WRKY33*. MPK3 is known to be involved in the up-regulation of *WRKY33* expression during *B. cinerea* infection (Mao et al., 2011), but this is believed to be mediated by kinase signalling. The effect of *MPK3* expression on *WRKY33* expression could be tested by comparing the expression of *WRKY33* in an inducible over-expressor of *MPK3* to that in a wildtype, during *B. cinerea* infection.

In addition, informative Bayesian priors were used to integrate literature knowledge into the SSMs. First, known direct regulation was included as a set of prior edges; which were then inferred by VBSSM. Secondly, a set of prior edges representing literature knowledge of indirect regulation were used in GRN structure inference by VBSSM. Half of these prior edges were inferred by VBSSM, although one negative prior edge was inferred to be positive (ARF2 $\rightarrow$ *WRKY33*). It is a well known feature of inference using BNs that there are cases where given certain datasets it is impossible to determine between similar network structures. These similar structures have been called 'equivalence classes', meaning sets of structures with identical probability given a dataset and prior (Pearl, 2000). It may therefore that certain local features of a network, such as direction of regulation and sign of regulation cannot always be inferred from a given dataset using BN approaches. Priors that were recovered represent those best aligned with the data and the SSM used. Finally, the results of Chapter 3 were used to inform additional prior edges, none of which were subsequently inferred by VBSSM. ANAC055 was inferred to regulate genes in all of the network models in Chapter 4, even those inferred with uninformative prior information.

While correct predictions and recovery of prior edges is encouraging, significant validation work remains. Mutant versus wildtype expression experiments during *B.*

cinerea can be used to validate/test these predictions. Additionally, ChIP can be used to confirm whether these predictions relate to direct transcriptional regulation.

5.2 Regulators of the defence response

5.2.1 Novel regulators of the defence response

In Chapter 2 novel regulators of the defence response were identified by reverse genetics. These TFs were selected for screening because they were differentially expressed, and were inferred to regulate other genes that were also differentially expressed, during *B. cinerea* infection. This is an extension of the approach of AbuQamar et al. (2006) who screened mutants of TFs up-regulated during infection, and similar to the approach of Windram (2010) who screened regulatory hubs inferred by VBSSM. Regulation was inferred by the application of TCAP to all genes differentially expressed during *B. cinerea* infection. A number of weak altered susceptibility phenotypes were discovered. Three independent knockouts of *ANAC072* were found to have decreased susceptibility to infection by *B. cinerea*. Two independent knockouts of *NUB* were found to have decreased susceptibility to infection by *B. cinerea*. A decreased susceptibility to infection by *B. cinerea* was seen in a knockout of *LBD41*, but no independent knockout was available to screen. An increased susceptibility to infection by *B. cinerea* was seen in a knockout of *RGL*, but was not seen in an independent knockout. The ability of this independently generated knockout to generate a functional transcript has not yet been experimentally tested. All novel altered phenotypes were weak in comparison to that of the MYB108 knockout, *bos1*, which was used as a positive control.

In conclusion, weak novel phenotypes were found with this approach. Weak phenotypes could have occurred either spuriously, because the TF contributes only slightly to defence, or because the TF contributes to the defence response in a partially redundant manner. A forward genetic approach might have allowed strong altered phenotypes to be focused on.

5.2.2 Qualitative model

A qualitative model of the structure of the GRN underlying the plant defence response was made by compilation of results from the literature. These results pertain to transcriptional regulation during various different contexts, mostly not during *B. cinerea* infection. This was used to provide hypotheses of regulation that could be occurring during *B. cinerea* infection. Some of these TF-target pairs were further validated in different contexts, by showing TF-promoter binding in Yeast or by showing activation/repression *in planta*. Validated predictions of the qualitative

model, as well as novel candidate regulators of the some of the genes in the model, are discussed below.

5.2.3 MYC2

MYC2 was not differentially expressed during *B. cinerea* infection and so its role in the defence response could not be modelled in Chapter 4. However, its role in the defence response of Arabidopsis to infection by *B. cinerea*, was studied experimentally in Chapter 3.

One of the known downstream targets of MYC2 is *WRKY33* (Dombrecht et al., 2007). This was previously not known to be regulated directly. In Chapter 3 MYC2 was shown to able to bind directly to a fragment of the promoter of *WRKY33* in Yeast. A MYC2 binding motif is found to be present near the start of this fragment. In addition, MYC2 was shown to activate the expression of a reporter fused to the promoter of *WRKY33* in a transiently transformed Arabidopsis leaves. This suggests a direct role for MYC2 in positively regulating the expression of *WRKY33*, in contrast with the negative role suggested by the data of Dombrecht et al. (2007). This can be reconciled by the experimental differences between the two studies, specifically negative regulation by MYC2 which is believed to be mediated by JAZ factors whose levels were probably low relative to MYC2 in the transient transactivation assay. The regulation of *WRKY33* expression by MYC2 could be tested further with mutant versus wildtype gene expression or ChIP experiments, performed on leaf samples infected with *B. cinerea*.

5.2.4 ARF2

ARF2 is a TF whose knockout mutant has decreased susceptibility to infection by *B. cinerea* (Youn-Sung Kim, in preparation). Its role in the defence response of Arabidopsis to infection by *B. cinerea* has not yet been studied in the literature. In this thesis, this role was studied bioinformatically and experimentally in Chapter 3, and with modelling in Chapter 4.

In Chapter 3 a high overlap was found between genes differentially expressed in an *ARF2* knockout (Vert et al., 2008) and during *B. cinerea* infection. This suggests that ARF2 regulates the pathogen-responsive expression of these genes, and also suggests that this regulation occurs in both uninfected seedlings and infected leaves. This suggests an important role for ARF2 in the regulation of pathogen-responsive gene expression, which could account for the enhanced resistance of its knockout to infection by *B. cinerea* (Youn-Sung Kim et al., in preparation). This role could be tested using mutant versus wildtype gene expression or ChIP experi-

ments, performed on leaf samples infected with *B. cinerea*.

One of the targets of ARF2 is *WRKY33*, i.e. *WRKY33* is differentially expressed in seedlings of a knockout of *ARF2* versus wildtype (Vert et al., 2008). In Chapter 3 ARF2 was shown to repress the expression of a reporter fused to the promoter, with mutated WRKY motifs, of *WRKY33* in a transiently transformed Arabidopsis leaves. This fits with the up-regulation of *WRKY33* expression a knockout of *ARF2* (Vert et al., 2008). The repression of *WRKY33* expression by ARF2 could be tested further with mutant versus wildtype gene expression or ChIP experiments, performed on leaf samples infected with *B. cinerea*.

In Chapter 4 *ARF2* was inferred to be regulated by ANAC055, by VBSSM with an uninformative prior. This is not known to occur, but could be tested with mutant versus wildtype gene expression or ChIP experiments, performed on leaf samples infected with *B. cinerea*. When an informative prior was used in VBSSM, to take into account the regulation of *WRKY33* expression by ARF2, it was inferred that ARF2 positively regulates *WRKY33* expression. This does not fit with the known role of ARF2 as a repressor of *WRKY33* expression (Vert et al., 2008) or with the finding that ARF2 could repress expression of a reporter fused to the promoter of *WRKY33* in Chapter 3. This could be the result of the models with ARF2 positively or negatively regulating the expression of *WRKY33* being in the same probabilistic 'equivalence classes', as discussed in Section 5.1.2.

5.3 Experimental analysis of transcriptional regulation

5.3.1 Yeast one-hybrid

In Chapter 3 Y1H has was used to identify novel interacting proteins of the promoters of TFs and physiological outputs with roles in the defence response of Arabidopsis to infection with *B. cinerea*. Y1H identified the direct interaction of WRKY33 with its own promoter, the fragment in which this interaction occurred overlaps with the fragment amplified by anti-WRKY33 ChIP-PCR by Mao et al. (2011). It also identified novel interactors of the promoters of *ARF2, LACS2, ORA59, PGIP1* and *WRKY33*. The impact of interactors of the *WRKY33* promoter, as well as ARF2, on the expression of a reporter fused to the *WRKY33* promoter was subsequently shown in a transactivation assay. This adds weight to the possibility that these interactors are involved in direct transcriptional regulation of *WRKY33*. These interactors could be tested further by mutating their inferred binding sites and testing for the abolition of interaction in a Y1H screen. Additionally, this binding could be tested *in planta* with ChIP, during *B. cinerea* infection, or in transactivation assays

with mutated binding sites in the promoter reporter plasmid.

5.3.2 Transactivation assays

In Chapter 3 a transactivation assay was used to characterise the effect of the interactors of the *WRKY33* promoter, and ARF2, on the expression of a reporter fused to *WRKY33* promoter fragments. This showed that WRKY25, WRKY33 and MYC2 positively, and ARF2 negatively, regulates the expression of *WRKY33*. The decrease in the reporter caused by over-expression of *ARF2* fits with the known role of ARF2 as a repressor of *WRKY33* expression (Vert et al., 2008). The activation of the reporter by *MYC2* over-expression suggests a positive role for MYC2 in regulating *WRKY33* expression, as discussed in Section 5.2.3 this makes sense in terms of the existing literature. Binding of these TFs to the *WRKY33* promoter during infection by *B. cinerea* could be tested with by ChIP.

The experimentally tractability of the transient transactivation assay used in Chapter 3 suggest that it could be a useful first step in the validation of the regulatory effect of novel Y1H interactors of other promoters. Another method, such as ChIP, will probably be necessary to conclusively demonstrate direct binding of TFs to promoters *in planta*.

5.4 Overall conclusion

In this thesis various bioinformatic, experimental and modelling approaches have been used to study the gene regulation underpinning the defence response of Arabidopsis to infection with *B. cinerea*. Over-representation analysis proved to be a useful way to interpret post-genomic datasets, such as the genes differentially expressed in various biological contexts, and the promoter sequences of co-expressed genes. In addition, a clustering method, TCAP, was developed to analyse time-delayed correlation in gene expression time series, this was shown to recover known and infer novel cases of transcriptional regulation. The inability of TCAP to recover combinatorial regulation was a motivation for applying VBSSM, but this was only tractable for smaller sets of genes. These sets were chosen based on literature knowledge and/or the results of binding motif over-presentation analysis, which allowed specific and therefore testable inferences to be made. However, most of the transcriptional regulation inferred by TCAP and VBSSM have not yet been experimentally tested.

Existing data on transcriptional regulation from the literature is sparse and may not be relevant to the context of *B. cinerea* infection, which means that its suit-

ability to assess the validity of these predictions is not currently known. Therefore, experiments performed during *B. cinerea* infection will be necessary to validate these predictions. These validation approaches typically test local features of GRN structure, and can be confounded by redundancy in some cases, meaning that validation of even a few local features of a GRN can be challenging. Because of this, inferred transcriptional regulation with an unknown false positive rate, and with no observed over-representation of known regulation, was considered too speculative to devote substantial time to test experimentally.

In order to tackle this, prior knowledge from the literature can be used to provide an initial qualitative model, which can be refined by experiment and analysis. This currently requires a trade-off between applying an unbiased approach, with respect to the literaure, and predictive accuracy. For example the literature bias of the qualitative approach led to validation of the ability of MYC2 and ARF2 to regulate the expression of *WRKY33*. This suggests that literature knowledge can be a good framework upon which to build future hypotheses. In the future, the increased availability and access to data will hopefully reduce the bias of this approach towards well studied TFs.

A similar trade-off can be made between the experimental throughput and context-specificity of validation work; in this thesis this was achieved by application of cloned TF library Y1H and transient transactivation assays. The throughput achieved provided many good candidates which can later be tested during *B. cinerea* infection. Cloned TF library Y1H screens also reduced literature bias by linking novel TFs to the defence response. For example WRKY25 was found to be able to bind to the promoter of *WRKY33*. WRKY25 was also shown to activate expression of a reporter fused to the *WRKY33* promoter in a transient transactivation assay. This suggests a possible role for WRKY25 in the regulation of *WRKY33* expression during *B. cinerea* infection. The combination of Y1H and transactivation assays proved useful in the identification of transcriptional regulators of *WRKY33*, and so could be applied to the promoters of other genes in the qualitative model to develop better knowledge of the local GRN structure. This could be useful both in testing modelling predictions, and in providing a potential structure of the GRN *de novo* which could then be modelled. TF-promoter interactions observed in Yeast and in transient transactivation assays could be followed up by ChIP, to study binding during infection and in the proper chromatin context, or by mutant versus wildtype microarray experiments to study regulation of gene expression. Both ChIP and microarray experiments performed on samples infected by *B. cinerea*, applied to the best candidates from Y1H and transient transactivation assays, could then provide

strong contextual validation of the qualitative model.

After development, partial validation and extension of the qualitative model, the key challenge was to combine it with a predictive quantitative model. This could then reveal hypotheses for the next round of experimental validation. Quantitative modelling of the GRN during *B. cinerea* infection was achieved by applying VBSSM, to the time series of the expression of these genes during infection and to informative priors based on the experimental evidence summarised in the qualitative model. For example, VBSSM inferred the central role of ANAC055 in regulating the defence response. This remains to be tested experimentally during *B. cinerea* infection.

This work shows that the scale of computational and experimental challenges involved in a Systems Biology approach to study context-specific GRNs in Arabidopsis is significant, which meant that cycles of inference and validation were hard to achieve within the timeframe of this project. In the future it is hoped that improvements in data availability, modelling approaches and the throughput of validation approaches will allow a tighter coupling of modelling and experimental validation in this area.

Bibliography

H Abe, K Yamaguchi-Shinozaki, T Urao, T Iwasaki, D Hosokawa, and K Shinozaki. Role of Arabidopsis MYC and MYB homologs in drought and abscisic acid-regulated gene expression. *Plant Cell*, 9:1859–1868, 1997.

H Abe, T Urao, T Ito, M Seki, K Shinozaki, and K Yamaguchi-Shinozaki. Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signaling. *Plant Cell*, 15(1):63–78, 2003.

R B Abramovitch, J C Anderson, and G B Marti. Bacterial elicitation and evasion of plant innate immunity. *Nat Rev Mol Cell Biol*, 7:601–611, 2006.

S AbuQamar, X Chen, R Dhawan, B Bluhm, J Salmeron, S Lam, R A Dietrich, and T Mengiste. Expression profiling and mutant analysis reveals complex regulatory networks involved in Arabidopsis response to Botrytis infection. *Plant J*, 48(1): 28–44, 2006.

S Aikawa, M J Kobayashi, A Satake, K K Shimizu, and H Kudoh. Robust control of the seasonal expression of the Arabidopsis FLC gene in a fluctuating environment. *Proc Natl Acad Sci USA*, 25(107):11632–11637, 2010.

J Allemeersch, S Durinck, R Vanderhaeghen, P Alard, R Maes, K Seeuws, T Bogaert, K Coddens, K Deschouwer, P Hummelen, M Vuylsteke, Y Moreau, J Kwekkeboom, A H M Wijfjes, S May, J Beynon, P Hilson, and M T R Kuiper. Benchmarking the CATMA microarray. a novel tool for Arabidopsis transcriptome analysis. *Plant Physiol*, 137:588–601, 2005.

U Alon. Network motifs: theory and experimental approaches. *Nat Rev Genet*, 8: 450–461, 2007.

J M Alonso, A N Stepanova, T J Leisse, C J Kim, H Chen, P Shinn, D K Stevenson, J Zimmerman, P Barajas, R Cheuk, C Gadrinab, C Heller, A Jeske, E Koesema, C C Meyers, H Parker, L Prednis, Y Ansari, N Choy, H Deen, M Geralt, N Hazari, E Hom, M Karnes, C Mulholland, R Ndubaku, I Schmidt, P Guzman, L Aguilar-Henonin, M Schmid, D Weigel, D E Carter, T Marchand, E Risseeuw, D Brogden,

A Zeko, W L Crosby, C C Berry, and J R Ecker. Genome-wide insertional mutagenesis of Arabidopsis thaliana. *Science*, 301:653–657, 2003.

E Andreasson, T Jenkins, P Brodersen, S Thorgrimsen, N H T Petersen, S Zhu, J-L Qiu, P Micheelsen, A Rocher, M Petersen, M-A Newman, H B Nielsen, H Hirt, I Somssich, O Mattsson, and J Mundy. The MAP kinase substrate MKS1 is a regulator of plant defense responses. *EMBO J*, 24(14):2579–2589, 2005.

M N Arbeitman, E E M Furlong, F Imam, E Johnson, B H Null, B S Baker, M A Krasnow, M P Scott, R W Davis, and K P White. Gene expression during the life cycle of Drosophila melanogaster. *Science*, 297(5590):2270–2275, 2002.

T Asai, G Tena, J Plotnikova, M R Willmann, W-L Chiu, L Gomez-Gomez, T Boller, F M Ausubel, and J Sheen. MAP kinase signalling cascade in Arabidopsis innate immunity. *Nature*, 415:977–983, 2001.

M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene Ontology: tool for the unification of biology. *Nat Genet*, 25:25–29, 2000.

K Audenaert, G De Meyer, and M M Höfte. Abscisic acid determines basal susceptibility of tomato to Botrytis cinerea and suppresses salicylic acid-dependent signalling mechanisms. *Plant Physiol*, 128:491–501, 2002.

G Badis, M F Berger, A A Philippakis, S Talukder, A R Gehrke, S A Jaeger, E T Chan, G Metzler, A Vedenko, X Chen, H Kuznetsov, C-F Wang, D Coburn, D E Newburger, Q Morris, T R Hughes, and M L Bulyk. Diversity and complexity in DNA recognition by transcription factors. *Science*, 324:1720–1723, 2009.

T Bailey, N Williams, C Misleh, and W Li. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res*, 34(Web Server issue): W369–W373, 2006.

T Bailey, M Boden, F Buske, M Frith, C Grant, L Clementi, J Ren, W Li, and W Noble. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*, 37(Web Server issue):W202–W208, 2009.

R Balasubramaniyan, E Hüllermeier, N Weskamp, and J Kämper. Clustering of gene expression data using a local shape-based similarity measure. *Bioinformatics*, 21 (7):1069–1077, 2004.

S Balazadeh, H Siddiqui, A D Allu, L P Matallana-Ramirez, C Caldana, M Mehrnia, M I Zanor, B Köhler, and B Mueller-Roeber. A gene regulatory network controlled by the NAC transcription factor ANAC092/AtNAC2/ORE1 during salt-promoted senescence. *Plant J*, 62(2):250–264, 2010.

S Bartnicki-Garcia. Cell wall chemistry, morphogenesis and taxonomy of fungi. *Annu Rev Microbiol*, 22:87–108, 1968.

K A Barton, A N Binns, A J M Matzke, and M-D Chilton. Regeneration of intact tobacco plants containing full length copies of genetically engineered T-DNA, and transmission of T-DNA to R1 progeny. *Cell*, 32(4):1033–1043, 1983.

L R Baugh, A A Hill, D K Slonim, E L Brown, and C P Hunter. Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome. *Development*, 130:889–900, 2003.

G W Beadle and E L Tatum. Genetic control of biochemical reactions in Neurospora. *Proc Natl Acad Sci USA*, 27(11):499–506, 1941.

M Beal, F Falciani, Z Ghahramani, C Rangel, and D L Wild. A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics*, 21(3):28–44, 2005.

N Bechtold and G Pelletier. In planta Agrobacterium-mediated transformation of adult Arabidopsis thaliana plants by vacuum infiltration. *Methods Mol Biol*, 82: 259–266, 1998.

C A Beelman and R Parker. Degradation of mRNA in Eukaryotes. *Cell*, 81:179–188, 1995.

R Bellman. *Adaptive control processes: a guided tour*. Princeton University Press, Princeton, New Jersey, 1961.

Y Benjamini and Y Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc B Met*, 57(1):289–300, 1995.

A Berr, E J McCallum, A Alioua, D Heintz, T Heitz, and W-H Shen. Arabidopsis histone methyltransferase SDG8 mediates induction of the jasmonate/ethylene-pathway genes in plant defense response to necrotrophic fungi. *Plant Physiol*, 154 (3):1403–1414, 2010.

M Berrocal-Lobo, A Molina, and R Solano. Constitutive expression of ETHYLENE-RESPONSE-FACTOR1 in Arabidopsis confers resistance to several necrotrophic fungi. *Plant J*, 29(1):23–32, 2002.

M Bessire, C Chassot, A-C Jacquat, M Humphry, S Borel, J M-C Petétot, J-P Métraux, and C Nawrath. A permeable cuticle in Arabidopsis leads to a strong resistance to Botrytis cinerea. *EMBO J*, 26(8):2158–2168, 2007.

M Bevan and S Walsh. The Arabidopsis genome: a foundation for plant research. *Genome Res*, 15:1632–1642, 2005.

M W Bevan, R B Flavell, and M-D Chilton. A chimaeric antibiotic resistance gene as a selectable marker for plant cell transformation. *Nature*, 304:184–187, 1983.

Z Bieniawska, C Espinoza, A Schlereth, R Sulpice, D K Hincha, and M A Hannah. Disruption of the Arabidopsis circadian clock is responsible for extensive variation in the cold-responsive transcriptome. *Plant Physiol*, 147(1):263–279, 2008.

J N Biraben. An essay concerning mankind's demographic evolution. *J Hum Evol*, 9(8):655–663, 1980.

R P Birkenbihl and I E Somssich. Transcriptional plant responses critical for resistance towards necrotrophic pathogens. *Frontiers in Plant Science*, 2, 2011.

R Blasberg and M K Gobbert. Parallel performance studies for a clustering algorithm. Technical report, number HPCF-2008-5, UMBC High Performance Computing Facility, University of Maryland, Baltimore County, 2008.

U Bodenhofer, A Kothmeier, and S Hochreiter. APCluster: an R package for affinity propagation clustering. *Bioinformatics*, 27(17):2463–2464, 2011.

T Boller and G Feli. A renaissance of elicitors: Perception of microbe-associated molecular patterns and danger signals by pattern-recognition receptors. *Annu Rev Plant Biol*, 60:379–406, 2009.

C E Bonferroni. *Teoria statistica delle classi e calcolo delle probabilita.* Libreria internazionale Seeber, 1936.

P Bowbrick. The causes of famine: A refutation of professor Sen's theory. *Food Policy*, 11(2):105–124, 1986.

E Breeze, E Harrison, S McHattie, L Hughes, R Hickman, Claire Hill, S Kiddle, Y Kim, C A Penfold, D Jenkins, C Zhang, K Morris, C Jenner, S Jackson, B Thomas, A Tabrett, Roxane Legaie, J D Moore, D L Wild, S Ott, D Rand, J Beynon, K Denby, A Mead, and V Buchanan-Wollaston. High-resolution temporal profiling of transcripts during Arabidopsis leaf senescence reveals a distinct chronology of processes and regulation. *Plant Cell*, 23:873–894, 2011.

R G Brown and P Y Hwang. *Introduction to Random Signals and Applied Kalman Filtering.* John Wiley and Sons, NY, 1997.

A Brutus, F Sicilia, A Macone, F Cervone, and G D Lorenzo. A domain swap approach reveals a role of the plant wall-associated kinase 1 (WAK1) as a receptor of oligogalacturonides. *Proc Natl Acad Sci USA*, 107(20):9452–9457, 2010.

Q Bu, H Jiang, C.B Li, Q Zhai, J Zhang, X Wu, J Sun, Q Xie, and C Li. Role of the Arabidopsis thaliana NAC transcription factors ANAC019 and ANAC055 in regulating jasmonic acid-signaled defense responses. *Cell Res*, 18:756–767, 2008.

I Cantone, L Marucci, F Iorio, M A Ricci, V Belcastro, M Bansal, S Santini, M Bernardo, D Bernardo, and M P Cosma. A Yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell*, 137(1): 172–181, 2009.

H Cao, J Glazebrook, J D Clarke, S Volko, and X Dong. The Arabidopsis NPR1 gene that controls systemic acquired resistance encodes a novel protein containing ankyrin repeats. *Cell*, 88(1):57–63, 1997.

J Carrera, G Rodrigo, A Jaramillo, and S Elena. Reverse-engineering the Arabidopsis thaliana transcriptional network under changing environmental conditions. *Genome Biol*, 10(9), 2009.

F Cervone, M.G Hahn, G De Lorenzo, A Darvill, and P Albersheim. Host-pathogen interactions: XXXIII. A plant protein converts a fungal pathogenesis factor into an elicitor of plant defense responses. *Plant Physiol*, 90(2):542–548, 1989.

C Chen and Z Chen. Isolation and characterization of two pathogen- and salicylic acid-induced genes encoding WRKY DNA-binding proteins from tobacco. *Plant Mol Biol*, 42:387–396, 2000.

M-D Chilton. A vector for introducing new genes into plants. *Sci Am*, 248:50–59, 1983.

M-D Chilton, R K Saiki, N Yadav, M P Gordon, and F Quetier. T-DNA from Agrobacterium Ti plasmid is in the nuclear DNA fraction of crown gall tumor cell. *Proc Natl Acad Sci USA*, 77(7):4060–4064, 1980.

A Chini, S Fonseca, G Fernndez, B Adie, J M Chico, O Lorenzo, G Garca-Casado, I Lpez-Vidriero, F M Lozano, M R Ponce, J L Micol, and R Solano. The JAZ family of repressors is the missing link in jasmonate signalling. *Nature*, 448(7154): 666–671, 2007.

S Chisholm, G Coaker, B Day, and B Staskawicz. Host-microbe interactions: Shaping the evolution of the plant immune response. *Cell*, 124(4):803–814, 2006.

I Ciolkowski, D Wanke, R P Birkenbihl, and I E Somssich. Studies on DNA-binding selectivity of WRKY transcription factors lend structural clues into WRKY-domain function. *Plant Mol Biol*, 68:81–92, 2008.

G Cleaskens and N L Hjort. *Model Selection and Model Averaging*. Cambridge University Press, Cambridge, UK, 2008.

S J Clough and A F Bent. Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. *Plant J*, 16(6):735–743, 1998.

A Coego, V Ramirez, M J Gil, V Flors, B Mauch-Mani, and P Vera. An Arabidopsis homeodomain transcription factor, OVEREXPRESSOR OF CATIONIC PEROXIDASE 3, mediates resistance to infection by necrotrophic pathogens. *Plant Cell*, 17:2123–2137, 2005.

P Collas. The current state of chromatin immunoprecipitation. *Mol Biotechnol*, 45 (1):87–100, 2010.

R S Cormack, T Eulgem, P J Rushton, P Köchner, K Hahlbrock, and I E Somssich. Leucine zipper-containing WRKY proteins widen the spectrum of immediate early elicitor-induced WRKY transcription factors in Parsley. *Biochim Biophys Acta*, 1576:92–100, 2002.

K Curvers, H Seifi, G Mouille, R D Rycke, B Asselbergh, A V Hecke, D Vanderschaeghe, H R H ofte, N Callewaert, F V Breusegem, and M M Hofte. ABA-deficiency causes changes in cuticle permeability and pectin composition that influence tomato resistance to Botrytis cinerea. *Plant Physiol*, 154(2):847–860, 2010.

S Datta and S Datta. Evaluation of clustering algorithms for gene expression data. *BMC Bioinformatics*, 7(Suppl 4):S17, 2006.

C O Daub, R Steuer, J Selbig, and S Kloska. Estimating mutual information using b-spline functions – an improved similarity measure for analysing gene expression data. *BMC Bioinformatics*, 5(118), 2004.

E H Davidson and D H Erwin. Gene regulatory networks and the evolution of animal body plans. *Science*, 311:796–800, 2006.

S de Pater, V Greco, K Pham, and J Memelink. Characterization of a zinc-dependent transcriptional activator from Arabidopsis. *Nucleic Acids Res*, 24(23):751–7, 1996.

A Decreux, A Thomas, B Spies, R Brasseur, P V Cutsem, and J Messiaen. In vitro characterization of the homogalacturonan-binding domain of the wall-associated

kinase WAK1 using site-directed mutagenesis. *Phytochemistry*, 67(11):1068–1079, 2006.

T P Delaney, S Uknes, B Vernooij, L Friedrich, K Weyman, D Negrotto, T Gaffney, M Gut-Rella, H Kessmann, E Ward, and J Ryals. A central role of salicylic acid in plant disease resistance. *Science*, 266:1247–1250, 1994.

A P Dempster, N M Laird, and D B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc*, 39(1):1–38, 1977.

K J Denby, P Kumar, and D J Kliebenstein. Identification of Botrytis cinerea susceptibility loci in Arabidopsis thaliana. *Plant J*, 38(3):473–86, 2004.

B Deplancke, D Dupuy, M Vidal, and A J M Walhout. A Gateway-compatible Yeast one-hybrid system. *Genome Res*, 14:2093–2101, 2004.

R Dhawan, H Luo, A M Foerster, S AbuQamar, H-N Du, S D Briggs, O M Scheid, and T Mengiste. HISTONE MONOUBIQUITINATION1 interacts with a subunit of the mediator complex and regulates defense against necrotrophic fungal pathogens in Arabidopsis. *Plant Cell*, 21:1000–1019, 2009.

B Dombrecht, G P Xue, S J Sprague, J A Kirkegaard, J J Ross, J B Reid, G P Fitt, N Sewelam, P M Schenk, J M Manners, and K Kazan. MYC2 differentially modulates diverse jasmonate-dependent functions in Arabidopsis. *Plant Cell*, 19: 2225–2245, 2007.

S Droby and A Lichter. Post-harvest Botrytis infection: Etiology, development and management. In *Botrytis: Biology, Pathology and Control*, pages 349–367. Springer Netherlands, 2004.

C Dubos, J L Gourrierec, A Baudry, G Huep, E Lanet, I Debeaujon, J-M Routaboul, A Alboresi, B Weisshaar, and L Lepiniec. MYBL2 is a new regulator of flavonoid biosynthesis in Arabidopsis thaliana. *Plant J*, 55:940–953, 2008.

W S Dynan and R Tjian. Control of eukaryotic messenger RNA synthesis by sequence-specific DNA-binding proteins. *Nature*, 316:774–778, 1985.

M B Eisen, P T Spellman, P O Brown, and D Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA*, 95:14863–14868, 1998.

C Espinosa-Sotoa, P Padilla-Longoriab, and E R Alvarez-Buylla. A gene regulatory network model for cell-fate determination during Arabidopsis thaliana flower development that is robust and recovers experimental gene expression profiles. *Plant Cell*, 16(11):2923–2939, 2004.

T Eulgem, P Rushton, S Robatzek, and I E Somssich. The WRKY superfamily of plant transcription factors. *Trends Plant Sci*, 5(5):199–206, 2000.

R E Evenson and D Gollin. Assessing the impact of the green revolution, 1960 to 2000. *Science*, 300(5620):758–762, 2003.

S Ferrari, J M Plotnikova, G Lorenzo, and F M Ausubel. Arabidopsis local resistance to Botrytis cinerea involves salicylic acid and camalexin and requires EDS4 and PAD2, but not SID2, EDS5 or PAD4. *Plant J*, 35(2):193–205, 2003a.

S Ferrari, D Vairo, F M Ausubel, F Cervone, and G De Lorenzo. Tandemly duplicated Arabidopsis genes that encode polygalacturonase-inhibiting proteins are regulated coordinately by different signal transduction pathways in response to fungal infection. *Plant Cell*, 15(1):93–106, 2003b.

S Ferrari, R Galletti, C Denoux, G D Lorenzo, F M Ausubel, and J Dewdney. Resistance to Botrytis cinerea induced in Arabidopsis by elicitors is independent of salicylic acid, ethylene, or jasmonate signaling but requires PHYTOALEXIN DEFICIENT3. *Plant Physiol*, 144:367–379, 2007.

B Feys, C E Benedetti, C N Penfold, and J G Turner. Arabidopsis mutants selected for resistance to the phytotoxin coronatine are male sterile, insensitive to methyl jasmonate, and resistant to a bacterial pathogen. *Plant Cell*, 6:751–759, 1994.

S Fields and O Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246, 1989.

G R Fink. Anatomy of a revolution. *Genetics*, 149(2):473–7, 1998.

R R Finkelstein, M L Wang, T J Lynch, S Rao, and H M Goodman. The Arabidopsis abscisic acid response locus ABI4 encodes an APETALA 2 domain protein. *Plant Cell*, 10:1043–1054, 1998.

B Frey and D Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.

M Fujita, Y Fujita, K Maruyama, M Seki, K Hiratsu, M Ohme-Takagi, L-S P Tran, K Yamaguchi-Shinozaki, and K Shinozaki. A dehydration-induced NAC protein, RD26, is involved in a novel ABA-dependent stress-signaling pathway. *Plant J*, 39:863–876, 2004.

R Galletti, S Ferrari, and G De Lorenzo. Arabidopsis MPK3 and MPK6 play different roles in basal and oligogalacturonide- or flagellin-induced resistance against Botrytis cinerea. *Plant Physiol*, 2011.

Y Galon, R Nave, J M Boyce, D Nachmias, M R Knight, and H Fromm. Calmodulin-binding transcription activator (CAMTA) 3 mediates biotic defense responses in Arabidopsis. *FEBS Lett*, 582(6):943–948, 2008.

M M Garner and A Revzin. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucl Acids Res*, 9(13):3047–3060, 1981.

A P Gasch, P T Spellman, C M Kao, O Carmel-Harel, M B Eisen, G Storz, D Botstein, and P O Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*, 11(12):4241–4257, 2000.

D Ghosh and A M Chinnaiyan. Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*, 18(2):275–286, 2002.

E Giraud, S Ng, C Carrie, O Duncan, J Low, C P Lee, O V Aken, A H Millar, M Murcha, and J Whelan. TCP transcription factors link the regulation of genes encoding mitochondrial proteins with the circadian clock in Arabidopsis thaliana. *Plant Cell*, 22(12):3921–3934, 2010.

A Gitter, Z Siegfried, M Klutstein, O Fornes, B Olivia, I Simon, and Z Bar-Joseph. Backup in gene regulatory networks explains differences between binding and knockout results. *Mol Sys Bio*, 5:276, 2009.

J Glazebrook. Contrasting mechanisms of defense against biotrophic and necrotrophic pathogens. *Annu Rev Phytopathol*, 43:205–227, 2005.

J Glazebrook and F M Ausubel. Isolation of phytoalexin-deficient mutants of Arabidopsis thaliana and characterization of their interactions with bacterial pathogens. *Proc Natl Acad Sci USA*, 91(19):8955–8959, 1994.

J Glazebrook, W Chen, B Estes, H-S Chang, C Nawrath, J-P Métraux, T Zhu, and F Katagiri. Topology of the network integrating salicylate and jasmonate signal transduction derived from global expression phenotyping. *Plant J*, 34(2):217–28, 2003.

M Godoy, J M Franco-Zorrilla, J Pérez-Pérez, J C Oliveros, O Lorenzo, and R Solano. Improved protein-binding microarrays for the identification of DNA-binding specificities of transcription factors. *Plant J*, 66:700–711, 2011.

E M Govrin and A Levine. The hypersensitive response facilitates plant infection by the necrotrophic pathogen Botrytis cinerea. *Curr Biol*, 10(13):751–7, 2000.

E M Govrin and A Levine. Infection of Arabidopsis with a necrotrophic pathogen, Botrytis cinerea, elicits various defense responses but does not induce systemic acquired resistance (SAR). *Plant Mol Biol*, 48:267–276, 2002.

T Graf and T Enver. Forcing cells to change lineages. *Nature*, 462:587–594, 2009.

R A Gútierrez, D E Shasha, and G M Coruzzi. Systems biology for the virtual plant. *Plant Physiol*, 138:550–554, 2005.

P Guzman and J R Ecker. Exploiting the triple response of Arabidopsis to identify ethylene-related mutants. *Plant Cell*, 2(6):513–523, 1990.

S P Gygi, Y Rochon, B R Franza, and R Aebersold. Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol*, 19(3):1720–1730, 1999.

T Hastie, R Tibshirani, and J H Friedman. *The Elements of Statistical Learning*. Springer-Verlag, 2001.

L E Hawker and R J Hendy. An electron-microscope study of germination of Conidia of Botrytis cinerea. *J Gen Microbiol*, 33:43–46, 1963.

P He, L Shan, N-C Lin, G B Martin, B Kemmerling, T Nürnberger, and J Sheen. Specific bacterial suppressors of MAMP signaling upstream of MAPKKK in Arabidopsis innate immunity. *Cell*, 125(3):563–575, 2006.

N Heard, C Holmes, D Stephens, and D Hand. Bayesian coclustering of Anopheles gene expression time series: Study of immune defense response to multiple experimental challenges. *P Natl Acad Sci USA*, 102(47):16939–16944, 2005.

N A Heard. Iterative reclassification in agglomerative clustering. *J Comput Graph Stat, doi:10.1198/jcgs.2011.09111.*, Ahead of print.

N A Heard, C C Holmes, and D A Stephens. A quantitative study of gene regulation involved in the immune response of Anopheline mosquitoes: An application of Bayesian hierarchical clustering of curves. *J AMSTAT*, 101:18–29, 2006.

D Heckerman, D Geiger, and D M Chickering. Learning Bayesian networks: the combination of knowledge and statistical data. *Mach Learn*, 20:197–243, 1995.

K Higo, Y Ugawa, M Iwamoto, and T Korenaga. Plant cis-acting regulatory DNA elements (PLACE) database. *Nucleic Acids Res*, 27:297–300, 1999.

A Honkela, C Girardot, E H Gustafson, Y-H Liu, E E M Furlong, N D Lawrence, and M Rattray. Model-based method for transcription factor target identification with limited data. *Proc Natl Acad Sci USA*, 107(17):7793–7798, 2010.

T Hruz, O Laule, G Szabo, F Wessendorp, S Bleuler, L Oertle, P Widmayer, W Gruissem, and P Zimmermann. Genevestigator V3: A reference expression database for the meta-analysis of transcriptomes. *Adv Bioinformatics*, page 420747, 2008.

F Jacob and J Monod. On the regulation of gene activity. *Cold Spring Harb Symp Quant Biol*, 26:193–211, 1961.

C A Janeway and R Medzhito. Innate immune recognition. *Annu Rev Immunol*, 20:197–216, 2002.

W Jarvis. *Botryotinia and Botrytis species. Taxonomy and pathogenicity.* Research Branch, Canada Dept. of Agriculture: obtainable from Information Division, Canada Dept. of Agriculture, 1977.

R A Jefferson, T A Kavanagh, and M W Bevan. GUS fusions: beta-glucuronidase as a sensitive and versatile gene fusion marker in higher plants. *EMBO J*, 6(13): 3901–3907, 1987.

W Johannsen. The genotype conception of heredity. *Am Nat*, 45(531):129–159, 1911.

J D G Jones and J L Dangl. The plant immune system. *Nature*, 444(7117):323–329, 2006.

R Kafri, M Springer, and Y Pilpel. Genetic redundancy: new tricks for old genes. *Cell*, 136(3):389–392, 2009.

M S Katari, S D Nowicki, F F Aceituno, D Nero, J Kelfer, L P Thompson, J M Cabello, R S Davidson, A P Goldberg, D E Shasha, G M Coruzzi, and R A Gútierrez. VirtualPlant: a software platform to support systems biology research. *Plant Physiol*, 152(2):500–515, 2010.

L Kaufman and P J Roussseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley, New York., 1990.

S J Kiddle, O P F Windram, S McHattie, A Mead, J Beynon, V Buchanan-Wollaston, K J Denby, and S Mukherjee. Temporal clustering by affinity propagation reveals transcriptional modules in Arabidopsis thaliana. *Bioinformatics*, 26(3):355–362, 2010.

E Koch and A J Slusarenko. Fungal pathogens of Arabidopsis thaliana (L.) Heyhn. *Bot Helv*, 100(2):257–268, 1990.

P J Krysan, J C Young, and M R Sussman. T-DNA as an insertional mutagen in Arabidopsis. *Plant Cell*, 11:2283–2290, 1999.

Z Lai, Y Li, F Wang, Y Cheng, B Fan, J-Q Yu, and Z Chen. Arabidopsis sigma factor binding proteins are activators of the WRKY33 transcription factor in plant defense. *Plant Cell*, Ahead of print, 2011a.

Z Lai, F Wang, Z Zheng, B Fan, and Z Chen. A critical role of autophagy in plant resistance to necrotrophic fungal pathogens. *Plant J*, 66(6):953–968, 2011b.

J Leemans, C Shaw, R Deblaere, H De Greve, J P Hernalsteens, M Maes, M van Montagu, and J Schell. Site-specific mutagenesis of Agrobacterium Ti plasmids and transfer of genes to plant cells. *J Mol Appl Genet*, 1(2):149–164, 1981.

J J Li and I Herskowitz. Isolation of ORC6, a component of the Yeast origin recognition complex by a one-hybrid system. *Science*, 262(5141):1870–1874, 1993.

S Li, Q Fu, L Chen, W Huang, and D Yu. Arabidopsis thaliana WRKY25, WRKY26, and WRKY33 coordinate induction of plant thermotolerance. *Planta*, 233(6): 1237–1252, 2011.

M Libault, J Wan, T Czechowski, M Udvardi, and G Stacey. Identification of 118 Arabidopsis transcription factor and 30 ubiquitin-ligase genes responding to chitin, a plant-defense elicitor. *Mol Plant Microbe In*, 20(8):900–911, 2007.

H Lillefors. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *J Am Stat Assoc*, 62:399–402, 1967.

B Lippok, R P Birkenbihl, G Rivory, J Brümmer, E Schmelzer, E Logemann, and I E Somssich. Expression of AtWRKY33 encoding a pathogen- or PAMP-responsive WRKY transcription factor is regulated by a composite DNA motif containing W box elements. *Mol Plant Microbe In*, 20(4):420–9, 2007.

Y Liu, M Koornneef, and W J Soppe. The absence of histone H2B monoubiquitination in the Arabidopsis hub1 (rdo4) mutant reveals a role for chromatin remodeling in seed dormancy. *Plant Cell*, 19:433–444, 2007.

Y G Liu, N Mitsukawa, T Oosumi, and R F Whittier. Efficient isolation and mapping of Arabidopsis thaliana T-DNA insert junctions by thermal asymmetric interlaced PCR. *Plant J*, 8:457–463, 1995.

F Llorente, P Muskett, A Sánchez-Vallet, G López, B Ramos, C Sánchez-Rodríguez, L Jordá, J Parker, and A Molina. Repression of the auxin response pathway increases Arabidopsis susceptibility to necrotrophic fungi. *Mol Plant*, 1(3):496–509, 2008.

J C W Locke, L Kozma-Bognar, P Gould, B Feher, E Kevei, F Nagy, M S Turner, A Hall, and A J Millar. Experimental validation of a predicted feedback loop in the multi-oscillator clock of Arabidopsis thaliana. *Mol Syst Biol*, 2(59):1–6, 2006.

O Lorenzo, R Piqueras, J J Sánchez-Serrano, and R Solano. ETHYLENE RESPONSE FACTOR1 integrates signals from ethylene and jasmonate pathways in plant defense. *Plant Cell*, 15(1):165–178, 2003.

O Lorenzo, J M Chico, J J Sánchez-Serrano, and R Solano. JASMONATE-INSENSITIVE1 encodes a MYC transcription factor essential to discriminate between different jasmonate-regulated defense responses in Arabidopsis. *Plant Cell*, 16(7):1938–1950, 2004.

S C Madeira and A L Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans Comput Biol Bioinform*, 1(1):24–45, 2004.

J M Manners, I A Penninckx, K Vermaere, K Kazan, R L Brown, A Morgan, D J Maclean, M D Curtis, B P Cammue, and W F Broekaert. The promoter of the plant defensin gene PDF1.2 from Arabidopsis is systemically activated by fungal pathogens and responds to methyl jasmonate but not to salicylic acid. *Plant Mol Biol*, 38(6):1071–1080, 1998.

G Mao, X Meng, Y Liu, Z Zheng, Z Chen, and S Zhang. Phosphorylation of a WRKY transcription factor by two pathogen-responsive MAPKs drives phytoalexin biosynthesis in Arabidopsis. *Plant Cell*, 23:1639–1653, 2011.

D Marbach, T Schaffter, C Mattiussi, and D Floreano. Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *J Comput Biol*, 16(2):229–239, 2009.

D Marbach, R J Prill, T Schaffter, C Mattiussi, D Floreano, and G Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. *Proc Natl Acad Sci USA*, 107(14):6286–6291, 2010.

C Martin and J Paz-Ares. MYB transcription factors in plants. *Trends Genet*, 13 (2):67–73, 1997.

V Matys, O V Kel-Margoulis, E Fricke, I Liebich, S Land, A Barre-Dirrie, I Reuter, D Chekmenev, M Krull, K Hornischer, N Voss, P Stegmaier, B Lewicki-Potapov, H Saxel, A E Kel, and E Wingender. TRANSFAC and its module TRANSCompel: Transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, 34:(Database issue) D108–D110, 2006.

P Matzinger. Friendly and dangerous signals: is the tissue in control? *Nat Immunol*, 8:11–13, 2007.

E Mazzucotelli, A M Mastrangelo, C Crosatti, D Guerra, A M Stanca, and L Cattivelli. Abiotic stress response in plants: When post-transcriptional and post-translational regulations control transcription. *Plant Sci*, 4(174):420–431, 2008.

C McClung. Comes a time. *Curr Opin Plant Biol*, 11:514–520, 2008.

J M McDowell and J L Dangl. Signal transduction in the plant immune response. *Trends Biochem Sci*, 25(2):79–82, 2000.

K C McGrath, B Dombrecht, J M Manners, P M Schenk, C I Edgar, D J. Maclean, W-R Scheible, M K Udvardi, and K Kazan. Repressor- and activator-type ethylene response factors functioning in jasmonate signaling and disease resistance identified via a genome-wide screen of Arabidopsis transcription factor gene expression. *Plant Physiol*, 139:949–959, 2005.

C Y McLean, P L Reno, A A Pollen, A I Bassan, T D Capellini, C Guenther, V B Indjeian, X Lim, D B Menke, B T Schaar, A M Wenger, G Bejerano, and D M Kingsley. Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature*, 471:216–219, 2011.

J Memelink. Regulation of gene expression by jasmonate hormones. *Phytochemistry*, 70:1560–1570, 2009.

J Meng, S J Gao, and Y Huang. Enrichment constrained time-dependent clustering analysis for finding meaningful temporal transcription modules. *Bioinformatics*, 25(12):1521–1527, 2009.

T Mengiste, X Chen, J Salmeron, and R Dietrich. The BOTRYTIS SUSCEPTIBLE1 gene encodes an R2R3MYB transcription factor protein that is required for biotic and abiotic stress responses in Arabidopsis. *Plant Cell*, 15:2551–2565, 2003.

T Meshi and M Iwabuchi. Plant transcription factors. *Plant Cell Physiol*, 36(8): 1405–1420, 1995.

R G Miller. *Simultaneous statistical inference, 2nd ed.* Springer Verlag, 1981.

P J Mitchell and R Tjian. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science*, 245:371–378, 1989.

N Mitsuda and M Ohme-Takagi. Functional analysis of transcription factors in Arabidopsis. *Plant Cell Physiol*, 50(7):1232–1248, 2009.

T H Morgan. Sex-limited inheritance in Drosophila. *Science*, 32:120–122, 1910.

E R Morrissey, M A Juárez, K J Denby, and N J Burroughs. On reverse engineering of gene interaction networks using timecourse data with repeated measurement. *Bioinformatics*, 26(18):2305–2312, 2010.

S Mukherjee and T Speed. Network inference using informative priors. *Proceedings of the National Academy of Sciences*, 105(38):14313–14318, 2008.

J M K Mulema and K J Denby. Spatial and temporal transcriptomic analysis of the Arabidopsis thaliana-Botrytis cinerea interaction. *Mol Biol Rep*, 39(4):4039–4049, 2012.

H J Muller. Artificial transmutation of the gene. *Science*, 66:84–87, 1927.

K Murphy and S Mian. Modelling gene expression data using dynamic Bayesian networks. Technical report, Univeristy of California, Berkeley, CA, 1999.

R Narsai, K A Howell, A H Millar, N O'Toole, I Small, and J Whelan. Genome-wide analysis of mRNA decay rates and their determinants in Arabidopsis thaliana. *Plant Cell*, 19:3418–3436, 2007.

C Needham, I Manfield, A Bulpitt, P Gilmartin, and D Westhead. From gene expression to gene regulatory networks in Arabidopsis thaliana. *BMC Syst Biol*, 3(1):85, 2009.

Y Nishizawa, A Kawakami, H Tadaaki, D-Y He, N Shibuya, and E Minami. Regulation of the chitinase gene expression in suspension-cultured rice cells by N-acetylchitooligosaccharides: differences in the signal transduction pathways leading to the activation of elicitor-responsive genes. *Plant Mol Biol*, 39:907–914, 1999.

T Oates, L Firoiu, and P R Cohen. Clustering time series with hidden Markov models and dynamic time warping. In *Proceedings of the IJCAI-99 Workshop on Neural, Symbolic and Reinforcement Learning Methods for Sequence Learning*, pages 17–21, 1999.

M Ohme-Takagi and H Shinshi. Ethylene-inducible DNA binding proteins that interact with an ethylene-responsive element. *Plant Cell*, 7(2):173–182, 1995.

M El Oirdi, T El Rahman, L Rigano, A E Hadrami, M C Rodriguez, F Daayf, A Vojnov, and K Bouarab. Botrytis cinerea manipulates the antagonistic effects between immune pathways to promote disease development in tomato. *Plant Cell*, 2011.

A N Olsen, H A Ernst, L L Leggio, and K Skriver. NAC transcription factors: structurally distinct, functionally diverse. *Trends Plant Sci*, 10(2):79–87, 2004.

A N Olson, H A Ernst, L L Leggio, and K Skriver. DNA-binding specificity and molecular functions of NAC transcription factors. *Plant Sci*, 169:785–797, 2005.

D A Orlando, C Y Lin, A Bernard, J Y Wang, J E S Socolar, E S Iversen, A J Hartemink, and S B Haase. Global control of cell-cycle transcription by coupled CDK and network oscillators. *Nature*, 453:944–947, 2008.

B Ou, K-Q Yin, S-N Liu, Y Yang, T Gu, J M W Hui, L Zhang, J Miao, Y Kondou, M Matsui, H-Y Gu, and L-J Qu. A high-throughput screening system for Arabidopsis transcription factors and its application to Med25-dependent transcriptional regulation. *Mol Plant*, 4:546–555, 2011.

P A Passarinho and S C D Vries. Arabidopsis chitinases: a genomic survey. *The Arabidopsis Book*, 1:e0023, 2002.

L Pauwels, G F Barbero, J Geerinck, S Tilleman, W Grunewald, A C Pérez, J M Chico, R V Bossche, J Sewell, E Gil, G García-Casado, E Witters, D Inzé, J A Long, G De Jaeger, R Solano, and A Goossens. NINJA connects the co-repressor TOPLESS to jasmonate signalling. *Nature*, 464:788–791, 2010.

J Pearl. *Causality: models, reasoning, and inference*, volume 47. Cambridge Univ Press, 2000.

C A Penfold and D L Wild. How to infer gene networks from expression profiles, revisited. *Interface Focus*, 1(6):857–870, 2011.

C A Penfold, V Buchanan-Wollaston, K J Denby, and D L Wild. Nonparametric bayesian inference for perturbed and orthologous gene regulatory networks. *Bioinformatics*, 28:i233–i241, 2012.

C Pieterse, A Leon-Reyes, S Ent, and S C M Wees. Networking by small-molecule hormones in plant immunity. *Nat Chem Biol*, 5(5):308–316, 2009.

J A Poland, P J Balint-Kurti, R J Wisser, R C Pratt, and R J Nelson. Shades of gray: the world of quantitative disease resistance. *Trends Plant Sci*, 14(1): 1360–1385, 2008.

G Povero, E Loreti, C Pucciariello, A Santaniello, D D Tommaso, G D Tommaso, D Kapetis, F Zolezzi, A Piaggesi, and P Perata. Transcript profiling of chitosan-treated Arabidopsis seedlings. *J Plant Res*, 124(5):619–629, 2011.

M Pré, M Atallah, A Champion, M de Vos, C M J Pieterse, and J Memelink. The AP2/ERF domain transcription factor ORA59 integrates jasmonic acid and ethylene signals in plant defense. *Plant Physiol*, 147:1347–1357, 2008.

R J Prill, D Marbach, J Saez-Rodriguez, P K Sorger, L G Alexopoulos, X Xue, N D Clarke, G Altan-Bonnet, and G Stolovitzky. Towards a rigorous assessment of systems biology models: The DREAM3 challenges. *PLoS one*, 5(2), 2010.

J L Pruneda-Paz, G Breton, A Para, and S A Kay. A functional genomics approach reveals CHE as a component of the Arabidopsis circadian clock. *Science*, 323: 1481–1485, 2009.

M Ptashne. How eukaryotic transcriptional activators work. *Nature*, 335:683–689, 1988.

L Pu and S Brady. Systems biology update: cell type-specific transcriptional regulatory networks. *Plant Physiol*, 152(2):411–419, 2009.

J Qian, M Dolled-Filhart, J Lin, H Yu, and M Gerstein. Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *J Mol Biol*, 314:1053–1066, 2001.

J Qian, J Lin, N M Luscombe, H Yu, and M Gerstein. Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics*, 22(13):1917–1926, 2003.

J-L Qiu, B K Fiil, K Petersen, H B Nielsen, C J Botanga, S Thorgrimsen, K Palma, M C Suarez-Rodriguez, S Sandbech-Clausen, J Lichota, P Brodersen, K D Grasser, O Mattsson, J Glazebrook, J Mundy, and M Petersen. Arabidopsis MAP kinase 4 regulates gene expression through transcription factor release in the nucleus. *EMBO J*, 27(16):2214–21, 2008.

V Ramirez, A Agorio, A Coego, J García-Andrade, M J Hernández, B Balaguer, P B F Ouwerkerk, I Zarra, and P Vera. MYB46 modulates disease susceptibility to Botrytis cinerea in Arabidopsis. *Plant Physiol*, 155(4):1920–1935, 2011.

K Ramonell, M Berrocal-Lobo, S Koh, J Wan, H Edwards, G Stacey, and Shauna Somerville. Loss-of-function mutations in chitin responsive genes show increased susceptibility to the powdery mildew pathogen Erysiphe cichoracearum. *Plant Physiol*, 138:1027–1036, 2005.

J W Reed. Roles and activities of Aux/IAA proteins in Arabidopsis. *Trends Plant Sci*, 6(9):420–425, 2001.

M H V V Regenmortel. Reductionism and complexity in molecular biology. *EMBO Rep*, 5(11):1016–1020, 2004.

D Reiss, N Baliga, and R Bonneau. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics*, 7:280, 2006.

D Ren, Y Liu, K-Y Yang, L Han, G Mao, J Glazebrook, and S Zhang. A fungal-responsive mapk cascade regulates phytoalexin biosynthesis in arabidopsis. *Proc Natl Acad Sci USA*, 105(14):5638–5643, 2008.

J M Ribault and D Hoisington. Marker-assisted selection: new tools and strategies. *Trends Plant Sci*, 3(6):236–239, 1998.

J L Riechmann and E M Meyerowitz. The AP2/EREBP family of plant transcription factors. *Biol Chem*, 379(6):633–646, 1998.

J L Riechmann and O J Ratcliffe. A genomic perspective on plant transcription factors. *Curr Opin Plant Biol*, 3(5):423–434, 2000.

J B Ristain. Tracking historic migrations of the irish potato famine pathogen, Phytophthora infestans. *Microbes Infect*, 4:1369–1377, 2002.

C W Roane. Trends in breeding for disease resistance in crops. *Annu Rev Phytopathol*, 11:463–486, 1973.

H Robbins. An empirical Bayes approach to statistics. *Proc Third Berkeley Symp on Math Statist and Prob*, 1:157–163, 1956.

R W Robinson. *Counting labeled acyclic digraphs, in F Harary (Ed.); New directions in the theory of graphs.* Academic press, Waltham, Massachusetts, US, 1973.

M W Rosegrant and S A Cline. Global food security: Challenges and policies. *Science*, 302(5652):1917–1919, 2003.

H Rosslenbroich and D Stuebler. Botrytis cinerea - history of chemical control and novel fungicides for its management. *Crop Prot*, 19:557–561, 2000.

H C Rowe and D J Kliebenstein. Complex genetics control natural variation in Arabidopsis thaliana resistance to Botrytis cinerea. *Genetics*, 180(4):2237–2250, 2008.

H C Rowe, J W Walley, J Corwin, E K F Chan, K Dehesh, and D J Kliebenstein. Deficiencies in jasmonate-mediated plant defense reveal quantitative variation in Botrytis cinerea pathogenesis. *PLoS Pathog*, 6(4), 2010.

S Roweis and Z Ghahramani. A unifying review of linear Gaussian models. *Neural Comput*, 11:305–345, 1999.

P J Rushton, J T Torres, M Parniske, P Wernert, K Hahlbrock, and I E Somssich. Interaction of elicitor-induced DNA-binding proteins with elicitor response elements in the promoters of Parsley PR1 genes. *EMBO J*, 15(20):5690–5700, 1996.

A Saha, J Wittmeyer, and B R Cairn. Chromatin remodelling: the industrial revolution of DNA around histones. *Nat Rev Mol Cell Biol*, 7:437–447, 2006.

D A Samac and D M Shah. Effect of chitinase antisense RNA expression on disease susceptibility of Arabidopsis plants. *Plant Mol Biol*, 25:587–596, 1994.

P R Sanders, J A Winter, A R Barnason, S G Rogers, and R T Fraley. Comparison of cauliflower mosaic virus 35S and nopaline synthase promoters in transgenic plants. *Nucl Acids Res*, 15(4):1543–1558, 1987.

S E Satchel and E W Nester. The genetic and transcriptional organization of the vir region of the A6 Ti plasmid of Agrobacterium tumefaciens. *EMBO J*, 5(7): 1445–1454, 1986.

A Schlumbaum, F Mauch, U Vögeli, and T Boller. Plant chitinases are potent inhibitors of fungal growth. *Nature*, 324:365–367, 1986.

B Schrammeijer, A Beijersbergen, K B Idler, L S Melchers, D V Thompson, and P J J Hooykaas. Sequence analysis of the vir region from Agrobacterium tumefaciens octopine Ti plasmid pTi15955. *J Exp Bot*, 51(347):1167–1169, 2000.

R Schuhegger, M Nafisi, M Mansourova, B L Petersen, C E Olsen, A Svatoš, B A Halkier, and E Glawischnig. CYP71B15 (PAD3) Catalyzes the Final Step in Camalexin Biosynthesis. *Plant Physiol*, 141:1248–1254, 2006.

E Segal, M Shapira, A Regev, D Pe'er, D Botstein, D Koller, and N Friedman. Module networks: identifying regulatory modules and their condition specific regulators from gene expression data. *Nat Genet*, 34:166–176, 2003.

C Sima, J Hua, and S Jung. Inference of gene regulatory networks using time-series data: A survey. *Curr Genomics*, 10(6):416–429, 2009.

K B Singh, R C Foley, and L O nate Sánchez. Transcription factors in plant defense and stress responses. *Curr Opin Plant Biol*, 5(5):430–436, 2002.

R Solano, A Stepanova, Q Chao, and J R Ecker. Nuclear events in ethylene signaling: a transcriptional cascade mediated by ETHYLENE-INSENSITIVE3 and ETHYLENE-RESPONSE-FACTOR1. *Genes Dev*, 12(23):3703–3714, 1998.

P T Spellman, G Sherlock, M Q Zhang, V R Iyer, K Anders, M B Eisen, P O Brown, D Botstein, and B Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, 9:3273–3297, 1998.

S H Spoel, A Koornneef, S M C Claessens, J P Korzelius, J A V Pelta, M J Mueller, A J Buchalad, J-P Métraux, R Brown, K Kazan, L C V Loona, X Dong, and C M J Pieterse. NPR1 modulates cross-talk between salicylate- and jasmonate-dependent defense pathways through a novel function in the cytosol. *Plant Cell*, 15:760–770, 2003.

O Stegle, K J Denby, D L Wild, Z G, and K M Borgwardt. A robust Bayesian two-sample test for detecting intervals of differential gene expression in microarray time series. *RECOMB 2009*, LNCS 5541:201–216, 2009.

H Steinhaus. Sur la division des corp materiels en parties. *Bull Acad Polon Sci*, 1: 801–804, 1956.

H U Stotz, Y Jikumaru, Y Shimada, E Sasaki, N Stingl, M J Mueller, and Y Kamiya. Jasmonate-dependent and COI1-independent defense responses against Sclerotinia sclerotiorum in Arabidopsis thaliana: Auxin is part of COI1-independent defense signaling. *Plant Cell Physiol*, Ahead of print, 2011.

Student. The probable error of a mean. *Biometrika*, 6(1):1, 1908.

D Swarbreck, C Wilks, P Lamesch, T Z Berandini, M Garcia-Hernandez, H Foerster, D Li, T Meyer, R Muller, L Ploetz, A Radenbaugh, S Singh, V Swing, C Tissier, P Zhang, and Eva Huala. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res*, 36(D1009-D1014), 2008.

H Szemenyei, M Hannon, and J A Long. TOPLESS mediates auxin-dependent transcriptional repression during Arabidopsis embryogenesis. *Science*, 319:1384–1386, 2008.

Y Tabei, S Kitade, Y Nishizawa, N Kikuchi, T Kayano, T Hibi, and K. Akutsu. Transgenic cucumber plants harboring a rice chitinase gene exhibit enhanced resistance to gray mold (Botrytis cinerea). *Plant Cell Reports*, 17:159–164, 1998.

Y C Tai and T P Speed. A multivariate empirical Bayes statistic for replicated microarray time course data. *Ann Statist*, 34(5):2387–2412, 2006.

S Tavazoie, J D Hughes, M J Campbell, R J Cho, and G M Church. Systematic determination of genetic network architecture. *Nat Genet*, 22:281–285, 1999.

A Thalamuthu, I Mukhopadhyay, X Zheng, and G C Tseng. Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, 22(19): 2405–2412, 2006.

B Thomma, K Eggermont, and K Tierens. Requirement of functional ethylene-insensitive 2 gene for efficient resistance of Arabidopsis to infection by Botrytis cinerea. *Plant Physiol*, 121:10931101, 1999.

B P Thomma, K Eggermont, I A Penninckx, B Mauch-Mani, R Vogelsang, B P Cammue, and W F Broekaert. Separate jasmonate-dependent and salicylate-dependent defense-response pathways in Arabidopsis are essential for resistance

to distinct microbial pathogens. *Proc Natl Acad Sci USA*, 95(25):15107–15111, 1998.

S B Tiwari, G Hagen, and T Guilfoyle. The roles of Auxin Response Factor domains in auxin-responsive transcription. *Plant Cell*, 15:533–543, 2003.

G Toledo-Ortiz, E Huq, and P H Quail. The Arabidopsis basic/helix-loop-helix transcription factor family. *Plant Cell*, 15(8):3921–3934, 2003.

L-S P Tran, K Nakashima, Y Sakuma, S D Simpson, Y Fujita, K Maruyama, M Fujita, M Seki, K Shinozaki, and K Yamaguchi-Shinozaki. Isolation and functional analysis of Arabidopsis stress-inducible NAC transcription factors that bind to a drought-responsive cis-element in the early responsive to dehydration stress 1 promoter. *Plant Cell*, 16:2481–2498, 2004.

V G Tusher, R Tibshirani, and G Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA*, 98(9):5116–5121, 2001.

T Ulmasov, G Hagen, and T J Guilfoyle. Dimerization and DNA binding of auxin response factors. *Plant J*, 19(3):309–319, 1999.

J van Kan. Licensed to kill: the lifestyle of a necrotrophic plant pathogen. *Trends Plant Sci*, 11(5):247–253, 2006.

P Veronese, H Nakagami, B Bluhm, and S AbuQamar. The membrane-anchored BOTRYTIS-INDUCED KINASE1 plays distinct roles in Arabidopsis Resistance to necrotrophic and biotrophic pathogens. *Plant Cell*, 18:257–273, 2006.

G Vert, C L Walcher, J Chory, and J L Nemhauser. Integration of auxin and brassinosteroid pathways by Auxin Response Factor 2. *Proc Natl Acad Sci USA*, 105(28):9829–9834, 2008.

J W Walley, S Coughlan, M E Hudson, M F Covington, R Kaspi, G Banu, S L Harmer, and K Dehesh. Mechanical stress induces biotic and abiotic stress responses via a novel cis-element. *PLoS Genet*, 3(10):1800–1812, 2007.

J W Walley, H C Rowe, Y Xiao, E W Chehab1, D J Kliebenstein, D Wagner, and K Dehesh. The chromatin remodeler SPLAYED regulates specific stress signaling pathways. *PLoS Pathog*, 4(12), 2008.

D Wang, N Amornsiripanitch, and X Dong. A genomic approach to identify regulatory nodes in the transcriptional network of systemic acquired resistance in plants. *PLOS Pathog*, 2(11):1042–1050, 2006.

L Wang, D Hua, J He, Y Duan, Z Chen, X Hong, and Z Gong. Auxin Response Factor2 (ARF2) and its regulated homeodomain gene HB33 mediate abscisic acid response in Arabidopsis. *PLoS Genet*, 7(7), 2011.

X Wang, B M V S Basnayake, H Zhang, G Li, W Li, N Virk, T, and F Song. The Arabidopsis ATAF1, a NAC transcription factor, is a negative regulator of defense responses against necrotrophic fungal and bacterial pathogens. *Mol Plant Microbe In*, 22(10):1227–1236, 2009.

Z Wang, P Yang, B Fan, and Z Chen. An oligo selection procedure for identification of sequence-specific DNA-binding activities associated with the plant defence response. *Plant J*, 16(4):515–522, 1998.

J H Ward. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc*, 58(301):236–244, 1963.

A V Werhli and D Husmeier. Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge. *Stat Appl Genet Mol Bio*, 6(1), 2007.

F Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bull*, 1(6): 80–83, 1945.

B Williamson, B Tudzynski, P Tudzynski, and J A L van Kan. Botrytis cinerea: the cause of grey mould disease. *Mol Plant Pathol*, 8(5):561–580, 2007.

O Windram. *Using a systems biology approach to elucidate transcriptional networks regulating plant defence.* PhD thesis, University of Warwick, 2010.

G A Wray. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet*, 8:206–216, 2007.

G A Wray, E Abouheif M W Hahn, J P Balhoff, M Pizer, M V Rockman, and L A Romano. The evolution of transcriptional regulation in Eukaryotes. *Mol Biol Evol*, 20(9):1377–1419, 2003.

H Wu, K Kerr, X Cui, and G Churchill. *MAANOVA: a software package for the analysis of spotted cDNA microarray experiments.* Springer, NY, 2003.

K Xu, X Xu, T Fukao, P Canlas, R Maghirang-Rodriguez, S Heuer, A M Ismail, J Bailey-Serres, P C Ronald, and D J Mackill. SUB1A is an ethylene-response-factor-like gene that confers submergence tolerance to rice. *Nature*, 442:705–708, 2006a.

X Xu, C Chen, B Fan, and Z Chen. Physical and functional interactions between pathogen-induced Arabidopsis WRKY18, WRKY40, and WRKY60 transcription factors. *Plant Cell*, 18:1310–1326, 2006b.

H Yang, S Yang, Y Li, and J Hua. The Arabidopsis BAP1 and BAP2 genes are general inhibitors of programmed cell death. *Plant Physiol*, 145:135–146, 2007.

C Yanhui, Y Xiaoyuan, H Kun, L Meihua, L Jigang, G Zhaofeng, L Zhiqiang, Z Yunfei, W Xiaoxiao, Q Xiaoming, S Yunping, Z Li, D Xiaohui, L Jingchu, D Xing-Wang, C Zhangliang, G Hongya, and Q Li-Jia. The MYB transcription factor superfamily of Arabidopsis: expression analysis and phylogenetic comparison with the rice MYB family. *Plant Mol Biol*, 60(1):107–124, 2006.

G Yona, W Dirks, S Rahman, and D M Lin. Effective similarity measures for expression profiles. *Bioinformatics*, 22(13):1616–1622, 2006.

X Yu, L Li, J Zola, M Aluru, H Ye, A Foudree, H Guo, S Anderson, S Aluru, P Liu, S Rodermel, and Y Yin. A brassinosteroid transcriptional network revealed by genome-wide identification of BESI target genes in Arabidopsis thaliana. *Plant J*, 65(4):634–646, 2011.

M Zander, S Camera, O Lamotte, J-P Metraux, and C Gatz. Arabidopsis thaliana class-II TGA transcription factors are essential activators of jasmonic acid/ethylene-induced defense responses. *Plant J*, 61(2):200–210, 2010.

A Zarei, A P Körbes, P Younessi, G Montiel, A Champion, and J Memelink. Two GCC boxes and AP2/ERF-domain transcription factor ORA59 in jasmonate/ethylene-mediated activation of the PDF1.2 promoter in Arabidopsis. *Plant Mol Biol*, 75:321–331, 2011.

J Zhang, D Guo, Y Chang, C You, X Li, X Dai, Q Weng, J Zhang, G Chen, X Li, H Liu, B Han, Q Zhang, and C Wu. Non-random distribution of T-DNA insertions at various levels of the genome hierarchy as revealed by analyzing 13,804 T-DNA flanking sequences from an enhancer-trap mutant library. *Plant J*, 49:947–959, 2007.

Z Zheng, S A Qamar, Z Chen, and T Mengiste. Arabidopsis WRKY33 transcription factor is required for resistance to necrotrophic fungal pathogens. *Plant J*, 48(4): 592–605, 2006.

J Zhu and M Q Zhang. SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics*, 15(7):607–611, 1999.

Z Zhu, F An, Y Feng, P Li, L Xue, M A, Z Jiang, J-M Kim, T K To, W Li, X Zhang, Q Yu, Z Dong, W-Q Chen, M Seki, J-M Zhou, and H Guo. Derepression of ethylene-stabilized transcription factors (EIN3/EIL1) mediates jasmonate and ethylene signaling synergy in Arabidopsis. *Proc Natl Acad Sci USA*, 108(30): 12539–12544, 2011.

Appendix A

Predicted co-regulated genes and their potential regulators

Table A.1: AP2-EREBP TFs differentially expressed during *B. cinerea* infection (see section 2.2.1)

(a) AP2-EREBP TFs differentially expressed during *B. cinerea* infection

CATMA id	AGI
CATMA1a00250	AT1G01250
CATMA1a03200	AT1G04370
CATMA1a05200	AT1G06160
CATMA1a21320	AT1G22190
CATMA1a22040	AT1G22985
CATMA1a26550	AT1G28370
CATMA1a41675	AT1G50640
CATMA1a44200	AT1G53170
CATMA1a60830	AT1G71520
CATMA1a61600	AT1G72360
CATMA1a67190	AT1G78080
CATMA2a20740	AT2G22200
CATMA2a24220	AT2G25820
CATMA2a29435	AT2G31230
CATMA2a31870	AT2G33710
CATMA2a33870	AT2G35700
CATMA2a36640	AT2G38340
CATMA2a37480	AT2G39250
CATMA2a40100	AT2G41710
CATMA2a43300	AT2G44840
CATMA2a44690	AT2G46310
CATMA3a13510	AT3G14230
CATMA3a14565	AT3G15210
CATMA3a15680	AT3G16280
CATMA3a20000	AT3G20310

(b) AP2-EREBP TFs differentially expressed during *B. cinerea* infection continued

CATMA id	AGI
CATMA3a23220	AT3G23220
CATMA3a23230	AT3G23230
CATMA3a23235	AT3G23240
CATMA3a43300	AT3G50260
CATMA4a17720	AT4G16750
CATMA4a18523	AT4G17490
CATMA4a18526	AT4G17500
CATMA4a19540	AT4G18450
CATMA4c42492	AT4G25490
CATMA4a34530	AT4G32800
CATMA4c42662	AT4G36900
CATMA4a39265	AT4G37750
CATMA4a41170	AT4G39780
CATMA5a04585	AT5G05410
CATMA5a06800	AT5G07580
CATMA5a09240	AT5G10510
CATMA5a10360	AT5G11590
CATMA5a11530	AT5G13330
CATMA5a15690	AT5G17430
CATMA5a16740	AT5G18450
CATMA5a16870	AT5G18560
CATMA5a43215	AT5G47230
CATMA5a47930	AT5G51990
CATMA5a57190	AT5G61590
CATMA5a60200	AT5G64750

Table A.2: Predicted targets of AP2-EREBP TFs during *B. cinerea* infection (see section 2.2.1)

(a) Predicted targets of AP2-EREBP TFs during *B. cinerea* infection

CATMA id	AGI
CATMA3A11240	AT3G12280
CATMA5A55730	AT5G59980
CATMA2A16330	AT2G17670
CATMA1A25110	AT1G26900
CATMA4A10090	AT4G10030
CATMA5A25530	AT5G27990
CATMA3A11640	AT3G12670
CATMA1a01140	AT1G02140
CATMA3A46820	AT3G53870
CATMA5a09790	AT5G11030
CATMA4A18960	AT4G17910
CATMA5c65156	AT5G66880
CATMA1a07630	AT1G08720
CATMA2b35970	AT2G37680
CATMA3A10270	AT3G11250
CATMA2A30690	AT2G32410
CATMA4A34560	AT4G32820
CATMA3A55510	AT3G62370
CATMA1a08590	AT1G09730
CATMA4A26890	AT4G25210
CATMA3c57251	AT3G17300
CATMA2A34560	AT2G36340
CATMA2a00870	AT2G01820
CATMA5a01050	AT5G01970
CATMA1A53810	AT1G64520
CATMA3A45000	AT3G52100
CATMA2a35470	AT2G37195
CATMA5a62980	AT5G67530

(b) Predicted targets of AP2-EREBP TFs during *B. cinerea* infection continued

CATMA id	AGI
CATMA1A64360	AT1G75010
CATMA5a09450	AT5G10710
CATMA4A23385	AT4G21710
CATMA5A24265	AT5G26830
CATMA3A38910	AT3G45890
CATMA1A14240	AT1G15240
CATMA2A30686	AT2G32400
CATMA1a06430	AT1G07360
CATMA1A38020	AT1G45160
CATMA1A16620	AT1G17590
CATMA5a09660	AT5G10910
CATMA3A24130	AT3G24200
CATMA2A16190	AT2G17510
CATMA3A52990	AT3G59990
CATMA5A22380	AT5G24740
CATMA4A27240	AT4G25550
CATMA3A43870	AT3G50860
CATMA3b42920	AT3G49870
CATMA3A53810	AT3G60830
CATMA1A10990	AT1G11960
CATMA3a00330	AT3G01340
CATMA1a19730	AT1G20693
CATMA3A22315	AT3G22320
CATMA1A11110	AT1G12060
CATMA4a06600	AT4G07410
CATMA1A18610	AT1G19580
CATMA1a00950	AT1G01960
CATMA5c64227	AT5G13850

Table A.3: WRKY TFs differentially expressed during *B. cinerea* infection (see section 2.2.1) and predicted targets

(a) WRKY TFs differentially expressed during *B. cinerea* infection

CATMA id	AGI
CATMA1c72251	AT1G80840
CATMA5a11290	AT5G13080
CATMA4a01430	AT4G01250
CATMA4a25630	AT4G23810
CATMA5a60235	AT5G64810
CATMA2a02260	AT2G03340
CATMA3a03670	AT3G04670
CATMA5a21650	AT5G24110
CATMA2a36760	AT2G38470
CATMA2a21820	AT2G23320
CATMA3a00955	AT3G01970
CATMA3a51720	AT3G58710
CATMA3a00120	AT3G01080
CATMA4a32590	AT4G30930
CATMA2b22910	AT2G24570
CATMA1a59120	AT1G69810
CATMA2a39050	AT2G40740
CATMA4a23750	AT4G22070
CATMA2a23310	AT2G25000
CATMA1a51390	AT1G62300
CATMA4a19190	AT4G18170
CATMA5a42350	AT5G46350
CATMA2a39060	AT2G40750
CATMA2a45730	AT2G47260
CATMA4a28220	AT4G26640
CATMA5a45500	AT5G49520
CATMA4a27990	AT4G26440
CATMA3c57822	AT3G56400

(b) Predicted targets of WRKY TFs during *B. cinerea* infection

CATMA id	AGI
CATMA2b16180	AT2G17500
CATMA1c72346	AT1G36622
CATMA1A65800	AT1G76600
CATMA1A28750	AT1G30700
CATMA1A59240	AT1G69930
CATMA2a30220	AT2G31945
CATMA3A25100	AT3G25250
CATMA4A23490	AT4G21830
CATMA3A17790	AT3G18250
CATMA1c71052	AT1G05575
CATMA2A45630	AT2G47190
CATMA1c71401	AT1G26380

Table A.4: NAC TFs differentially expressed during *B. cinerea* infection (see section 2.2.1) and predicted targets in SplineCluster cluster 27

(a) NAC TFs differentially expressed during *B. cinerea* infection

CATMA id	AGI
CATMA1a00725	AT1G01720
CATMA1a01200	AT1G02220
CATMA1a31100	AT1G32770
CATMA1a31380	AT1G33060
CATMA1a43920	AT1G52890
CATMA1a50150	AT1G61110
CATMA1a58800	AT1G69490
CATMA1c72195	AT1G77450
CATMA2a01350	AT2G02450
CATMA2a15760	AT2G17040
CATMA2a22760	AT2G24430
CATMA2a25690	AT2G27300
CATMA2c47571	AT2G33480
CATMA2a41400	AT2G43000
CATMA3a03030	AT3G04060
CATMA3a03040	AT3G04070
CATMA3a09500	AT3G10500
CATMA3a14530	AT3G15170
CATMA3a14910	AT3G15500
CATMA3a42560	AT3G49530
CATMA3a49500	AT3G56530
CATMA3a50145	AT3G57150
CATMA4a10370	AT4G10350
CATMA4a28990	AT4G27410
CATMA4a30160	AT4G28500
CATMA4a37230	AT4G35580
CATMA4a38710	AT4G37130
CATMA5c64139	AT5G08790
CATMA5a08150	AT5G09330
CATMA5a11390	AT5G13180
CATMA5a16550	AT5G18270
CATMA5a35200	AT5G39610
CATMA5a59330	AT5G63790
CATMA5a61670	AT5G66300

(b) Predicted targets of NAC TFs during *B. cinerea* infection from SplineCluster cluster 27

CATMA id	AGI
CATMA1A22110	AT1G23040
CATMA5a06655	AT5G07440
CATMA2A41240	AT2G42810
CATMA3b41885	AT3G48890
CATMA3A53140	AT3G60130
CATMA1A26690	AT1G28480
CATMA5A45250	AT5G49280
CATMA1A11620	AT1G12640
CATMA4A13970	AT4G13790
CATMA3A46350	AT3G53400
CATMA3a09320	AT3G10320
CATMA1A30460	AT1G32120
CATMA2A31305	AT2G33150
CATMA4A20060	AT4G18950
CATMA3a08380	AT3G09520
CATMA4A24680	AT4G22920
CATMA4a08153	AT4G08390
CATMA4A21020	AT4G19810
CATMA1A60390	AT1G71100
CATMA5A24910	AT5G27520
CATMA1c71319	AT1G21310
CATMA4a15083	AT4G14680
CATMA2A16440	AT2G17760
CATMA4c42085	AT4G03370
CATMA5A21620	AT5G24090
CATMA5A11730	AT5G13500
CATMA3A45470	AT3G52540
CATMA5A41900	AT5G45900
CATMA1a08380	AT1G09510
CATMA4a05150	AT4G04620
CATMA3A13276	AT3G14050
CATMA4A23670	AT4G21980
CATMA2c47379	AT2G23170
CATMA1a07793	AT1G08920

Table A.5: Predicted targets of NAC TFs in SplineCluster cluster 38

(a) Predicted targets of NAC TFs from SplineCluster cluster 38

CATMA id	AGI
CATMA3A56445	AT3G63260
CATMA3c57703	AT3G51430
CATMA5A56290	AT5G60580
CATMA3b55995	AT3G62830
CATMA2a41960	AT2G43540
CATMA4c42131	AT4G05590
CATMA3A50820	AT3G57785
CATMA1a08835	AT1G09960
CATMA3A18110	AT3G18520
CATMA2A38710	AT2G40420
CATMA5A43190	AT5G47200
CATMA1A64610	AT1G75270
CATMA3A10350	AT3G11330
CATMA2A28075	AT2G29700
CATMA5A42370	AT5G46380
CATMA3A41140	AT3G48140
CATMA5c64332	AT5G20650
CATMA4A31200	AT4G29580
CATMA3a53030	AT3G60020
CATMA4A32985	AT4G31300
CATMA3A48370	AT3G55390
CATMA3A11040	AT3G12100
CATMA2c47433	AT2G26230
CATMA5c64177	AT5G11960

(b) Predicted targets of NAC TFs from SplineCluster cluster 38 continued

CATMA id	AGI
CATMA5A18550	AT5G20120
CATMA1A43580	AT1G52550
CATMA5c64040	AT5G03290
CATMA2a00905	AT2G01850
CATMA3A54820	AT3G61680
CATMA1b11820	AT1G12820
CATMA1a08010	AT1G09180
CATMA5a09850	AT5G11090
CATMA4c42470	AT4G23530
CATMA5A22760	AT5G25050
CATMA1a25200	AT1G27000
CATMA1a25410	AT1G27170
CATMA5A25120	AT5G27710
CATMA2A38006	AT2G39780
CATMA3A44890	AT3G51990
CATMA5A59710	AT5G64250
CATMA1A60470	AT1G71180
CATMA1A25330	AT1G27100
CATMA4a01610	AT4G01410
CATMA4A14220	AT4G14010
CATMA3c57155	AT3G11200
CATMA1c71362	AT1G23100
CATMA1c71165	AT1G12200

Appendix B

Network inference applied to time series missing first timepoint

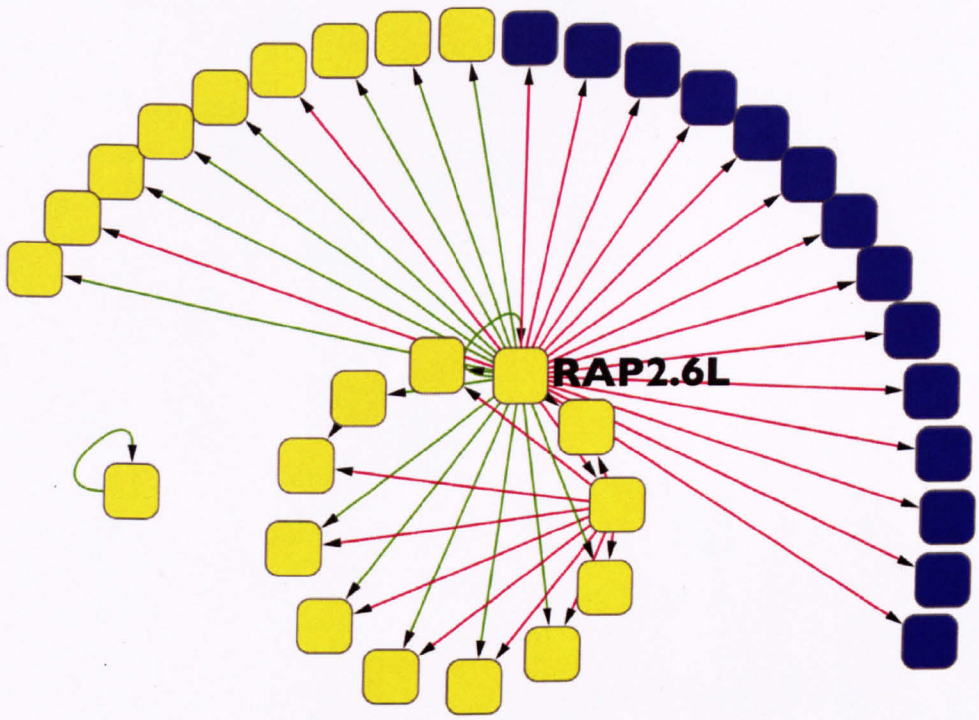


Figure B.1: Network inference sensitivity to dataset tested for genes in Tables A.1–A.2. Figure shows gene regulation inferred by VBSSM, with 9 hidden states and a threshold z-score of 3, when applied to all time-points except the first in the time series of gene expression during *B. cinerea*. Blue nodes are co-expressed (co-clustered) genes and contain the known binding sequence of the AP2-ERE BP TF family. The yellow nodes indicate members of the AP2-ERE BP TF family. Green arrows indicate inferred positive regulation. Red arrows indicate predicated negative regulation. This can be compared to the inferred network structure obtained using the full dataset, as shown in Figure 2.1(c).

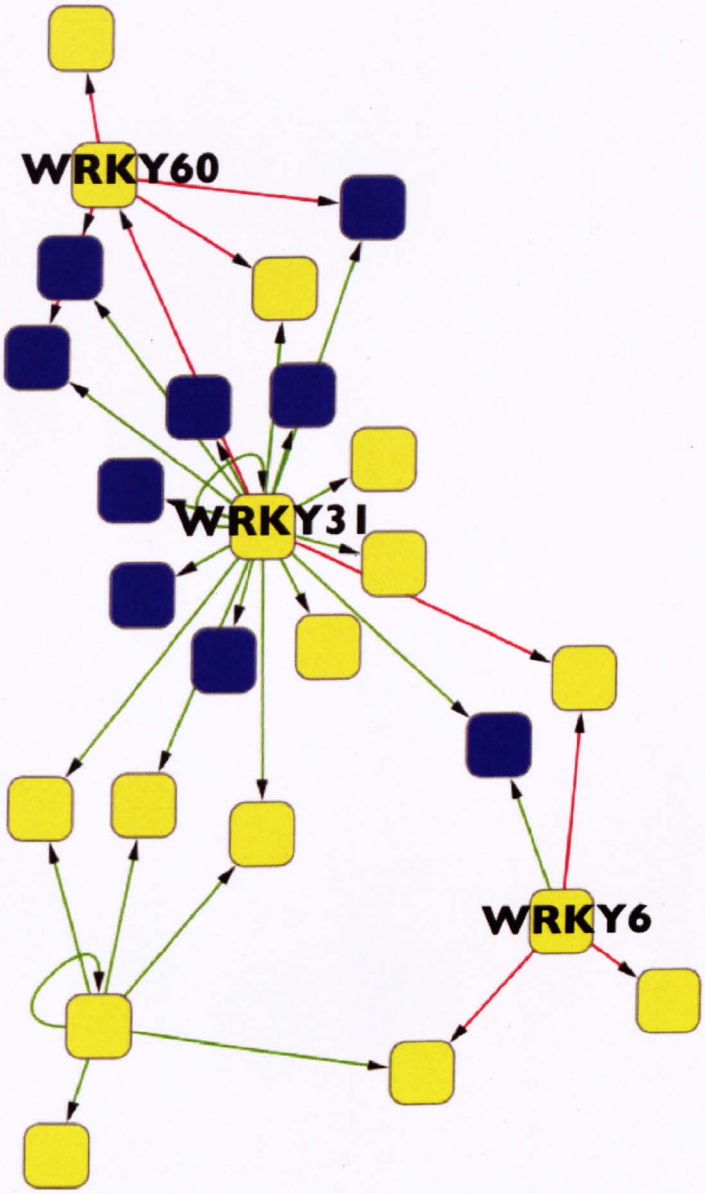


Figure B.2: Network inference sensitivity to dataset tested for genes in Table A.3. Figure shows gene regulation inferred by VBSSM, with 6 hidden states and a threshold z-score of 3, when applied to all time-points except the first in the time series of gene expression during *B. cinerea*. Blue nodes are co-expressed (co-clustered) genes and contain the known binding sequence of the WRKY TF family. The yellow nodes indicate members of the WRKY TF family. Green arrows indicate inferred positive regulation. Red arrows indicate predicated negative regulation. This can be compared to the inferred network structure obtained using the full dataset, as shown in Figure 2.2(c).

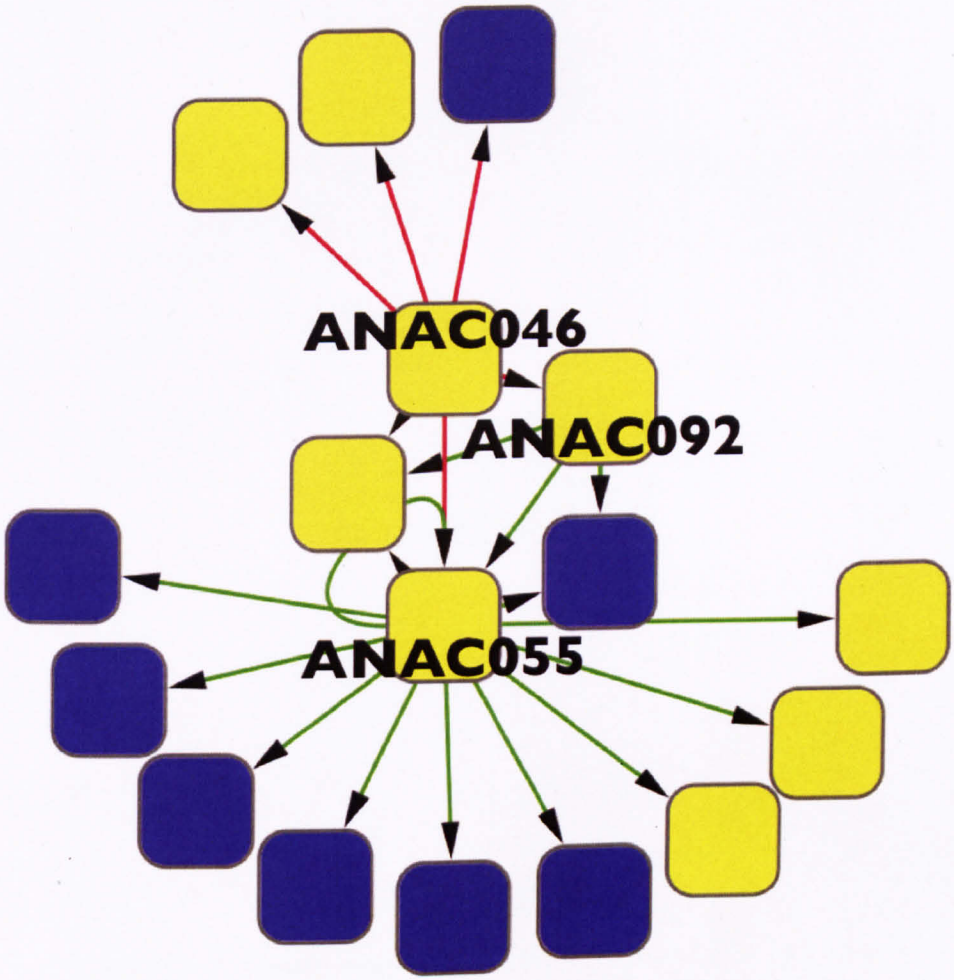


Figure B.3: Network inference sensitivity to dataset tested for genes in Table A.4. Figure shows gene regulation inferred by VBSSM, with 7 hidden states and a threshold z-score of 3, when applied to all time-points except the first in the time series of gene expression during *B. cinerea*. Blue nodes are co-expressed (co-clustered) genes and contain the known binding sequence of the NAC TF family. The yellow nodes indicate members of the NAC TF family. Green arrows indicate inferred positive regulation. Red arrows indicate predicated negative regulation. This can be compared to the inferred network structure obtained using the full dataset, as shown in Figure 2.3(c).

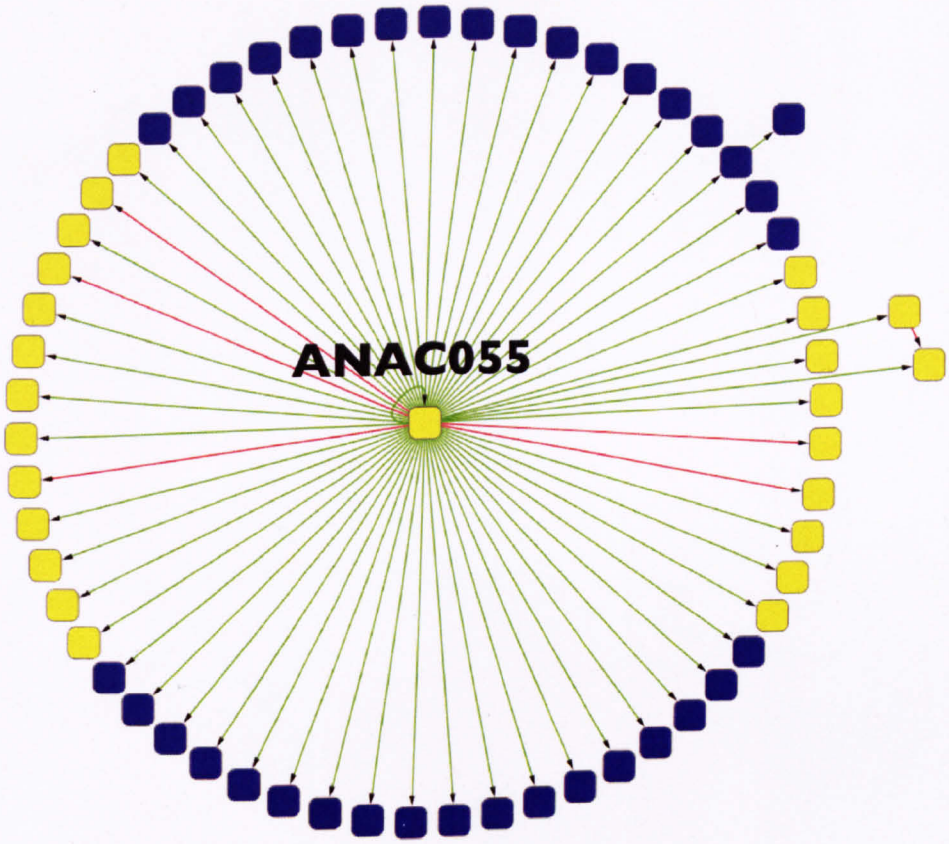


Figure B.4: Network inference sensitivity to dataset tested for genes in Table A.5. Figure shows gene regulation inferred by VBSSM, with 9 hidden states and a threshold z-score of 3, when applied to all time-points except the first in the time series of gene expression during *B. cinerea*. Blue nodes are co-expressed (co-clustered) genes and contain the known binding sequence of the NAC TF family. The yellow nodes indicate members of the NAC TF family. Green arrows indicate inferred positive regulation. Red arrows indicate predicated negative regulation. This can be compared to the inferred network structure obtained using the full dataset, as shown in Figure 2.4(c).

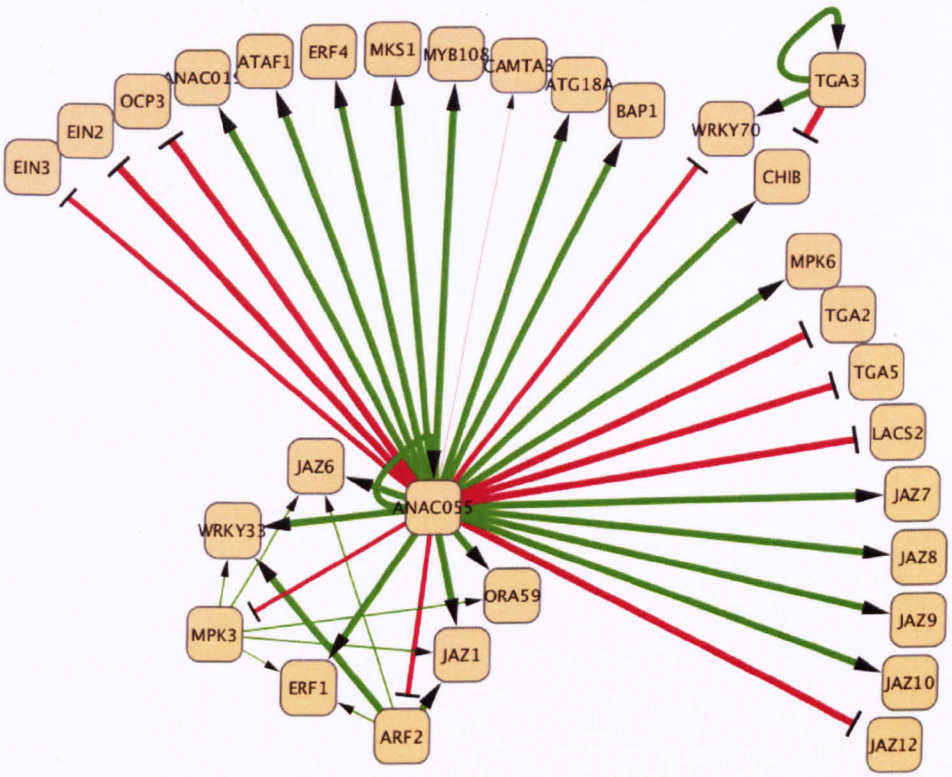


Figure B.5: Sensitivity to dataset tested for application with uninformative prior. Network structure inferred by VBSSM from the expression of the genes shown in Figure 3.2 and the differentially expressed JAZs, using 20 initialisations and 3 hidden states. Green arrows indicate positive regulation and red arrows indicate negative regulation. The thickness of the arrows correspond to the number of initialisations that led to that inferred edge. This can be compared to the inferred network structure obtained using the full dataset, as shown in Figure 4.1.

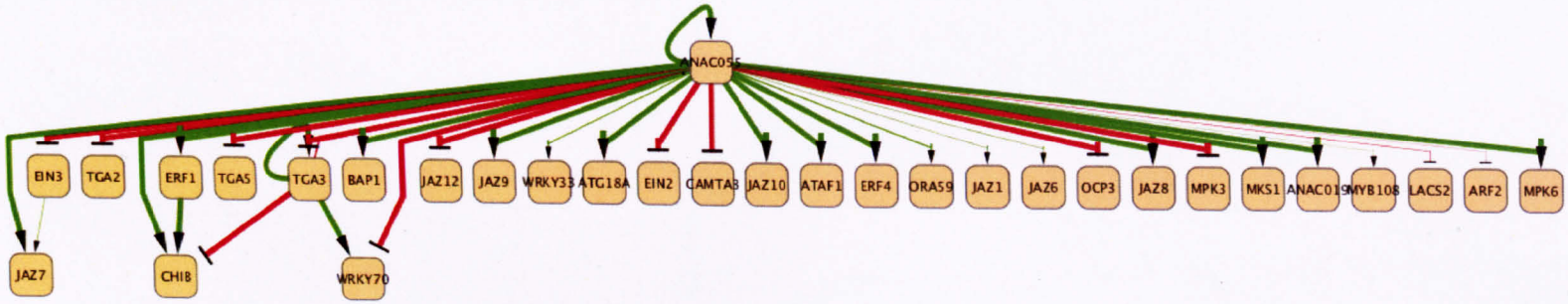


Figure B.6: Sensitivity to dataset tested for application with a prior representing literature on direct regulation. Network structure inferred by VBSSM from the expression of the genes shown in Figure 3.2 and the differentially expressed JAZs, using 20 initialisations and 3 hidden states. Green arrows indicate positive regulation and red arrows indicate negative regulation. The thickness of the arrows correspond to the number of initialisations that led to that inferred edge. A prior was used to reflect knowledge of transcriptional regulation, each with a prior weight of 0.5 standard deviations. This can be compared to the inferred network structure obtained using the full dataset, as shown in Figure 4.2.

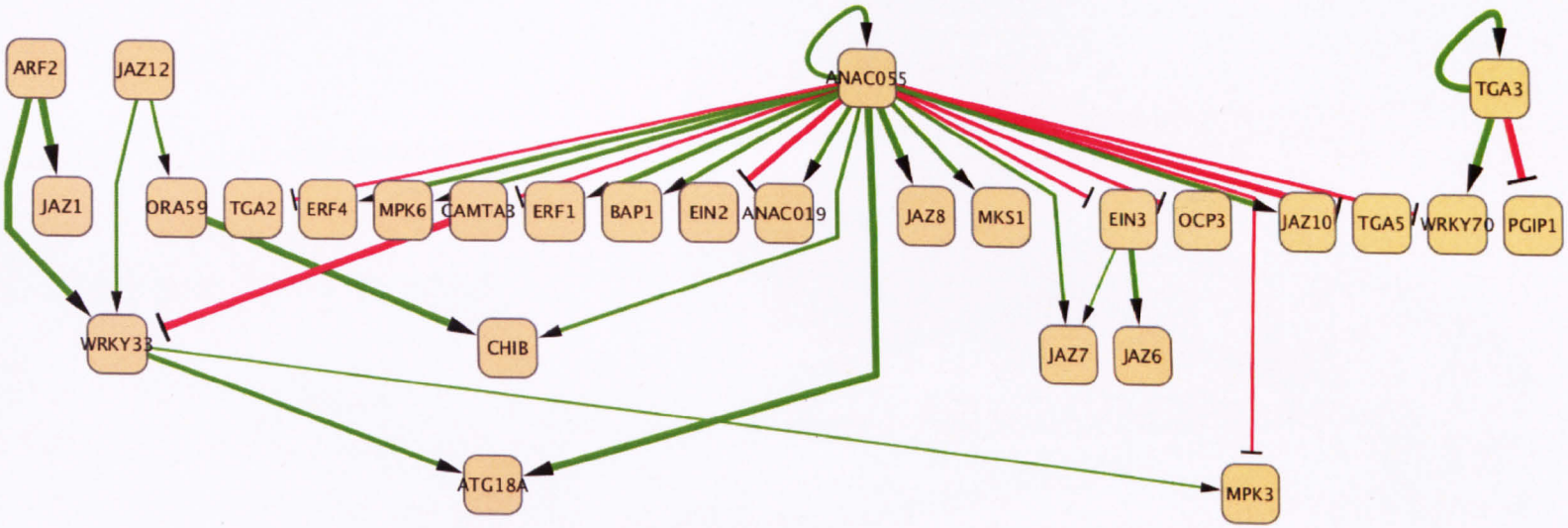


Figure B.7: Sensitivity to dataset tested for application with a prior representing literature on both direct and indirect regulation. Network structure inferred by VBSSM from the expression of the genes shown in Figure 3.2 and the differentially expressed JAZs, using 20 initialisations and 4 hidden states. Green arrows indicate positive regulation and red arrows indicate negative regulation. The thickness of the arrows correspond to the number of initialisations that led to that inferred edge. This can be compared to the inferred network structure obtained using the full dataset, as shown in Figure 4.3.

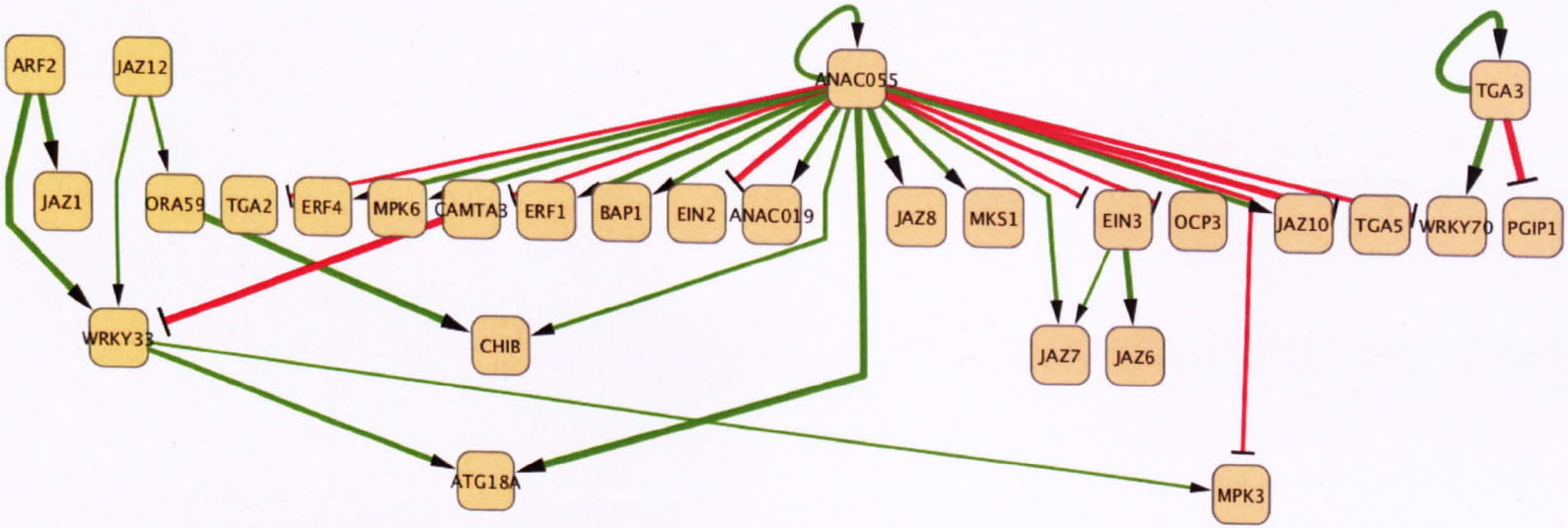
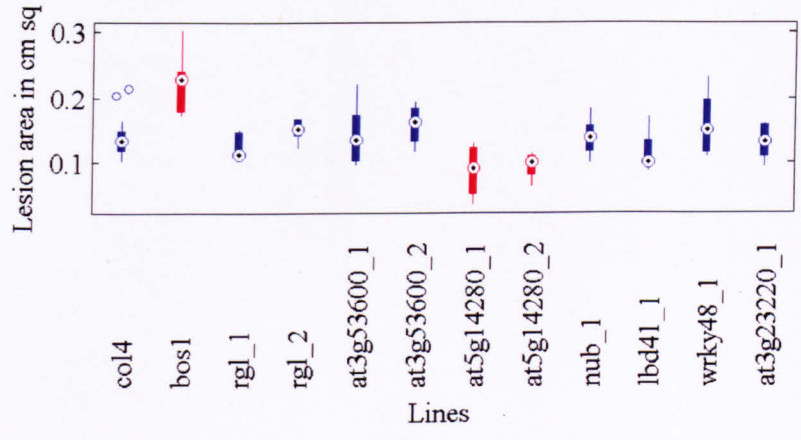


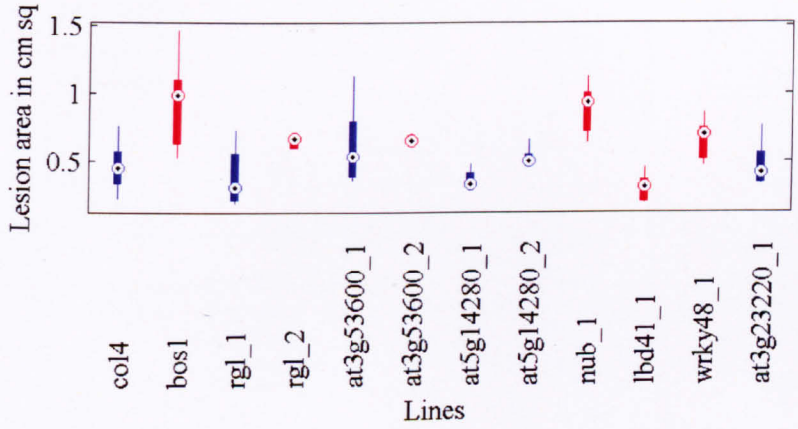
Figure B.8: Sensitivity to dataset tested for application with a prior representing the results of the previous, and literature on both direct and indirect regulation. Network structure inferred by VBSSM from the expression of the genes shown in Figure 3.2 and the differentially expressed JAZs, using 20 initialisations and 4 hidden states. Green arrows indicate positive regulation and red arrows indicate negative regulation. The thickness of the arrows correspond to the number of initialisations that led to that inferred edge. This can be compared to the inferred network structure obtained using the full dataset, as shown in Figure 4.4.

Appendix C

B. cinerea susceptibility screens



(a) 48 hpi



(b) 75 hpi

Figure C.1: *B. cinerea* susceptibility screen of TDNA knockout lines of predicted regulators of the defence response. Box plots are coloured according to the outcomes of two hypotheses tests, a t-test and a Mann-Whitney-Wilcoxon (MWW) test, both testing the hypothesis that the distribution of the lesion areas of the knockout line is different from that of a wildtype line (Col4). If the alternative hypothesis is rejected in both tests at the 5% significance level then the boxplot is coloured blue. If the alternative hypothesis is accepted in both tests at the 5% significance level then the boxplot is coloured red.

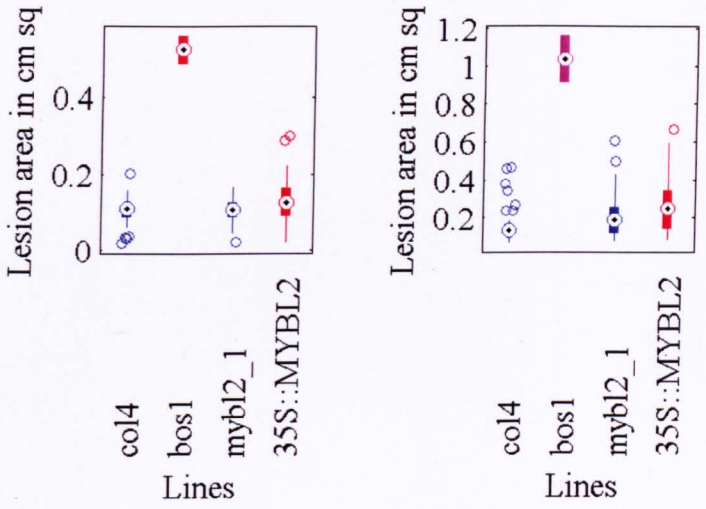
Table C.1: Reverse genetics screen of susceptibility to *B. cinerea*. Here the results of a single screen are shown, with lesion area measured at the timepoints indicated and compared to wildtype lesions by hypothesis testing. T-tests and Mann-Whitney-Wilcoxon tests are used to compare *B. cinerea* infection lesions on wildtype and mutants lines. P-values below 5% are coloured red to signify significance. The Lillefor test is used to test the hypothesis that the data is distributed normally.

(a) 48 hpi

Line Name	Number of replicates	MWW p-value	Lillefor p-value	T-test p-value
Col4	20	NA	0.046162	NA
bos1	10	$3.867e-05$	≥ 0.5	$3.9787e-05$
rgl_1	5	0.29168	0.099748	0.25323
rgl_2	5	0.13457	≥ 0.5	0.24102
at3g53600_1	5	0.91877	≥ 0.5	0.86018
at3g53600_2	4	0.27765	≥ 0.5	0.32598
at5g14280_1	5	0.011867	≥ 0.5	0.037907
at5g14280_2	5	0.0012402	≥ 0.5	0.0024184
nub_1	5	0.97288	≥ 0.5	0.9619
lbd41_1	5	0.082968	0.23512	0.19197
wrky48_1	5	0.61011	≥ 0.5	0.45007
at3g23220_1	5	0.81191	≥ 0.5	0.60074

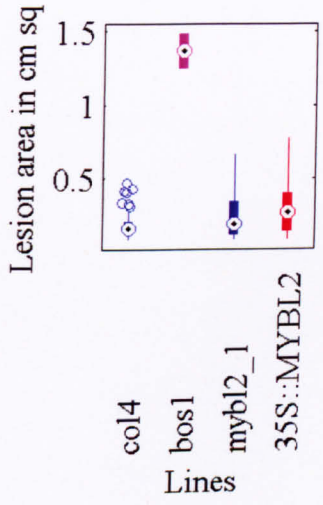
(b) 75 hpi

Line Name	Number of replicates	MWW p-value	Lillefor p-value	T-test p-value
Col4	20	NA	≥ 0.5	NA
bos1	10	0.00011816	0.27467	0.00035273
rgl_1	5	0.32458	0.406	0.56626
rgl_2	5	0.010846	0.38915	$7.7325e-05$
at3g53600_1	5	0.26231	≥ 0.5	0.3064
at3g53600_2	4	0.018151	≥ 0.5	0.00010341
at5g14280_1	5	0.29224	0.066001	0.080818
at5g14280_2	5	0.39577	0.10428	0.2298
nub_1	5	0.0015828	≥ 0.5	0.0048771
lbd41_1	5	0.032354	≥ 0.5	0.02306
wrky48_1	5	0.045056	≥ 0.5	0.043938
at3g23220_1	5	0.9729	0.48124	0.89998



(a) 44 hpi

(b) 60 hpi



(c) 70 hpi

Figure C.2: *B. cinerea* susceptibility screen of TDNA knockout lines of predicted regulators of the defence response. Box plots are coloured according to the outcomes of two hypotheses tests, a t-test and a Mann-Whitney-Wilcoxon (MWW) test, both testing the hypothesis that the distribution of the lesion areas of the knockout line is different from that of a wildtype line (Col4). If the alternative hypothesis is rejected in both tests at the 5% significance level then the boxplot is coloured blue. If the alternative hypothesis is accepted in the MWW test at the 5% significance level, and the alternative hypothesis is rejected in the t-test at the 5% significance level, then the boxplot is coloured purple. If the alternative hypothesis is accepted in both tests at the 5% significance level then the boxplot is coloured red.

Table C.2: Reverse genetics screen of susceptibility to *B. cinerea*. Here the results of a single screen are shown, with lesion area measured at the timepoints indicated and compared to wildtype lesions by hypothesis testing. T-tests and Mann-Whitney-Wilcoxon tests are used to compare *B. cinerea* infection lesions on wildtype and mutants lines. P-values below 5% are coloured red to signify significance. The Lillefor test is used to test the hypothesis that the data is distributed normally.

(a) 44 hpi

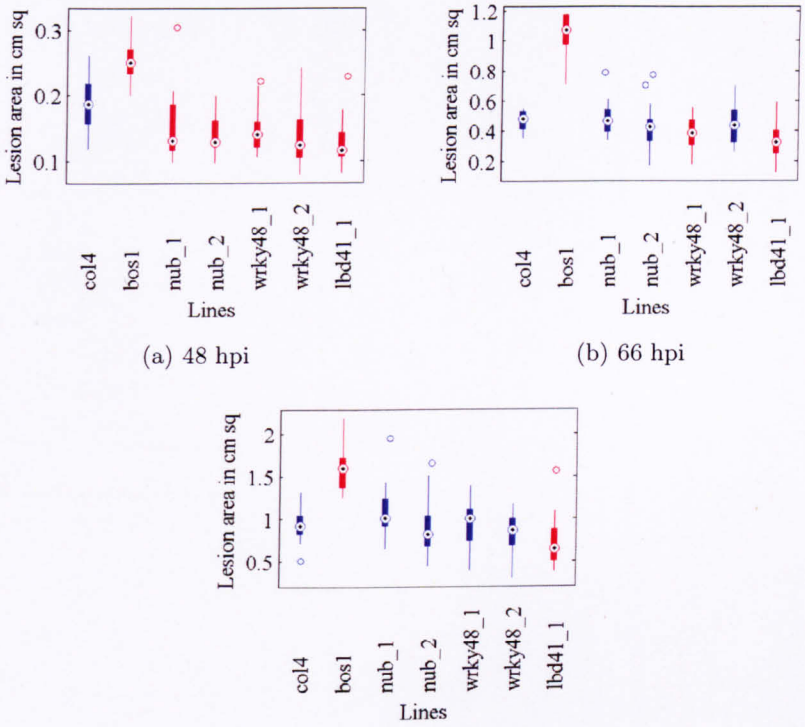
Line Name	Number of replicates	MWW p-value	Lillefor p-value	T-test p-value
Col4	39	NA	0.075474	NA
mybl2_1	40	0.99217	0.45916	0.99514
35S::MYBL2	38	0.020984	0.048891	0.011299
bos1	2	0.019692	NaN	0.047777

(b) 60 hpi

Line Name	Number of replicates	MWW p-value	Lillefor p-value	T-test p-value
Col4	40	NA	≤ 0.001	NA
mybl2_1	40	0.056715	0.062306	0.056129
35S::MYBL2	37	0.0001342	0.32425	0.00033116
bos1	2	0.01961	NaN	0.083961

(c) 70 hpi

Line Name	Number of replicates	MWW p-value	Lillefor p-value	T-test p-value
Col4	39	NA	≤ 0.001	NA
mybl2_1	40	0.58289	0.0064913	0.14051
35S::MYBL2	35	0.027183	0.10329	0.0052491
bos1	2	0.019751	NaN	0.057854



(a) 48 hpi

(b) 66 hpi

(c) 82 hpi

Figure C.3: *B. cinerea* susceptibility screen of TDNA knockout lines of predicted regulators of the defence response. Box plots are coloured according to the outcomes of two hypotheses tests, a t-test and a Mann-Whitney-Wilcoxon (MWW) test, both testing the hypothesis that the distribution of the lesion areas of the knockout line is different from that of a wildtype line (Col4). If the alternative hypothesis is rejected in both tests at the 5% significance level then the boxplot is coloured blue. If the alternative hypothesis is accepted in both tests at the 5% significance level then the boxplot is coloured red.

Table C.3: Reverse genetics screen of susceptibility to *B. cinerea*. Here the results of a single screen are shown, with lesion area measured at the timepoints indicated and compared to wildtype lesions by hypothesis testing. T-tests and Mann-Whitney-Wilcoxon tests are used to compare *B. cinerea* infection lesions on wildtype and mutants lines. P-values below 5% are coloured red to signify significance. The Lillefor test is used to test the hypothesis that the data is distributed normally.

(a) 48 hpi

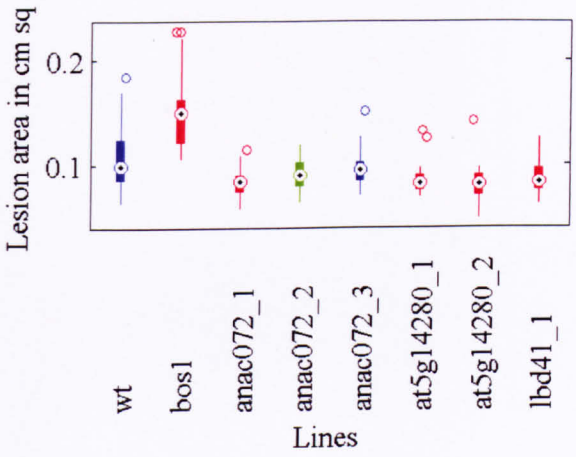
Line Name	Number of replicates	MWW p-value	Lillefor p-value	T-test p-value
Col4	20	NA	≥ 0.5	NA
bos1	8	0.0012277	≥ 0.5	0.00073679
nub_1	20	0.010142	0.043466	0.02286
nub_2	20	0.0012684	0.026725	0.00043596
wrky48_1	20	0.0055449	0.1487	0.0040037
wrky48_2	20	0.0010604	0.24064	0.00047413
lbd41_1	20	$7.3432e-05$	0.10431	$2.23e-05$

(b) 66 hpi

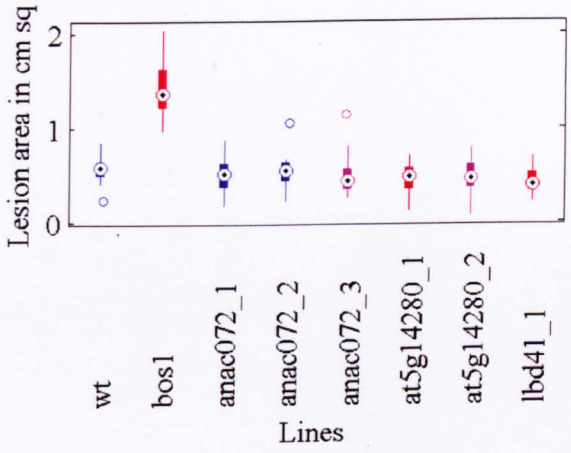
Line Name	Number of replicates	MWW p-value	Lillefor p-value	T-test p-value
Col4	20	NA	0.28287	NA
bos1	8	$5.2533e-05$	0.075546	$1.4494e-05$
nub_1	20	0.91381	≥ 0.5	0.82977
nub_2	20	0.083372	0.34052	0.22011
wrky48_1	20	0.0060356	≥ 0.5	0.0034293
wrky48_2	20	0.37926	≥ 0.5	0.28898
lbd41_1	20	$6.6e-05$	≥ 0.5	$3.6721e-05$

(c) 82 hpi

Line Name	Number of replicates	MWW p-value	Lillefor p-value	T-test p-value
Col4	20	NA	≥ 0.5	NA
bos1	8	$8.106e-05$	0.3785	0.00019659
nub_1	20	0.088337	0.37783	0.063507
nub_2	20	0.15165	0.020539	0.53354
wrky48_1	20	0.59786	0.22643	0.96121
wrky48_2	20	0.17193	≥ 0.5	0.11897
lbd41_1	20	0.0060403	0.028016	0.01468



(a) 40 hpi



(b) 56 hpi

Figure C.4: *B. cinerea* susceptibility screen of TDNA knockout lines of predicted regulators of the defence response. Box plots are coloured according to the outcomes of two hypotheses tests, a t-test and a Mann-Whitney-Wilcoxon (MWW) test, both testing the hypothesis that the distribution of the lesion areas of the knockout line is different from that of a wildtype line (Col4). If the alternative hypothesis is rejected in both tests at the 5% significance level then the boxplot is coloured blue. If the alternative hypothesis is accepted in the t-test at the 5% significance level, and the alternative hypothesis is rejected in the MWW test at the 5% significance level, then the boxplot is coloured green. If the alternative hypothesis is accepted in the MWW test at the 5% significance level, and the alternative hypothesis is rejected in the t-test at the 5% significance level, then the boxplot is coloured purple. If the alternative hypothesis is accepted in both tests at the 5% significance level then the boxplot is coloured red.

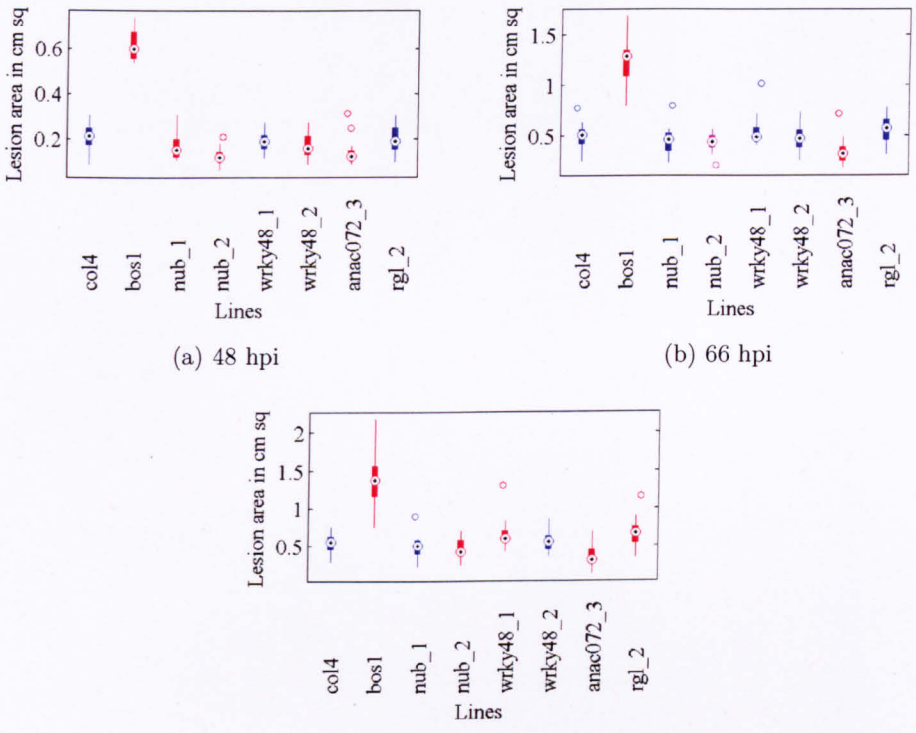
Table C.4: Reverse genetics screen of susceptibility to *B. cinerea*. Here the results of a single screen are shown, with lesion area measured at the timepoints indicated and compared to wildtype lesions by hypothesis testing. T-tests and Mann-Whitney-Wilcoxon tests are used to compare *B. cinerea* infection lesions on wildtype and mutants lines. P-values below 5% are coloured red to signify significance. The Lillefor test is used to test the hypothesis that the data is distributed normally.

(a) 40 hpi

Line Name	Number of replicates	MWW p-value	Lillefor p-value	T-test p-value
col0	20	NA	0.13824	NA
bos1	20	0.00087392	0.078784	0.00046005
anac072_1	20	0.0070812	≥ 0.5	0.0049985
anac072_2	20	0.12966	≥ 0.5	0.03794
anac072_3	20	0.3571	0.1095	0.18873
at5g14280_1	20	0.0067315	0.004495	0.010637
at5g14280_2	20	0.0048728	0.20109	0.0032848
lbd41_1	20	0.011383	0.010328	0.011184

(b) 56 hpi

Line Name	Number of replicates	MWW p-value	Lillefor p-value	T-test p-value
col0	20	NA	0.20768	NA
bos1	20	$6.7765e-08$	0.45169	$1.178e-12$
anac072_1	20	0.15161	0.24165	0.15342
anac072_2	20	0.42484	0.13864	0.48001
anac072_3	20	0.01435	0.0097506	0.11157
at5g14280_1	20	0.022257	≥ 0.5	0.01315
at5g14280_2	20	0.049842	≥ 0.5	0.058021
lbd41_1	20	0.0012845	0.49034	0.00093675



(a) 48 hpi

(b) 66 hpi

(c) 72 hpi

Figure C.5: *B. cinerea* susceptibility screen of TDNA knockout lines of predicted regulators of the defence response. Box plots are coloured according to the outcomes of two hypotheses tests, a t-test and a Mann-Whitney-Wilcoxon (MWW) test, both testing the hypothesis that the distribution of the lesion areas of the knockout line is different from that of a wildtype line (Col4). If the alternative hypothesis is rejected in both tests at the 5% significance level then the boxplot is coloured blue. If the alternative hypothesis is accepted in the MWW test at the 5% significance level, and the alternative hypothesis is rejected in the t-test at the 5% significance level, then the boxplot is coloured purple. If the alternative hypothesis is accepted in both tests at the 5% significance level then the boxplot is coloured red.

Table C.5: Reverse genetics screen of susceptibility to *B. cinerea*. Here the results of a single screen are shown, with lesion area measured at the timepoints indicated and compared to wildtype lesions by hypothesis testing. T-tests and Mann-Whitney-Wilcoxon tests are used to compare *B. cinerea* infection lesions on wildtype and mutants lines. P-values below 5% are coloured red to signify significance. The Lillefor test is used to test the hypothesis that the data is distributed normally.

(a) 48 hpi

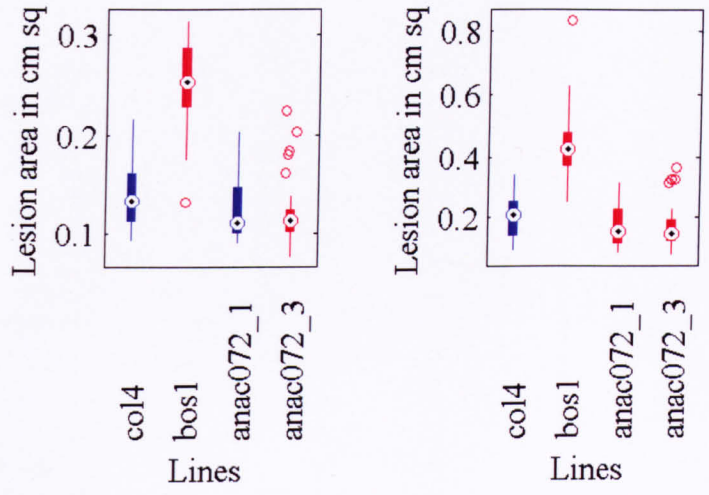
Line Name	Number of replicates	MWW p-value	Lillefor p-value	T-test p-value
Col4	20	NA	≥ 0.5	NA
bos1	10	$1.2009e-05$	0.36983	$2.1444e-11$
nub_1	20	0.017933	0.30127	0.017908
nub_2	20	$8.2797e-05$	0.32749	$1.2258e-05$
wrky48_1	16	0.27904	≥ 0.5	0.32402
wrky48_2	20	0.034841	0.11385	0.036242
anac072_3	20	0.0010119	≤ 0.001	0.00050581
rgl_2	20	0.63588	0.096132	0.67873

(b) 66 hpi

Line Name	Number of replicates	MWW p-value	Lillefor p-value	T-test p-value
Col4	20	NA	≥ 0.5	NA
bos1	20	$6.786e-08$	≥ 0.5	$1.0506e-12$
nub_1	20	0.11664	0.37348	0.15595
nub_2	20	0.048255	≥ 0.5	0.055339
wrky48_1	16	0.71409	0.037102	0.35231
wrky48_2	20	0.5338	≥ 0.5	0.55027
anac072_3	20	0.00011582	≥ 0.5	$9.5058e-05$
rgl_2	20	0.13321	≥ 0.5	0.14674

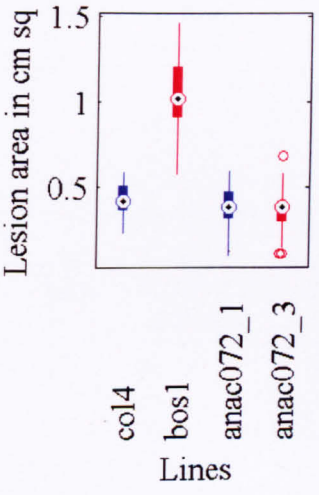
(c) 72 hpi

Line Name	Number of replicates	MWW p-value	Lillefor p-value	T-test p-value
Col4	20	NA	≥ 0.5	NA
bos1	20	$6.7956e-08$	≥ 0.5	$1.4302e-10$
nub_1	20	0.23932	0.031889	0.31353
nub_2	20	0.037255	0.046554	0.04391
wrky48_1	16	0.038453	0.011893	0.041914
wrky48_2	20	0.88171	≥ 0.5	0.74357
anac072_3	20	$8.286e-05$	0.16403	$1.2895e-05$
rgl_2	20	0.02844	≥ 0.5	0.020252

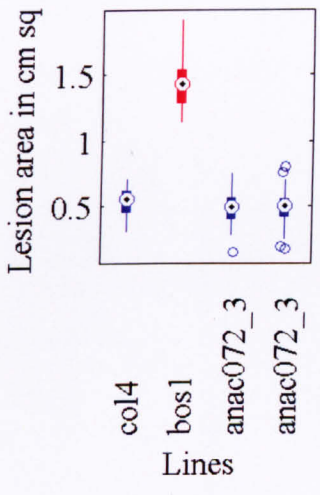


(a) 36 hpi

(b) 48 hpi



(c) 60 hpi



(d) 72 hpi

Figure C.6: *B. cinerea* susceptibility screen of TDNA knockout lines of predicted regulators of the defence response. Box plots are coloured according to the outcomes of two hypotheses tests, a t-test and a Mann-Whitney-Wilcoxon (MWW) test, both testing the hypothesis that the distribution of the lesion areas of the knockout line is different from that of a wildtype line (Col4). If the alternative hypothesis is rejected in both tests at the 5% significance level then the boxplot is coloured blue. If the alternative hypothesis is accepted in both tests at the 5% significance level then the boxplot is coloured red.

Table C.6: Reverse genetics screen of susceptibility to *B. cinerea*. Here the results of a single screen are shown, with lesion area measured at the timepoints indicated and compared to wildtype lesions by hypothesis testing. T-tests and Mann-Whitney-Wilcoxon tests are used to compare *B. cinerea* infection lesions on wildtype and mutants lines. P-values below 5% are coloured red to signify significance. The Lillefor test is used to test the hypothesis that the data is distributed normally.

(a) 36 hpi

Line Name	Number of replicates	MWW p-value	Lillefor p-value	T-test p-value
Col4	40	NA	0.03723	NA
bos1	40	$2.7837e-13$	≥ 0.5	$1.545e-21$
anac072_1	40	0.067425	≤ 0.001	0.10052
anac072_3	40	0.0023446	≤ 0.001	0.0085559

(b) 48 hpi

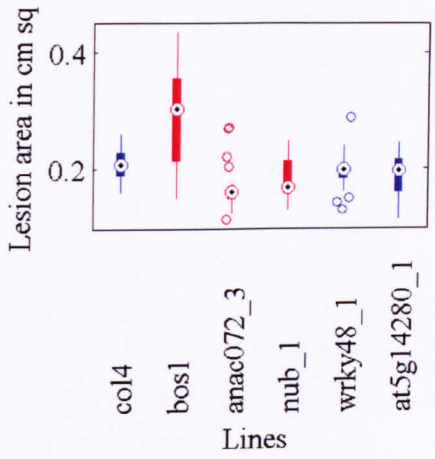
Line Name	Number of replicates	MWW p-value	Lillefor p-value	T-test p-value
Col4	40	NA	0.11328	NA
bos1	40	$3.0169e-14$	0.047431	$3.001e-19$
anac072_1	40	0.019612	0.013162	0.030129
anac072_3	40	0.0067466	0.0013583	0.011196

(c) 60 hpi

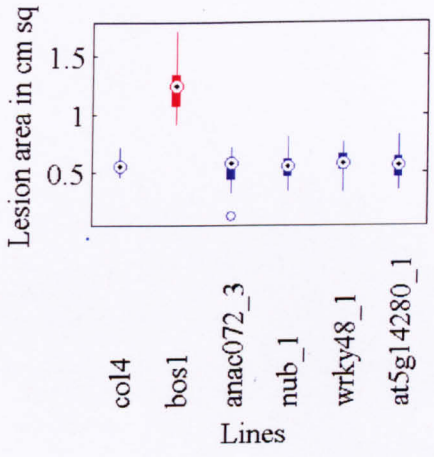
Line Name	Number of replicates	MWW p-value	Lillefor p-value	T-test p-value
Col4	40	NA	0.0522	NA
bos1	40	$1.5423e-14$	≥ 0.5	$4.9772e-24$
anac072_1	40	0.095923	≥ 0.5	0.084768
anac072_3	40	0.015711	0.011754	0.0073969

(d) 72 hpi

Line Name	Number of replicates	MWW p-value	Lillefor p-value	T-test p-value
Col4	40	NA	0.25998	NA
bos1	40	$1.4321e-14$	0.28848	$2.2097e-35$
anac072_1	40	0.089424	≥ 0.5	0.089538
anac072_3	40	0.060575	0.031389	0.070435



(a) 48 hpi



(b) 72 hpi

Figure C.7: *B. cinerea* susceptibility screen of TDNA knockout lines of predicted regulators of the defence response. Box plots are coloured according to the outcomes of two hypotheses tests, a t-test and a Mann-Whitney-Wilcoxon (MWW) test, both testing the hypothesis that the distribution of the lesion areas of the knockout line is different from that of a wildtype line (Col4). If the alternative hypothesis is rejected in both tests at the 5% significance level then the boxplot is coloured blue. If the alternative hypothesis is accepted in the t-test at the 5% significance level, and the alternative hypothesis is rejected in the MWW test at the 5% significance level, then the boxplot is coloured green. If the alternative hypothesis is accepted in the MWW test at the 5% significance level, and the alternative hypothesis is rejected in the t-test at the 5% significance level, then the boxplot is coloured purple. If the alternative hypothesis is accepted in both tests at the 5% significance level then the boxplot is coloured red.

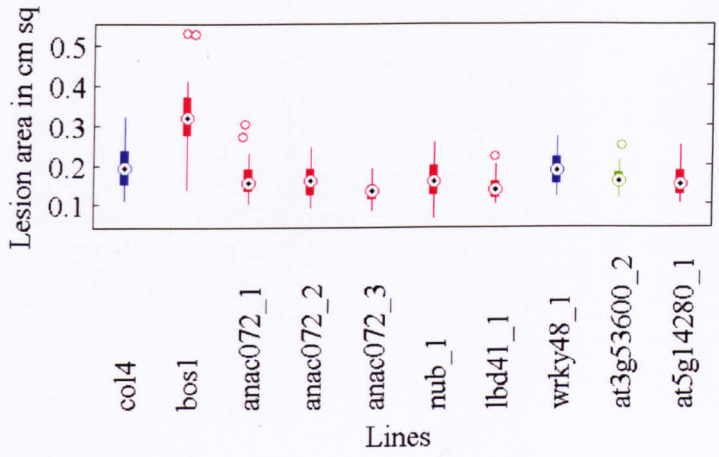
Table C.7: Reverse genetics screen of susceptibility to *B. cinerea*. Here the results of a single screen are shown, with lesion area measured at the timepoints indicated and compared to wildtype lesions by hypothesis testing. T-tests and Mann-Whitney-Wilcoxon tests are used to compare *B. cinerea* infection lesions on wildtype and mutants lines. P-values below 5% are coloured red to signify significance. The Lillefor test is used to test the hypothesis that the data is distributed normally.

(a) 48 hpi

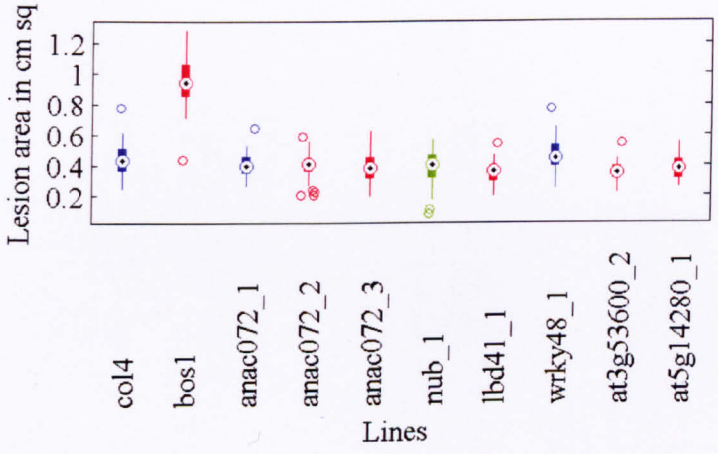
Line Name	Number of replicates	MWW p-value	Lillefor p-value	T-test p-value
Col4	20	NA	≥ 0.5	NA
bos1	20	0.013816	0.084788	0.0021926
anac072_3	20	0.00087392	≤ 0.001	0.0015009
nub_1	19	0.014978	0.041417	0.013914
wrky48_1	20	0.21313	0.063113	0.21868
at5g14280_1	20	0.11657	≥ 0.5	0.055965

(b) 72 hpi

Line Name	Number of replicates	MWW p-value	Lillefor p-value	T-test p-value
Col4	20	NA	0.15982	NA
bos1	20	$6.7956e-08$	≥ 0.5	$2.3951e-13$
anac072_3	20	0.73524	0.10603	0.29836
nub_1	19	0.54574	≥ 0.5	0.55091
wrky48_1	20	0.41703	≥ 0.5	0.47527
at5g14280_1	20	0.86043	≥ 0.5	0.70193



(a) 50 hpi



(b) 65 hpi

Figure C.8: *B. cinerea* susceptibility screen of TDNA knockout lines of predicted regulators of the defence response. Box plots are coloured according to the outcomes of two hypotheses tests, a t-test and a Mann-Whitney-Wilcoxon (MWW) test, both testing the hypothesis that the distribution of the lesion areas of the knockout line is different from that of a wildtype line (Col4). If the alternative hypothesis is rejected in both tests at the 5% significance level then the boxplot is coloured blue. If the alternative hypothesis is accepted in the t-test at the 5% significance level, and the alternative hypothesis is rejected in the MWW test at the 5% significance level, then the boxplot is coloured green. If the alternative hypothesis is accepted in both tests at the 5% significance level then the boxplot is coloured red.

Table C.8: Reverse genetics screen of susceptibility to *B. cinerea*. Here the results of a single screen are shown, with lesion area measured at the timepoints indicated and compared to wildtype lesions by hypothesis testing. T-tests and Mann-Whitney-Wilcoxon tests are used to compare *B. cinerea* infection lesions on wildtype and mutants lines. P-values below 5% are coloured red to signify significance. The Lillefor test is used to test the hypothesis that the data is distributed normally.

(a) 50 hpi

Line Name	Number of replicates	MWW p-value	Lillefor p-value	T-test p-value
Col4	40	NA	≥ 0.5	NA
bos1	40	$2.2165e-09$	0.38111	$8.3731e-11$
anac072_1	40	0.015903	0.073957	0.013194
anac072_2	38	0.0055234	0.25652	0.0015343
anac072_3	39	$8.7869e-08$	≥ 0.5	$1.1425e-08$
nub_1	38	0.0059661	0.051921	0.0028877
lbd41_1	40	$1.4206e-06$	0.31427	$3.7208e-07$
wrky48_1	40	0.64409	0.23104	0.49823
at3g53600_2	12	0.11526	0.14691	0.036445
at5g14280_1	40	0.00094531	0.001407	0.00042062

(b) 65 hpi

Line Name	Number of replicates	MWW p-value	Lillefor p-value	T-test p-value
Col4	40	NA	0.42151	NA
bos1	40	$7.8876e-14$	≥ 0.5	$1.9115e-25$
anac072_1	40	0.095939	≥ 0.5	0.091045
anac072_2	38	0.027494	0.016284	0.028725
anac072_3	40	0.042785	≥ 0.5	0.038584
nub_1	38	0.085507	0.16094	0.028794
lbd41_1	40	$4.4042e-05$	≥ 0.5	$2.4222e-05$
wrky48_1	40	0.50055	0.48574	0.44993
at3g53600_2	12	0.010703	0.42669	0.0064083
at5g14280_1	40	0.0037722	0.1156	0.003104

Appendix D

Primers for cloning of promoter fragments

Table D.1: Oligonucleotides for Y1H promoter fragments with restrictions sites.

(a)

WRKY33 promoter

Fragment 1	
Forward oligos	5'-gggggagctcCCTGACATCTCAATAAGAACATTTATGG CTAC-3'
Reverse oligos	5'-ccccactagtGATTAGTATTTAGAAGTGAGTTTGTGAG-3'
Fragment 2	
Forward oligos	5'-gggggagctcCCGATACGGATACAAAATAGTTTGATAA TC-3'
Reverse oligos	5'-ccccactagtCATCATCTTCATATGTCTCGTTCTGACA CG-3'
Fragment 3	
Forward oligos	5'-gggggagctcCTCACAAACTCACTTCTAAATACTAATC-3'
Reverse oligos	5'-ccccactagtGTCACATATGAAGAAGAGTAGTTTCTGA GAAG-3'
Fragment 4	
Forward oligos	5'-gggggagctcCGTGTCAGAACGAGACATATGAAGATGA TG-3'
Reverse oligos	5'-ccccactagtACGAAAAATGGAAGTTTGTTTTATAAAA GACC-3'

(b)

LACS2 promoter

Fragment 1	
Forward oligos	5'-gggggagctcTCCTGATTATGACAGTGACTGTGAGCTG GT-3'
Reverse oligos	5'-ccccactagtCGGTGGTGAAGTTTGGAGATTGTGGTTA TG-3'
Fragment 2	
Forward oligos	5'-gggggagctcGACGATCTAGTGTTAACCCAGAGAATTC TA-3'
Reverse oligos	5'-ccccactagtGAGTAAAGAGTAATTGGACCAAACGTAG AC-3'
Fragment 3	
Forward oligos	5'-gggggagctcGGGCTGACCTTGTAATAATATAGGGAGC AG-3'
Reverse oligos	5'-ccccactagtAACTTCAACTTTCGGATGAGAGAAAGAG GC-3'

Table D.2: Oligonucleotides for Y1H promoter fragments with restrictions sites continued.

(a)

PGIP1 promoter

Fragment 1	
Forward oligos	5'-gggggagctcAATCACTTATCTCAATAGAGCCGTTTGT GA-3'
Reverse oligos	5'-ccccactagtGTGTTAGTGATACATATACATACTATATA GTGAGTG-3'
Fragment 2	
Forward oligos	5'-gggggagctcAACGAAACCAAAGCATTTAGACTTGGC GTG-3'
Reverse oligos	5'-ccccactagtAATGTATACTGAGGCAATGTCTTCACCA TC-3'
Fragment 3	
Forward oligos	5'-gggggagctcCCTCCCAAAAGAAAGAATAAAAAGGTG TGG-3'
Reverse oligos	5'-ccccactagtGTTTATAATGGGCACTATGAAAGCCACT AGAC-3'

(b)

ARF2 promoter

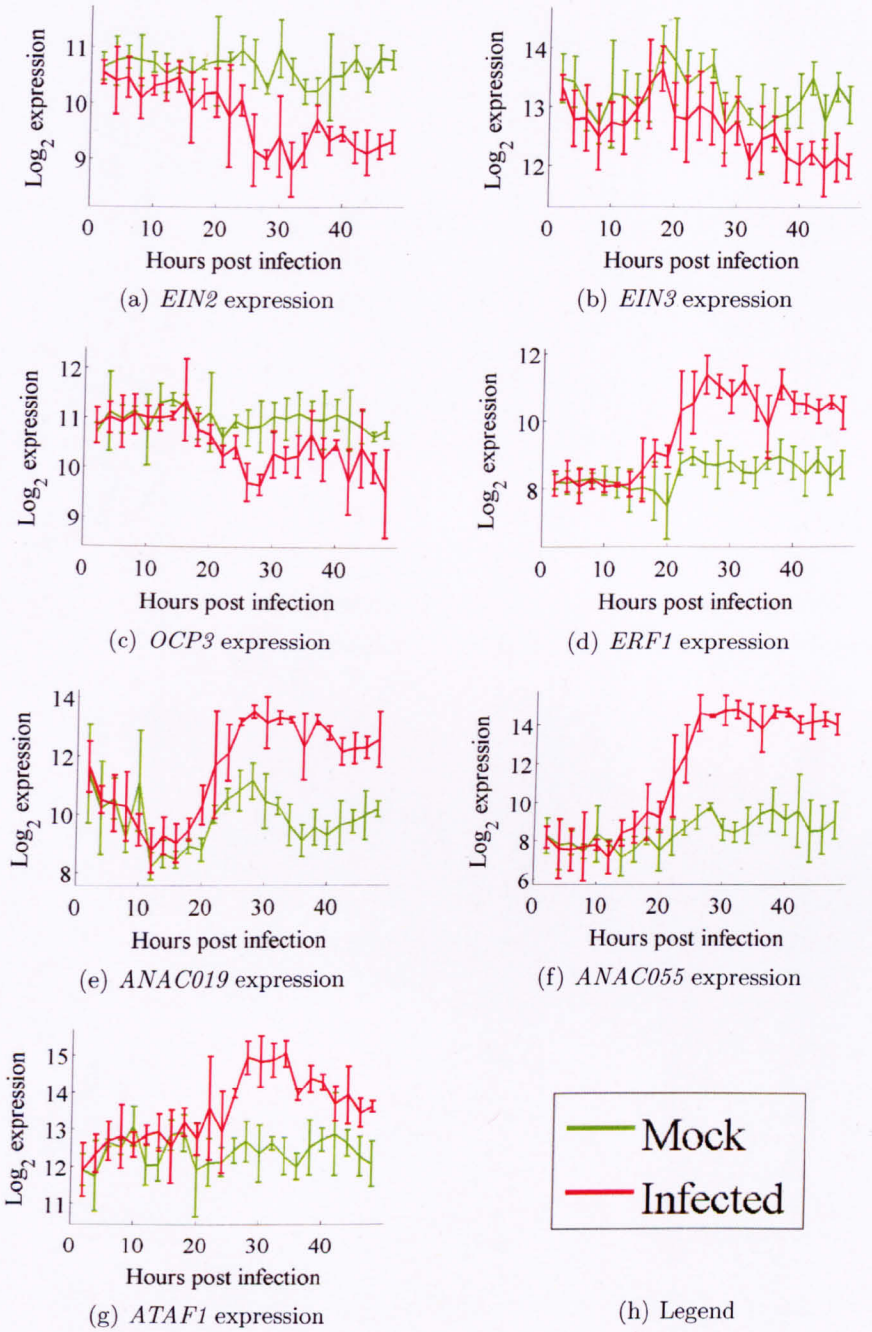
Fragment 1	
Forward oligos	5'-gggggaattcGATTCACGAGACGAAAATTCCTAGAGG CGC-3'
Reverse oligos	5'-ccccaagcttacgcgtGTAGAGGGGATTAGCAAGTAAGA AAGGCTGC-3'

Table D.3: Oligonucleotides for Y1H Gateway promoter fragments.

ORA59 promoter	
Fragment 1	
Forward oligos	5'-aaaaaagcaggcttcGTGCAATTGATCACTATATTAGTTG AACTG-3'
Reverse oligos	5'-caagaaagctgggtcGTGTCTAAGTGGCACTAAGTTTGG G-3'
Fragment 2	
Forward oligos	5'-aaaaaagcaggcttcCCGCCTTAGTTTCTGACAGAGTTT CGACTC-3'
Reverse oligos	5'-caagaaagctgggtcGAGTGTATGACGTACGGCGGCGTA TTCCCG-3'
Fragment 3	
Forward oligos	5'-aaaaaagcaggcttcCTGTTCTGTCGAGTTGTTGCTTGT TGAGCC-3'
Reverse oligos	5'-caagaaagctgggtcTGTGGGCAAAATAGGTCAAACATG CGGC-3'
Generic oligos	
Forward oligos	5'-GGGGACAAGTTTGTACAAAAAAGCAGGCT-3'
Reverse oligos	5'-GGGGACCACTTTGTACAAGAAAGCTGGGT-3'

Appendix E

Additional expression profiles



(a) *EIN2* expression

(b) *EIN3* expression

(c) *OCP3* expression

(d) *ERF1* expression

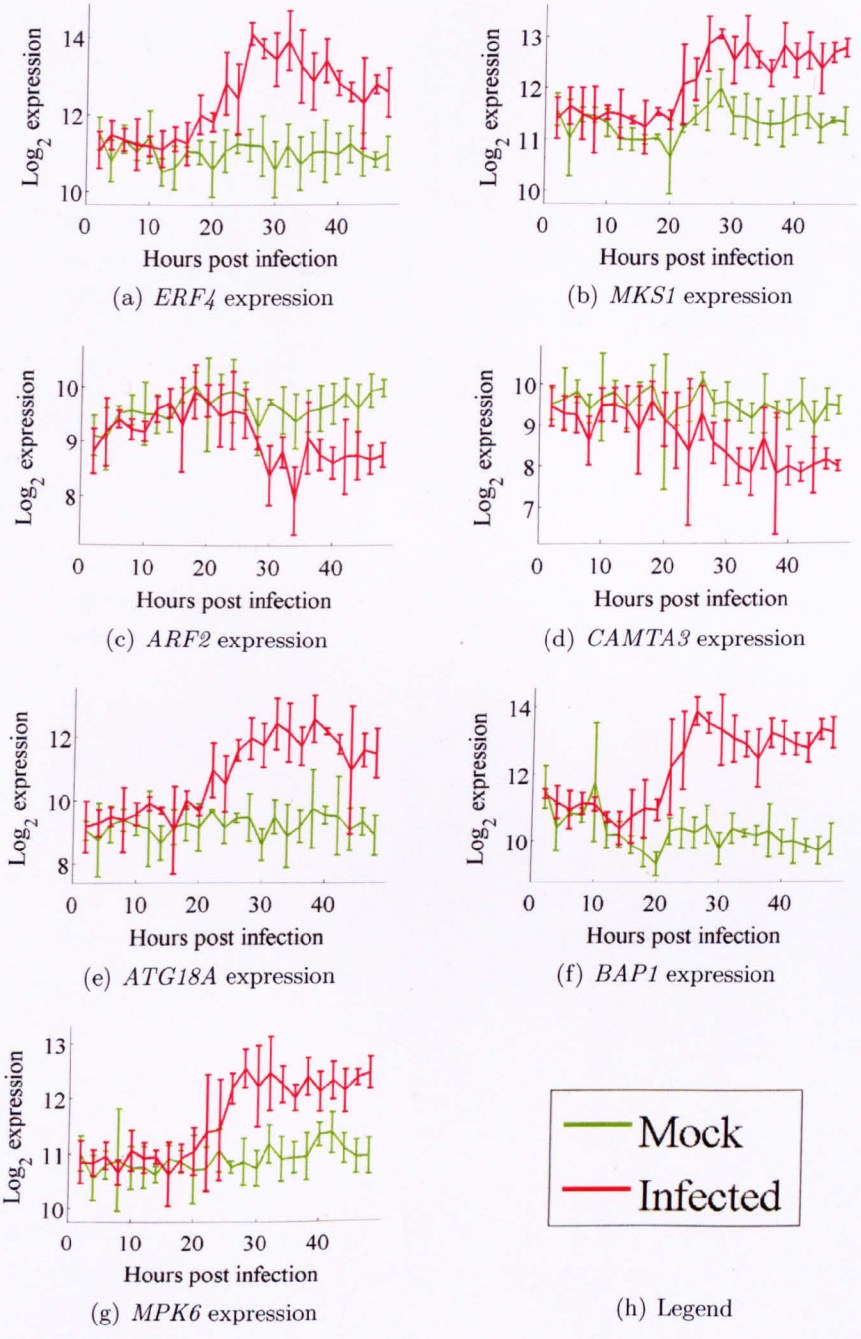
(e) *ANAC019* expression

(f) *ANAC055* expression

(g) *ATAF1* expression

(h) Legend

Figure E.1: Expression of genes in mock and *B. cinerea* infected leaves. (Expression data from the experiment introduced in Section 2.2.1. Experiment will be published in Denby et al., manuscript in preparation, and is also presented in Windram, 2010). Lines show the mean expression profile, while bars represent standard deviations.



(a) *ERF4* expression

(b) *MKS1* expression

(c) *ARF2* expression

(d) *CAMTA3* expression

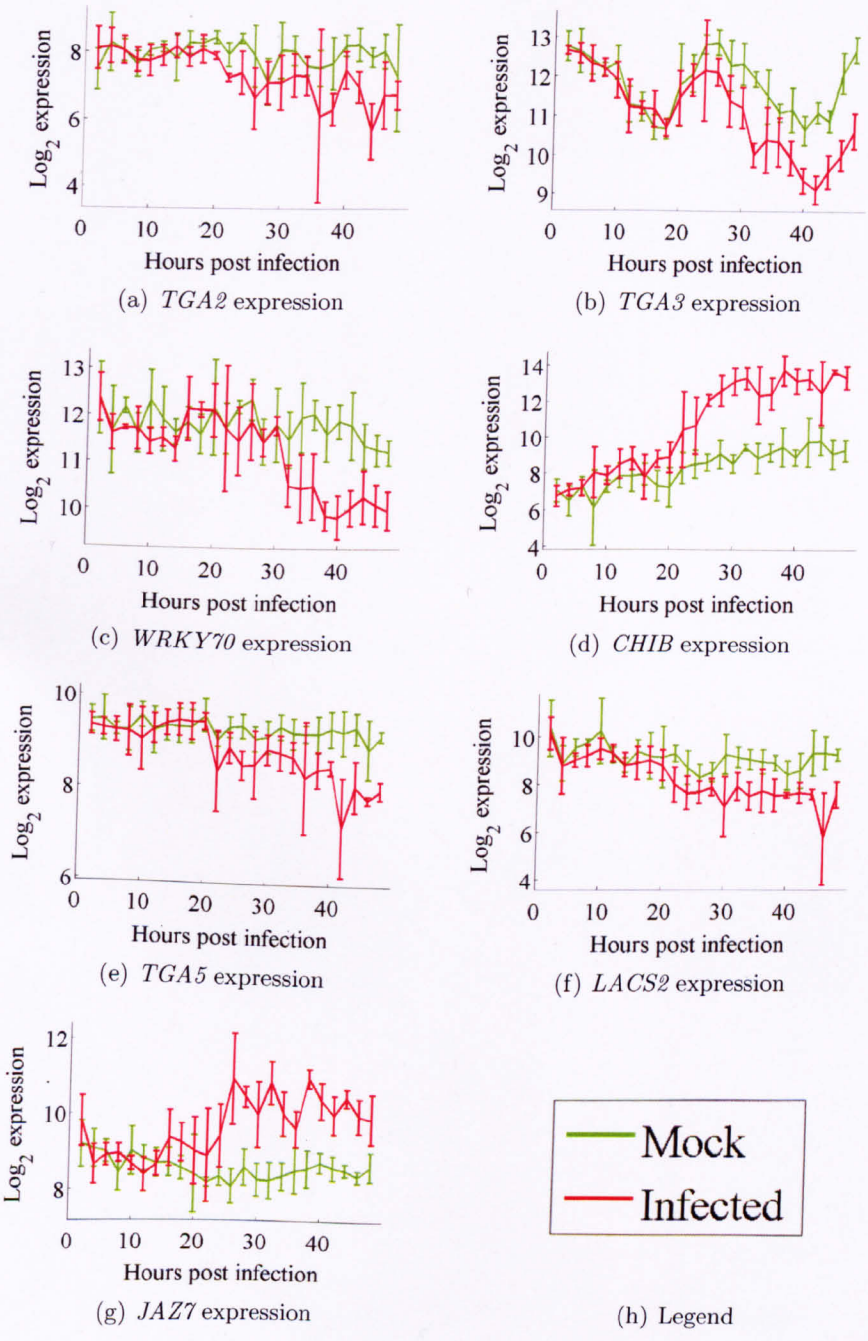
(e) *ATG18A* expression

(f) *BAP1* expression

(g) *MPK6* expression

(h) Legend

Figure E.2: Expression of genes in mock and *B. cinerea* infected leaves. (Expression data from the experiment introduced in Section 2.2.1. Experiment will be published in Denby et al., manuscript in preparation, and is also presented in Windram, 2010). Lines show the mean expression profile, while bars represent standard deviations.



(a) *TGA2* expression

(b) *TGA3* expression

(c) *WRKY70* expression

(d) *CHIB* expression

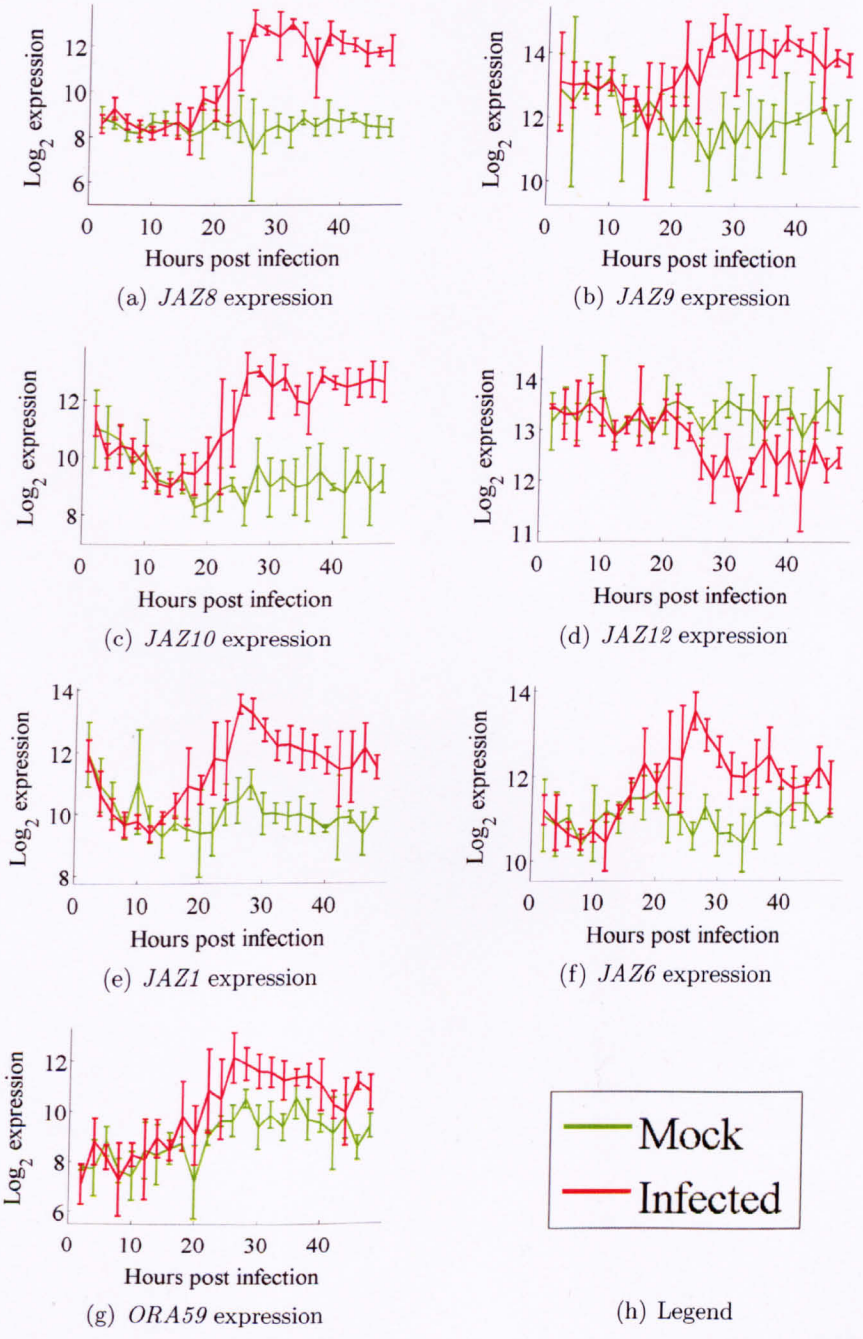
(e) *TGA5* expression

(f) *LACS2* expression

(g) *JAZ7* expression

(h) Legend

Figure E.3: Expression of genes in mock and *B. cinerea* infected leaves. (Expression data from the experiment introduced in Section 2.2.1. Experiment will be published in Denby et al., manuscript in preparation, and is also presented in Windram, 2010). Lines show the mean expression profile, while bars represent standard deviations.



(a) *JAZ8* expression

(b) *JAZ9* expression

(c) *JAZ10* expression

(d) *JAZ12* expression

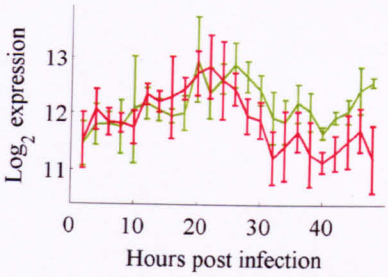
(e) *JAZ1* expression

(f) *JAZ6* expression

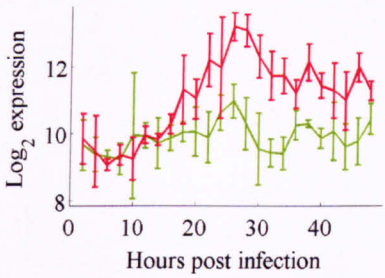
(g) *ORA59* expression

(h) Legend

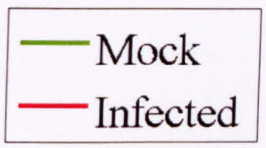
Figure E.4: Expression of genes in mock and *B. cinerea* infected leaves. (Expression data from the experiment introduced in Section 2.2.1. Experiment will be published in Denby et al., manuscript in preparation, and is also presented in Windram, 2010). Lines show the mean expression profile, while bars represent standard deviations.



(a) *MPK3* expression



(b) *WRKY33* expression



(c) Legend

Figure E.5: Expression of genes in mock and *B. cinerea* infected leaves. (Expression data from the experiment introduced in Section 2.2.1. Experiment will be published in Denby et al., manuscript in preparation, and is also presented in Windram, 2010). Lines show the mean expression profile, while bars represent standard deviations.