THE UNIVERSITY OF
WARWICK

Library Declaration and Deposit Agreement

1. STUDENT DETAILS

Please complete the following:

Full name: ……………………………… PHILIP JOHN LAW ………………………………

University ID number: ……………… 085 9963 ………………………………

2. THESIS DEPOSIT

2.1 I understand that under my registration at the University, I am required to deposit my thesis with the University in BOTH hard copy and in digital format. The digital version should normally be saved as a single pdf file.

2.2 The hard copy will be housed in the University Library. The digital version will be deposited in the University's Institutional Repository (WRAP). Unless otherwise indicated (see 2.3 below) this will be made openly accessible on the Internet and will be supplied to the British Library to be made available online via its Electronic Theses Online Service (EThOS) service.
[At present, theses submitted for a Master's degree by Research (MA, MSc, LLM, MS or MMedSci) are not being deposited in WRAP and not being made available via EthOS. This may change in future.]

2.3 In exceptional circumstances, the Chair of the Board of Graduate Studies may grant permission for an embargo to be placed on public access to the hard copy thesis for a limited period. It is also possible to apply separately for an embargo on the digital version. (Further information is available in the *Guide to Examinations for Higher Degrees by Research*.)

2.4 *If you are depositing a thesis for a Master's degree by Research, please complete section (a) below. For all other research degrees, please complete both sections (a) and (b) below:*

(a) Hard Copy

I hereby deposit a hard copy of my thesis in the University Library to be made publicly available to readers (please delete as appropriate) EITHER immediately OR ~~after an embargo period of ……………………… months/years as agreed by the Chair of the Board of Graduate Studies~~.

I agree that my thesis may be photocopied. YES / ~~NO~~ *(Please delete as appropriate)*

(b) Digital Copy

I hereby deposit a digital copy of my thesis to be held in WRAP and made available via EThOS.

Please choose one of the following options:

EITHER My thesis can be made publicly available online. YES / ~~NO~~ *(Please delete as appropriate)*

OR My thesis can be made publicly available only after…..[date] (Please give date)
~~YES / NO~~ *(Please delete as appropriate)*

OR My full thesis cannot be made publicly available online but I am submitting a separately identified additional, abridged version that can be made available online.
~~YES / NO~~ *(Please delete as appropriate)*

OR My thesis cannot be made publicly available online. ~~YES / NO~~ *(Please delete as appropriate)*

3. **GRANTING OF NON-EXCLUSIVE RIGHTS**

Whether I deposit my Work personally or through an assistant or other agent, I agree to the following:

Rights granted to the University of Warwick and the British Library and the user of the thesis through this agreement are non-exclusive. I retain all rights in the thesis in its present version or future versions. I agree that the institutional repository administrators and the British Library or their agents may, without changing content, digitise and migrate the thesis to any medium or format for the purpose of future preservation and accessibility.

4. **DECLARATIONS**

(a) I DECLARE THAT:

I am the author and owner of the copyright in the thesis and/or I have the authority of the authors and owners of the copyright in the thesis to make this agreement. Reproduction of any part of this thesis for teaching or in academic or other forms of publication is subject to the normal limitations on the use of copyrighted materials and to the proper and full acknowledgement of its source.

The digital version of the thesis I am supplying is the same version as the final, hard-bound copy submitted in completion of my degree, once any minor corrections have been completed.

I have exercised reasonable care to ensure that the thesis is original, and does not to the best of my knowledge break any UK law or other Intellectual Property Right, or contain any confidential material.

I understand that, through the medium of the Internet, files will be available to automated agents, and may be searched and copied by, for example, text mining and plagiarism detection software.

(b) IF I HAVE AGREED (in Section 2 above) TO MAKE MY THESIS PUBLICLY AVAILABLE DIGITALLY, I ALSO DECLARE THAT:

I grant the University of Warwick and the British Library a licence to make available on the Internet the thesis in digitised format through the Institutional Repository and through the British Library via the EThOS service.

If my thesis does include any substantial subsidiary material owned by third-party copyright holders, I have sought and obtained permission to include it in any version of my thesis available in digital format and that this permission encompasses the rights that I have granted to the University of Warwick and to the British Library.

5. **LEGAL INFRINGEMENTS**

I understand that neither the University of Warwick nor the British Library have any obligation to take legal action on behalf of myself, or other rights holders, in the event of infringement of intellectual property rights, breach of contract or of any other right, in the thesis.

Please sign this agreement and return it to the Graduate School Office when you submit your thesis.

Student's signature: .....[redacted]............................... Date: ....03/03/2014....

Quantitative analysis of time-series microarray data, with application to investigating responses to environmental stresses in Arabidopsis

by

Philip John Law

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Systems Biology Doctoral Training Centre

November 2013

Contents

List of Tables

List of Figures

Acknowledgements

Foremost, I would like to thank my supervisors, Andrew Mead and Vicky Buchanan-Wollaston, for their patience and guidance throughout this project. I would also like to thank the members of my advisory committee (Jay Moore, Miriam Gifford, Julia Brettschneider and Simon Spencer) for all their suggestions and assistance.

I would like to thank all the people who helped me with my lab work: Justyna Prusinska, Claire Hill, Alison Jackson, Jack Grundy, Ana Mendes, Emily Breeze and Tina Payne. To all my friends and colleagues that I've met over the course of the last four years - Alex Jironkin, Jo Hulsmans, Rob Smith, Mark Walsh, Dafyd Jenkins, Adam Talbot, John Lynch, Max Newbert, Chris Barnes, Kate Richardson, Damon Daniels - thank you for all the enthusiastic discussions and entertainment.

Finally, I would like to thank my good friend Chris Macdonald for all the support he has provided me over the years.

Declaration

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.

Abstract

High-throughput technologies have made it possible to perform genome-scale analyses to investigate a variety of research areas. From these analyses, vast amounts of potentially noisy data is generated which could obscure the underlying signal.

In this thesis, a high-throughput regression analysis approach was developed, where a variety of linear and nonlinear models were fitted to gene expression profiles from time course experiments. These models included the logistic, Gompertz, exponential, critical exponential, linear+exponential, Gaussian, and hyperbolic functions. The fitted parameters from these models reflect aspects of the model shape, and are thus biologically interpretable. Investigating the fitted parameters allowed for the interpretation of the gene expression profiles in terms of the underlying biology, such as the time of initial expression. This provides a potentially more mechanistic approach to study the genetic responses to stimuli. This analysis was applied to three time series gene expression experiments - a *Saccharomyces cerevisiae* time course as a validation of the method, and two time course experiments on *Arabidopsis thaliana* investigating stress responses to the senescence process, and pathogen infection by *Botrytis cinerea*.

A cluster analysis, named ShapeCluster, was developed as an application of the fitted models. Using this analysis, it was possible to cluster on aspects of the shape of the expression profiles using different combinations of parameters. This added flexibility to the analysis and allowed for the investigation of the data in multiple ways. Specifically, performing the cluster analysis on a specific parameter permitted the identification of genes that are co-regulated, or participate in response to the biological stress in question. Several methods of producing clusters with combinations of parameters, namely simultaneous parameter clustering, sequential meta-clustering, and cross meta-clustering, provided additional means of interrogating the data. Clusters from these methods were assessed for significance through the use of over-represented annotation terms and motifs, and found to produce biologically relevant sets of genes.

Experiments using quantitative-PCR and luciferase transcriptional reporters were designed to determine the response to a combined Botrytis and senescence stress. A predicted model was identified by fitting a factor model to the experimental data, and identifying the most significant model effects. This model removed noise from the biological data, and confirmed that the effects of the two stresses was additive.

In cross-sectional data, each sample is obtained from separate individuals (plants), and thus may be different biological ages. An iterative, cross-validation multivariate regression approach was developed, termed time shifting, to estimate the true biological age of the replicate samples, and it was shown that the approach resulted in better model fits for a large proportion of the genes.

In this thesis, a number of novel analytical approaches for obtaining information from gene expression microarray datasets were developed. These analyses provided biologically oriented descriptions of individual gene expression profiles, allowing for the modelling and greater interpretation of profiles obtained from time-series experiments. Through careful choice of appropriate models, such statistical regression approaches allow for an improved comparison of gene expression profiles, and may provide an improved understanding of common regulatory mechanisms between genes.

Abbreviations

ABA	-	Abscisic acid
ANOVA	-	Analysis of variance
BHC	-	Bayesian Hierarchical Clustering
BHI	-	Biological Homogeneity Index
DAS	-	Days after sowing
df	-	Degrees of freedom
ET	-	Ethylene
GO	-	Gene Ontology
hpi	-	Hours post-infection
JA	-	Jasmonic acid
LoF	-	Lack of fit
luc	-	Luciferase
MS	-	Mean square
qPCR	-	Quantitative polymerase chain reaction
R^2	-	Coefficient of determination
R^2_a	-	Adjusted R^2
R^2_{LoF}	-	Lack of fit adjusted R^2
RMS	-	Residual mean square
SA	-	Salicylic acid
SED	-	Standard error of differences between means
SS	-	Sum of squares
TF	-	Transcription factor

Glossary

Cross meta-clustering - Clustering performed independently on two parameters, and the overlapping genes identified

Function - The mathematical relationship describing the relationship between observed measurements and a set of variables

Meta-clustering - Performing the cluster analysis on clusters

Model - The models are the various curves that were fitted to the data using the regression analysis, e.g. logistic or Gompertz

Sequential meta-clustering - Clustering is performed based on one parameter, and each of these clusters are subsequently clustered on a second parameter

Shape - The shape refers to the pattern of the curve, which has been fitted to one of the models, and possesses describable physical characteristics, e.g. increasing/decreasing

Simultaneous clustering - Clustering on multiple parameters, where the distance between two genes is determined by a Euclidean distance between their parameters

Temporal displacement - The difference between the observed time point and the estimated biological time point for a replicate gene

Temporal marker genes - Set of genes used to perform the temporal displacements

Time shift - The movement of replicates based on the temporal displacements

1. Introduction

The world's population continues to grow, and this coupled with changes in the climate, mean that the production of food crops must intensify. It has been estimated that food production needs to increase by 60% of the 2005 level in order to meet the demand for food in 2050 (FAO, 2012). While farming techniques such as the application of fertiliser or pesticides aid in improving yields, these can have serious environmental impacts (Oerke, 2006). In addition, given the limited area of land for agriculture, other means of increasing yields are required (Oerke, 2006). Furthermore, current climate change prediction models suggest that average surface air temperatures will rise by approximately 3°C in the next 50–100 years, which could have an enormous impact on the agricultural system (Solomon *et al.*, 2007). Thus it is desirable to understand the processes that occur in a plant's response to environmental stresses. Using this information, it will be possible to develop crops that are resilient to these stresses by exploiting the innate ability of the plants to protect themselves.

A stress is defined as a stimulus that leads to the disruption of normal metabolic processes, causing a change in the biological system, potentially leading to reduced yield, or even death in extreme cases (Bijlsma and Loeschcke, 2005; Oerke, 2006; Rao *et al.*, 2006). Plants are sessile organisms, and thus unlike animals, are unable to relocate to escape from stresses. They are constantly exposed to environmental stresses, both biotic (such as infection by pathogens, or predation by herbivores) and abiotic (such as heat, cold, drought or salt) (Oerke, 2006; Naika *et al.*, 2013; Rao *et al.*, 2006), and have developed intricate mechanisms to detect and respond to these stimuli (Atkinson and Urwin, 2012; Naika *et al.*, 2013). These stresses can all reduce the yield in crop plants (Atkinson and Urwin, 2012; Mittler and Blumwald, 2010), and it has also been suggested that stresses may also impact the nutritional value of food crops (Andre *et al.*, 2008; Jansen *et al.*, 2008). Therefore, it is vital to understand the principles of these stress responses in order to determine how plants adapt to stress tolerance, and so develop plants that are resistant, or able to adapt, to a variety of these stresses so that necessary crop yields can be maintained in a changing environment (Naika *et al.*, 2013).

1.1. Responses of plants to stress

Plants are constantly exposed to a variety of environmental stresses, both to biotic and abiotic, and thus need to be able to respond efficiently to these stresses. Many plant hormones including salicylic acid (SA), jasmonic acid (JA), ethylene (ET), abscisic acid (ABA), auxin, gibberellic acid, cytokinins and brassinosteroids have been implicated in producing a stress response (reviewed in Bari and Jones, 2009 and Glazebrook, 2005).

Stresses cause a change in concentration of these molecules, which in turn activate signalling cascades, ultimately leading to the activation of genes involved in stress responses. In addition to forming part of a signal transduction network, they are involved in a number of biological functions, including regulation of growth, development, and reproduction. SA, JA, ET and ABA are the most well characterised molecules that are known to be involved in signal transduction networks that influence stress responses, and will be discussed in greater detail below.

Biotic pathogens can be broadly grouped into two categories: biotrophic and necrotrophic pathogens, each of which have different modes of attack. As a result, plants have different mechanisms for dealing with the pathogens (reviewed in Glazebrook, 2005 and Bari and Jones, 2009). Biotrophic pathogens do not cause host cell death, and cause minimal cell damage in order to extract nutrients from the host cells. These pathogens include the oomycete, *Hyaloperonospora parasitica* and fungal pathogens, *Golovinomyces orontii* and *Erysiphe pisi*. In contrast, necrotrophic pathogens kill host tissue by producing cell wall degrading enzymes or toxins, leading to host tissue maceration, and thus obtains resources from the cell remnants. Pathogens of this type include *Botrytis cinerea* and *Alternaria brassicicola*.

1.1.1. Salicylic acid (SA)

SA is typically associated with resistance against biotrophic pathogens, mediated through the expression of *NONEXPRESSER OF PR GENES 1* (*NPR1*) and *PATHOGENESIS-RELATED GENE 1* (*PR1*) (Zhou *et al.*, 2000). In uninfected cells, NPR1 exists as an inactive oligomeric complex, and is primarily found in the cytoplasm. Increased cellular concentrations of SA induces the release of monomeric NPR1. These active NPR1 monomers are translocated into the nucleus where they physically interact with members of the TGACG-binding motif family of basic region/leucine zipper (bZIP) transcription factors (TFs), and this complex binds to the PR1 promoter (reviewed in Dong, 2004; Johnson *et al.*, 2003; Pajerowska-Mukhtar *et al.*, 2013; Spoel *et al.*, 2009). In addition, a number of other TFs, including SUPPRESSOR OF SNI1 2 (SSN2), RADIATION SENSITIVE 51D (RAD51D), BREAST CANCER 2A (BRCA2A), and activating WRKY transcription factors are recruited onto the PR1 promoter. Repressor proteins, such as SUPPRESSOR OF NPR1 INDUCIBLE 1 (SNI1), NON-INDUCIBLE IMMUNITY 1 (NIM1)-INTERACTING (NIMIN) proteins, and repressive WRKY factors are dissociated from the PR1 promoter (reviewed in Pajerowska-Mukhtar *et al.*, 2013). These events subsequently result in the activation of *PR1* gene expression and defence responses.

As previously mentioned, one of the most common groups of TFs associated with the SA response is the WRKY family. It has been shown that a large proportion of the genes belonging to this family are differentially expressed in response to SA treatment (Dong *et al.*, 2003), and are often associated with resistance against biotrophic pathogens (reviewed in Singh *et al.*, 2002). In addition, some members of the family act to regulate the SA signalling response (van Verk *et al.*, 2011). The name of this family is derived from the conserved WRKYGQK amino acid sequence within the TF sequences, and these TFs tend to bind to sequences containing the W-box motif (TGAC), many of which are upstream of stress response genes (Eulgem *et al.*, 2000; Rushton *et al.*, 2010).

SA has also been found to be involved in a number of other biological functions, including senescence, plant development and photosynthesis (Morris *et al.*, 2000; Rivas-San Vicente and Plasencia, 2011).

1.1.2. Jasmonic acid (JA)

Where SA is typically associated with defence against biotrophic pathogens, JA is generally involved in signalling pathways that confer resistance to necrotrophic pathogens. As would be expected from pathogens that cause cell death, JA is also associated with the wounding response, particularly through the action of various jasmonate Zim domain (JAZ) proteins (Katsir *et al.*, 2008). In the inactive state, a key positive regulator of JA responses, MYC2, is repressed by JAZ proteins (Chini *et al.*, 2007, 2009), with an additional layer of repression present due to the binding of the NINJA and TOPLESS (TPL) proteins (Pauwels *et al.*, 2010). Biotic stresses result in the accumulation of JA, which acts as a "molecular glue" between CORONATINE INSENSITIVE 1 (COI1) and the JAZ repressors. In the presence of (*3R,7S*)-jasmonoyl-L-isoleucine (JA-Ile) or coronatine (COR), this forms a stable complex that interacts with the E3 ubiquitin ligase, SCF^{COI1}. This results in proteosomal degradation of the JAZ repressor, subsequently releasing the NINJA-TPL complex, and relieving the repression from MYC2, facilitating the activation of JA-responsive genes, such as the defensin *PLANT DEFENSIN 1.2* (*PDF1.2*, *PR-12*) and the thionin *THIONIN 2.1* (*THI2.1*, *PR-13*) (reviewed in Robert-Seilaniantz *et al.*, 2011).

Other TFs that are associated with JA signalling include members of the APETALA 2/ETHYLENE RESPONSE FACTOR (AP2/ERF) domain TF family, such as ETHYLENE RESPONSE FACTOR 1 (ERF1), and OCTADECANOID-RESPONSIVE ARABIDOPSIS AP2/ERF 59 (ORA59). However, it is currently unknown whether these TFs directly interact with the JAZ proteins, or if they form part of another signalling response pathway (Pré *et al.*, 2008).

In addition to regulating biotic stress responses, JA has also been implicated in drought responses, controlling cell growth and proliferation, as well as inducing the expression of genes that are involved in the production of stress-associated metabolites including glucosinolates, phenylpropanoids and anthocyanins (Pauwels *et al.*, 2008; Sasaki-Sekimoto *et al.*, 2005). JA has also been found to increase as leaves senescence (Breeze *et al.*, 2011), suggesting that as a plant ages, JA down-regulates growth genes and up-regulates stress response genes (Pauwels *et al.*, 2008). Other functions of JA include involvement in salt and osmotic stress (Lehmann *et al.*, 1995; Xu *et al.*, 1994).

1.1.3. Ethylene (ET)

Like JA, ET is largely involved in the responses to necrotrophic pathogens. ET is detected by receptors, such as ETHYLENE RESPONSE 2 (ETR2), ETHYLENE RE-

SPONSE SENSOR 1 and 2 (ERS1 and ERS2) and ETHYLENE INSENSITIVE 4 (EIN4), which are found on the endoplasmic reticulum. In the absence of ET, these receptors maintain the negative regulator CONSTITUTIVE TRIPLE RESPONSE 1 (CTR1) in an active form, which represses ET signalling components, such as EIN2 and EIN3. Under these conditions, EIN3 is degraded by the proteasome-mediated degradation pathway through the interaction with EIN3-BINDING F BOX PROTEIN 1 and 2 (EBF1 and EBF2) (reviewed in Robert-Seilaniantz *et al.*, 2011 and van Loon *et al.*, 2006).

In the presence of ET, CTR1 is inactivated, which removes the repression of EIN2, allowing it to interact with EBF1 and EBF2, consequently preventing the degradation of EIN3. EIN3 is stabilised by ET and has been found to play a major role in photosynthetic, developmental and defence pathways (Zhong *et al.*, 2009). EIN2 enhances the expression of ETHYLENE INSENSITIVE 3-LIKE (EIL) TFs, which then activate ERF family members such as ERF1 (Solano *et al.*, 1998). These ERF family members target the GCC-sequence which is found in many stress related genes (Okamuro *et al.*, 1997). This family of proteins is important for the resistance to biotic stresses, as it is has been shown that ERF mutants show altered susceptibility, depending on pathogen (reviewed in van Loon *et al.*, 2006). For example, ERF1 over-expressors exhibit enhanced resistance to *Botrytis cinerea*, but increased susceptibility to *Pseudomonas syringae* pv tomato (Lorenzo *et al.*, 2003).

ET levels increase during the senescence process, due to the up-regulation of ET biosynthetic genes as the plant ages (van der Graaff *et al.*, 2006), and it has been shown that *ein2* mutants exhibit delayed senescence (Oh *et al.*, 1997). ET is also been found to be involved in plant growth and development, response to wounding, dehydration, and cold and salt stress (Morgan and Drew, 1997).

1.1.4. Abscisic acid (ABA)

The role of ABA in biotic stresses is not as clear as the other plant hormones, as ABA can result in both positive and negative effects (reviewed in Asselbergh *et al.*, 2008, Robert-Seilaniantz *et al.*, 2011 and Ton *et al.*, 2009). The response appears to depend on the timing of the infection, where ABA induction before infection increases resistance to pathogens by closing the stomata, thus restricting pathogen entry into the plant (Melotto *et al.*, 2006). However, if ABA induced signalling is activated during *P. syringae* pv tomato infection, this results in increased susceptibility, possibly due to ABA signalling interfering with the other hormone responses (de Torres-Zabala *et al.*, 2007).

ABA induces the expression of genes that contain the conserved ABA response element (ABRE) ((C/T)ACGTGGC), which allows them to interact with a group of bZIP TFs known as ABRE-binding factors (ABFs). ABF2, ABF3 and ABF4 have been shown to be key regulators of the ABA signalling response, and activate gene expression when under drought stress (Yoshida *et al.*, 2010). In addition to the ABRE motif, ABA-regulated genes contain the binding sites for other stress related TFs, including MYC2 and MYB2. These TFs have been shown to be involved in ABA-mediated induction of genes such as *RESPONSIVE TO DESSICATION 22 (RD22)* and *ALCOHOL DEHYDROGENASE 1* (*ADH1*) (Abe *et al.*, 2003).

ABA was originally linked to water deprivation, which induces ABA biosynthesis, and has since also been implicated in the regulation of plant growth and development, and osmotic and salt stress where it is thought to help protect plants against cellular damage from water loss (reviewed in Cutler *et al.*, 2010). Application of exogenous ABA induces premature senescence (Gepstein and Thimann, 1980), and like ET, ABA accumulates in senescencing leaves due to the up-regulation of ABA biosynthetic genes during senescence (Breeze *et al.*, 2011; Buchanan-Wollaston *et al.*, 2005; van der Graaff *et al.*, 2006).

1.1.5. Crosstalk between the signalling pathways

The above signalling pathways do not exist in isolation, but rather are closely associated in order to tune the defence response to a particular environmental stress. Crosstalk between the various signalling pathways thus provides plants with the ability to regulate the responses for different stresses (reviewed in Pieterse *et al.*, 2009; Robert-Seilaniantz *et al.*, 2011; van Loon *et al.*, 2006).

JA and ET are often found to be synergistic, where it has been shown that there is a large overlap of differentially expressed genes after treatment with JA and ET (Schenk *et al.*, 2000). In addition, it has been found that activation of *PDF1.2*, a key indicator of JA-signalling, requires both ET and JA signalling components (Penninckx *et al.*, 1996). It is thought that the main convergence points between these two signalling pathways involve ERF1 and ORA59 (Lorenzo *et al.*, 2003; Pré *et al.*, 2008).

Conversely, SA signalling is antagonistic to the JA- and ET-signalling pathways. It has been found that biotrophic pathogens trigger SA signalling, which suppresses JA- and ET-signalling through the action of TGA and WRKY TFs (Li *et al.*, 2004; Ndamukong *et al.*, 2007).

With abiotic stresses, there is not as much information as the ABA signalling pathway is more complicated. In general, it appears that ABA interferes with the SA-, JA-, and ET- signalling pathways, increasing susceptibility to both biotrophic and necrotrophic pathogens (Anderson *et al.*, 2004; Flors *et al.*, 2008; Mohr and Cahill, 2007).

1.2. The PRESTA project

As described previously, it is becoming increasingly important to understand the molecular effects of environmental stresses on a plant, in order to exploit the plant's innate ability to adapt to incoming stresses. *Arabidopsis thaliana* is a model plant species and is commonly used for determining the function of genes and the general understanding of plant biology. Its genome was sequenced in 2000 by an international consortium (The Arabidopsis Genome Initiative, 2000), and much work has been done since then to determine the function of the genes (Garcia-Hernandez *et al.*, 2002; Lamesch *et al.*, 2012).

To specifically investigate the effect of environmental stresses, the PRESTA (Plant Responses to Environmental STress in Arabidopsis) project has performed a number of high resolution time series microarray experiments in order to elucidate the signalling networks to explain plant responses to a variety of biotic and abiotic stresses. There are several datasets that include long (16 hours light) and short day (8 hours light) senescence, *Botrytis cinerea* infection, *Pseudomonas syringae* pv tomato infection, drought stress, and high light. The gene expression analyses were performed using CATMA version 3 and 4 microarrays (Sclep *et al.*, 2007), which contained over 32 500 probes. These probes mapped to approximately 24 000 unique gene models. Thus, the dynamic changes in gene expression levels for effectively the entire Arabidopsis genome have been determined in response to a number of different stresses, and biological interpretation regarding the co-expression and co-regulation of genes associated with a particular stress responses can be ascertained. Of particular interest are the genes that are key regulators of multiple environmental stresses.

Outcomes of this project include identifying a transcription factor that enhances resistance to drought and infection to virulent *P. syringae* pv tomato DC3000 and *Hyaloperonospora arabidopsidis* (Bechtold *et al.*, 2013) as well as identifying a local network around a group of transcription factors that are involved in stress responses (Hickman *et al.*, 2013). In addition, a number of theoretical advancements were made, including a text-mining analysis (Hassani-Pak *et al.*, 2010), a clustering approach for identifying potentially co-regulated genes (Kiddle *et al.*, 2010), a tool to identify conversed regulatory regions between different plant species (Baxter *et al.*, 2012), the reverse-engineering of gene regulatory networks (Penfold and Wild, 2011; Penfold *et al.*, 2012), providing an interface for motif finding and analysis (Brown *et al.*, 2013), and a tool to estimate transcription activation and repression points (Jenkins *et al.*, 2013). Finally, the papers by Breeze *et al.* (2011) and Windram *et al.* (2012) are two experimental reports detailing the changes in gene expression of Arabidopsis plants in response to long day senescence and *Botrytis cinerea* infection, respectively. These are both very high resolution datasets with a large number of time points (11 and 24 for the senescence and Botrytis experiments, respectively) and provide an in-depth exploration of the transcriptional changes that take place under the respective stresses. These stresses will be discussed in greater detail below.

1.2.1. Senescence

Senescence in green plants is a complex process and the process is mostly clearly illustrated in autumn, when the leaves on trees begin to change from green to vibrant yellows, reds and oranges, before finally turning brown and falling off. The process of senescence is the final stage of leaf development and is controlled by numerous environmental signals. Senescence is a result of carefully organised changes within the cellular structure, metabolism and gene expression, and ultimately leads to the death of the leaf. However, this process is essential for the overall fitness of the plant, particularly in monocarpic plants (plants that reproduce once and die at the end of the reproductive phase), where nutrients are moved from the leaves to the seeds (reviewed in Buchanan-Wollaston *et al.*, 2003 and Lim *et al.*, 2007). In non-annual plants, the nutrients are stored until they are needed for the following season's growth phase. The degeneration of the cells begins with the chloroplast and continues with the degradation of all the cellular macronutrients (such as lipids and proteins). The mitochondria and nucleus remain intact until the final stages. Therefore, if it was possible to control when crops undergo senescence, it may be possible to improve crop yields, in addition to reducing post-harvest yellowing and the concomitant decrease in the nutrient content of leafy vegetables.

The senescence process results in the maturation of crops such as maize or rice. However, premature senescence occurs when plants undergo an environmental stress, which may lead to reduced yield and quality of crops (Wright, 1999). Here, the plant "sacrifices" parts of itself to improve the chances of survival, for example those parts that are under attack by a pathogen (Tanaka *et al.*, 2005). This ensures that the nutrients in the distressed areas will not be completely lost. Abiotic factors that induce senescence include drought stresses and extreme temperature, whereas biotic factors include pathogen infection and shading by other plants. All the above hormone signalling responses (SA, JA, ET and ABA signalling pathways) have been implicated in senescence (Breeze *et al.*, 2011; Buchanan-Wollaston *et al.*, 2005; Morris *et al.*, 2000; van der Graaff *et al.*, 2006).

An interesting aspect of leaf senescence is that it can be reversed (Rivero *et al.*, 2007; Zavaleta-Mancera *et al.*, 1999). Using methods such as cytokinin (CK) treatments (Zavaleta-Mancera *et al.*, 1999), it is possible to induce plants that have yellowed and have moved the majority of its nutrients out of the leaf, to become green again. The promoter of *SENESCENCE ASSOCIATED GENE 12* (*SAG12*) was fused to the coding sequence of isopentenyltransferase (IPT), an enzyme involved in CK biosynthesis, and plants with this gene construct were found to exhibit delayed senescence, as well as tolerance to excessive water (flooding) stress. These plants exhibited chlorophyll retention as well as higher biomass and carbohydrate content as compared to wildtype plants (Huynh *et al.*, 2005). In addition, Rivero *et al.* (2007) showed that transgenic plants expressing IPT from a senescence enhanced promoter were able to survive severe drought stress and exhibited vigorous growth after re-watering. In both cases, the expression of the IPT gene linked to a senescence related promoter resulted in increased synthesis of CKs in the plant, suggesting that these CKs are important in delaying senescence, as well as protecting against osmotic stresses (flooding and drought stresses).

The senescence process is extremely complex, resulting in a large number of transcriptomic changes (Breeze *et al.*, 2011; Buchanan-Wollaston, 1997; van der Graaff *et al.*, 2006). One of the main groups of TFs that have been implicated as a regulator in the senescence response is the NAC family (Balazadeh *et al.*, 2010; Breeze *et al.*, 2011; Hickman *et al.*, 2013; Meng *et al.*, 2009), where mutations of members of this TF family result in altered senescence patterns. Members of the WRKY and ARF TF families

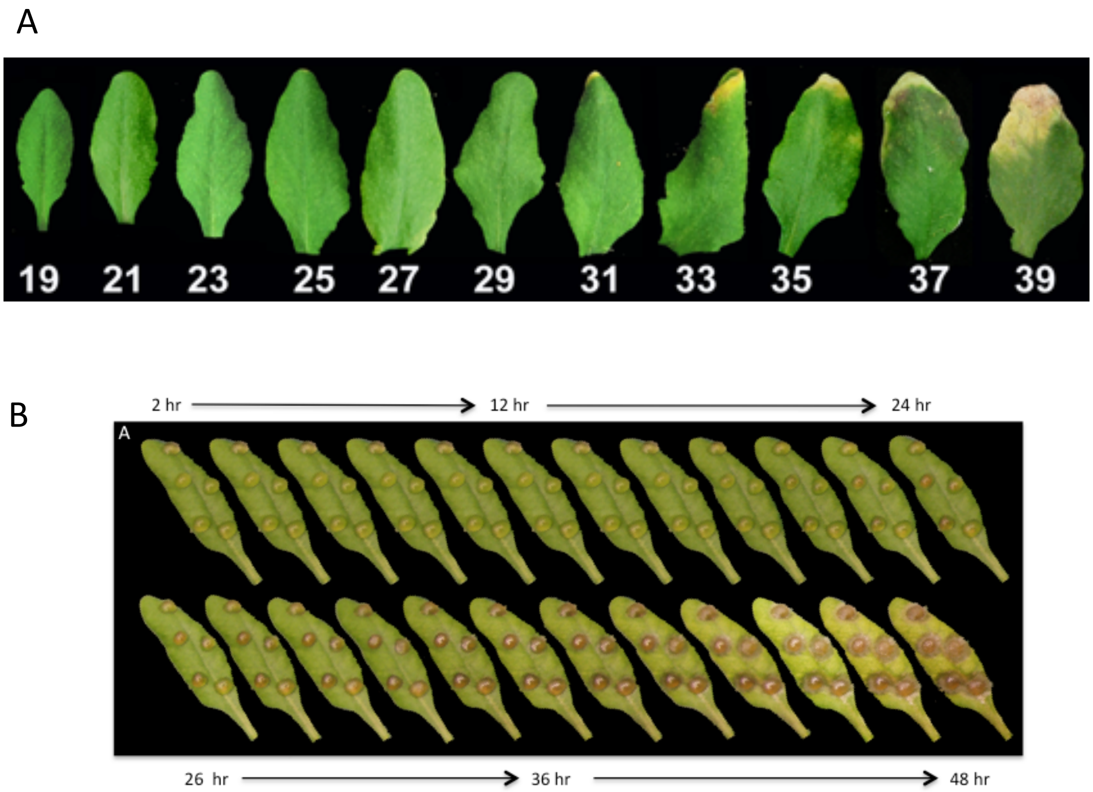


Figure 1.1: Sample leaves from the PRESTA datasets. (A) Examples of the sampled leaves in the senescence experiment. The numbers indicate the age of the plant (days after sowing). (B) Images of a leaf showing the progression of Botrytis infection. Each leaf image indicates a time difference of two hours. Figures obtained from Breeze *et al.* (2011) and Windram *et al.* (2012).

have also been found to be involved in regulation of the senescence process.

The senescence dataset (Breeze *et al.*, 2011) consists of 22 time points, taken every second day for 22 days from 19 to 39 days after sowing (DAS). The seventh leaf of each sample plant was sampled at 7 and 14 hours into the light period (morning and afternoon samples, respectively), and 4 biological replicate samples were taken at each time point, where each biological replicate (leaf) was obtained from a different plant. In addition, each biological replicate had 4 technical replicates (2 dye swaps) in the microarray experiment (Figure 1.1A). The microarray experiment was designed using a loop design in order to enable the most efficient extraction of data from the two-colour microarray system. The data from the microarrays was analysed using an adapted version of the MAANOVA package (Wu *et al.*, 2003) to perform the quality assurance, normalisation and mixed model analyses.

1.2.2. *Botrytis cinerea*

The grey mould, *Botrytis cinerea*, is a necrotrophic fungus that infects a wide range of host plants, including fruits, vegetables and even cut flowers. Necrotrophic organisms kill host cells with the use of phytotoxins to promote host cell death, and use the remains for further growth (van Kan, 2006; Williamson *et al.*, 2007). With a broad host range of over 200 different plant species, Botrytis is considered one of the most significant fungal plant pathogens (Dean *et al.*, 2012), and is considered as a model for necrotrophic pathogens (van Kan, 2006). Botrytis is difficult to control as it possesses a variety of modes of attack, and can survive for extended periods of time in soil (Williamson *et al.*, 2007). As a result of all these factors, Botrytis has significant economic impact, both due to crop loss because of infection (both pre- and post-harvest), and the cost of using fungicides to control infections (Dean *et al.*, 2012). Thus, increasing the knowledge of the molecular effects of Botrytis infection could aid in developing new mechanisms with which to combat the disease (van Kan, 2006). Interestingly, Botrytis is useful in wine production, where it can result in noble rot in the grapes, which can then be used to produce sweet wines (Dean *et al.*, 2012).

Infection by Botrytis results in a large scale reorganisation of the Arabidopsis transcriptome (Windram *et al.*, 2012). A large number of different TF families become up-regulated in response to Botrytis infection, including members of the WRKY, ERF, NAC, and MYB families (Lorenzo *et al.*, 2003; Windram *et al.*, 2012). Much of this is due to JA- and ET-signalling, which are generally found to be involved in response to necrotrophic pathogens. Indeed, it has been shown that JA is important for resistance to Botrytis (Govrin and Levine, 2000; Thomma *et al.*, 1998), particularly with *PDF1.2*, a key JA-signalling gene shown to be expressed upon fungal attack. However, there is a significant amount of overlap involving other hormones, such as ABA. In addition, there is crosstalk between the different hormone signalling pathways where TGA3, belonging to a TF family typically found in SA-dependent signalling, has been shown to be important for resistance to Botrytis (Ferrari *et al.*, 2003; Windram *et al.*, 2012).

In the Botrytis time series experiment (Windram *et al.*, 2012), the seventh leaf for each replicate plant was initially infected with an inoculum made up of 100 000 spores/ml of *Botrytis cinerea* pepper isolate suspended in half strength grape juice (Denby *et al.*, 2004), at 6 hours into the light period, and samples were taken every 2 hours from 2 to 48 hours after infection (24 time points). Controls were obtained at each time point by inoculating leaves with droplets of grape juice only. Several droplets of the inoculum were placed on each leaf to ensure complete coverage. Like the senescence time series experiment, this experiment used 4 biological replicates at each time point, and an average of 3 technical replicates for each biological replicate in the microarray experiment (Figure 1.1B). The design of the microarray experiment for the Botrytis time course was also based on a loop design, and again the data was analysed using an adapted version of the MAANOVA package (Wu *et al.*, 2003) to perform the quality assurance, normalisation and mixed model analyses.

1.3. Multiple stress responses

In the field, combinations of environmental stresses often occur simultaneously, and this can reduce the yield in agricultural crops. In general, most studies have investigated transcriptome changes in response to a single stress, and different patterns may be obtained in response to each stress (Rasmussen *et al.*, 2013). However, simultaneous exposure to combinations of stresses results in a complex set of responses to allow the plant to respond appropriately to environmental conditions. To develop transgenic crops that are resilient to these stresses, the interactions between the various stress responses need to be better understood. While the effects on development and nutrient allocation have been previously studied (reviewed in Mooney *et al.*, 1991), the genetic interactions that occur in response to combinations of stresses are generally not well understood.

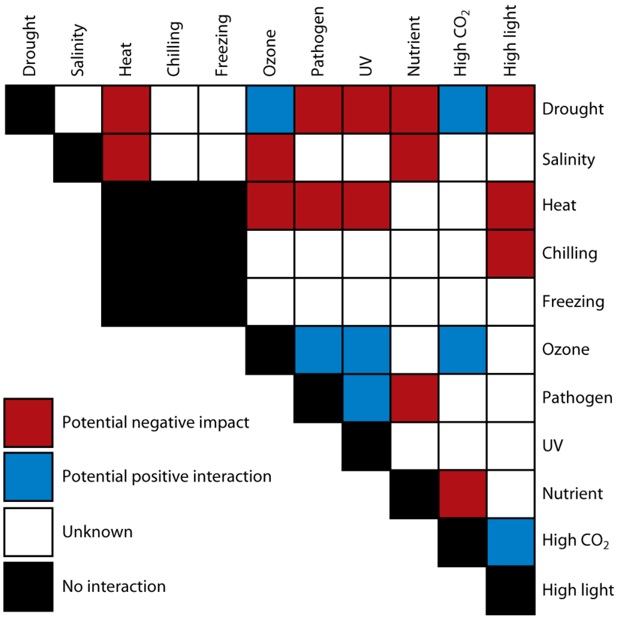
The most commonly studied combination of stresses is the effect of different abiotic stresses with various pathogens (biotic stresses) (reviewed in Mittler and Blumwald, 2010). It has been found that abiotic stresses generally resulted in increased susceptibility to biotic stresses, possibly due to the plant sacrificing the infected parts in order to ensure it survives the potentially fatal abiotic stress (Atkinson and Urwin, 2012). A summary of the effects of other combinations of environmental stresses are shown in Figure 1.2. This figure shows the stress interactions that have a negative effect on crop productivity, such as nutrient stress and drought, in addition to some environmental effects that may be beneficial to yield, such as high CO_2 and ozone.

Rasmussen *et al.* (2013) performed both single and combined stress treatments on a variety of environmental stresses, including cold, heat, high-light, salt, and flagellin treatments, and were not able to predict the majority of combined responses from the single stress responses. However, these predictions were only performed on a single time point for each stress. In a similar experiment by Prasch and Sonnewald (2013), heat and drought stresses were found to increase the susceptibility of Arabidopsis plants to turnip mosaic virus, and different sets of genes were found to be activated in the combined stresses as compared to the single stresses.

1.4. Methods for analysing gene expression

High throughput technologies, such as microarrays, provide a means to detect expression levels of multiple genes at once, allowing researchers to observe the transcriptional changes that occur within a plant in response to a stress (Naika *et al.*, 2013; Rao *et al.*, 2006). With the sequencing of the *Arabidopsis thaliana* genome (The Arabidopsis Genome Initiative, 2000), it has been possible to investigate the genetic responses of Arabidopsis to environmental stresses on a whole genome scale. These findings may potentially be applied to other economically valuable plants in the *Brassica* genus, including oilseed rape, cauliflower, and broccoli (Paterson *et al.*, 2001).

The transcriptome of an organism refers to the level of mRNA molecules (transcripts) that are being expressed by a cell at a specific point in time, and this will change depending on factors including cell type, developmental processes, or environmental stimuli. One of the most commonly used techniques to study gene expression is microarrays. Microarrays are typically glass slides which contain DNA molecules attached at fixed locations called spots or probes, with each spot representing a gene or gene fragment (Causton *et al.*, 2003). mRNA extracted from the samples are reverse transcribed to



	Drought	Salinity	Heat	Chilling	Freezing	Ozone	Pathogen	UV	Nutrient	High CO$_2$	High light
Drought	⬛	⬜	🟥	⬜	⬜	🟦	🟥	🟥	🟥	🟦	🟥
Salinity		⬛	🟥	⬜	⬜	🟥	⬜	⬜	🟥	⬜	⬜
Heat			⬛	⬛	⬛	🟥	🟥	🟥	⬜	⬜	🟥
Chilling			⬛	⬛	⬛	⬜	⬜	⬜	⬜	⬜	🟥
Freezing			⬛	⬛	⬛	⬜	⬜	⬜	⬜	⬜	⬜
Ozone						⬛	🟦	🟦	⬜	🟦	⬜
Pathogen							⬛	🟦	🟥	⬜	⬜
UV								⬛	⬜	⬜	⬜
Nutrient									⬛	🟥	⬜
High CO$_2$										⬛	🟦
High light											⬛

🟥 Potential negative impact

🟦 Potential positive interaction

⬜ Unknown

⬛ No interaction

Figure 1.2: “Stress matrix” showing the effects of various combinations of environmental stresses. Interactions that have a deleterious effect on crop production are shown in red, and those with a potentially positive effect are shown in blue. Figure obtained from Mittler and Blumwald (2010).

cDNA and labelled, typically with fluorescent Cyanine dyes. The labelled cDNA is hybridised to the spots on the slide, and the amount of cDNA that hybridises to a particular spot is relative to the expression level of the gene represented by that spot. There are typically thousands of spots on one microarray slide, and this provides researchers the opportunity to observe the changes in the entire transcriptome at a given time, in response to a given stimulus. Issues with microarrays include variation due to non-specific hybridisation, different hybridisation affinities between the spots, and differences in the labelling efficiency (Marioni *et al.*, 2008).

While microarrays are a popular technique, other high-throughput techniques also exist to study gene expression. These include serial analysis of gene expression (SAGE) (Roulet *et al.*, 2002), cap analysis of gene expression (CAGE) (Shiraki *et al.*, 2003), and massively parallel signature sequencing (MPSS) (Lu *et al.*, 2006). Whereas microarrays rely on hybridisation, these technologies are based on sequencing tags and quantifying the number of tags to infer the level of gene expression. With the advent of cheaper, high-throughput DNA sequencing technologies, techniques such as RNA-Seq have been developed, and can provide a more accurate representation of transcript levels (reviewed in Wang *et al.*, 2009). An RNA-Seq analysis is performed by producing a library of cDNA fragments from the extracted RNA, and sequencing each of these fragments. These sequence reads are generally aligned to a reference sequence, and the number of reads that align to a particular transcript provide an indication of the gene expression. Advantages of RNA-Seq over microarrays include requiring less total RNA, lower background noise, and a larger range of detectable expression levels. In addition, a complete reference genome is not required, so it is possible to detect new, unannotated genes, as well as splice variants. Some disadvantages include the potential presence of ribosomal RNA, and highly expressed genes can be problematic as the majority of tags would be mapped to them. RNA-Seq is still a relatively expensive technology to use, particularly for large scale studies, although as sequencing costs continue to decrease, it is expected that RNA-Seq will replace microarrays for the analysis of gene expression. Despite these difference between microarrays and RNA-Seq, it has been shown that the gene expression values obtained from these two methods correlated well (Marioni *et al.*, 2008), as well as complemented each other (Kogenaru *et al.*, 2012).

An issue with microarrays and RNA-Seq methods is that they are limited by the detection of low abundance transcripts, such as transcription factors (Caldana *et al.*, 2007). As a result, quantitative real-time polymerase chain reaction (qPCR) is commonly used to validate the gene expression levels determined by these methods (Kogenaru *et al.*, 2012). This methodology uses fluorescent reporter molecules (such as SYBR Green or TaqMan) together with primer pairs specific to a gene of interest to detect the levels of a target gene during the cycles of a PCR reaction. By measuring the fluorescence at each cycle, it is possible to quantify the amount of transcript present. Gene expression levels are determined using a threshold value, where the higher the gene expression of the target gene at the start of the reaction, the fewer cycles that will be required to reach this threshold (Bustin *et al.*, 2005; Nolan *et al.*, 2006). This methodology provides a sensitive, precise, and accurate assay, which is more flexible than microarrays, as it is possibly to easily add or remove target genes (Czechowski *et al.*, 2004). Although qPCR is generally used for a small number of genes, it is possible to scale up the experiment by generating primer pairs for a large library of genes to screen (Caldana *et al.*, 2007; Czechowski *et al.*, 2004). However, the number of genes that are screened are typically an order of magnitude smaller when compared to microarrays (thousands as opposed to tens of thousands). In addition, the cost of the reagents required becomes prohibitive.

A technology that has recently been garnering interest is the NanoString nCounter gene expression assay (Geiss *et al.*, 2008). This platform aims to fit the niche between the large-scale genome-wide (e.g. microarrays) and more focussed (qPCR) gene expression analyses (Kulkarni, 2011; Malkov *et al.*, 2009). This method uses a capture probe to bind to the target mRNA at the 5'-end and contains an affinity tag (such as biotin), as well as a reporter probe to bind to the 3' end, which contains a series of fluorophores. The order of these fluorophores acts as a unique identifier for the gene of interest. After removing excess probes, an image is acquired, and the number of each of the coloured tags is determined, providing an indication of the gene expression of the genes of interest in the sample. Advantages are that it can be used with any nucleic acid sequence, so RNA does not need to be amplified or reverse transcribed, and the process can be multiplexed with up to 800 different genes (Geiss *et al.*, 2008; Kulkarni, 2011). However, as it is a relatively new technology, the cost is still high.

The previous technologies require the extraction of the genetic material from the samples, and then quantifying them using various techniques. One technology that does not require this is the use of a luciferase transcriptional reporter (Brasier *et al.*, 1988; Millar *et al.*, 1992, 1995). By attaching the coding sequence of the firefly luciferase enzyme to the promoter region of a gene of interest, and transforming plants to constitutively express this construct, it is possible to produce a sensitive, non-invasive, real-time reporter of gene expression in a plant (Brasier *et al.*, 1988; Millar *et al.*, 1992). Using a low light camera, it is possible to identify the spatial and temporal activity of the gene of interest. The main disadvantage in this technology is the requirement of the expensive imaging equipment, as well as the large amount of time required to generate the transgenic lines (Millar *et al.*, 1992, 1995).

Microarrays and other similar technologies have made it possible to observe changes in the expression of a multitude of genes, under a range of conditions, and at different time points. The identification of patterns in the gene expression may be indicative of a variety of biological phenomena, from disease and developmental states, to responses to stimuli (Causton *et al.*, 2003; Wang *et al.*, 2009). While a single microarray is valuable for quantifying the level of gene expression at a single point in time, it is possible to determine the expression levels of genes over a number of time points (Storey *et al.*, 2005). In this way, these time series experiments can be used to observe the dynamics of the system to determine the different genes that are up- or down-regulated in response to some stimuli (Androulakis *et al.*, 2007; Bar-Joseph, 2004; Kiddle *et al.*, 2010).

1.5. Clustering data

Clustering is a generic term that is used for the grouping of objects into groups, or clusters (Jain *et al.*, 1999). The aim of the analysis is to identify groups of observations that are similar to each other, but distinct from all other groups. This form of analysis has been applied in many research areas, including medicine (e.g. Everitt and Hothorn, 2010), marketing science (e.g. Dolnicar, 2003), image analysis (e.g. Everitt and Bullmore, 1999) and biological data (e.g. Eisen *et al.*, 1998). Even within these different fields, the applications of clustering can be vast, from pattern analysis and grouping, to machine-leaning and data mining. This diverse range of application areas reflects the broad appeal and usefulness of the approach in the exploration of a large dataset (Everitt and Hothorn, 2010; Jain *et al.*, 1999). It should be noted that clustering is an unsupervised form of classification. This means that the data points are unlabelled, and there is little to no prior information regarding the data. This provides the researcher with a large amount of flexibility as to the way the analysis proceeds, but also requires that the researcher not make any invalid assumptions about the data (Jain *et al.*, 1999).

The most common clustering algorithms are the hierarchical methods, where objects are compared in a pairwise manner. These methods may be divided into agglomerative and divisive methods. Agglomerative methods produce clusters in a “bottom-up” fashion where each object starts in its own cluster, and pairs of objects or clusters are successively merged together. Divisive methods are the opposite, where all the objects start in a single global cluster, and are divided up as the process continues. In addition, it is also possible to obtain “fuzzy” clusters, where each object contains a probability of belonging to each cluster (Kaufman and Rousseeuw, 2009).

Hierarchical clustering relies on determining the dissimilarity, or distance, between two objects. The most common method to do this is to use a Euclidean or Manhattan city-block distance. To group objects into clusters, a linkage method is used to determine if a pair of genes or clusters should be merged (in the case of agglomerative clustering) or divided (in divisive clustering). The most common linkage methods include single linkage which identifies objects with the smallest distance, complete linkage which identifies objects with the largest distance, and average linkage which compares the average distances between all pairs of objects, with one object from each cluster. (Izenman, 2008; Jain and Dubes, 1988)

Many hierarchical clustering algorithms create a dendrogram as a representation of the relation of the observations to each other, where the branches are the clusters. Typically this dendrogram is cut at some arbitrary cut-off in order to form the desired clusters. Alternatively, the dendrogram may be cut in such a way that a predetermined number of clusters in formed, but this requires prior information as to the structure of the clustering. Langfelder *et al.* (2008) present a more objective approach where the number of clusters are automatically determined by detecting patterns which are indicative of an underlying subcluster structure, such as the minimisation of the intra-cluster distance, and the maximisation of the inter-cluster distance.

Eisen *et al.* (1998) were among the first to apply a clustering algorithm to gene expression data. The hypothesis was that genes which clustered together (co-expressed genes) would have similar functions, be involved in the same metabolic pathway, and possibly co-regulated by the same transcription factors (Williams and Bowles, 2004). By using hierarchical clustering, with an uncentred correlation distance and centroid linkage, yeast microarray data from various stages in the organism's life cycle, and under different abiotic conditions was analysed (Eisen *et al.*, 1998).

Partitioning (non-hierarchical) methods have also been applied to cluster gene expression data, each with its own assumptions and biases. Some examples include the use of self-organising maps to identify similar sized sets of genes (Tamayo *et al.*, 1999), and the use of a small set of marker genes in an algorithm called gene shaving (Hastie *et al.*, 2000). Other popular clustering algorithms include k-means clustering and Partitioning Around Medoids (PAM), where in both cases, k representative objects (genes) are randomly selected, and each other object is assigned to the most similar representative object (Kaufman and Rousseeuw, 2009).

Using a variety of clustering methods, researchers have been able to identify groups of genes that participate in similar signalling and metabolic pathways (Williams and Bowles, 2004), sets of genes with similar expression profiles were found across multiple datasets, under different temporal, environmental and genetic conditions (Mentzen and Wurtele, 2008), and co-expressed genes could be differentiated between various environmental stresses (Weston *et al.*, 2008).

While the identification of an appropriate clustering algorithm is important, the assessment of the resultant clusters is equally important (Jain and Dubes, 1988). Different clustering algorithms can potentially generate very different clusters, even on the same dataset. As a result, a cluster may be an arbitrary shape and size and it can be difficult in determining the significance of the clusters. In addition, the interpretation of clusters is highly subjective, as no precise definition of what a cluster should be exists (Causton *et al.*, 2003; D'haeseleer, 2005). If the true grouping of a set of genes is known, the clusters can be assessed using this. However, these "gold standards" rarely exist (Jain and Dubes, 1988; Yeung *et al.*, 2001). Generally, the results from a clustering is assessed though visual inspections, often with the use of prior biological information (Androulakis *et al.*, 2007).

1.6. Clustering time series data

Time series data refers to experiments when responses were obtained at multiple time points. This type of data may be cross-sectional or longitudinal. Cross-sectional data indicates that each data point was obtained from an independent sample (e.g. leaf samples at different times), whereas in longitudinal data, subsequent data points are obtained from the same individual (e.g. blood samples for a patient at different times).

Traditional clustering algorithms, such as those described above, are not suitable for time series data as they assume that the observations are independent of each other and disregard the temporal relationship between the observations (Bar-Joseph, 2004; Cooke *et al.*, 2011; Heard *et al.*, 2006; Lin *et al.*, 2008). In time series data, there is an obvious dependence of each observation on the past observations (Ernst *et al.*, 2005; Ma *et al.*, 2006).

Numerous techniques have been developed to take this temporal information into account, and include the use of Bayesian-based hierarchical clustering algorithms (Cooke *et al.*, 2011; Heard *et al.*, 2006), smoothing spline clustering (Ma *et al.*, 2006), hidden Markov model (HMM) algorithms (Oh *et al.*, 2013; Schliep *et al.*, 2003), and curve fitting using smoothing spline clustering models (Déjean *et al.*, 2007; Ernst *et al.*, 2005; Ma *et al.*, 2006). Ernst *et al.* (2005) used an approach where a set of possible gene expression profile combinations over a series of time points are calculated and the statistically significant profiles identified. In addition, many algorithms exist to specifically identify circadian patterns in the time series data. These include using time-domain (Straume, 2004) and frequency-domain methods (Chudova *et al.*, 2009; Yang and Su, 2010), as well as using a Fourier series approximation of periodic gene expression (Wang *et al.*, 2012). Multiple techniques have also been integrated to improve the grouping accuracy, as performed by Déjean *et al.* (2007) where spline smoothing was performed to reduce the amount of noise prior to hierarchical clustering. These techniques can be computationally expensive (Bar-Joseph, 2004; Ma *et al.*, 2006). For example, the approach by Ernst *et al.* (2005) is only tractable for experiments with small numbers of time points, due to the exponential growth in the number of permutations with an increase in time points, and HMM clustering tends to overfit the data when there are less than ten time points (Ma *et al.*, 2006). Techniques that utilise splines (a segmented polynomial function) are popular as they are able to model any unknown shape with relative amount of ease, without requiring any prior information about the structure of the data (Heard *et al.*, 2006; Ma *et al.*, 2006; Seber and Wild, 1989). However, these methods calculate a smoothed curve through the data, which may not be biologically relevant. In addition, they are sensitive to noise and can overfit the data (Bar-Joseph, 2004), meaning the model is overly complex, and describes random noise. Most clustering methods, including those described above, often group observations together based on the profiles appearing the same. However, in a biological system, these methods may not identify genes which regulate each other. Kiddle *et al.* (2010) presented a method of clustering time series expression data using an affinity propagation algorithm to identify profiles with a transient correlation or time delay (possibly indicating the presence of a regulator that is expressed before the cluster of genes it regulates), as well as inverted profiles (possibly indicating negative regulation).

1.7. Regression models to analyse time series gene expression profiles

Gene expression time courses are hard to interpret, are often noisy, and can have measurements from irregular intervals (Chechik and Koller, 2009). These gene expression profiles can be analysed through the use of both parametric and non-parametric models. Parametric models have a predefined structure, with a fixed number of parameters, and these parameters often have physical interpretations. Non-parametric models, on the other hand, do not have a specified model structure, and may have no parameters, such as kernel smoothers, or a flexible (theoretically infinite) number of parameters, such as spline smoothers (Eilers and Marx, 1996). However, the values of the parameters are not directly interpretable in terms of the data. Most of the algorithms described in the previous section used non-parametric models. While these non-parametric techniques such as splines may provide a greater deal of flexibility, the regression approach provides a defined set of interpretable parameters, which can be used to provide more information regarding the underlying system.

Parametric regression analysis is a common technique that has been applied to multiple fields of science, including ecology (e.g. Dalbiès-Dulout and Doré, 2001 and Schoolfield *et al.*, 1981), analytical chemistry (e.g. Watkins and Venables, 2006), and medical statistics (e.g. Woolcock *et al.*, 1984) where a specific model is fitted to some data. In all these cases, a parametric model is used to describe the relationship between the response and the predictor. In addition, many other analyses (both parametric and non-parametric) take only the mean of the replicates into account, effectively ignoring the between-replicate variation. It has been shown that the inclusion of the replicate information can greatly improve the analyses (Cooke *et al.*, 2011). A regression approach is able to use this information as an indicator of the quality of the fitted model in the form of deviations of the functions to the data, or the level of uncertainty in the parameter estimates (Motulsky and Ransnas, 1987).

Since the expression of genes in cells occurs continuously, it can be represented as a continuous function, namely a curve (Ma *et al.*, 2006). Regression analyses can thus be used to obtain a better explanation into the function and mode of operation of genes, by using the parameters to provide insights into differences between sets of genes, or indicating when particular events occur (Eastwood *et al.*, 2008; Ma *et al.*, 2006; Seber and Wild, 1989; Watkins and Venables, 2006). For example, Watkins and Venables (2006) used the fitted values for a parametric model to identify the optimal separation point and pH of four related carboxylic acids. Similarly, Eastwood *et al.* (2008) used the critical-exponential model to describe the expression changes in a number of genes, and used the fitted parameters to identify the time of maximal transcript level, and identify the asymptotic response level.

A regression model defines how the value of a dependent variable Y changes in response to changes in the independent variable X, with one or more model parameters (Motulsky and Christopoulos, 2004). The X term is also called the regressor, and the values of X are fixed settings during an experiment, i.e. the times of sampling. In contrast, the value of Y is determined by the underlying system processes and is predicted using values of X (Ratkowsky, 1990; Seber and Wild, 1989). The values of the unknown parameters can then be estimated by observing the values of Y for a given set of X (Seber and Wild, 1989). In terms of gene expression, the values of Y would be the level of gene expression for a gene, for a given set of sample time points (X).

The broad aim of regression is to describe the relationship between the response variable Y and the regressor X (Ritz and Streibig, 2008). In the simplest case, this would be a linear response, where a straight line is used to model the response variable (Bates and Watts, 1988; Draper and Smith, 1981; Rawlings *et al.*, 1998). However, they are generally not complex enough for gene expression data, and so more complex, nonlinear models are often used (Bates and Watts, 1988). The predictor and response variables are related through some predefined function f, which can be estimated by a set of p parameters $\beta = (\beta_1, \ldots, \beta_p)$, and a general form can be written as $y = f(x, \beta)$ (Motulsky and Christopoulos, 2004; Ratkowsky, 1990; Ritz and Streibig, 2008; Seber and Wild, 1989). The most common method to determine the best set of parameters for a given function is to use a least squares process to minimise the residual sum of squares ($SS_{residual}$).

$$SS_{residual}(\beta) = \sum_{i=1}^{n}(y_i - f(x_i, \beta))^2.$$

This means that for a given set of observed data points, $y = y_1, y_2, \ldots, y_n$, the sum of the squared vertical differences between the data points and the fitted line using the parameters β is minimised, and the set of parameters that give the best estimates is denoted $\hat{\beta}$. In linear regression, this process can be performed using algebraic solutions, and as such, the functions are computationally simple. To perform a nonlinear regression analysis, starting parameter estimates are required. An iterative process is then performed, whereby at each step a new set of parameters is determined based on the data, the model, and the current parameter values. Ideally, this would result in the optimal parameters being approached in a stepwise manner (Motulsky and Christopoulos, 2004; Ritz and Streibig, 2008). Since many parameters for nonlinear curves have physical interpretations, it is possible to find sensible starting values by plotting the data and predicting the starting values from aspects of the shape of the data (Motulsky and Christopoulos, 2004; Seber and Wild, 1989). However, for high-throughput analyses, where there are a large number of responses, this would be a time consuming and arduous process. In these cases, self-starter functions may be used to provide estimates of the starting values based on shape characteristics of the data (Ritz and Streibig, 2008).

A common algorithm to perform the parameter optimisations is the Gauss-Newton method, which uses Newton's optimisation algorithm to adjust the given starting values at each iteration in order to find a set of parameter values that results in the lowest residual sum of squares (Myers, 1990). There can be issues with the Gauss-Newton algorithm where incremental changes can be poorly estimated, and a large number of iterations may be needed resulting in slow convergence (Motulsky and Ransnas, 1987). In addition, good starting values are essential, otherwise the optimisation procedure may move in the wrong direction. Other popular methods include the steepest descent algorithm, the Marquardt-Levenberg algorithm (Bates and Watts, 1988; Draper and Smith, 1981; Motulsky and Christopoulos, 2004; Rawlings *et al.*, 1998; Ritz and Streibig, 2008), and the Nelder-Mead algorithm (Nelder and Mead, 1965).

Given a fitted curve, it is possible to determine how well a particular model fitted the data. Statistics that describe this aspect of the fit are called the model goodness-of-fit (Motulsky and Ransnas, 1987), and some are described below. For any regression model, it is possible to partition the total variability such that

$$SS_{total} = SS_{regression} + SS_{residual}$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1.1}$$

where SS_{total} is the total sum of squares, $SS_{regression}$ is the regression sum of squares, $SS_{residual}$ is the residual sum of squares, n is the number of time points, $\bar{y}$ is the mean of the samples and $\hat{y_i}$ is the predicted value of the fitted curve at time i. This relationship is known as the law of total variance, or the decomposition of variability (Johnson and Bhattacharyya, 2001; Myers, 1990). The SS_{total} describes the total deviation of the data about the mean, whereas the $SS_{regression}$ describes how much variability in the data is explained by the model (a measure of how well the model fits the data), and the $SS_{residual}$ describes the amount of error between the fitted model and the data (the unexplained variability).

If replicate samples (multiple y observations at a given x-value) were obtained, it is possible to extend equation (1.1) to include the replicative observations at each x-value. Assuming equal replication at each x-value,

$$\sum_{i=1}^{n}\sum_{j=1}^{r}(y_{ij}-\bar{y})^2=\sum_{i=1}^{n}\sum_{j=1}^{r}(\hat{y}_{ij}-\bar{y})^2+\sum_{i=1}^{n}\sum_{j=1}^{r}(y_{ij}-\hat{y}_i)^2 \tag{1.2}$$

where r is the number of replicates per time point.

It is then possible to further decompose the $SS_{residual}$ into two components - the pure error and the lack-of-fit sums of squares (Brook and Arnold, 1985).

$$SS_{residual}=SS_{pure\,error}+SS_{lack-of-fit}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{r}(y_{ij}-\hat{y}_i)^2=\sum_{i=1}^{n}\sum_{j=1}^{r}(y_{ij}-\bar{y}_i)^2+\sum_{i=1}^{n}r(\bar{y}_i-\hat{y}_i)^2. \tag{1.3}$$

The pure error sum of squares is defined as the sum of squared differences between the average y-value at each x-value ($\bar{y}_i$, the local average of replicate samples) and each observed value at that x-value, and describes the variation that is due to the between-replicate variation, pooled across the time points (pure random error). The lack-of-fit sum of squares is a sum of squared differences between the local average and the fitted value at the same x-value, weighted by the number of replicates. This value describes the amount of variation due to lack of model fit (variation between the replicate means and fitted curve). If the lack-of-fit sum of squares makes up a large proportion of the $SS_{residual}$, this suggests that the fitted model is insufficient to fit the data and indicates a poor fit to the mean y-values.

The R^2 value (*coefficient of determination*) is a commonly used statistic to determine the goodness-of-fit of a fitted curve, with values ranging between 0 and 1, where 1 indicates that the fitted line passes perfectly through the data points (Motulsky and Christopoulos, 2004). This value indicates the proportion of the variance that is accounted for by the fitted curve.

$$R^2=1-\frac{SS_{residual}}{SS_{total}}.$$

While the above statistics are useful for determining the goodness-of-fit for a model fit, they are generally poor indicators for model selection (Burnham and Anderson, 2002). Thus, when selecting which of the various regression models best fitted the data, *Akaike's Information Criterion (AIC)* is commonly used. This statistic can be used to compare both nested and non-nested models (nested models are those that are mathematically related, where the terms of a smaller model occur in a larger model, such as the exponential and critical exponential curves). The AIC is calculated using aspects of maximum likelihood theory, information theory and entropy of information (Motulsky and Christopoulos, 2004), and is calculated as follows:

$$AIC=N.\log\left(\frac{SS_{residual}}{N}\right)+2p$$

where N is the total number of data points and p is the number of parameters. The AIC equation takes into account both the goodness-of-fit and the number of parameters required, thus attempting to prevent overfitting with overly complex models (Ritz and Streibig, 2008; Motulsky and Christopoulos, 2004). The AIC statistic thus describes how likely the model is to be correct, given the data. The value of the AIC is not easily interpretable, and AIC values of models from different datasets cannot be compared (Burnham and Anderson, 2002; Motulsky and Christopoulos, 2004). However, the relative differences between the AIC values of a set of models from the same dataset can describe how much better one model fits the data than another model to the same data (Burnham and Anderson, 2002). Variations of this statistic include the Bayesian Information Criterion.

1.8. Biological significance of sets of genes

The great difficulty with cluster analysis, is that often there is no known answer. Many statistical methods exist to determine the accuracy of the clustering, such as the Rand Index and correlation-type methods (Savage *et al.*, 2009). However, these methods generally require prior information regarding how the clusters should be formed.

One method to analyse a set of differentially expressed genes is the use of gene-set enrichment analysis (GSEA) (Clark and Ma'ayan, 2011; Huang *et al.*, 2009; Subramanian *et al.*, 2005). This method requires prior information that the genes in a set are functionally related, such as genes that encode proteins that are involved in similar metabolic pathways. By comparing the differences in the expression levels of the genes in the set between two treatments, the statistical significance of the set can be calculated (Subramanian *et al.*, 2005; Thomas *et al.*, 2011). This type of approach is not easily applied in an exploratory analysis, such as when using unsupervised clustering, since the genes that are grouped together may not be known to function together.

Thus, algorithms have been developed to determine if the genes in the clusters are biologically meaningful and functionally related through the use of annotation information (Huang *et al.*, 2009). Generally, the over-representation of annotation terms is used to identify the function of a set of genes. The principle behind this, is that genes that are involved in the same or similar biological functions would have a greater proportion of the relevant annotation terms than would be expected to be found by chance. The over-representation is commonly quantified through the use of statistical methods, such as the Chi-square and Fisher's exact tests, or using the Binomial and Hypergeometric probability distributions (Huang *et al.*, 2009). However, the drawback to these annotation based metrics is annotation bias, where certain genes that are more studied have more information about them. It is estimated that in Arabidopsis, over half of the genes are annotated purely based on sequence similarity to other Arabidopsis genes. In addition, a further 30% of genes do not have any sequence similarity with a gene of known function (Quanbeck *et al.*, 2012). This implies that only around 20% of genes in the genome have been biochemically characterised.

One of the most common methods of determining biological significance of a cluster is by calculating the level of over-representation of Gene Ontology (GO) terms. GO provides a means to represent biological knowledge by using a structured language and precisely defined, interrelated terms (Ashburner *et al.*, 2000). GO is a controlled vocabulary, structured as a directed acyclic graph, where terms are connected with directed links from less specific to more specific annotations. GO consists of three separate graphs which describes the gene products in terms of the molecular function, biological process, and subcellular location (Berardini *et al.*, 2004). Over-representation is determined by counting the occurrence of each GO term in a cluster, and performing a hypergeometric test based on the number of times the term occurs in the rest of the dataset (Falcon and Gentleman, 2007).

These over-representation tests result in a test statistic to measure the level of over-representation. A p-value is produced from this to obtain an interpretation of the significance of the result. Since a large number of annotation terms are tested, a multiple hypothesis correction, such as Bonferroni or False Discovery Rate is generally performed. However, multiple hypothesis corrections would merely adjust the p-values to provide a ranking of the GO terms, and this process is not straightforward as the tests are heavily correlated (Lewin and Grieve, 2006; Zhong and Xie, 2007). It has been suggested that the p-values rather play a role in an exploratory procedure and suggest what the most significant terms are, as opposed to a statistical solution to the problem (Huang *et al.*, 2009).

An alternative method to determine biological significance is to use text mining. The use of text mining to aid in the functional interpretation of gene lists is an increasingly popular approach, particularly to exploit the large amounts of information contained in information repositories, such as PubMed (Blaschke *et al.*, 2001; Chaussabel and Sher, 2002; Glenisson *et al.*, 2004; Leong and Kipling, 2009). With the increasing availability of scientific literature, it seems reasonable to be able to leverage the information contained in this data. By applying natural language processing and other text mining techniques, it is possible to identify connections between genes that may not previously been identified. Numerous methods have been developed to use text mining to integrate data in order to interpret experimental results (Chaussabel and Sher, 2002; Glenisson *et al.*, 2004; Hassani-Pak *et al.*, 2010; Leong and Kipling, 2009).

While it can be useful to determine what annotations are over-represented within a cluster, it can also be interesting to determine if some of the genes are involved in the same, or related, metabolic pathways. It is possible to obtain this information from resources such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Ogata *et al.*, 1999) or MAPMAN (Thimm *et al.*, 2004). Both of these tools consist of databases with mappings of genes to various metabolic pathways, as well as provide users with a graphical representations of the genes that are involved in similar metabolic processes. KEGG is a database resource that links genomic information to metabolic pathway maps (Kanehisa and Goto, 2000; Kanehisa *et al.*, 2006), while MAPMAN provides a similar functionality and displays gene expression data onto diagrams of metabolic pathways (Thimm *et al.*, 2004; Usadel *et al.*, 2009).

It is generally thought that genes with similar expression profiles could be regulated by common transcription factors (TFs), and thus would have a similar promoter structure

(Jenkins *et al.*, 2013). Gene promoters are the sequences that are generally upstream of the transcription start site, which contain binding sites for TFs. These binding sites are short DNA sequences (motifs) where TFs can bind to the DNA and regulate the expression of the gene. These binding sites are conserved between TF families, for example, WRKY TFs have been found to bind to the TTGAC motif (Eulgem *et al.*, 2000). Promoter sequences are commonly analysed by searching for these motifs that are over-represented in the promoter sequences of the genes in the cluster. A popular technique for determining this is MEME (Multiple Expectation maximization for Motif Elicitation) (Bailey *et al.*, 2006, 2009). This tool identifies common motifs, which may be DNA-binding sites, that are present in the upstream regions of the genes in the cluster. This tool uses statistical modelling techniques to select motifs, and thus does not require any prior information. However, many of these motifs may not be biologically relevant. In contrast, other tools such as TRANSFAC (Matys *et al.*, 2003; Wingender *et al.*, 2000) and PLACE (Higo *et al.*, 1999) are databases which contain information on a large number of TFs, and their DNA-binding sites, across a variety of organisms. These databases contain biologically validated information from published experiments, which is more useful in determining if a specific motif is known to be involved in the stimuli being applied. It is also possible to investigate whether a particular DNA-binding sequence is evolutionarily conserved across a number of different organisms, possibly suggesting that it is vital for the expression of the gene (Baxter *et al.*, 2012).

1.9. Project aims

The overall aim of the project was to develop a statistical analysis approach that could be used to model gene expression profiles in a single stress, and use these models to predict the effect of a combined stress response. To this end, a mechanistic basis for gene expression models was developed by using a parametric regression analysis, where models were fitted to the data from the Arabidopsis long day senescence experiment (Breeze *et al.*, 2011), and response of Arabidopsis to *Botrytis cinerea* infection (Windram *et al.*, 2012). The use of the regression models to describe the shape of the response allows for a more precise description of the underlying expression profile by reducing the amount of noise in the raw data. In addition, these models contain biologically meaningful and easily interpretable parameters, and this information can be used to provide a new means of investigating gene sets based on particular features (parameters) of the gene expression profiles (Chapter 3). These biologically oriented descriptions of individual gene expression profiles thus allow for improved modelling and greater interpretation of profiles obtained from time-series experiments.

The regression models were applied by clustering the genes based on the fitted parameters. This provided a means of identifying genes that have shape characteristics aspects in common, such as starting position, or rate of change. In this way, the underlying mechanisms behind the responses could be directly interrogated. As a proof of concept of this approach, the cluster analysis was applied to a well annotated organism, namely yeast (Chapter 4), before being applied to the aforementioned PRESTA datasets (Chapter 5). In both cases, it was shown that by investigating a particular aspect of the expression profile, it was possible to extract more biological information regarding the function of the set of genes than traditional analysis methods.

Plants activate a specific and unique stress response when subjected to a combination of multiple stresses. Thus by using the fitted regression models, the effect of multiple stresses was predicted. These predictions were validated using biological techniques, such as qPCR and luciferase transcriptional reporters (Chapter 6).

Finally, an investigation was undertaken to identify a previously unexplored source of biological variability, namely that of biological age, where individual replicate samples may be developmentally dissimilar. A time shifting analyses was developed to provide an estimate of the "true" biological time associated with gene expression responses, and as a consequence, generated a more detailed time-course for the response (more time-points with fewer biological replicates) that could be used to generate better gene network models (Chapter7). Using the regression analyses, the effect of the time shifting could be evaluated.

Through careful choice of appropriate models, such statistical regression approaches allow for an improved comparison of gene expression profiles, and may provide an improved understanding of common regulatory mechanisms between genes. The development of these new tools provided a better assessment of the mechanisms underlying stress responses, and could assist in future breeding of Brassica and other crop species with improved yields in a changing climate.

2. Applied methods

2.1. Computational methods

2.1.1. Models used in regression analysis

When performing a regression analysis, it is necessary to define the model function with which to fit to the data. The functions used are summarised in Table 2.1, along with a description of the parameters, and an example figure of what the function looks like. The models are discussed in detail in Section 3.2.

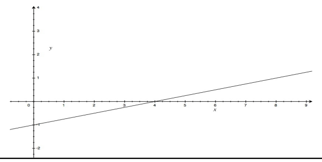
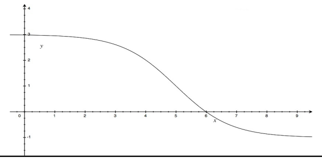
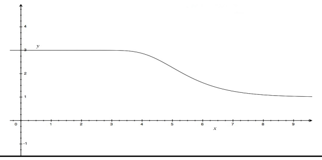
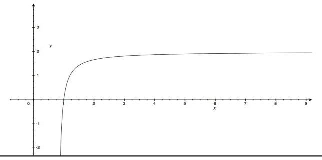
The models used were the linear, logistic, Gompertz, Gaussian, exponential, critical exponential, linear+exponential, and hyperbola. The linear model describes a straight line response, the logistic and Gompertz models describe a sigmoid response, the Gaussian describes a peaked response, the exponential describes a response with a rapid rate of growth or decay, the critical exponential describes an exponential response, with a peak and asymptote, the linear+exponential describes a exponential response with a linear component, and the hyperbola represents a response similar to the exponential, but with a faster rate of change.

2.1.2. Third party packages used for development of regression and clustering scripts

The regression and clustering analysis pipelines were developed using Python and R. Python is a scripting programming language (Van Rossum and Drake Jr, 1995), and was primarily used for data processing, including parsing the data files, performing basic analyses, and storing the results in a database. The R statistics package was used to perform the statistical analyses (R Development Core Team, 2011). The most used functions were the `lm` function to fit linear models, `nls` function for nonlinear models, and the `aov` function to perform analysis of variance calculations.

In addition to using the core libraries of these programming languages, a number of third-party packages were also used to aid in the development of the analyses. These packages and the versions used are listed in Table 2.2.

Table 2.1: Summary of the models used in the regression analysis, showing the equation for the curves, the interpretation of some of the parameters, as well as an example of the model is shown. The value of y indicates the level of gene expression, and the value of t is an indication of time.

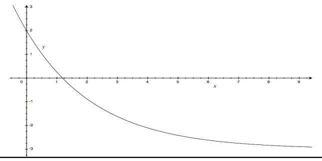
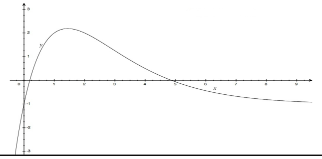
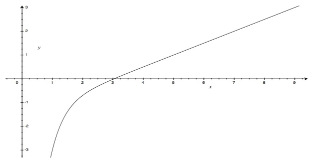
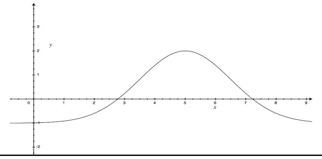
Model name	Equation	Parameter characteristics	Example of model shape
Linear	$y = c + m \cdot t$	- rate of change (m) - y-intercept (c)	
Logistic	$y = a + \frac{b}{1+\exp((m-t)/s)}$	- time of max growth (m) - max growth rate (s) - range (b) - asymptote (a)[1]	
Gompertz[2]	$y = a + b \cdot \exp(-\exp(s \cdot (t - m)))$	- time of max growth (m) - max growth rate (s) - range (b) - asymptote (a)[3]	
Hyperbola	$y = \frac{a \cdot (t-c)}{b+(t-c)}$	- rate of change (b) - upper asymptote (a) - horizontal shift (c)	

[1] Which asymptote depends on the sign of s. If $s<0$, a=right asymptote; if $s>0$, a=left asymptote.

[2] Due to the asymmetricity of the Gompertz curve, there are two possible versions of the curve - one where the growth rate is faster on the left of the midpoint, and one where the growth rate is faster to the right of the midpoint.

[3] Which asymptote depends on the sign of s. If $s>0$, a=right asymptote; if $s<0$, a=left asymptote.

Table 2.1 (cont)

Model name	Equation	Parameter characteristics	Example of model shape
Exponential	$y = a + b \cdot \exp(-r \cdot t)$	- rate of change (r) - asymptote (a) - concavity (b)[4]	
Critical exponential	$y = a + (b + c \cdot x) \cdot \exp(-r \cdot t)$	- rate of change (r) - asymptote (a) - size of peak (b) - concavity (c)[5]	
Linear + exponential	$y = a + b \cdot \exp(-r \cdot t) + c \cdot t$	- rate of change of exponential portion (r) - rate of change of linear portion (c) - related to vertical shift (a) - concavity (b) [4]	
Gaussian	$y = a + b \cdot \exp(\frac{-(t-m)^2}{2 \cdot s^2})$	- time of max/min (m) - spread around m (s) - range (b) - base level (a)	

[4] Concavity dependant on the signs of b and r.

[5] Concavity dependant on the signs of c and r

Table 2.2: Table of packages used in the development of the regression and clustering software.

Package Name	Function	Version
car	Companion to Applied Regression. R package required for the deltaMethod function	2.0-16
clValid	Provides the BHI calculation	0.6-4
dynamicTreeCut	R package used to dynamically determine where to cut the clustering dendrogram	1.21
goStats	GO over-representation tests	2.20.0
Graphviz	Package used for laying out graphs. Accessed through Python with pygraphviz (1.0)	2.28.0
nltk	Natural Language Toolkit. Python package used for the analysis of word counts	2.0.1rc1
org.At.tair.db	GO annotation data for Arabidopsis	2.6.4
org.Sc.sgd.db	GO annotation data for yeast	2.6.4
Python	Primary scripting language to parsing and working with data	2.7.2
R	Statistics calculations	2.14.0
ReportLab	Python package used to generate PDF reports of the outputs	2.5
RPy	Linking language between Python and R	2.2.1
sqlalchemy	Python module to connect to the database	0.7.3
SQLite	Database engine	3.7.9

2.1.3. Formation of clusters

In a cluster analysis, a distance matrix is created illustrating how dissimilar the fitted parameters are for a given set of genes. By applying the `hclust` function in R to this distance matrix, it is possible to perform a hierarchical clustering to group the most similar genes together. During the clustering procedure, the distances between clusters are calculated using a linkage function. In this case, the average linkage distance was used, where the distance between two clusters is calculated as the average distance between all pairs of genes in the first and second clusters (D'haeseleer, 2005). Mathematically, this is represented as

$$L(X,Y) = \frac{1}{N_X \cdot N_Y} \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} d(x_i, y_j)$$

where L is the linkage function between clusters X and Y, N_X and N_Y are the number of genes in clusters X and Y respectively, and *d(x,y)* is the distance between a gene in cluster X and a gene in cluster Y.

Typically in hierarchical clustering, clusters are determined using a predefined, and often arbitrary, dendrogram height cut-off. This means that the dendrogram tree structure is cut at a fixed level of similarity, and each branch below this height is designated as a cluster. Instead, here the *Dynamic Tree Cut* package was used to identify the clusters in a more objective manner (Langfelder *et al.*, 2008). This algorithm uses a bottom-up approach, where preliminary clusters are initially identified using the dendrogram. These clusters are determined by ensuring that the following criteria are met: minimum cluster size, minimisation of within-cluster distance, maximisation of between-cluster distance, and that the tip of the branch (the "cluster core") is tightly connected. Following this, a "branch pruning" is performed where any unassigned genes are assigned to the closest cluster, while still maintaining the above criteria. This final step of assigning objects makes the algorithm a hybrid of hierarchical clustering and partitioning around medoids (PAM) (Langfelder *et al.*, 2008). Thus, this algorithm provides an objective and automated approach for determining the clusters.

The parameters used were a cut distance of 0.1 (90% similarity), using the Dynamic Hybrid algorithm with a deep split level of 2, allowing PAM, and not requiring the PAM stage to be constrained by the dendrogram. The deep split level ranges from 0 to 3, and influences the number of clusters, where larger values produces more clusters with a greater within-cluster distance, and separated by smaller between-cluster distance.

2.1.4. Gene Ontology (GO) analysis

The Gene Ontology (GO) is one of the most commonly used methods to annotate genes, and provides a structured and controlled vocabulary to describe the molecular function, biological process, and subcellular location of a gene product (Ashburner *et al.*, 2000). Functional analysis on a cluster of genes is often calculated based on the probability of whether the group of genes associated with a specific GO term occurs more often than would be expected. To do this, a background distribution of the GO terms is used, and in these analyses, the GO terms associated with all the unique gene models on the microarray slides were used (Section 1.2). The probability of finding a number of a specific GO term, given its background distribution, can be modelled by the hypergeometric distribution and is calculated as

$$P(X = k) = \frac{\binom{N_t}{k}\binom{N - N_t}{D - k}}{\binom{N}{D}} \tag{2.1}$$

where N is the total number of genes, N_t is the size of the cluster containing GO term t, k is the number of genes with GO term t, and D is the total number of genes with GO term t. A p-value is calculated from this by determining the probability of identifying at least k genes with term t, and is calculated as

$$p(t) = 1 - \sum_{i=0}^{k-1} P(X = i). \tag{2.2}$$

The smaller the p-value, the higher the significance of the over-representation.

Due to the hierarchical nature of the GO annotations, there is a large degree of overlap in the annotations, since the child GO term inherits all its parent's annotations. Thus, to identify over-represented GO terms, the `GOstats` R package (Falcon and Gentleman, 2007) was used where the hypergeometric test is first performed on the terms with no children (the leaves of the GO annotation tree). If a term was found to be significant, the parents of this term were removed from further testing, and this process is repeated until all the terms have been tested (Falcon and Gentleman, 2007).

The GO annotation terms were obtained from BioConductor, namely the `org.At.tair.db` package for Arabidopsis (Carlson, 2012*a*), which uses the TAIR (The Arabidopsis Information Resource) annotations (20110910 date stamp) (Lamesch *et al.*, 2012), and the `org.Sc.sgd.db` for yeast (Carlson, 2012*b*), which uses the SGD (Saccharomyces Genome Database) annotations (20110924 date stamp) (Cherry *et al.*, 2012).

After performing a GO analysis, a multiple hypothesis correction, such as Bonferroni or False Discovery Rate is generally performed. However, multiple hypothesis corrections would merely adjust the p-values to provide a ranking of the GO terms, and this process is not straightforward as the tests are heavily correlated (Lewin and Grieve, 2006; Zhong and Xie, 2007). Thus, in these analyses, a multiple hypothesis correction was not performed, as it is not the ranking of the most significant terms that is of interest, but rather if there are sets of related terms that may be over-represented. That is, the GO terms play an advisory role to suggest what the most significant terms are, as opposed to a means to make a decision, such as in clinical studies (Huang *et al.*, 2009).

2.1.5. Word over-representation

As an alternative to using GO to determine biological significance, it is possible to use word lists. GO over-representation tests can be affected by annotation bias, where terms appear to be over-represented simply due to the fact that the background frequency of the terms is artificially under-estimated (Leong and Kipling, 2009).

To this end, the words used in the functional annotations for each gene were used to determine if any were over-represented. Only the unique words in each gene's annotation were used to avoid false positives due to repeated words, and punctuation and individual numbers were ignored. In addition, two- and three-word phrases (bigrams and trigrams, respectively) were identified to provide a greater level of context.

Similar to the GO analysis, a simple hypergeometric test was performed to determine if any words or phrases were over-represented in the annotations of the gene. The p-values were calculated in the same manner as in Equations (2.1) and (2.2), except t is a word or phrase from the annotation, N is the total number of words or phrases, N_t is the size of the cluster containing annotation term t, k is the number of genes with annotation term t, and D is the total number of genes with annotation t.

All frequency calculations were determined using the Python Natural Language Toolkit (NLTK) (Bird *et al.*, 2009), and the p-values were calculated using the `phyper` function in R. Annotations were obtained from TAIR (TAIR10 annotations) (Lamesch *et al.*, 2012) and SGD (25/02/2013 annotations) (Cherry *et al.*, 2012) for Arabidopsis and yeast, respectively.

2.1.6. Pathway information

It can be useful to determine if some of the genes are involved in the same, or related, metabolic pathways. It is possible to obtain this information from resources such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Ogata *et al.*, 1999) or MAPMAN (Thimm *et al.*, 2004).

In this analysis, the mappings of genes to metabolic pathway categories in MAPMAN were extracted, and used to determine if any genes in a cluster encode proteins that are involved in any of the metabolic pathways that are defined.

2.1.7. Promoter analysis

In addition to the above annotation information, it is also possible to determine if there are any over-represented upstream regulatory sequences. This could possibly indicate the presence of a transcription factor which is able to regulate multiple genes.The promoter analysis was performed as described in Breeze *et al.* (2011). In brief, 351 binding motifs in plants were obtained from the TRANSFAC (Matys *et al.*, 2003) and PLACE (Higo *et al.*, 1999) databases. Motifs were identified from sequences 500 bp upstream of the transcription start site of each gene in the cluster. For each motif, the frequency in the cluster was computed, and compared to the frequency in the Arabidopsis genome. A hypergeometric test was used to provide a description of the significance of the presence of a motif.

2.1.8. Biological Homogeneity Index

The Biological Homogeneity Index (BHI) is a metric developed by Datta and Datta (2006), which determines how homogeneous the annotations within a cluster are. That is, it determines the number of shared annotations between genes. It is thought that a clustering algorithm which generates biologically meaningful clusters will group genes that have a similar biological function. As a result, this metric can be used to compare a number of different clustering algorithms applied to the same dataset. The BHI value is calculated as

$$BHI = \frac{2}{k}\sum_{j=1}^{k}\frac{1}{n_j(n_j-1)}\sum_{x \neq y \in D_j} I(C(x) = C(y))$$

where the k is the number of clusters, n_j is the number of annotated genes in cluster D_j, *C(x)* and *C(y)* are the GO terms for genes x and y respectively, and the indicator function, *I*, is equal to 1 if there are any common terms between genes x and y, and 0 if not (Datta and Datta, 2006). The BHI value ranges between 0 and 1, where 0 indicates that there are no common annotations within genes in a cluster, and 1 indicates that all the genes in each cluster are annotated as having the same function. By using the GO terms associated with the genes in the cluster, it is thus possible to determine if there are a large number of genes that have the same GO term, and thus, be involved in the same biological function.

The BHI function was applied by using the `clValid` R package (Brock *et al.*, 2008), and uses the same annotation sources as the `GOstats` package described above, namely the `org.Sc.sgd.db` BioConductor annotation database for yeast (Carlson, 2012*b*), and the `org.At.tair.db` database for Arabidopsis (Carlson, 2012*a*).

2.1.9. Fitting a penalised spline

For the time shifting analysis, a penalised spline was used for fitting the gene expression profiles, using an implementation in R, as described by Eilers and Marx (Eilers and Marx, 1996, 2010). The number of segments used was 5, using a degree of 3 for the B-splines, and an order of difference penalty equal to 2. A range of lambda values were used, ranging from 0.1 to 1000, on a logarithmic scale. This was calculated as $lambda = 10^i$ where $i \in \{-1; -0.8; \ldots; 2.8; 3\}$. The lambda value used is the value which minimised the cross-validation value that is returned from the spline fitting procedure.

2.2. Experimental methods

2.2.1. Multiple stress analysis

2.2.1.1. Plant growth

Arabidopsis seeds (wildtype Col-0) were stratified in 0.1% w/v agarose at 4ºC for 72 hours in the dark. Stratified seeds were sown in pre-watered Arabidopsis soil mix (6:1:1 ratio of Levington F2S compost:sand:vermiculate) in 4 cm pots (P24, Plankpak). Pots were covered with cling film and placed in a growth chamber to germinate. The covers were removed 7 days after sowing (DAS) and the seedlings thinned out so that there was one plant per pot. Plants were grown in 16 hour light conditions (120µmol photons.m^{-2}.s^{-1}) at 20ºC, 60% humidity, and 350 ppm CO_2.

2.2.1.2. Botrytis inoculation

Two weeks prior to use, the *Botrytis cinerea* pepper strain (Denby *et al.*, 2004) spores were germinated and cultured on sterile tinned apricot halves (Tesco) in deep Petri dishes, and incubated at 25ºC in complete darkness. Prior to infection, the spores were harvested and suspended in sterile water, and filtered through glass wool cloth. The inoculum was prepared finally in half-strength sterile grape juice (Tesco) and the concentration adjusted to 100 000 spores/ml. The spore concentration was measured using a hemocytometer.

Harvested leaves were placed on 0.8% w/v plant agar (Duchefa Biochemie) in propagator trays. Several 0.6 µl droplets of the above inoculum were placed on each leaf (between 3-6 spots, depending on the size of the leaf) ensuring coverage of the leaf. The propagator trays were covered with lids, and incubated under the same conditions as plant growth (Section 2.2.1.1), except increasing the relative humidity to 90%. Mock inoculations were performed by applying droplets of half-strength grape juice to the leaves.

2.2.1.3. RNA extraction

RNA extractions were performed by homogenising the leaf tissue (approximately 1 g) with 1ml TRIzol (Invitrogen) in a pre-chilled Dremel drill. Samples were incubated at room temperature for 5 minutes to allow for the dissociation of nucleoprotein complexes before the addition of 200 µl chloroform. The samples were shaken vigorously by hand for 15 seconds and incubated for a further 3 minutes at room temperature. The samples were then centrifuged at 8 000 x*g* for 15 minutes at 4ºC. The upper aqueous phase was transferred to a fresh 1.5 ml Eppendorf tube, followed by the addition of 500 µl isopropanol, and incubated for 2 hours at -20ºC. The samples were centrifuged at 8 000 x*g* for 20 minutes at 4ºC. The RNA pellets were washed with 1 ml 70% v/v ethanol (made up with diethylpyrocarbonate-treated [DEPC] water), followed by centrifugation at 8 000 x*g* for 10 minutes at 4ºC. The supernatant was completely removed and the pellets allowed to air dry for 5 minutes before re-suspension in 100µl DEPC-treated water. Each sample was purified using the RNAeasy purification kit (QIAgen), according to manufacturer's instructions. The samples were eluted in 50 µL DEPC-treated water. Total RNA concentrations were determined using a Nanodrop ND-1000 spectrophotometer (Thermo Scientific) using a 1 µl sample. Total RNA integrity was determined using a 2100 BioAnalyzer with the RNA 6000 Nano LabChip kit (Agilent), according to manufacturer's instructions. RNA samples were stored at -80ºC.

2.2.1.4. cDNA synthesis

Using the RNA concentrations determined from the Nanodrop above, the appropriate volume of RNA solution was used to ensure a starting quantity of 1000 ng RNA. Prior to cDNA synthesis, any DNA in the sample was removed by treating the RNA sample with 1 µl RQ1 DNaseI (Promega), 1 µl 10xRQ1 DNaseI buffer, and enough water to make up the solution to 10 µl. This sample was incubated at 37ºC for 30 minutes, before the addition of 1 µl RQ1 DNase Stop Solution (Promega). The sample was incubated at 65°C for 10 minutes to inactivate the DNase.

The first strand synthesis was performed by the addition of 1 µl 50 uM oligo(dT)$_{18}$ primers and 1 µl 10 mM dNTPs to the above solution. The sample was incubated at 65ºC for 5 minutes to allow the primers to anneal to the RNA. Following the incubation, 4 µl 5X First Strand Buffer (Invitrogen), 2 µl 0.1 M DTT (Invitrogen) and 1 µl RNaseOUT Recombinant Ribonuclease Inhibitor (Invitrogen) was added to the sample, and incubated at 42ºC for 2 minutes. Finally 1 µl SuperScript II Reverse Transcriptase (Invitrogen) was added, and allowed to incubate at 42ºC for 50 minutes, followed by 70ºC for 15 minutes to inactivate the enzyme. cDNA samples were stored at -20ºC.

2.2.1.5. qPCR analysis

Primers for the qPCR analysis were designed using Primer3Plus (Untergasser *et al.*, 2007), using the qPCR setting. The cDNA samples were 10x diluted prior to the qPCR analysis and the qPCR was performed on a CFX384 Touch Real-Time PCR Detection platform (Bio-Rad). One 384 well qPCR plate was used for each gene being analysed. Each sample was made up of 1 µl cDNA template, 0.5 µl each of the forward and reverse primers (5 µM) for the gene of interest, 5 µl SsoAdvanced SYBR Green Supermix (Bio-Rad), and 3 µl water. Each sample was performed in triplicate as technical replicates. For each biological replicate, the position of each harvest on the qPCR plate was randomised, and within each of the harvests, the position of the treatments were randomised. A no-template control was included for each replicate to ensure no contamination was present.

The qPCR run was set at 95ºC for 3 minutes, 45 cycles of 95ºC for 10 seconds, and 55ºC for 30 seconds. After each of these cycles, the fluorescence in each well was determined. At the end of the run, a melt curve analysis was performed by applying 95ºC for 10 seconds, and then running a temperature gradient from 65ºC to 95ºC in 0.5ºC increments every 5 seconds. The fluorescence was measured after each temperature change.

2.2.1.6. Analysis of qPCR data

Following the analysis of the plates, the threshold cycle number (Ct) was calculated using the `qpcR` package in R (Ritz and Spiess, 2008). This package allowed for the dynamic calculation of these values, instead of a simple threshold. In this package, a four-parameter logistic curve was fitted to the fluorescence curves, and any reactions which failed this fitting process were discarded as poor or incomplete reactions. Using these fitted curves, it was possible to calculate the Ct values by determining the maximum of the second derivative of the fitted curve (Luu-The *et al.*, 2005). The three technical replicates were fitted simultaneously to provide an overall level of confidence of the biological replicate. In addition, the package was used to calculate the efficiencies of the primers for each reaction.

The relative difference in gene expression was calculated as the difference between the Ct value of the gene of interest, and the Ct of the reference gene. Thus

$$\Delta Ct = Ct_{gene} - Ct_{reference}.$$

2.2.2. Real time transcription investigation using a luciferase reporter

2.2.2.1. Amplification of promoter sequences

Promoter fragments were amplified by PCR from Col-0 genomic DNA using the primers are listed in Table B.2. The PCR was performed using 5 µl KOD Master Mix (Novagen), 1 µl each of forward and reverse primers, and 1 µl template DNA (50ng/ml), and made up to 10 µl with sterile water. The following PCR thermocycler was used: 2 minutes at 95ºC followed by 11 cycles of 95ºC for 15 seconds, 55ºC for 15 seconds, and 68ºC for 2 minutes. To make the promoter sequences suitable for Gateway cloning, a second PCR was performed using generic Gateway primers (Gateway sequence primers in Table B.2) to attach the *attB* and *attP* sites. 1 µl each of the forward and reverse primers was added to the above mixture, together with 25 µl KOD master mix, and made up to 50 µl with sterile water. The PCR program for this amplification was 95ºC for 2 minutes, followed by 5 cycles of 95ºC for 15 seconds, 45ºC for 15 seconds and 68ºC for 2 minutes; followed by 35 cycles of 95ºC for 15 seconds, 55ºC for 15 seconds, and 68ºC for 2 minutes, ending with 68ºC for 5 minutes.

PCR products were run on an analytical 1% w/v TAE (40mM Tris, 20mM acetic acid, and 1mM EDTA, pH 8.0) agarose gel to check for clean amplification of products. Successful products were purified using a QIAquick PCR Purification Kit (QIAgen) according to manufacturer's instructions, and eluted in 15 µl sterile water. The concentration of the purified promoter fragments was determined using a Nanodrop ND-1000 spectrophotometer, and the samples were stored at -20ºC.

2.2.2.2. Cloning into donor vector (pDONR/Zeo)

Once purified, the promoter fragment was cloned into the donor vector, pDONR/Zeo, using the Gateway BP recombination system (Hartley *et al.*, 2000). From the Nanodrop concentrations obtained above, the amount of purified promoter fragment required to ensure a concentration of 150ng/µl was determined. The BP reaction was performed by adding this volume of the promoter to 1µl pDONR/Zeo (150ng/µl, Invitrogen), 1 µl Gateway BP Clonase II enzyme mix (Invitrogen), and made up to 5 µl with TE buffer (pH 8). This solution was incubated overnight, at 25ºC. 1 µl of this solution was added to 10 µl α-Select Gold Efficiency cells (Bioline), incubated on ice for 30 minutes, heat shocked at 42ºC for 30 seconds, and incubated on ice for 2 minutes. 250 µl SOC (Super Optimal broth with Catabolite repression) medium was added to the solution, and incubated at 37ºC for an hour with shaking. 100 µl of the cells were plated onto low salt LB (Lysogeny Broth) agar plates containing zeocin (50 µg/ml, Invitrogen), and incubated overnight at 37ºC.

2.2.2.3. Colony PCR

To verify that the bacterial colonies contained the vector, a colony PCR was performed. This process allows for the screening of the inserted plasmids directly from the bacterial cells. 10 individual colonies were picked using pipette tips, and suspended in 100 µl sterile water. 1 µl of the colony/water suspension was added to 2 µl 10x PCR buffer (Invitrogen), 0.6 µl $MgCl_2$ (50mM, Invitrogen), 0.6 µl of each of the colony PCR primers (Table B.2, 5 µM), 0.4 µl dNTPs (10 mM), 0.15 µl *Taq* DNA Polymerase (Invitrogen), and made up to 20 µl with sterile water. The PCR program for this amplification was 94ºC for 2 minutes, followed by 25 cycles of 94ºC for 30 seconds, 55ºC for 1 minute and 72ºC for 1 minute, ending with 72ºC for 3 minutes.

The PCR products were verified using a 1% w/v TAE agarose gel. Colonies which tested positive for the vector were used to inoculate 5 ml LB media containing zeocin (50 µg/ml), and incubated overnight at 37ºC, shaking at 200 rpm to produce cultures. The vectors were extracted and purified from the cultures using a QIAprep Spin Miniprep Kit (QIAgen) according to manufacturer's instructions and eluted with 20 µl sterile water.

In order to ensure that the correct promoter was inserted into the vector, a sequence verification was performed. Sequencing was performed using the colony PCR primers (Table B.2) in LIGHTrun sequencing (GATC Biotech).

2.2.2.4. Cloning into destination vector (pBGWL7)

The destination vector used was pBGWL7 destination vector (Karimi *et al.*, 2005), which contained the luciferase coding sequence and used a Basta selection, a glufosinate herbicide. The cloning process was similar to the donor vector cloning procedure (Section 2.2.2.2), except that the Gateway LR Clonase II enzyme mix (Invitrogen) was used, and the selection was spectinomycin (100 µg/ml) using high salt LB media. Following the bacterial transformation process, 10 colonies from the transformation were patched onto high salt LB agar plates containing spectinomycin (100 µg/ml), as well as plates containing zeocin (50 µg/ml). These plates were incubated at 37ºC overnight. Only colonies that only grew on the spectinomycin-containing plates were selected, and the presence of the vector was confirmed using the colony PCR method as described above (Section 2.2.2.3), using spectinomycin as the selection.

2.2.2.5. Agrobacterium transformation

1µg of the destination vector was incubated with 100 µl *Agrobacterium tumefaciens* cells (GV3101) on ice for 5 minutes. The cells were incubated in liquid nitrogen for 5 minutes, followed by a heat shock in a 37ºC water bath for 5 minutes. 100 µl low salt LB media was added, and this was incubated at 28ºC with shaking for 2 hours. 100 µl of the cells were plated onto low salt LB agar plates containing gentomycin (50 µg/ml) and spectinomycin (100 µg/ml). The plates were allowed to incubate for 2 days at 28ºC.

Arabidopsis plants were transformed using the floral dip method (Clough and Bent, 1998). In brief, a culture of the Agrobacterium cells was prepared, and Arabidopsis plants with developing floral tissue were dipped into this culture to perform the transformation. These plants were allowed to grow to seed.

2.2.2.6. Selection of transformants with Basta

To determine which of the seeds from the transformation contained the inserted vector, the seeds were grown with a quarter-strength Murashige and Skoog basal medium (1.1 g/L, Sigma) together with the Basta selection (40 mg/L) on Rockwool. 10 seedlings that successfully germinated were transplanted to soil and grown to seed. These seeds were harvested, and again sown on Rockwool containing Basta. Seeds which germinated with a 3:1 ratio of growth:death, were identified as being heterozygous transformants (containing one insertion site). Approximately 40 of these seedlings were selected and transplanted to soil for the luciferase experiment.

2.2.2.7. Investigation of real time transcription

Leaf 7 was detached from plants of different ages, namely 28, 32 and 36 days after sowing (DAS), with 24 replicate samples for each of the plant ages - 12 each for the mock and infected inoculations. These leaves were distributed amongst four 12 cm square Petri dishes, containing 0.8% w/v plant agarose, where 3 leaves from each treatment and each harvest time were placed on a plate. The treatments and harvests were randomised within the plates.

A solution of luciferin (25 µg/µl) was made up with 0.01% v/v Triton X-100 and filter sterilised using a 0.2 µm filter. After spraying this solution onto the leaves, the plates were covered and placed under constant light for 24 hours to eliminate any existing luciferase. Following this, the Petri dishes were placed under a liquid nitrogen cooled CCD camera (Princeton Instruments), using a light intensity of 15 µmol photons.m^{-2}.s^{-1}. Images were obtained every 2 hours using a 20 minute exposure. A baseline level of fluorescence was obtained for 6 hours, and following this, the leaves were treated with the Botrytis and mock inocula, as described in Section 2.2.1.2. Multiple 0.6 µl droplets of the inocula were applied to the leaves to ensure complete coverage. Images were analysed using ImageJ software in order to quantify the fluorescence by measuring the mean grey value of the leaves, which was calculated as the sum of the grey values of the pixels in the selection divided by the number of pixels (Abramoff *et al.*, 2004). These values were background subtracted by removing the the mean grey value of a region on the plate where there were no leaves.

3. Analysing gene expression data using parametric regression models

3.1. Regression analyses applied to gene expression data

The use of regression analyses to fit a specific model is a technique that has been applied to data in multiple fields of science, including agronomy (Lobell *et al.*, 2005), analytical chemistry (Watkins and Venables, 2006), and medical statistics (Woolcock *et al.*, 1984). In all these cases, a parametric regression model was used to describe the relationship between the response and the predictor. In addition, the regression approach provided a defined set of interpretable parameters, which could be used to provide more information regarding the underlying system.

A regression model defines how the value of a dependent variable Y changes in response to changes in the independent variable T, with one or more model parameters (Motulsky and Christopoulos, 2004). The T term is also called the regressor, and the values of T are fixed settings during an experiment, i.e. the times of sampling. In contrast, the value of Y is determined by the underlying system processes and is predicted using values of T (Ratkowsky, 1990; Seber and Wild, 1989). The values of the unknown parameters can then be estimated by observing the values of Y for a given set of T values (Seber and Wild, 1989). In terms of gene expression time-course studies, the values of Y_i would be the level of gene expression for gene i for a given set of sampling time points (T).

Since the expression of genes in cells occurs continuously, it can be represented as a continuous function, namely a curve (Ma *et al.*, 2006). The application of functions with interpretable parameters to time-course gene expression data has been previously proposed by Eastwood *et al.* (2008) and Chechik and Koller (2009). In both of these cases, parametric models were fitted to gene expression data, and the fitted parameters were used to analyse the responses for different genes. Through this approach, models with relatively few parameters were fitted as a function of a single explanatory variable (the time of sampling). The parameters, or functions of parameters, from a regression analyses can thus be used to obtain a better explanation into the function and mode of operation of genes. For example, they can be used to provide insights into differences between sets of genes, or indicating when particular events occur (Eastwood *et al.*, 2008; Ma *et al.*, 2006; Seber and Wild, 1989; Watkins and Venables, 2006). In addition, many gene expression analysis methods only take the mean of the replicates into account, effectively ignoring the between-replicate variation. A regression approach, on the other hand, is able to use this replicate information as an indicator of the quality of the fitted model in the form of deviations of the data about the fitted functions, or the level of uncertainty in the parameter estimates (Motulsky and Ransnas, 1987).

In order to use a regression analysis, it is necessary to first select a model to fit the data. This function should be selected where the effect of the parameters on the model is understood and the physical model it represents is known, or it could result in misleading results or misinterpretation. In nonlinear regression, good initial estimates for the function parameters need to be selected. The choice of starting values will have a large impact on the performance of the regression optimisation, with poor starting values potentially leading to a local minimum, or resulting in a failure to converge (Motulsky and Ransnas, 1987). Since many parameters for nonlinear curves are related to observable physical features, it is possible to find starting values by plotting the data and estimating the starting values from aspects of the shape of the observed response data (Motulsky and Christopoulos, 2004; Seber and Wild, 1989).

Techniques such as splines and Gaussian process regression are able to describe gene expression data well. However, these techniques do not necessarily provide an easy interpretation of the data. In contrast a regression analysis is more predictive, providing interpretable parameters to analyse aspects of the gene expression.

3.2. Applying regression models to analyse function and shape

To perform nonlinear regression, it is necessary to provide an equation of the curve for the model. While there are many different functions with similar shapes, a number of models were selected to provide a broad range of response shapes, such as sigmoid or peaked functions. In addition, these models were selected to have interpretable parameters that may be used to better understand the processes underlying the observed response. These parameters influence aspects of the model shape, and include rates of change, or influential time points (such as the midpoint). Thus, the parameters are representative of the physical world, and in terms of gene expression may imply time points that genes are being activated, or the change in rates of transcription. The models selected were the linear, logistic, Gompertz, exponential, critical exponential, linear+exponential, Gaussian and hyperbolic functions. A summary of the models used is shown in Table 2.1, and the models are described in detail below.

The *linear* ($y = c + m \cdot t$) model is a simple straight line. This model has two parameters, namely the parameter c representing the gene expression at time t=*0*, and m representing the rate of change of gene expression over time. Linear regressions are computationally cheap to perform and can easily be solved algebraically (Ratkowsky, 1990). Polynomial functions, such as the quadratic or cubic curves, can also be used to describe the data, and the same techniques can also be used with these models. However, the parameters from these functions do not have a physical interpretation (Motulsky and Christopoulos, 2004), and are generally considered too restrictive to model real world phenomena (Abraham *et al.*, 2003). The sign of m will influence if the

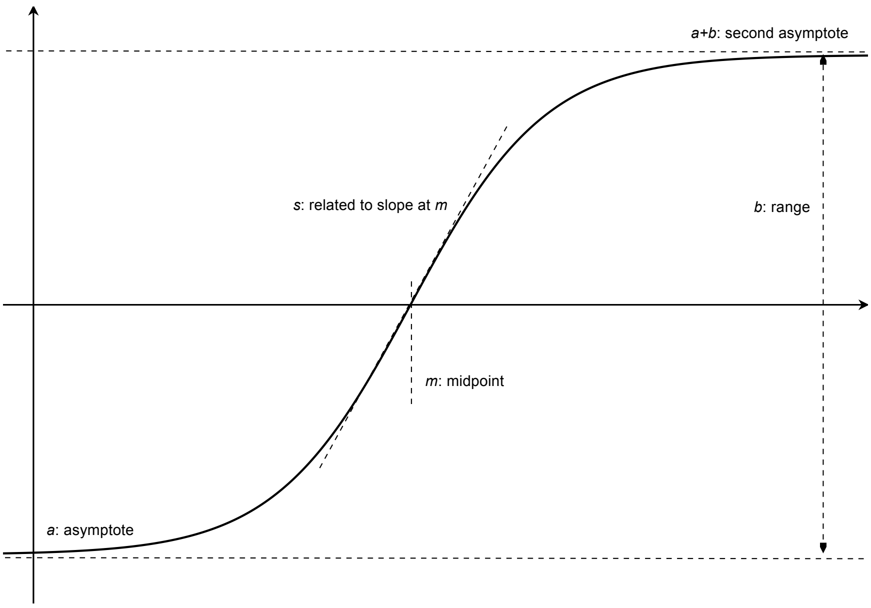


Figure 3.1: Example of a logistic curve ($y = a + \frac{b}{1+\exp((m-t)/s)}$) with the parameters illustrating the aspects of the curve they influence.

shape is increasing or decreasing, where if *m<0*, the curve is decreasing, and if *m>0,* the curve is increasing.

The *logistic model* ($y = a + \frac{b}{1+\exp((m-t)/s)}$) is representative of the typical sigmoid curves that are used in a variety of fields, particularly to model a growth-response relationship in fields including biology, environmental sciences, medicine, pharmacology, and toxicology (Ritz and Streibig, 2008). In particular, the logistic function represents a symmetric sigmoid curve characterised by a rapid growth rate in the beginning, slowing down to a constant growth rate, before finally approaching the asymptotic maximum value. The parameter a is the initial value, b is the distance between asymptotes (range), m is the time (t-value) at which maximum growth is reached, and s is related to the slope at m. A decreasing response is also possible with a maximum initial value followed by rapid decrease to a minimum asymptote. If s and b have the same sign, the curve is increasing, whereas if they are of opposite signs, it represents a decreasing response (Ratkowsky, 1990). Figure 3.1 shows an example of a logistic curve including how the parameters influence the shape of the curve.

The *Gompertz model* ($y = a + b \cdot \exp(-\exp(s \cdot (t - m)))$) is a sigmoid curve that is similar to the logistic except that is not symmetrical around the midpoint, and is also commonly used in dose response analyses (Ritz and Streibig, 2005). As with the logistic curve, the maximum growth rate occurs at *t=m*, s is related to the slope at m, a is the initial value, and the range is b (Ratkowsky, 1990; Seber and Wild, 1989). Since the Gompertz function is asymmetrical, there are two forms to the curve, affected by the sign of s. If *s>0*, the change in growth rate is faster to the right, and this form of the curve was termed the "Gompertz1" models. Alternatively, if *s<0*, the change in growth rate is faster on the left of the midpoint, and these graphs were termed the "Gompertz2" models. The differences between these curves are illustrated in Figure

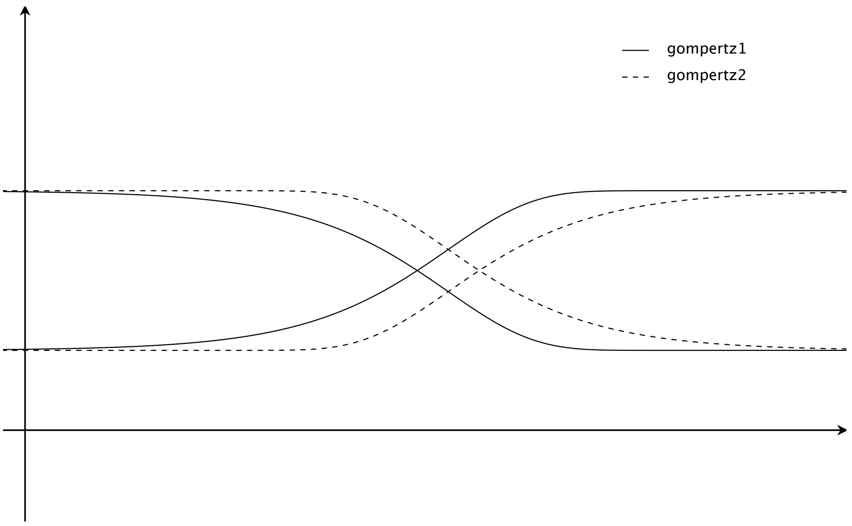


Figure 3.2: Graph showing the different forms of the Gompertz curve. The curves with the solid lines are those that had a faster change in growth rate after the midpoint, and were termed as the Gompertz1 graphs. The figures with the dashed lines had a faster growth rate to the left of the midpoint and were termed Gompertz2 graphs.

3.2. The direction of the Gompertz graph is determined by the sign of the s and b parameters - if s and b are the same sign, the function is decreasing, and if they are opposite signs, the function is increasing.

The *exponential model* ($y = a + b \cdot \exp(-r \cdot t)$) represents a growth or decay response with increasing values of t. The parameter a represents the asymptote of the curve, b is the range of the response between the asymptotic value and the value at t=*0*, and r relates to the rate of change in growth or decay (Ratkowsky, 1990). In addition, r affects the "direction" of the asymptote. If r>*0*, the asymptote is towards the right ($y \rightarrow a$ as $t \rightarrow +\infty$) (Figures 3.3A and C), and if r<*0*, the asymptote is on the left (Figures 3.3B and D). The sign of the b parameter affects the concavity of the curve. If b<*0*, the curve is concave ("bulges" out) (Figures 3.3A and B), and if b>*0*, it is convex (Figure 3.3C and D). Thus, with different combinations of these two parameters, it is possible to have four different forms of the curve, namely concave increasing, concave decreasing, convex increasing, and convex decreasing.

The *critical exponential model* ($y = a + (b + c \cdot t) \cdot \exp(-r \cdot t)$) is an extension of the exponential curve, and is characterised by an exponential growth, leading to a single maximum or minimum and an asymptote. The b parameter determines the height of the peak from the asymptote, c determines the magnitude of the decline from the maximum to the asymptote, b/c is the time to the maximum, a is the asymptotic value, and *a+b* is the value at t=*0* (Eastwood *et al.*, 2008). Again, the value of r greatly influences the shape of the curve. If r>*0*, the curve first begins with exponential growth or decay, followed by an asymptote (Figure 3.4A). If r<*0*, the curve first begins with the asymptote on the left, followed by the exponential portion (Figure 3.4B). The sign of c influences whether the nature of the turning point is a maximum or a minimum (whether the "peak" will point up or down), and also depends on the sign of r. If both parameters are the same sign, the function will increase to a peak, and if the parameters

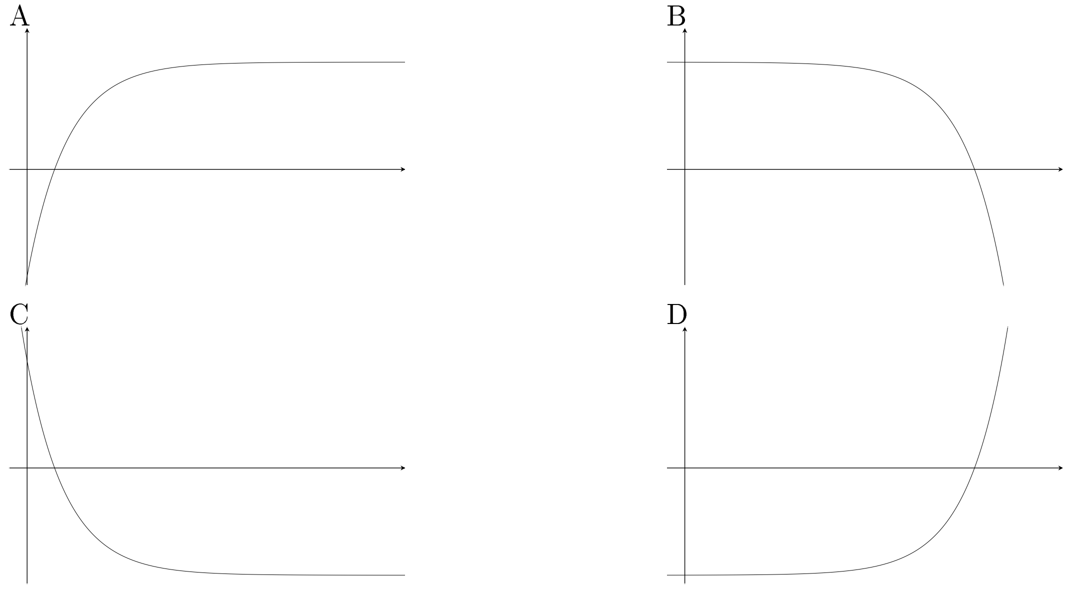


Figure 3.3: Examples of different forms of the exponential model, showing the flexibility in fitting shapes by altering the signs of the parameters. (A): concave increasing (*b<0, r>0*); (B): concave decreasing (*b<0, r<0*); (C): convex decreasing (*b>0, r>0*); (D): convex increasing (*b>0, r<0*). In all cases, the *a* parameter refers to the asymptote.

are opposite signs, the decrease to a trough (Figure 3.4).

The *linear+exponential model* ($y = a + b \cdot \exp(-r \cdot t) + c \cdot t$) is another extension of the exponential curve, consisting of an exponential curve on one side of its turning point and a straight line on the other. Like the critical exponential function, it describes two processes, and is the sum of two simpler functions. In this case, the model represents exponential growth, followed by linear growth. The parameter c is the gradient of the linear portion of the curve, and a and b influences the y-intercept ($a + b = y_0$). If *r>0*, the linear portion of the graph tends toward the right side (Figure 3.5A). In contrast, if *r<0* the linear portion of the graph will tend toward the left side of the graph (Figure 3.5B). Like the exponential curves, the b parameter affects the concavity of the

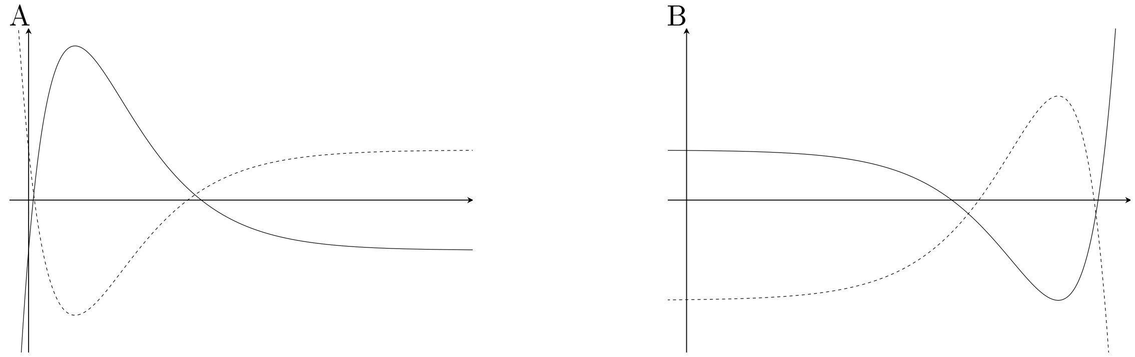


Figure 3.4: Graph showing the different forms of the critical exponential model. The shape of the model changes depending on the sign of the parameters. (A) shows curves where *r>0*, and (B) shows curves where *r<0*. In both cases, the solid line if where *c>0*, and the dotted line is where *c<0*.

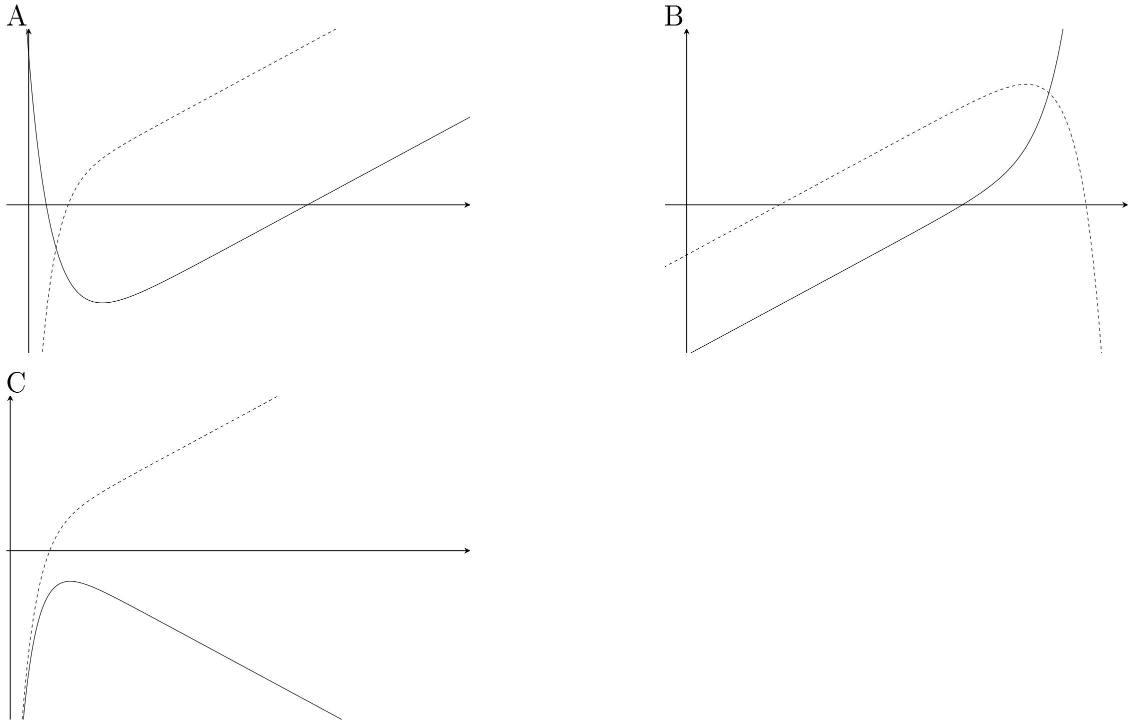


Figure 3.5: Graph showing some of the different forms of the linear+exponential curve. (A) shows curves where $r>0$, and (B) shows curves where $r<0$. In both cases, the solid line is where $b>0$, and the dotted line is where $b<0$. In all these graphs, $c>0$. (C) shows the effect of $c<0$ (solid line). The parameter values were the same in both curves, save for the sign of c.

exponential part - if $b>0$, the exponentials are concave, and if $b<0$, the exponentials are convex (Figure 3.5, dotted lines). The signs of the r and b parameters determine if the exponential portion of the function is increasing or decreasing - if they are the same sign, the exponential portion is decreasing, and if they are opposite signs, the exponential part is increasing. Thus, between the b, c and r parameters, eight possible shapes are possible. Figure 3.5 shows some of the possible shapes for this model. If the c is positive, possible shapes are exponential decreasing followed by linear increasing (Figure 3.5A, solid line), exponential increasing followed by linear increasing (Figure 3.5A, dotted line), linear increasing and exponential increasing (Figure 3.5B, solid line), and linear increasing and exponential decreasing (Figure 3.5B, dotted line). Similarly, these same shapes are possible, except with a decreasing linear shape. Figure 3.5C shows one such case.

The *Gaussian* model ($y = a + b \cdot \exp(\frac{-(t-m)^2}{2 \cdot s^2})$) is the characteristic symmetric bell-shaped curve, and is widely used in statistics and mathematics to describe normal distributions. It is characterised by parameters where a is the initial value, b is the range of the response ($a+b$ is thus the maximum value), m is the midpoint of the graph (mean) and s describes the spread of the function around the midpoint (standard deviation). The s parameter is also related to the duration of a response, where a larger spread indicates a greater duration of the gene expression taking place. The sign of b affects the shape of the curve, where if $b>0$, the curve increases to a maximum, whereas

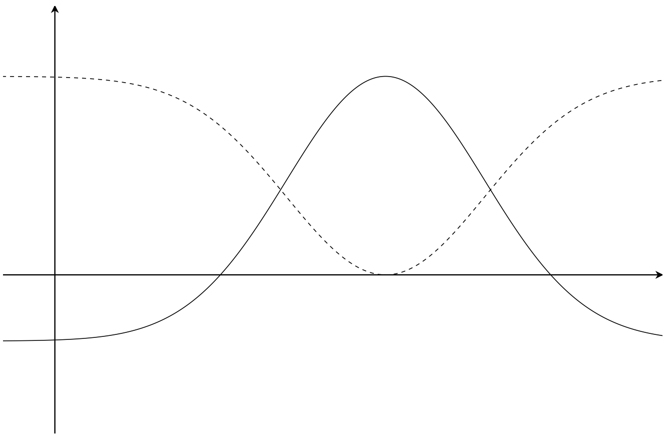


Figure 3.6: Graph showing the different forms of the Gaussian curves. The curve with the solid line had a positive *b* parameter value, while the curve with the dotted line had a negative *b* parameter value.

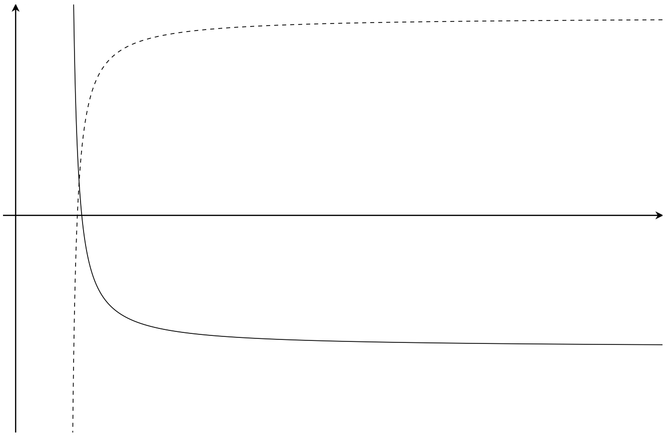


Figure 3.7: Graph showing the different forms of the hyperbolic model. The curve with the solid line had a positive *b* parameter value, while the curve with the dotted line had a negative *b* parameter value.

if $b<0$, the curve decreases to a minimum. These shapes are shown in Figure 3.6.

The function used to model a *hyperbolic* model ($y = \frac{a \cdot (t-c)}{b+(t-c)}$) is based on the Michaelis–Menten model: $y = \frac{V \cdot x}{k+x}$. This function is commonly used in biochemistry for modelling enzyme kinetic reactions, where the parameters V and k which refer to the maximum velocity of the enzymatic reaction, and the concentration of substrate required for half the maximum velocity, respectively. This equation was reparametrised to include an additional parameter c to allow for a horizontal shift in the function, thus no longer requiring the intercept to pass through the origin. In general, a represents the upper asymptote, and $b+c$ is the point on the time axis at which $\frac{a}{2}$ has been obtained (Ratkowsky, 1990; Motulsky and Christopoulos, 2004). The sign of b affects the direction of the curve, where if $b>0$, the curve is increasing to the asymptote, whereas if $b<0$, the curve is decreasing. These shapes are shown in Figure 3.7.

These shapes were selected to describe a variety of responses that would be expected in a biological process. For example, the sigmoid curves (logistic and Gompertz) indicate a response that increases (or decreases) from one equilibrium position to another. In terms of gene expression, a would indicate the starting gene expression level, m the time at which maximum expression occurs, and s would be related to the rate of expression. Similarly, the exponential response represents a rapid change in gene expression. Thus, it is possible to interpret these shapes in terms of the parameters as well as the processes that they represent.

3.3. Initial parameter selection and self-starter functions

In order to reliably perform non-linear regression, a model is required as well as initial estimates for the model parameters. Typically, these initial parameters would be estimated by plotting the data and identifying key values which relate to the parameters, or if the range of the parameters are known, perform a grid search on a subset of values (Ritz and Streibig, 2008). However, since regression analyses in this study are performed over the large Arabidopsis datasets (Section 1.2), it needs to be automated, and to this end, self-starter functions were used. Self-starter functions are pieces of code that automate the parameter search for starting values. These functions are specific to a particular nonlinear model, which can then be used to calculate starting values for the model for any given dataset. The self-starter may not always result in a successful convergence, but in general it should provide estimates of that parameters that are close enough to allow the estimation algorithms to converge. Several collections of self-starters exist, including several in the base R installation, as well as in third-party packages, such as the `drc` and `HydroMe` packages. However, these self-starters are for specialised models in a specific field, or for a different parameterisation of the same shape. Thus for the majority of the selected models, self-starters were developed and used in the fitting process. The process of estimating the initial parameters for each shape is described below.

3.3.1. Sigmoid curves - logistic and Gompertz

Both the logistic and Gompertz self-starter functions were adapted from existing sources. The logistic self-starter function was derived from the `SSfpl` function in the built-in `stats` package in R (R Development Core Team, 2011), and is a simple re-parameterisation from the original $y = a + \frac{b-a}{1+\exp((m-t)/s}$ to $y = a + \frac{b}{1+\exp((m-t)/s)}$ (i.e. changing the range value to a single parameter).

The Gompertz self-starter function was adapted from the `gompertz` function in the `drc` package (Ritz and Streibig, 2005). Similarly to the logistic self-starter, the function was re-parameterised so that the range value is a single parameter - $y = a + b \cdot \exp(-\exp(s \cdot (t-m)))$. As the Gompertz model is asymmetric, there are multiple shapes possible, depending on the value of the parameters. Specifically, if s>*0*, this means the curve exhibits accelerated growth towards the right asymptote (slower initial growth). In contrast, if s<*0*, this represents the form where there is accelerated growth from the left asymptote. The original `gompertz` function in the `drc` package only took the Gompertz1 form (growth rate is faster on the left of the midpoint) into account, so an additional self-starter function was added to detect the Gompertz2 shape (growth rate is faster on the right of the midpoint).

3.3.2. Exponential

A parametrisation of the standard exponential equation was identified by Ratkowsky (1990) that expressed the equation in terms of expected value parameters:

$$y = y_1 + (y_2 - y_1)\frac{1 - k^{m-1}}{1 - k^{n-1}}$$

where $m - 1 = \frac{(n-1)(T-T_1)}{(T_2-T_1)}$, n is the number of data points, and k and r are related in the following manner: $r = k^{(n-1)/(T_2-T_1)}$. Another parametrisations was also shown such that

$$y = y_1 + (y_2 - y_1)\frac{(1 - [(y_2 - y_3)/(y_3 - y_1)]^q)}{1 - [(y_2 - y_3)/(y_3 - y_1)]^2}$$

where $q = \frac{2(T-T_1)}{(T_2-T_1)}$, and y_1, y_2 and y_3 correspond to the y-values at $T = T_1$, $T = T_2$ and $T = (T_1 + T_2)/2$, respectively. T_1 and T_2 are the first and last observed values (time points) in the dataset, respectively. From these two parametrisations, it was assumed that the denominators of both these equations were equal, implying that $k^{n-1} \approx [(y_2 - y_3)/(y_3 - y_1)]^2$. Using this assumption, and the association of k and r described above, it was possibly to estimate $r \approx [(y_2 - y_3)/(y_3 - y_1)]^{2/(T_2-T_1)}$.

Once this approximate solution of r was found, the values of a and b could be easily identified. By using the equation, $y_i = a + b \cdot \exp(-r \cdot t_i)$, where y_i corresponds to the point $T = T_i$, and using the first and last data pairs ($(t_1; y_1)$ and $(t_2; y_2)$, respectively), it is possible to solve for a and b such that

$$b = \frac{y_2 - y_1}{\exp(-r \cdot t_2) - \exp(-r \cdot t_1)}$$

and $a = y_1 - b \cdot \exp(-r \cdot t_1)$.

3.3.3. Critical exponential

Since the r parameter is the primary parameter that influences the shape of the curve, it is the most important to identify first. To find an approximate value for the parameters, the data was divided into two parts, separated by the maximum absolute y-value. The absolute maximum was taken to ensure that the curves with a dip instead of a peak were also identified. Thus, the data was divided from $(t_1; y_1)$ to $(t_{max}; y_{max})$, and $(t_{max}; y_{max})$ to $(t_n; y_n)$, where t_{max} is the time point of the maximum absolute y-value in the dataset. The data was further divided at half way between the first value and the maximum value (called *mid1*), and between the maximum value and the last value (*mid2*) (Figure 3.8). The difference between the y-values of *mid1* and the first value, and the last value and *mid2* were calculated (*diff1* and *diff2*, respectively) and compared. If *diff1* was greater than *diff2*, it implied the curve had a faster growth rate on the left

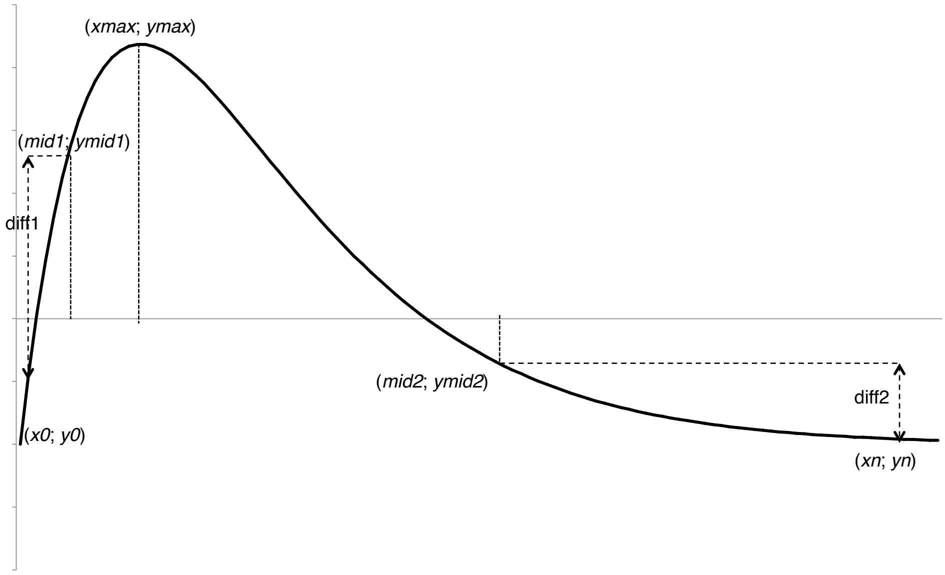


Figure 3.8: Illustration of the self-starter process for the critical exponential function. The data was first divided by the maximum value, and then further subdivided into two halves (*mid1* and *mid2*). The difference between the first data point and *mid1* and *mid2* and the last data point were calculated (*diff1* and *diff2*, respectively). If the first segment of the graph has a faster growth rate (*diff1*>*diff2*), this meant that the asymptote was on the right side of the graph, and thus *r*>*0*. Conversely, if *diff2*>*diff1*, the graph has the asymptote on the left hand side, and *r*<*0*. The other parameters could then be estimated based on the aspects of the curves they influence.

side and tails towards an asymptote on the right side, meaning *r>0*, and r was therefore set to 0.2 (Figure 3.4A). This were arbitrarily set as a "push in the right direction". The other parameters could then be estimated where a was approximately the last value (the asymptotic value), and b approximately equal to the difference between the first y-value and a (since $a + b \approx y_1$). The converse was true if *diff2* was greater than *diff1*, so r was set to -0.2 (Figure 3.4B, and a and b were approximately equal to the first value. The c parameter is the difference between the maximum value and the asymptotic value.

3.3.4. Linear+exponential

To find approximate starting values for the linear+exponential curve, the fact that a portion of the curve is linear was taken advantage of. To do this, a similar approach to the critical exponential was performed where the data was divided at the maximum absolute y-value, i.e. the two datasets were from $(x_1; y_1)$ to $(x_{max}; y_{max})$ and $(x_{max}; y_{max})$ to $(x_n; y_n)$, where x_{max} is the x-value for the maximum absolute y-value in the dataset. If the function was monotonic, the data was divided in half.

Once again, the primary parameter influencing the shape of the curve was the r parameter. As shown in Figure 3.5, the side of the exponential portion is determined by the sign of r. To determine an estimate for this parameter, a linear regression was performed on each section of data to determine which was more linear. This comparison was performed using the value of the R^2 value from the regression. The data points at $(x_{max}; y_{max})$ were used in both linear regressions. If the first section was more linear (Figure 3.9B), it implied that *r<0*, and r was approximated to -0.2; and *vice versa* if the second section was more linear, r was set to 0.2 (Figure 3.9A),. The rest of the parameters could then be estimated with a and c approximately equal to the intercept and slope of the linear regression, respectively. The b parameter affects the concavity of the graph, and is estimated as $b \approx y_1 - a$.

3.3.5. Gaussian

The Gaussian curve is estimated using various aspects of the curve. The m parameter is the time point where the maximum absolute y-value occurs, and was calculated using the `which.max` function in R. The a parameter is the average between the first and last y-values to estimate the base level, the b parameter (range) is the difference between the base level and maximum absolute y-value, and the s parameter (t-spread around m) is estimated as the difference between the m estimate and the time point where half the maximum response occurs ($y = a + \frac{b}{2}$). Since there are two time points where this occurs, the first was selected.

3.3.6. Hyperbola

Like the logistic self-starter, the self-starter for the hyperbola was a re-parameterisation of the `SSmicmen` function the built-in `stats` package, $y = \frac{Vm \cdot input}{K + input}$. An additional parameter, c, was added to allow the function to shift on the time axis. Since c is the

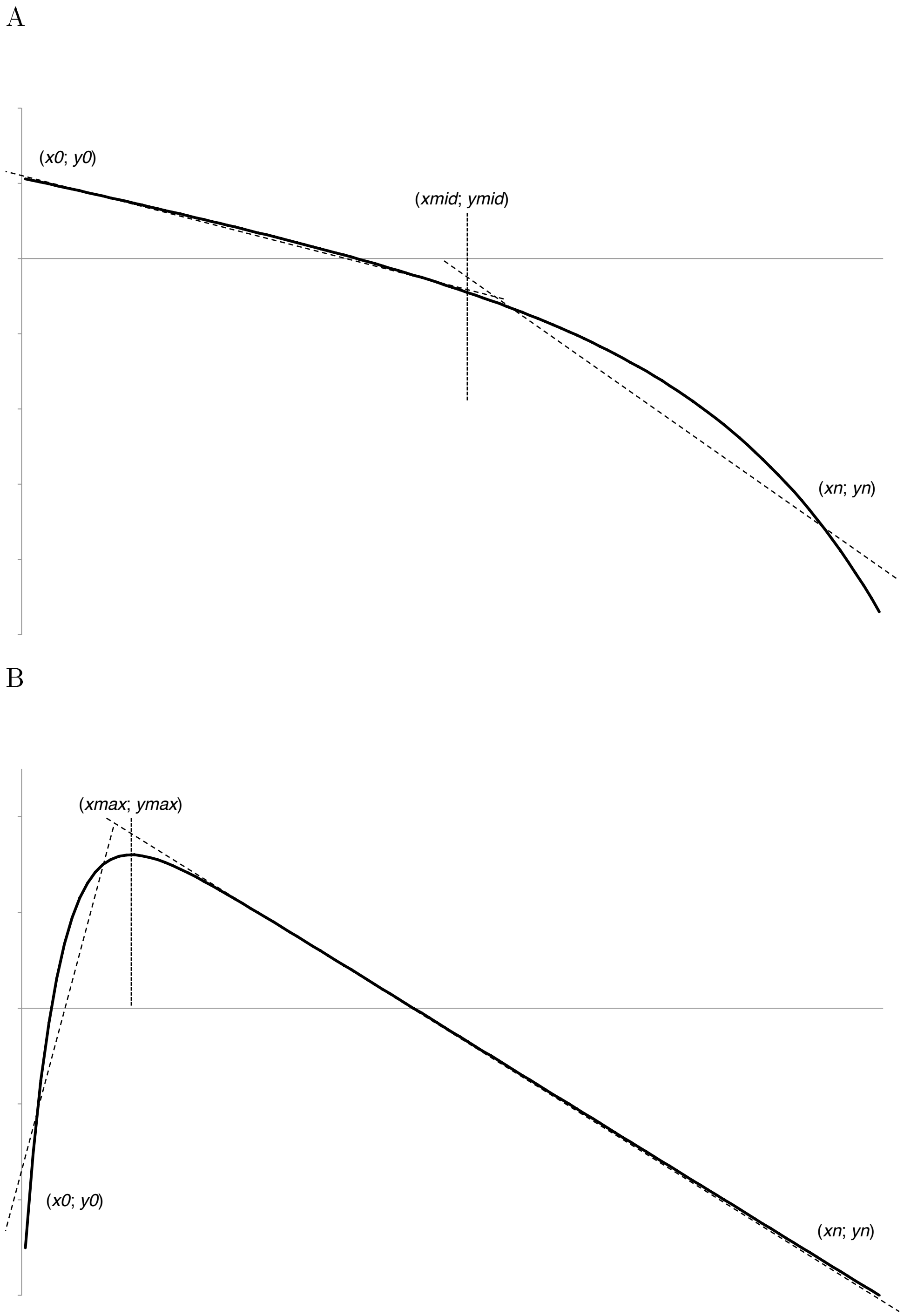


Figure 3.9: Illustration of the self-starter process for the linear+exponential function. The data was divided by the maximum (B), or in half if the function was monotonic (A). A linear regression was performed on each segment to determine which portion was more linear. If the first segment was more linear, $r<0$ (A), and if the second segment was more linear, $r>0$ (B).

time point where $y = 0$, this value was estimated as the time point where the y-value is closest to 0.

3.4. Statistics to determine goodness-of-fit

As described in Section 1.7, for a given regression model, it is possible to determine the model goodness-of-fit using the decomposition of variability. Thus

$$SS_{total} = SS_{regression} + SS_{residual}$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.1}$$

where SS_{total} is the total sum of squares, $SS_{regression}$ is the regression sum of squares, $SS_{residual}$ is the residual sum of squares, n is the number of time points, $\bar{y}$ is the mean of the samples and $\hat{y_i}$ is the predicted value of the fitted curve at time i. The SS_{total} describes the total deviation of the data about the mean, whereas the $SS_{regression}$ describes how much variability in the data is explained by the model (a measure of how well the model fits the data), and the $SS_{residual}$ describes the amount of error between the fitted model and the data (the unexplained variability).

These sums of squares can be used to calculate the respective mean squares, which provide a measure of the variance of the data. The mean squares for each of the sum of squares can be calculated by dividing by the respective number of degrees of freedom. For example, the residual mean square ($MS_{residual}$) can be calculated as

$$MS_{residual} = \frac{SS_{residual}}{df_{residual}}$$

where $df_{residual}$ is the residual degrees of freedom. The total mean square (MS_{total}) and regression mean square ($MS_{regression}$) can be calculated in an analogous fashion.

If replicate samples (multiple y observations at a given x-value) were obtained, it is possible to extend equation 3.1 to include the replicated observations at each x-value. Assuming equal replication at each x-value,

$$\sum_{i=1}^{n}\sum_{j=1}^{r}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{n}\sum_{j=1}^{r}(\hat{y}_{ij} - \bar{y})^2 + \sum_{i=1}^{n}\sum_{j=1}^{r}(y_{ij} - \hat{y}_i)^2 \tag{3.2}$$

where r is the number of replicates per time point.

It is then possible to further decompose the $SS_{residual}$ into two components - the pure error and the lack-of-fit sums of squares (Brook and Arnold, 1985).

$$SS_{residual} = SS_{pure\,error} + SS_{lack-of-fit}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{r}(y_{ij} - \hat{y}_i)^2 = \sum_{i=1}^{n}\sum_{j=1}^{r}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{n} r(\bar{y}_i - \hat{y}_i)^2. \tag{3.3}$$

The pure error sum of squares is defined as the sum of squared differences between the average y-value at each x-value ($\bar{y}_i$, the local average of replicate samples) and each observed value at that x-value, and describes the error that is due to the between-replicate variation, pooled across the time points (pure random error). The lack-of-fit sum of squares is a sum of squared differences between the local average and the fitted value at the same x-value, weighted by the number of replicates. This value describes the amount of variation due to lack of model fit (variation between replicate means and fitted curve). If the lack-of-fit sum of squares makes up a large proportion of the $SS_{residual}$, this suggests that the fitted model is insufficient to fit the data and indicates a poor fit to the mean y-values.

It is possible to calculate the lack-of-fit sum of squares through the fitting of a saturated model, so named as a parameter is allowed for each time point, and is thus saturated with parameters. This model describes the expected response at each time point, and represents a model that is formulated with no assumptions with regards to the response shape and time dependence of the data points. From the saturated model, an analysis of variance (ANOVA) table with the sums of squares is obtained. As in equation 3.1, the saturated model total sum of squares can be decomposed into two parts. However, in this case, the residual sum of squares is the pure error sum of squares. By using these values, together with equation 3.3, the lack-of-fit sum of squares ($SS_{lack-of-fit}$) for the regression model can be calculated.

$$SS_{total} = SS_{saturated} + SS_{pure\,error}$$

$$SS_{lack-of-fit} = SS_{residual} - SS_{pure\,error}. \tag{3.4}$$

Similarly, the lack-of-fit degrees of freedom can be calculated as the difference in the degrees of freedom in the regression and saturated model residuals ($df_{lack-of-fit} = df_{residual} - df_{pure\,error}$), and the *lack-of-fit mean square* can then be determined as

$$MS_{lack-of-fit} = \frac{SS_{lack-of-fit}}{df_{lack-of-fit}}. \tag{3.5}$$

The R^2 value (*coefficient of determination*) is a commonly used statistic to determine the goodness-of-fit of a fitted curve, with values ranging between 0 and 1, where 1 indicates that the fitted line passes perfectly through the data points (Motulsky and Christopoulos, 2004). This value indicates the proportion of the variance that is accounted for by the fitted curve.

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}}.$$

However, this R^2 value will incorrectly increase with an increase in the number of parameters, meaning models with more parameters would appear to have a better fit, potentially leading to over-fitting. Thus, in order to take this into account, an *adjusted R^2 value* (R_a^2) is used.

$$R_a^2 = 1 - \frac{MS_{residual}}{MS_{total}}. \tag{3.6}$$

The R_a^2 provides an indication of the proportion of the variance in the data which can be explained by the model, and thus indicates which models had a good fit to the data points.

Using the lack-of-fit information described above (equation 3.5), it is possible to calculate the goodness-of-fit of the curve without the variation due to the biological replicates. This statistic was termed the R_{LoF}^2 ("*lack-of-fit adjusted* R^2") and can be calculated in an analogous fashion to the R_a^2 value (Equation 3.6):

$$R_{LoF}^2 = 1 - \frac{MS_{lack-of-fit}}{MS_{saturated}}$$

where $MS_{saturated}$ is the regression mean square from the saturated model. This R_{LoF}^2 is effectively the same as if the regression had been performed only using the replicate means at each time point. In this way, it describes how much variance one model has explained.

Another simple statistic for the goodness of fit of the model is the *F-test*. The F-test provides an indication of whether the fitted model is the appropriate shape given the data. However, it can sometimes be difficult to interpret the significance of the F-test as the test is sensitive to variation in the data (Ritz and Streibig, 2008). The F statistic is calculated as

$$F = \frac{MS_{regression}}{MS_{residual}}$$

where the significance level is determined using an $F(df_{regression}, df_{residual})$ distribution.

While the R_a^2 is a useful statistic for determining the goodness-of-fit for a model fit, it is generally a poor indicator for model selection (Burnham and Anderson, 2002). Thus, when selecting which of the various regression models best fitted the data, *Akaike's Information Criterion (AIC)* was used. While it is also possible to use a F-test to compare models, this approach cannot be used when comparing non-nested models. The AIC is calculated using aspects of maximum likelihood theory, information theory and entropy of information (Motulsky and Christopoulos, 2004), and is calculated as follows:

$$AIC = N.\log\left(\frac{SS_{residual}}{N}\right) + 2p$$

where N is the total number of data points and p is the number of parameters. The AIC equation takes into account both the goodness-of-fit and the number of parameters required, thus attempting to prevent overfitting with overly complex models (Ritz and Streibig, 2008; Motulsky and Christopoulos, 2004). The AIC statistic thus describes how likely the model is to be correct, given the data. The value of the AIC is not easily interpretable, and AIC values of models from different datasets cannot be compared (Burnham and Anderson, 2002; Motulsky and Christopoulos, 2004). However, the rel-

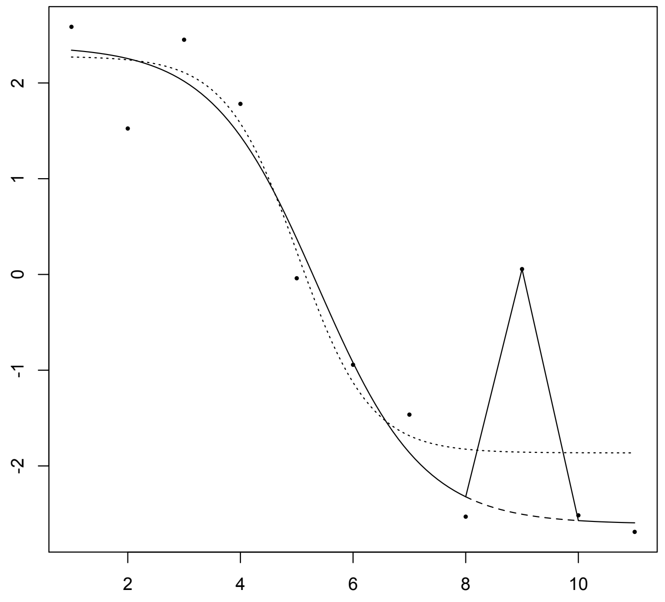


Figure 3.10: Illustration of an example of a spike. An outlying replicate is shown at time point 9, and removal of this point improved the model fit. The fitted curve is shown as the dotted line, and the curve with the spike point removed is shown as the solid line.

ative differences between the AIC values of a set of models from the same dataset can describe how much better one model fits the data than another model to the same data (Burnham and Anderson, 2002).

The application of these statistics is described below in Sections 3.6 and 3.8.

3.5. Spike detection

In some gene expression profiles, a phenomenon was observed where the profile followed one of the models, except for a sharp increase or decrease of expression at a single time point. This occurrence was termed a spike, and may result from a rapid and sudden activation or repression of a gene.

To determine if a gene exhibited a spike, a brute force leave-one-out-cross-validation method was employed. Here, all the observations from a given time point were removed, and the regression analysis performed again on the reduced dataset. If the resulting fit was better than the fit of the full dataset (R^2_{LoF} of the reduced dataset $>$ R^2_{LoF} of the original dataset) and the mean of the removed replicate data was significantly different from the predicted value of the reduced dataset, this was marked as a potential spike. Significance was determined using the nonparametric Mann-Whitney test, comparing the mean of the removed replicate data points to the predicted value at the removed time point. The removed time point was considered significant if the p-value of the test was less than 0.05. This process was performed for each time point, and the time point with the smallest p-value (the most significant difference) was retained as a potential spike. This is shown in Figure 3.10. The spike is only a representation of how the spike occurs, since the exact mechanisms behind this spike are unknown. Thus, the spike is drawn as increasing from the previous time point's expression level, and decreasing to the following time point's value.

This methodology is based on the concept of leverage and studentised residuals, where certain observations may significantly influence the fitted model (Cook, 1982; Fox, 2002; St Laurent and Cook, 1993, 1992).

3.6. Filtering model fits by goodness-of-fit and standard error

Prior to further analyses, it is necessary to determine which of the models had a good fit. As described above and in Chapter 2, a range of models with diverse shapes is fitted to the expression profiles, and thus it would be expected that not every model would fit to the data. To determine if the model fitted the data well, the statistics in Section 3.4 were used.

In determining which of the models had the best fit, the AIC was used. As recommended by Burnham and Anderson (2002), any models with a difference in AIC values (ΔAIC) less than 2 to the smallest AIC value were retained. As a result, a gene expression profile may be fitted by more than one model shape. However, since the fit may not necessarily be a good fit, a number of goodness-of-fit statistics were used to determine the quality of fit, namely the the R_a^2, R_{LoF}^2, and F-test (Section 3.4). These values provide a quick and simple means to gauge the quality of the fit.

In addition to filtering by general fit quality, the quality of the estimated parameters was determined through the use of the standard error of estimation of the fitted parameters. For a given fitted parameter, if the standard error is large, this implies that there was insufficient information to accurately determine the parameter value, possibly as a result of extrapolation. Thus, a standard t-test was performed to determine if the parameters were significantly different from 0. For an estimated parameter $\hat{\beta}$, a t-statistic is calculated as follows:

$$t_{\hat{\beta}} = \frac{\hat{\beta}}{\text{stderr}(\hat{\beta})},$$

where $stderr(\hat{\beta})$ is the standard error from the least-squares estimation of parameter $\hat{\beta}$. The critical value from the t-distribution was calculated using the residual degrees of freedom (*total observations - number parameters - 1*) and $\alpha = 0.05$. If $t_{\hat{\beta}}$ is less than the critical value, this implies that the parameter is poorly estimated, and the fitted model was not included in further analyses.

3.7. Deriving new parameters and estimating their standard error

While the parameters from some models are directly informative, it may be desirable to calculate other parameter estimates which describe alternative aspects of the curves. For example, for the sigmoid models, the values of the function increase (or decrease) from asymptotic value to another. It is thus possible to find the time point (x-value) when 5% of the difference between asymptotes has been reached (denoted the *5per* parameter), indicating the point at which the increase (or decrease) is starting. In terms of gene expression, this may indicate the point at which the gene is being activated or repressed. To calculate the *5per* parameter, this would be the point where $y = a + 0.05 \cdot b$. Thus, using the equation for the logistic function described in Section 3.2, the time point where this value occurs is

$$5per = m - s \cdot \log\left(\frac{1}{0.05} - 1\right).$$

In order to calculate the standard error of estimation of these derived parameters for the filtering (Section 3.6), the delta method was used (Fox, 2002, 2008; Ritz and Streibig, 2008). The delta method assumes that the joint distribution of the parameter estimates is approximately normal, implying that any derivation of the parameter estimates will also be normally distributed. Therefore, estimates of parameters that are functions of the original parameters can be calculated (Fox, 2008; Ritz and Streibig, 2008).

Several other parameters have been derived and are summarised in Table 3.1. These include the *grad* parameter for the sigmoid curves and describes the maximum rate of change. This value is calculated as the gradient of the curve (first derivative of the function) at the point of the maximum change in growth rate, m. For the logistic curve, this would be

$$grad = \frac{b}{4s}.$$

The *turnp* parameter is used for the critical exponential model to indicate the time point where the maximum or minimum peak on the curve occurs. Therefore, this is the time point where the first derivative of the critical exponential function is equal to 0. Thus

$$turnp = \frac{c - r \cdot b}{c \cdot r}.$$

Finally, for the linear+exponential, the *linpnt* is defined as the time point at which the curve begins to become linear. This point could be calculated as the time point where the first derivative of the critical exponential function is equal to 0. However, in some cases, these is no turning point, and the function is monotonic (for example, Figure 3.5A, dotted). Thus, in order to maintain consistency across all the shapes, this parameter was estimated as the point at which the gradient of the linear+exponential function becomes close to the c parameter (the gradient of the linear portion). Therefore, to the

Table 3.1: Table of the derived parameter values and their descriptions.

Derived parameter name	Valid models	Description
5per	Logistic, Gompertz	Time point (x-value) at which 5% of the maximum growth or decay has occurred.
grad	Logistic, Gompertz	Maximum rate of change of expression. Defined as $f'(m)$.
turnp	Critical exponential	Time point (x-value) where the maximum or minimum peak occurs. Defined as the x-value where $f'(x) = 0$.
linpnt	Linear+exponential	Time point (x-value) where the graph starts to become linear. Defined as the x-value where $f'(x) \to c$.

time point at which the gradient of the function is 95% of the linear gradient is

$$linpnt = \frac{log(\frac{b \cdot r}{0.05})}{r}.$$

3.8. Algorithmic development

As described in Section 1.2, the primary datasets used in the analyses were from the PRESTA long day senescence (Breeze *et al.*, 2011) and *Botrytis cinerea* infection (Windram *et al.*, 2012) time series experiments. In brief, the senescence morning sample dataset consisted of 11 time points, taken every second day for 22 days, and the Botrytis dataset consisted of 24 time points, taken every 2 hours for 48 hours. The gene expression analysis was performed using CATMA version 3 microarrays (Sclep *et al.*, 2007), which contained over 32 500 probes, where multiple probes mapped to the same gene. Thus, a unique set of 23 802 best probes was identified that provided a more comprehensive set of genes to represent the Arabidopsis genome. In both sets of data, the data was previously analysed by other members of the PRESTA group and the mean values for each biological replicate was estimated from the observed technical replicates, using MAANOVA (Wu *et al.*, 2003), an analysis tool for microarray data. The data used was standardised to have a mean of zero per gene across all samples, but the standard deviation was not scaled, as the differences in the amplitude were of interest.

The regression analysis (Figure 3.11) was performed on the data from the above datasets. In order to efficiently analyse the fits of the above linear and nonlinear models to gene expression data, a number of scripts were developed to fit the various models to the data, calculate the goodness-of-fit, store the fits in a database, and analyse the statistics for the fitted curves. The scripts used the R statistics language for the statistical analyses, together with the Python programming language for data processing. A number of other third party packages were used in the development of the analysis pipeline and are listed in Table 2.2.

For each of the gene expression responses, it was attempted to fit a series of nonlinear

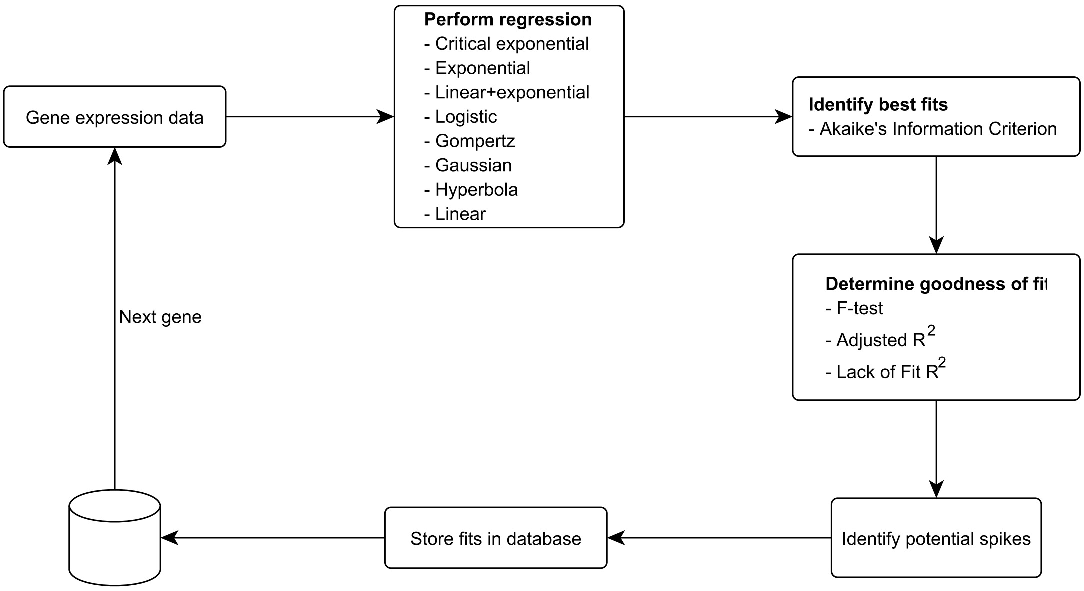
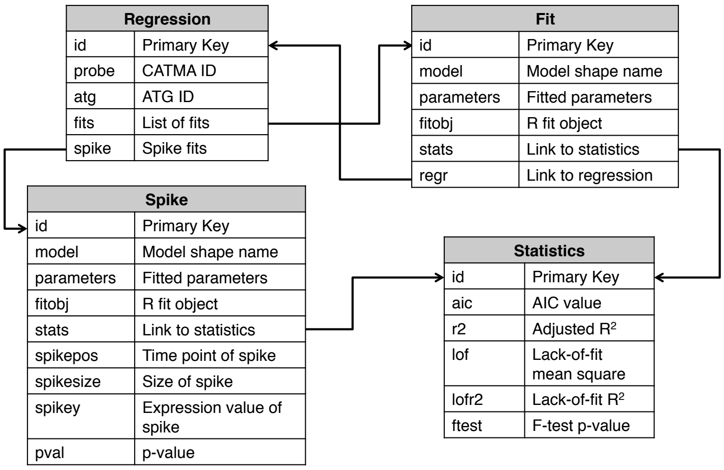


Figure 3.11: Flow diagram showing the process for the regression calculations to determine the best curve fit for each gene. For each gene, the expression data is fitted to all the models, and the best fits are selected using AIC. Fit statistics are calculated for each fit to determine the goodness-of-fit, and potential spikes are identified. All this information is stored in an SQLite database (Figure 3.12).

and linear functions (Section 2.1), with the use of custom self-starter functions (Section 2.2). As described in Section 3.6, the best model fits to each gene expression profile were determined using the AIC, and retaining any models with a difference in AIC values (ΔAIC) less than 2 to the smallest AIC value. Statistics to determine the goodness-of-fit were also calculated, namely the adjusted R^2, lack-of-fit R^2 (R^2_{LoF}), and the F-test p-value (Section 3.4). Finally, for any genes which exhibited a poor fit (arbitrarily defined as having a $R^2_{LoF} < 0.5$), an analysis was performed to determine the presence of a potential spike which would explain the poor fit the spike detection (Section 3.5).

All the results of this analysis, including gene information, best fit information and fit statistics, were stored in an SQLite database. A schematic describing the database structure is shown in Figure 3.12. The database consisted of four tables - the Regression table, the Fit table, the Statistics table, and the Spike table. Each of these tables contains a unique identifier (primary key) to use in cross referencing between the tables. The *Regression* table contains the annotation information about the gene, including the CATMA and ATG identifiers. The “fit” field in this table acts as a foreign key (a field in one table that is used to uniquely identify a row in a different table) to the Fit table. Since an expression profile may fit more than one model, there can be multiple entries for this field. The *Fit* table contains details regarding the model that fitted the data, such as model fit and fit parameters, as well as the R fit object (from `nls` or `lm` for nonlinear or linear regression, respectively), in case further analyses were required. This table also contained a foreign key to the Statistics table to link the statistics for this fit. The *Statistics* table contains all the goodness-of-fit statistics, such as the R^2_a and R^2_{LoF}.



Regression	
id	Primary Key
probe	CATMA ID
atg	ATG ID
fits	List of fits
spike	Spike fits

Fit	
id	Primary Key
model	Model shape name
parameters	Fitted parameters
fitobj	R fit object
stats	Link to statistics
regr	Link to regression

Spike	
id	Primary Key
model	Model shape name
parameters	Fitted parameters
fitobj	R fit object
stats	Link to statistics
spikepos	Time point of spike
spikesize	Size of spike
spikey	Expression value of spike
pval	p-value

Statistics	
id	Primary Key
aic	AIC value
r2	Adjusted R^2
lof	Lack-of-fit mean square
lofr2	Lack-of-fit R^2
ftest	F-test p-value

Figure 3.12: Schematic diagram showing the structure of the tables and a description of the fields in the regression database. The *Regression* table contains the information about the gene, the *Fit* table contains details regarding the model that fitted the data, the *Statistics* table contains all the goodness-of-fit statistics, and the *Spike* table contains the fit information for models with a potential spike. The arrows indicate foreign keys linking between the tables.

In addition, the AIC value and lack-of-fit mean squares were stored. The *Spike* table is similar to the Fit table, in that it contains information about the regression fit and a link to the relevant statistics field. In addition to this, it also contain details regarding the identified spike, such as the position (time point) and the magnitude of the spike.

3.9. Results

The regression analysis was performed on the 23 802 genes in the senescence and Botrytis time series datasets (Section 1.2). For the senescence dataset, the data consisted of morning and afternoon samples, and only the morning samples were used in this analysis. The Botrytis dataset consisted of responses for mock and infected treatments, and both were used. The results for the analysis of each of these datasets were stored in separate databases.

Figure 3.13 show the plots of the three main statistics used to determine goodness-of-fit in the regression, namely R_a^2, R_{LoF}^2, and F-test p-value for the senescence (Figure 3.13A) and Botrytis data (Figure 3.13B). These statistics all provided different information about the fit, so each of these were plotted against the other to determine the overall quality of the fits. The R_a^2 value describes how well the fit explains the data (including the between-replicate variability), the R_{LoF}^2 value describes the fit without the between-replicate variability, and the F-test provides an indication whether the fitted shape is appropriate. For the R_a^2 and R_{LoF}^2 values, the closer the value is to 1, the better the fit. For the F-test, the closer the value is to 0, the more likely that the fitted shape is appropriate. Since these p-values are often close to 0, the log of the p-values were plotted in the figures below. Thus, as the p-value tended towards 0 (becomes more significant), the more negative the plotted value becomes.

As would be expected, the lower p-values from the F-test are associated with higher values from the R_a^2 and R_{LoF}^2 values. In these plots, there is a greater amount of spread in the plot against R_{LoF}^2, as opposed to plotted against R_a^2 (Figure 3.13, column 1 vs. column 2). As described above, the R_a^2 calculation takes the between-replicate variability in account, whereas this has been removed in the R_{LoF}^2 value. This means that for a given F-test p-value, the R_{LoF}^2 value is greater than the R_a^2 value. This can also be seen in the plots of the R_a^2 against the R_{LoF}^2 (Figure 3.13, column 3), where a strong correlation between the two values can be seen, albeit a nonlinear relationship. This is caused by a relatively large amount of between replicate variability in the datasets, which is decreasing the value of the R_a^2 statistic. This effect is much more evident in the senescence dataset, likely due to the greater amount of biological variability in the older plants. In addition, in the plots of the F-test vs. R_a^2 (Figure 3.13, column 2), three distinct lines can be seen. These lines correspond to the number of parameters that each model possesses (two, three or four parameters). This indicates that for a given F-test p-value, models with more parameters (more complex models) would have a larger R_a^2 value.

Figure 3.14 shows the cumulative number of genes that pass each of the individual thresholds for the senescence (Figure 3.14A) and Botrytis data (Figure 3.14B). As with the above figures, the datasets have very similar patterns. The R_a^2 and R_{LoF}^2 both show an approximately linear drop off in numbers of genes as the stringency increases. With the F-test p-value, there are significantly fewer genes with a p-value greater than 0.1.

From these graphs and those of Figure 3.13, thresholds of $R_a^2 > 0.6$, $R_{LoF}^2 > 0.6$, and F-test p-value < 0.05 (log_{10}(p-value) $<$ -1.3) were established that were thought to be stringent enough to remove inadequate fits, but still flexible enough to allow a sufficient proportion of the genes through. Thus a set of genes with good fits was determined that could be used for further analyses.

The breakdown of fits according to model shape are shown in Figure 3.15. From these figures it can be seen that approximately 50% of the genes (12830 genes) in the senescence dataset, and 35% of the genes (8728 genes) in the Botrytis dataset had good fits for one or more models, according to the thresholds described above. In both datasets, the most common shapes were the linear, Gaussian and sigmoid curves (logistic and the two Gompertz variants). However, the majority of the linear model fits exhibited a flat expression profile (fitted parameters close to 0), and thus resulted in poor fit statistics. The exponential-type models fitted a fair amount of genes, but generally did not have a very good fit, particularly in the Botrytis dataset. This is due to a similar situation as the linear model fits, where the fitted parameters were all close to 0. Finally, the hyperbola shape did not match many genes.

As mentioned previously in Section 3.4, all regression fits with a $\Delta AIC < 2$ relative to the best fit, were retained. This means that an expression profile may be fitted by more than one model. Figure 3.15 showed only the counts of the models that fitted the best (i.e. had the smallest AIC value). Table 3.2 shows the number of genes with a secondary model fit for a given “best fit” model for the senescence dataset, and the

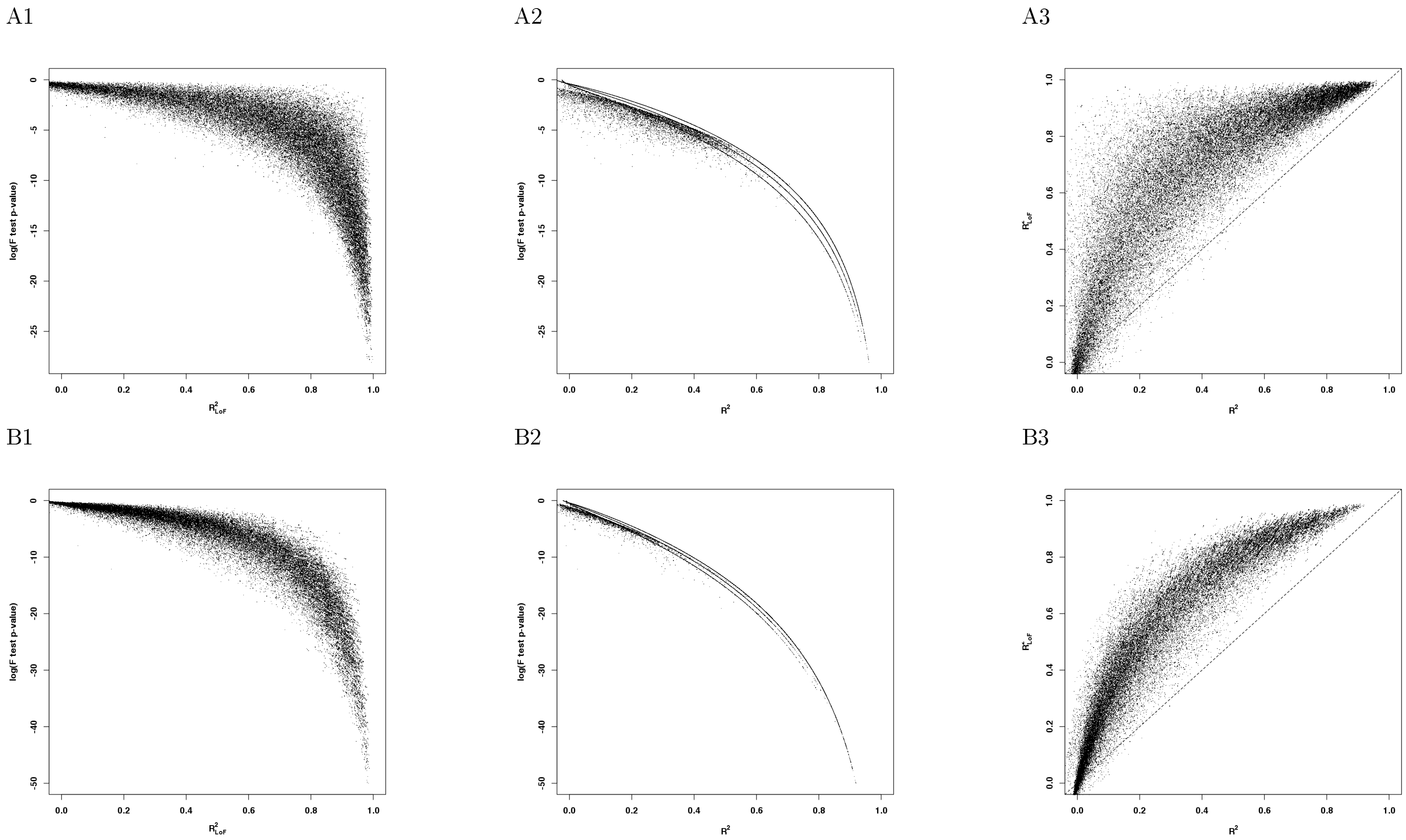
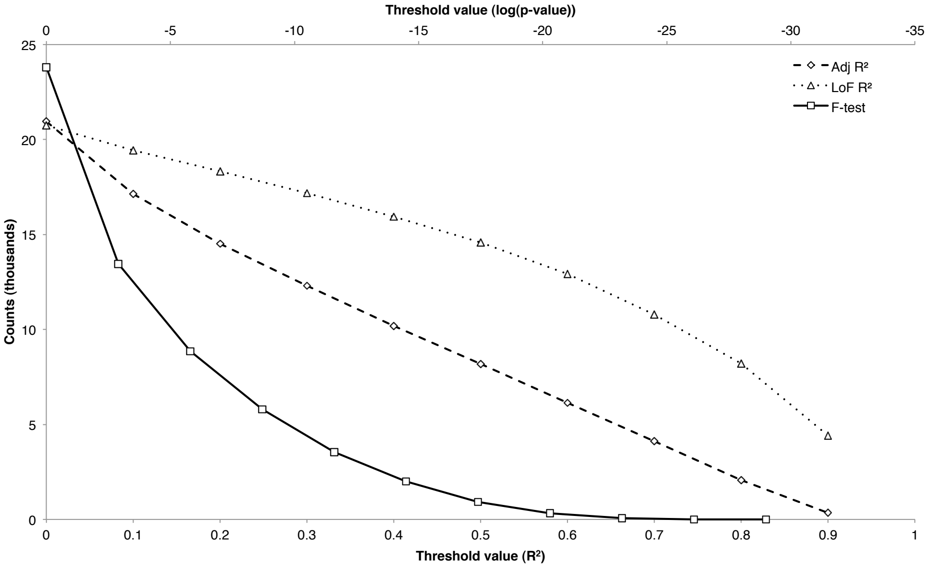


Figure 3.13: Plots of goodness-of-fit statistics for the senescence (row A) and Botrytis infected (row B) datasets. Column 1: R^2_{LoF} vs. F-test p-value; column 2: R^2_a vs. F-test p-value; column 3: R^2_a vs. R^2_{LoF}. High values for the R^2_a and R^2_{LoF}, and small values for the F-test p-value indicate a good fit. For better illustration purposes, the plotted values for the F-test p-values are given as log_{10}(p-value).

A



B

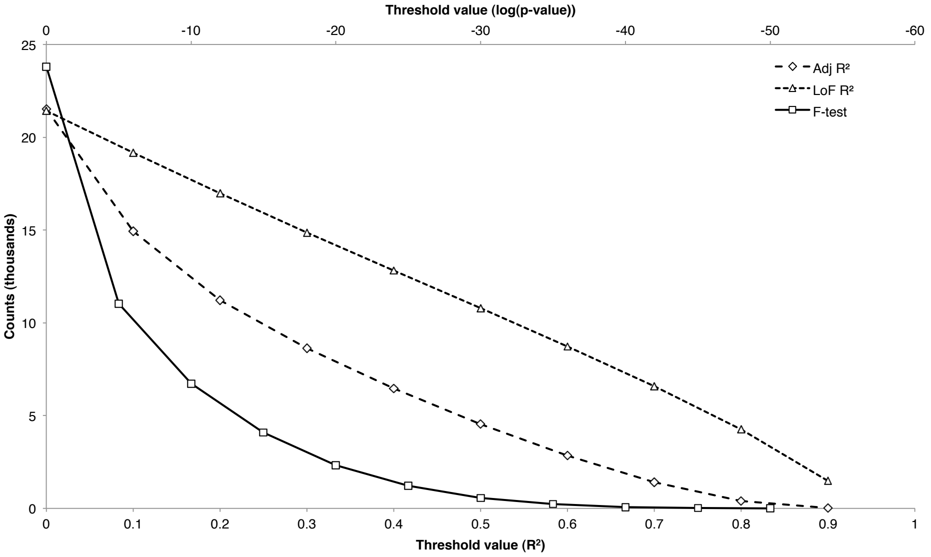


Figure 3.14: Cumulative counts of the number of genes that passed the various threshold levels for each of the goodness-of-fit statistics. Shown are the number of genes with a statistic greater than the respective cut-off value for the (A) senescence data, and the (B) Botrytis infected data. Values to the right indicate better fits. F-test values are read off using the top axis.

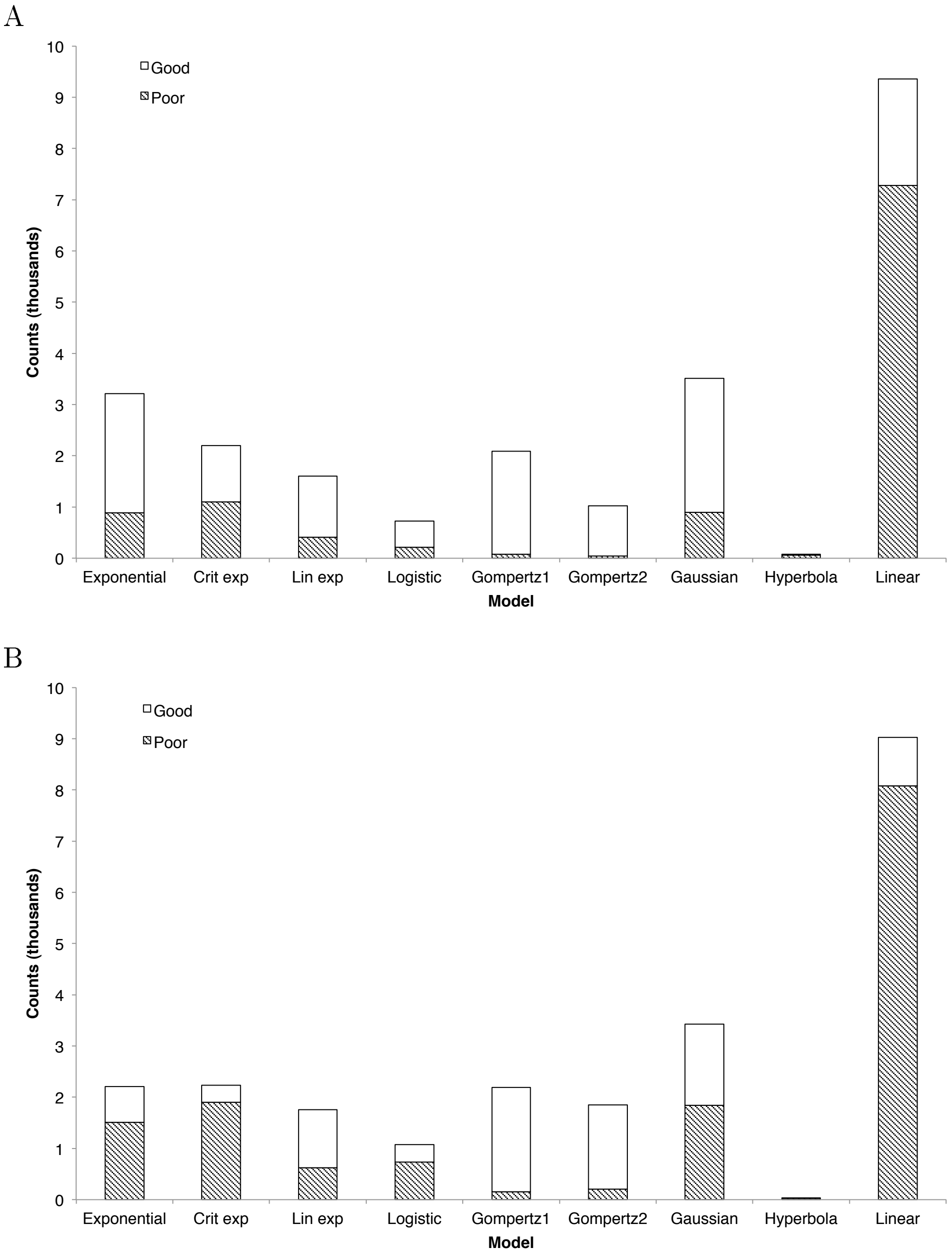


Figure 3.15: Break down of the fitted models by shape for the senescence (A) and Botrytis infected (B) datasets. The white portions of the bars indicate the number of genes that had a good fit according to the thresholds described previously ($R_a^2 > 0.6$, $R_{LoF}^2 > 0.6$, and F-test p-value < 0.05).

equivalent results for the Botrytis dataset are shown in Table 3.3. As would be expected, the sigmoid functions (logistic and the two Gompertz forms) often occurred together. The exponential and linear functions also had a high level of co-occurrence. A large proportion of genes only exhibited a single fit to the models.

Some examples of fits to the various datasets are shown in Figure 3.16, where a selection of different shapes are displayed. The fitted curves are displayed over the observed data points. It can be seen that the selected curve shapes are able to describe the general structure of the expression profile. Figures 3.16C and D show examples of genes where multiple shapes fitted to the same expression profile.

Although many of the gene expression profiles had a good fit to at least one of the given models, there were a number that did not. Figure 3.17 shows some examples of genes that did not match well, and possessed poor goodness-of-fit statistics. Figure 3.17A shows an example where the goodness-of-fit statistics were poor due to a flat, unchanging gradient, and Figure 3.17B shows an example of an expression profile that is inadequately described by the available models. This is particularly evident in the Botrytis data where the genes show circadian patterns.

While the goodness-of-fit statistics could be used to identify the quality of the fit to the data, an additional source of information is the standard error of estimation of the parameters for each of the models. The estimated value for a parameter may be unusual if a model is fitted to the data without sufficient information regarding the model's complete shape, for example in the sigmoid models, the second asymptote is not apparent. Using the standard error as a filter therefore provides an additional level of confidence that the fitted shape is appropriate. As described in Section 3.6, by using the standard error of estimation of the fitted parameters, it is possible to determine the level of confidence of the parameter estimates. Figure 3.18 shows the distribution of the fitted parameter values for the logistic model, after applying the goodness-of-fit thresholds, as well as the standard error of estimation checks. By doing this, it helps remove out any of the unusual fits, and acts as an additional filter to find genes with good fits for further analyses.

In addition to identifying the general expression profile, it is also possible to detect spikes in the profiles. A spike is defined as a large increase or decrease in gene expression at a single time point that deviates from the general fitted curve. Figure 3.19 shows some examples of spikes detected. With the use of the spikes, it is possible to find models that fit the expression profiles, save for one time point that has a different expression level to samples from neighbouring time points. This may be indicative of a cellular event where the gene expression is quickly being activated and then deactivated (or *vice versa*), indicating some other underlying activation mechanism.

Figure 3.20 shows the number of genes with a spike detected at each time point in the senescence and Botrytis datasets. Since the method for detecting the spikes is based on a leave-one-out methodology, it would be expected that time points with a large amount

Table 3.2: Table of number of best fits against the various secondary fits for the senescence dataset. The number of genes with a secondary fit of a particular shape (column) is given in the intersection with the best fit of the shape of interest (row). Genes which only had a fit to the model in question are shown on the diagonal. The proportion of genes with that particular shape is given in brackets.

		Secondary fit								
		Exponential	Crit exp	Lin+exp	Logistic	Gompertz1	Gompertz2	Gaussian	Hyperbola	Linear
Best fit	Exponential	951 (0.139)	652 (0.096)	531 (0.078)	896 (0.131)	995 (0.146)	976 (0.143)	635 (0.093)	21 (0.003)	1167 (0.171)
	Crit exp	112 (0.045)	1420 (0.574)	251 (0.101)	104 (0.042)	66 (0.027)	51 (0.021)	192 (0.078)	4 (0.002)	273 (0.110)
	Lin+exp	170 (0.080)	290 (0.136)	823 (0.385)	172 (0.081)	172 (0.081)	146 (0.068)	215 (0.101)	2 (0.001)	146 (0.068)
	Logistic	85 (0.069)	34 (0.028)	35 (0.029)	193 (0.158)	288 (0.235)	306 (0.250)	176 (0.144)	1 (0.001)	107 (0.087)
	Gompertz1	433 (0.102)	60 (0.014)	106 (0.025)	1518 (0.358)	333 (0.079)	969 (0.228)	591 (0.139)	5 (0.001)	227 (0.054)
	Gompertz2	162 (0.072)	48 (0.021)	79 (0.035)	794 (0.353)	598 (0.266)	106 (0.047)	323 (0.144)	0 (0.000)	137 (0.061)
	Gaussian	240 (0.048)	168 (0.033)	292 (0.058)	692 (0.137)	635 (0.126)	572 (0.113)	1992 (0.395)	1 (0.000)	453 (0.090)
	Hyperbola	17 (0.189)	4 (0.044)	3 (0.033)	0 (0.000)	0 (0.000)	0 (0.000)	3 (0.033)	34 (0.378)	29 (0.322)
	Linear	4535 (0.389)	377 (0.032)	303 (0.026)	577 (0.050)	523 (0.045)	439 (0.038)	693 (0.059)	100 (0.009)	4104 (0.352)

Table 3.3: Table of number of best fits against the various secondary fits for the Botrytis infected datasets. The number of genes with a secondary fit of a particular shape (column) is given in the intersection with the best fit of the shape of interest (row). Genes which only had a fit to the model in question are shown on the diagonal. The proportion of genes with that particular shape is given in brackets.

B

		Secondary fit								
		Exponential	Crit exp	Lin+exp	Logistic	Gompertz1	Gompertz2	Gaussian	Hyperbola	Linear
Best fit	Exponential	552 (0.129)	681 (0.159)	456 (0.106)	445 (0.104)	422 (0.099)	398 (0.093)	374 (0.087)	3 (0.001)	952 (0.222)
	Crit exp	186 (0.070)	1345 (0.504)	213 (0.080)	133 (0.050)	69 (0.026)	56 (0.021)	187 (0.070)	2 (0.001)	480 (0.180)
	Lin+exp	184 (0.122)	134 (0.089)	504 (0.333)	109 (0.072)	63 (0.042)	96 (0.063)	131 (0.087)	1 (0.001)	290 (0.192)
	Logistic	79 (0.035)	101 (0.045)	51 (0.023)	753 (0.334)	307 (0.136)	637 (0.283)	147 (0.065)	0 (0)	177 (0.079)
	Gompertz1	109 (0.030)	84 (0.023)	38 (0.010)	1546 (0.425)	457 (0.126)	1095 (0.301)	191 (0.053)	0 (0)	114 (0.031)
	Gompertz2	159 (0.048)	101 (0.030)	85 (0.026)	1312 (0.395)	882 (0.265)	345 (0.104)	241 (0.073)	1 (0)	198 (0.060)
	Gaussian	147 (0.036)	180 (0.044)	184 (0.044)	301 (0.073)	238 (0.058)	172 (0.042)	2531 (0.612)	0 (0)	383 (0.093)
	Hyperbola	7 (0.194)	4 (0.111)	2 (0.056)	0 (0)	0 (0)	0 (0)	0 (0)	11 (0.306)	12 (0.333)
	Linear	4222 (0.370)	883 (0.077)	708 (0.062)	541 (0.047)	320 (0.028)	420 (0.037)	677 (0.059)	70 (0.006)	3577 (0.313)

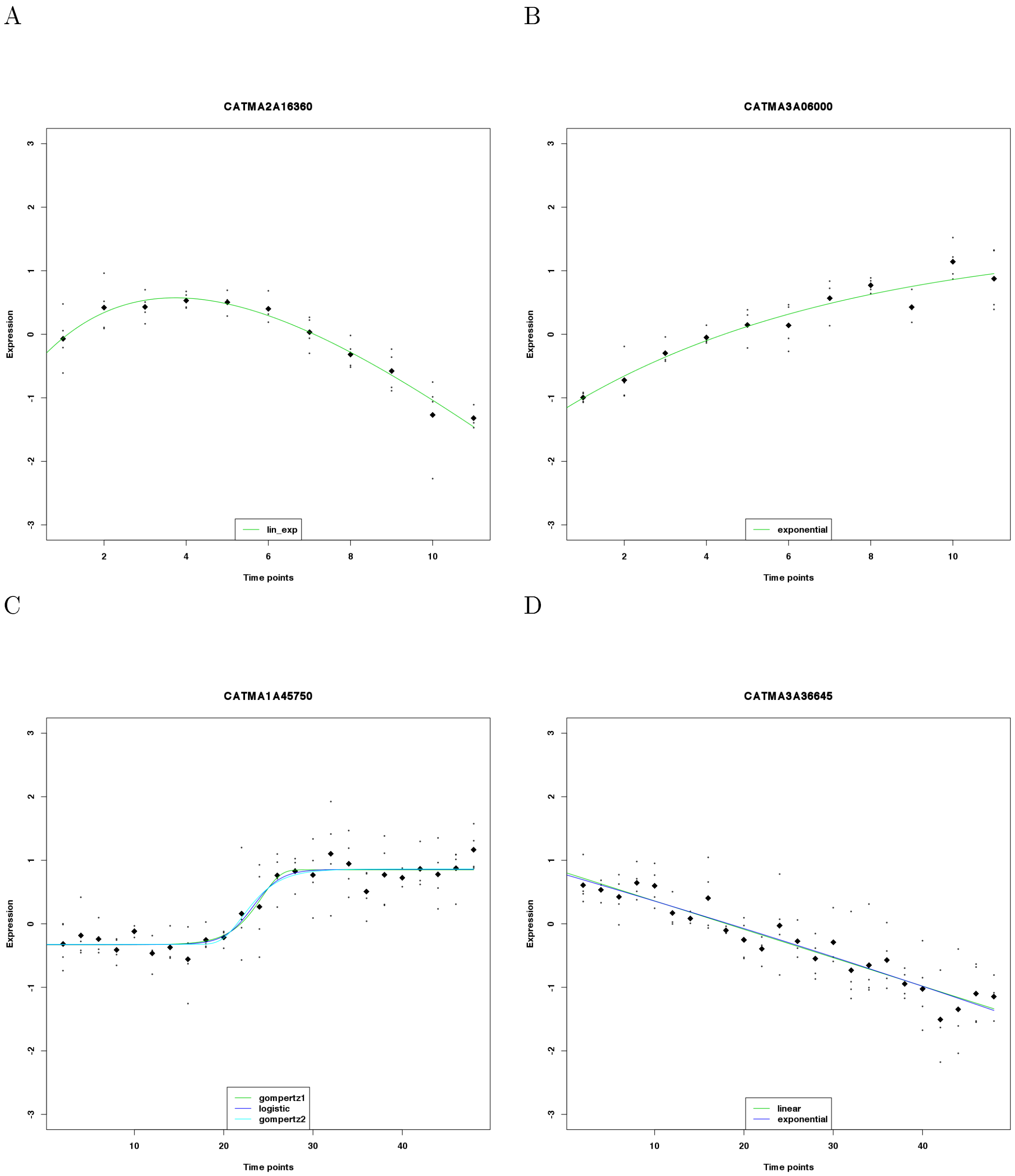


Figure 3.16: Examples of regression fits to the gene expression profiles. The top row represents samples from the senescence dataset, where (A) is an example of a linear+exponential fit, and (B) shows an exponential fit. The bottom row shows profiles from the Botrytis dataset where (C) fitted the sigmoid curves, and (D) fitted linear and exponential. The small dots on the graphs represent the replicate samples, and the black diamonds represent the replicate means.

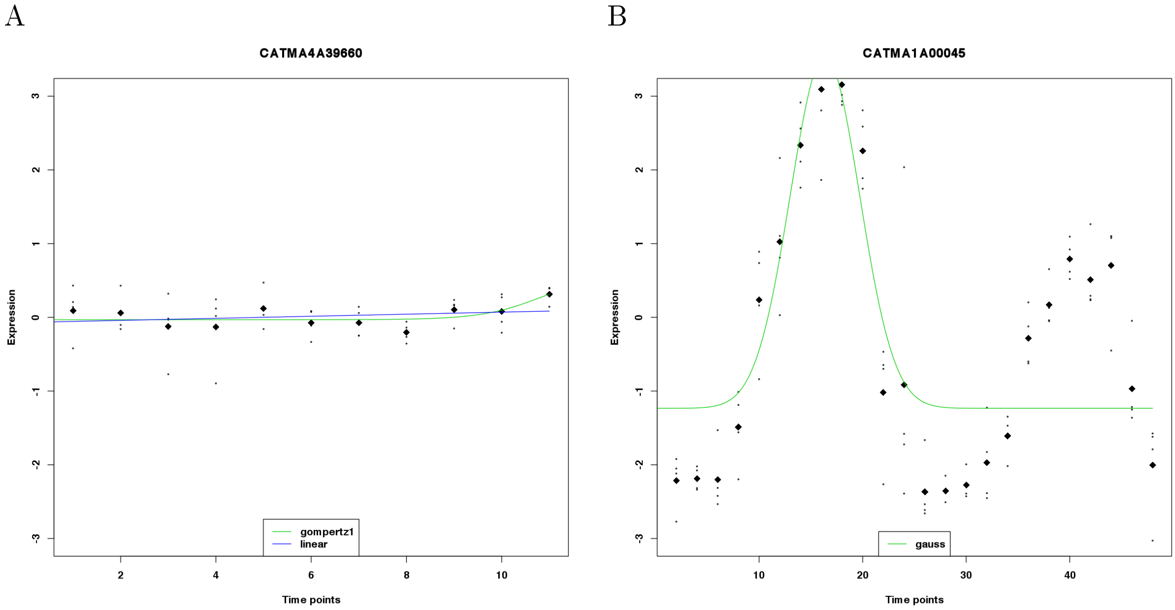


Figure 3.17: Examples of genes with poor fits, where (A) shows an example where the fitted parameters were close to 0, resulting in poor goodness-of-fit statistics; and (B) shows an example of an expression profile that is more complex than the models used in the regression analysis.

of leverage would be picked up more often. This can be seen in the figure where the time points at the beginning and end of the two time series have a larger proportion of detected spikes. In total, there were not a large proportion of genes that were detected with spikes. 6723 (28% of genes) in the senescence dataset and 3178 (13% of genes) in the Botrytis dataset.

In the time series papers by Breeze *et al.* (2011) and Windram *et al.* (2012), sets of genes were identified as being differentially expressed through a variety of statistical approaches, including MAANOVA (Wu *et al.*, 2003) and a Gaussian process two-sample test (Stegle *et al.*, 2010). These genes were compared to the genes that were identified as possessing good fits to the expression profiles. Using the filtering methods described above (R^2_a, R^2_{LoF}, F-test p-value, and standard error of the parameter estimates) a set of 8216 and 5303 genes were respectively found in the senescence and Botrytis datasets, which possessed a good fit to one or more of the models. Figure 3.21 shows the degree of overlap between the regression method and these established statistical methods for identifying differentially expressed genes. In the senescence dataset, approximately equal number of genes were found by both methods, and the majority of genes that were identified as being differentially expressed also had a good fit to the regression models. However, there were a relatively large number of genes that were uniquely identified by each of the two methods. In contrast, in the Botrytis dataset, there were almost twice as many genes found by the differential expression analyses as the regression approach. However, the majority of the genes possessing a good model fit were also identified as being differentially expressed.

A

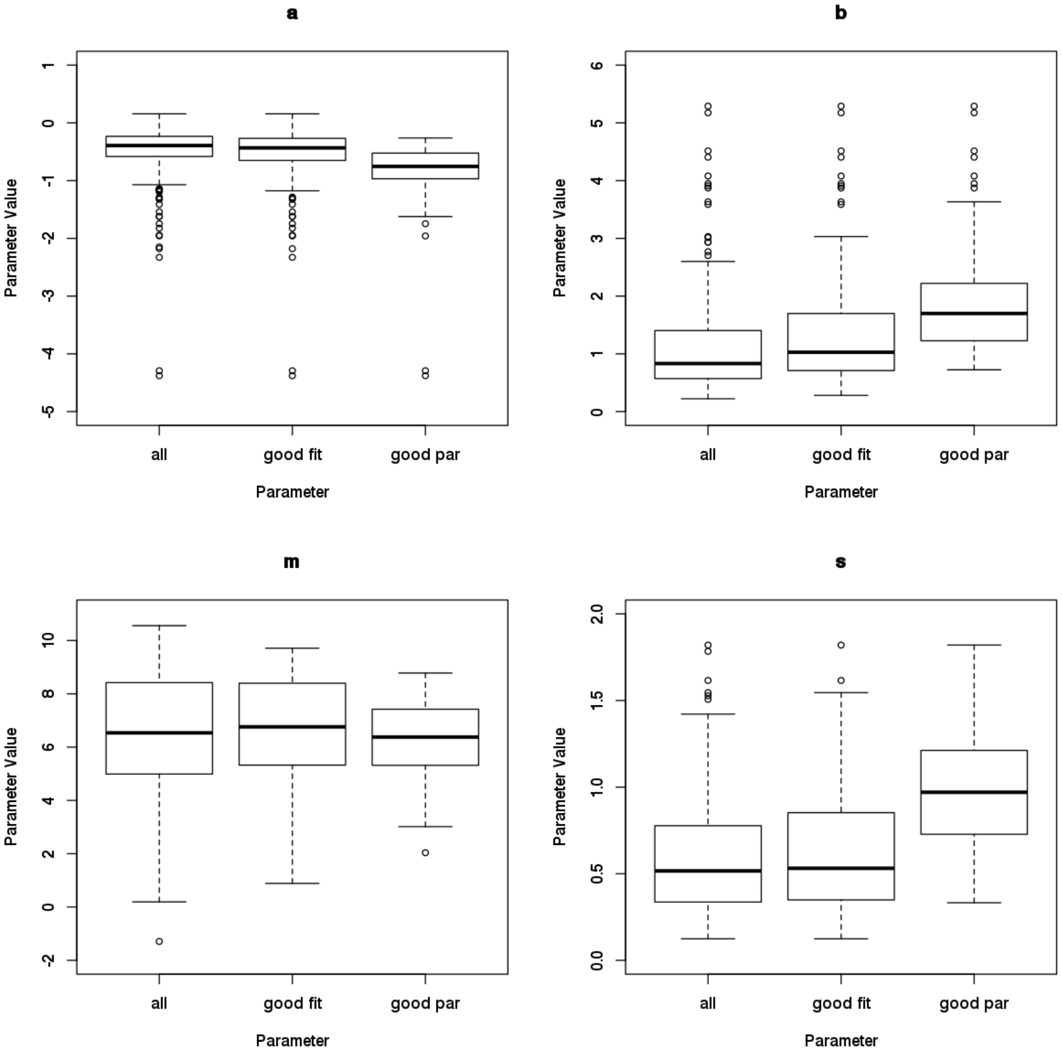


Figure 3.18: Distribution of the parameters for the logistic function in the senescence dataset, showing the (A) increasing, and (B) decreasing functions (overleaf). The distribution of each parameter is shown for the full dataset (left), filtered by the goodness-of-fit thresholds (centre), and filtered by the goodness-of-fit thresholds as well as the standard error of estimation (right).

B

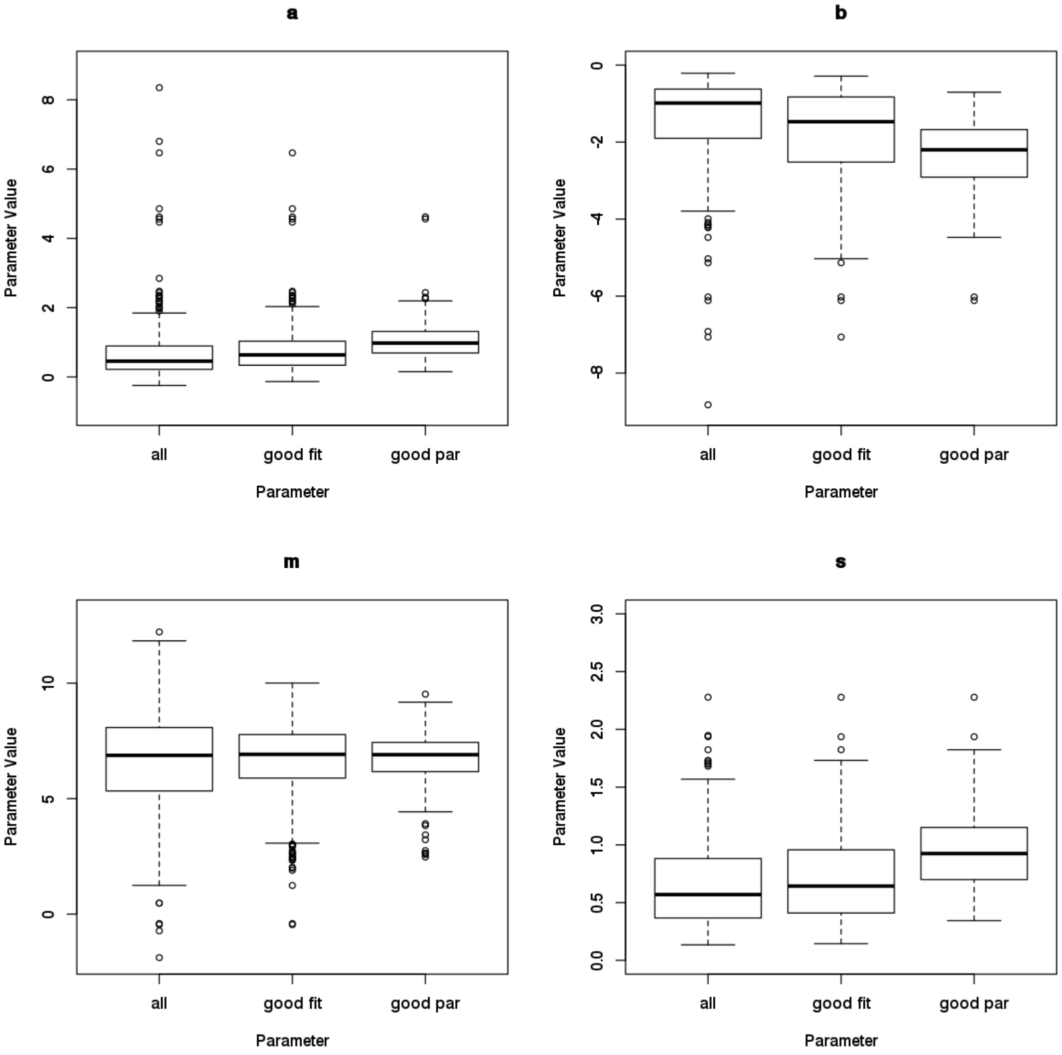


Figure 3.18: (cont.) Distribution of the parameters for the logistic function in the senescence dataset, showing the (B) decreasing functions. The distribution of each parameter is shown for the full dataset (left), filtered by the goodness-of-fit thresholds (centre), and filtered by the goodness-of-fit thresholds as well as the standard error of estimation (right).

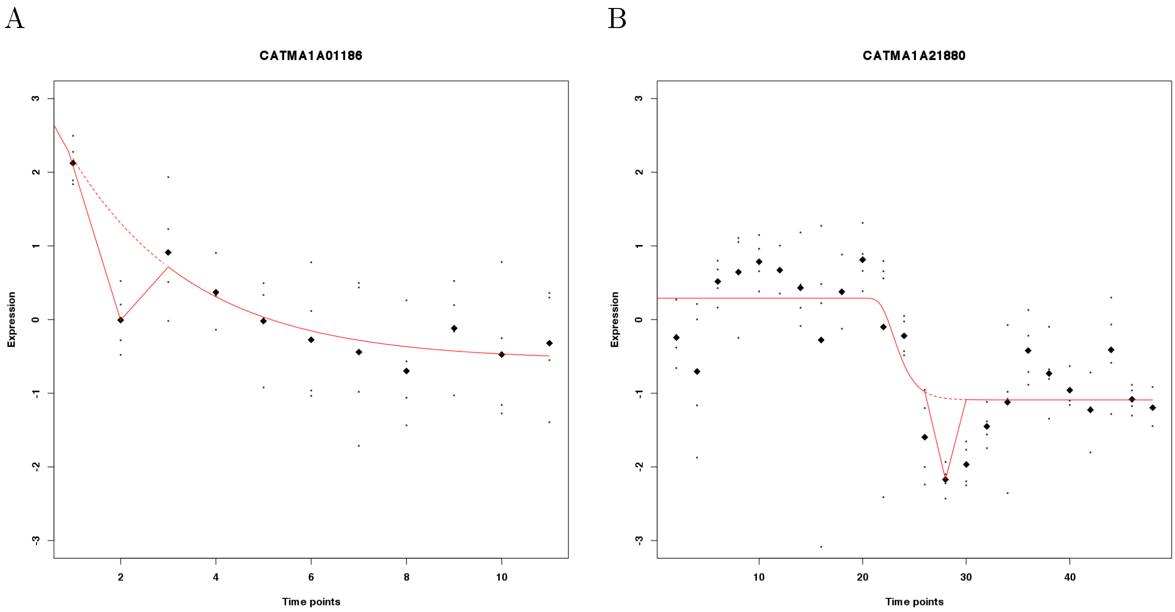


Figure 3.19: Some examples of spikes from the senescence dataset (A), and the Botrytis dataset (B). The dotted portion of the line represents the fitted curve without the spike. The small dots on the graphs represent the replicate samples, and the black diamonds represent the replicate means.

3.10. Discussion

In this analysis, a regression approach was used to fit a selection of models to gene expression profiles, and thus obtain biologically interpretable parameters to aid in the identification of functionally related genes. Eight distinct shapes were used to fit the expression profiles, and these shapes were able to fit to a large proportion of the genes. Ordinarily with nonlinear regression, starting values for the regression would be estimated using a graphical exploration, or through the use of a grid search of potential parameter values (Ritz and Streibig, 2008). However, in this case, there were 2 datasets with over 23 000 gene expression profiles in each. Thus a more automated approach was needed. Self-starter functions were developed to estimate starting values, and were integrated into an analysis pipeline to fit each of the selected models to each gene expression profile, and determine the best fits. All the relevant data was stored in a database for further analysis.

Through the use of goodness-of-fit statistics, the quality of the fits were determined. These statistics included the R_a^2, R_{LoF}^2, and F-test p-value, and each of these provided a different indication of the fit. Investigating the overall trends in these statistics, threshold values were determined in order to filter out the models with poor fits. The thresholds determined were $R_a^2 > 0.6$, $R_{LoF}^2 > 0.6$, and F-test p-value < 0.05, although these can easily be changed to increase or decrease the stringency as desired. The number of genes that fitted each shape was calculated, and it was found in both the senescence and Botrytis datasets that the predominant shape was the linear response. Investigation of some of these fits revealed that the genes exhibited a low level of expression, and had a flat, unchanging response over time, thus having both a gradient

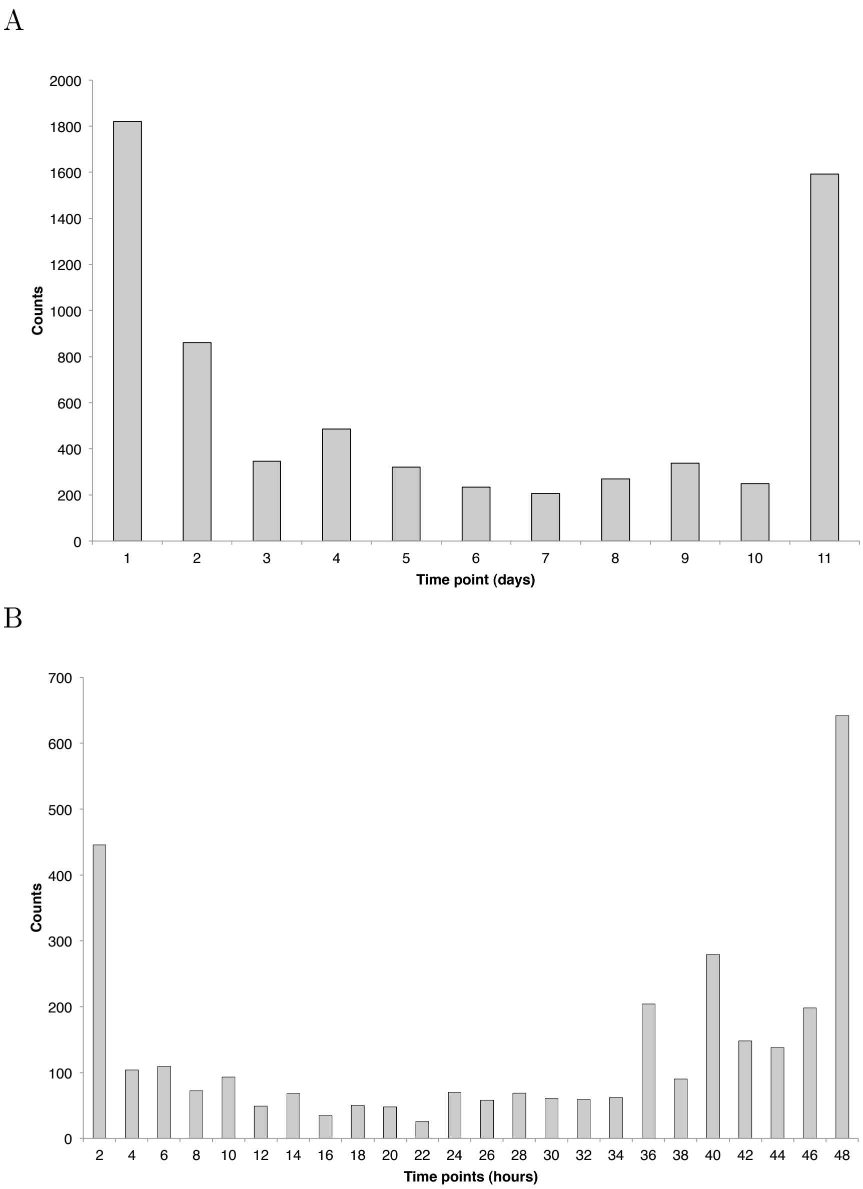


Figure 3.20: Figures showing the distribution of the spikes in the senescence (A) and Botrytis (B) datasets. Shown are the number of spikes detected at each of the time points.

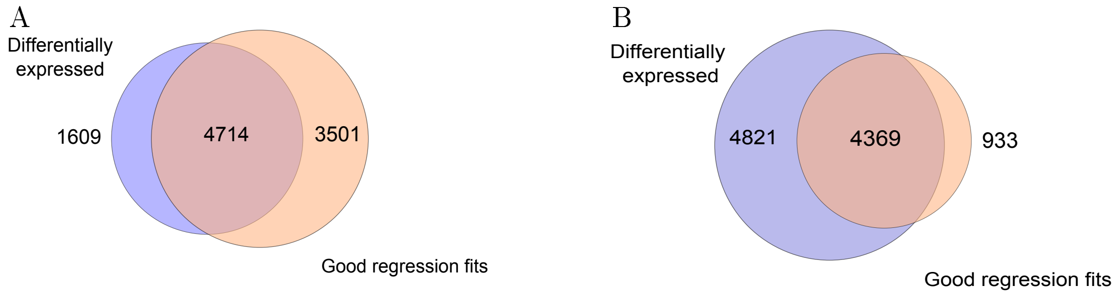


Figure 3.21: Venn diagrams showing the degree of overlap between the genes identified as being differentially expressed through statistical means (blue circles), and the genes with a good regression fit (orange circles). The senescence dataset is shown in (A) and the Botrytis dataset in (B)

and intercept close to 0. The other commonly occurring shapes were the logistic and the two forms of the Gompertz curve, i.e. the sigmoid shapes. This is to be expected as they follow the anticipated change in gene expression, where a gene is activated in response to some stimuli, which results in an increase (or decrease) in expression until a new steady state level is achieved. With regard to the exponential type curves, a similar situation to the linear fits was found where the expression profiles were flat and unchanging, so the fitted parameters were close to 0. Finally, the hyperbolic shape did not fit many genes at all, most likely due to the exponential curve providing a better fit.

These results seem to suggest that while many of the gene expression profiles could be adequately described by the selected shapes, there were still some that were not. Further investigation would be needed to identify and parameterise the missing model shapes. Here, techniques such as splines have the advantage, as they are more flexible and thus able to handle unusual profile shapes. However, the purpose of this analysis is to obtain more information from the expression profiles than merely their shapes, namely additional information regarding the underlying mechanisms for the given expression profile. Through the use of the fitted parameters and goodness-of-fit statistics, a more exploratory approach was developed to aid in the analysis of the data.

The thresholds described above were used to identify genes that had a good fit to the data, and were selected based on the number of genes that passed a given threshold. This was in an attempt to maximise the number of genes included in further analyses, while still maintaining a level of stringency. However, these thresholds are still ultimately arbitrary, and may be raised or lowered to make them more or less stringent, respectively. In addition, the thresholds do not inform about the amount of error in the parameter estimates. This means that although the fit may be of high quality, the standard error of the parameters may be relatively high, indicating that there is insufficient data to accurately predict the parameter value. For example, a gene expression profile may look like half of a Gaussian curve, and the Gaussian profile would fit it reasonable well. However there would be high errors associated with the asymptote parameter estimate, indicating that some components of the shape were from extrapolation of the dataset. Thus, fitted curves should be further filtered by investigating the errors associated with the parameter values.

The inclusion of spikes provided a means of extending the fitting process to include more unusual shapes. At present only one spike is permitted per expression profile. An extension would be to allow multiple spikes, particularly those that are adjacent, indicative of a dip as opposed to a spike. Other extensions could include the use of piecewise regression or a broken stick model, where different portions of the expression profile are fitted by multiple shapes. Using a leave-one-out methodology to find the spikes biases the analysis to identify genes that are at the beginning or end of the time series. That is, the time points which possess a large amount of leverage on the regression fit. Of particular interest are the genes with spikes in the middle of the time series. It would be interesting to determine if there is some biological, or possibly technical, reason for a set of genes to have spikes at a particular time point.

While it was possible to fit a variety of models to a large number of expression profiles, many of these fits were poor quality. As mentioned previously, it was found that many of these poor fits were largely unresponsive across the time series. Thus is may be possible to use the regression and goodness-of-fit assessments as a means of identifying differentially expressed genes. When the list of good fits was compared to the differentially expressed gene lists from statistical analyses such as MAANOVA, it was found to be largely consistent. In the case of senescence, many more genes were found to be differentially expressed by the regression analysis, although this could be adjusted by making the default thresholds more stringent. In the Botrytis dataset, the regression analysis found half as many genes as the differential expression analyses. However, the majority of the genes were found by both methods. Thus, the regression analysis could act as a means of filtering out genes for further investigation. The genes that were found to be differentially expressed but not possess a good model fit could be due to the presence of circadian genes that cannot currently be accurately detected by this regression approach. A way to possibly identify these genes would be to attach a sine term to the regression models as an additional parameter to overlay oscillatory behaviour. Alternatively, a Fourier analysis could be used to identify the diurnal signal. Nonetheless, in both datasets most of the differentially expressed genes were identified as being good fits and could be used as a simple means to identify differentially expressed genes.

The chapters that follow will build on the use of the fitted models, demonstrating a variety of applications.

4. Using fitted parameter values to group genes

4.1. Introduction

In the previous chapter, a regression approach was described whereby linear and non-linear functions were fitted to time series gene expression profiles, in order to obtain parameters which may be used to interpret the underlying biology, such as timing events or rates of change. In this chapter, an application of these fitted models will be described, namely clustering genes based on the fitted parameter values.

Clustering is a generic term that is used for the division of objects on which multivariate data has been measured into groups, or clusters (Jain *et al.*, 1999). The aim of the analysis is to identify groups of genes that have similar profiles to each other, but are distinct from all other groups. Eisen *et al.* (1998) were among the first to apply a clustering algorithm to gene expression data. The hypothesis was that genes which clustered together (co-expressed genes) would have similar functions, be involved in the same metabolic pathway, and possibly be co-regulated by the same transcription factors (Rasmussen *et al.*, 2013; Williams and Bowles, 2004).

In traditional clustering analyses, expression profiles are grouped together across the entire time-series. In this analysis, a different approach was taken whereby the profiles were clustered using the fitted parameters. Since each parameter influences the shape of the various models, by clustering genes on a single parameter, this makes it possible to focus in on a specific aspect of the expression profile, and find genes with that aspect in common. In this way, the parameters could be related to some underlying biological process, and were used to identify genes that are potentially functionally related in a mechanistic manner. For example, it was possible to identify groups of genes that are activated at the same time point suggesting that these genes have a common regulator. In addition, it was possible to find genes with similar rates of change in expression, suggesting that the genes that are activated earlier may be regulating those that are activated later. Thus, this makes it possible to analyse the sets of genes using biologically oriented descriptions of individual gene expression profiles, enabling easier interpretation of the profiles and providing a means of understanding the common regulatory mechanisms between genes.

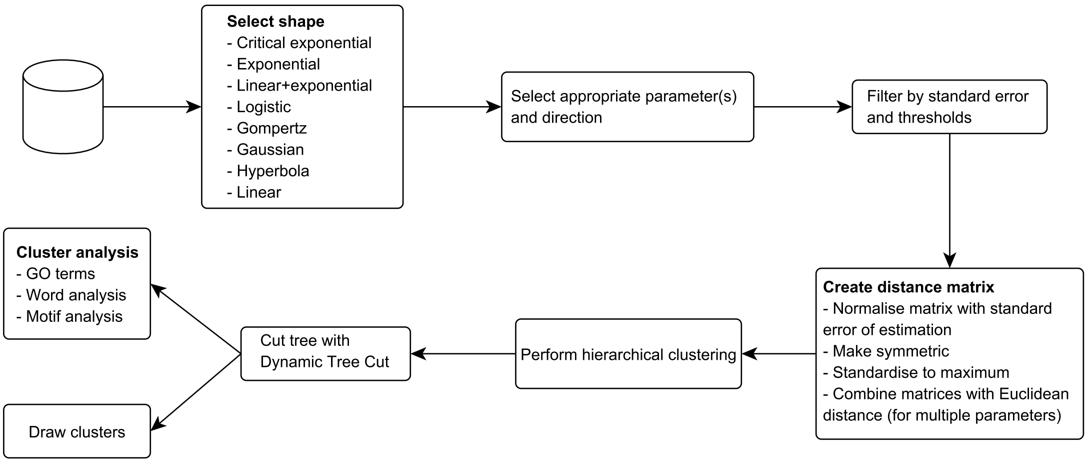

Select shape
- Critical exponential
- Exponential
- Linear+exponential
- Logistic
- Gompertz
- Gaussian
- Hyperbola
- Linear
Select appropriate parameter(s) and direction
Filter by standard error and thresholds
Create distance matrix
- Normalise matrix with standard error of estimation
- Make symmetric
- Standardise to maximum
- Combine matrices with Euclidean distance (for multiple parameters)
Perform hierarchical clustering
Cut tree with Dynamic Tree Cut
Cluster analysis
- GO terms
- Word analysis
- Motif analysis
Draw clusters

Figure 4.1: Schematic diagram of ShapeCluster's clustering process. Fit information for a given model shape is extracted from the database, and the parameter values extracted. A distance matrix is calculated from this data, and used in a hierarchical clustering. Once the clusters are formed, the biological significance of the clusters is determined through the use of annotation over-representation.

4.2. Method development

4.2.1. ShapeCluster - clustering genes using shape and parameters

In order to identify sets of genes with similar fitted parameter values, a methodology to cluster fitted expression profiles using the results from the regression analysis (Chapter 3) was developed, and referred to as ShapeCluster.

In contrast to conventional clustering analyses, ShapeCluster analyses the fitted curves rather than the observed data. The algorithm operates in a two step process: first, sets of genes are identified based on the particular model that best fitted the gene expression profiles; and second, the similarity of genes based on one or more of the biologically interpretable parameters is determined. These parameters may be the actual fitted parameters from the regression, or derived parameters (Section 3.7). A distance matrix is created from the parameter values, and a simple hierarchical clustering analysis is applied to this. The clusters are then determined using the Dynamic Tree Cut algorithm (Section 2.1.3, Langfelder *et al.* 2008). Once the clusters have been identified, the biological significance of the clusters is determined through the use of annotation metrics, such as Gene Ontology (GO) terms (Section 2.1.4-2.1.8). This process is described in greater detail below, and is summarised in Figure 4.1.

4.2.2. Formation of the distance matrix for clustering

After the expression profiles are fitted by a model, it is possible to cluster the fitted parameters. After selecting a model and a parameter to cluster on, a matrix of scaled differences between each gene was calculated. This was determined by taking the absolute difference between the parameter value for each gene, and dividing by the standard error for the parameter of the first gene. So, if the distance between gene A and gene

B for the parameter m was to be determined, this would be calculated as

$$distance_{AB;m} = \frac{|m_A - m_B|}{stderr_{m:A}}.$$

The calculation for $distance_{BA;m}$ would be the same except dividing by the standard error of m for gene B. A square matrix of distances can thus be calculated. To make the matrix symmetric, the average of the distances for the two genes in question was calculated.

$$sym\,dist_{AB;m} = sym\,dist_{BA;m} = \frac{distance_{AB;m} + distance_{BA;m}}{2}.$$

To cluster based on multiple parameters, the Euclidean distance between the individual parameter distance matrices was calculated. For example, the distance between genes A and B for parameters m and s would be calculated as

$$distance_{AB;m,s} = \sqrt{(sym\,dist_{AB;m})^2 + (sym\,dist_{AB;s})^2}. \tag{4.1}$$

This form of clustering with multiple parameters was termed *simultaneous clustering*, as the distance between multiple parameters is determined in one calculation. This is to distinguish it from the meta-clustering approaches described below in Section 4.2.3, where multiple clustering parameters are dealt with independently.

The matrix was then normalised by dividing by the maximum value, thus resulting in the highest value in the matrix always equalling 1. Using the `hclust` function and this distance matrix, a hierarchical clustering was performed to group the most similar genes together. Clusters were formed using the average linkage distance method and the `Dynamic Tree Cut` package (Section 2.1.3, Langfelder *et al.* 2008).

4.2.3. Meta-clustering

In addition to the simple clustering approach described above, several extensions for multiple parameters were developed. In the previously described process, differences between multiple parameters were combined together into a single metric using a Euclidean distance. However, this could result in a loss of some of the individual underlying structure as described by each parameter, particularly when the parameters occurred on different scales. These alternate forms of clustering could provide more biologically relevant clusters by preserving this structure, and so aid in more effectively identifying co-expressed and potentially co-regulated sets of genes.

One alternative method to analyse the clusters is to re-cluster the members of each initial cluster using additional information. Two different approaches were taken. The first is a "sequential" clustering (Figure 4.2A). In this clustering, the shapes are clustered first on a single parameter. Following this, the members of each of the clusters are further clustered based on a second parameter. Thus, by clustering the members of the original clusters, more refined sets of genes are identified. An example is shown in Figure 4.2A. After selecting a cluster and a parameter to cluster on, a set of clusters is formed (left). Following the arrows, it can be seen that each of these clusters is then subsequently clustered using the second parameter, resulting in a new set of clusters (right).

The second clustering method is a "cross" clustering (Figure 4.2B). In this case, the clustering on each of the two parameters is performed independently, thus resulting in two sets of clusters. The genes contained within each cluster in one set is compared to the genes contained in every other cluster in the other set, in a pair-wise manner (cross classification of the cluster memberships). If a gene is contained in a pair of comparisons, the gene is assigned to this intersection of two clusters. From Figure 4.2B, it can be seen that this results in a matrix of subclusters. The margins represent the clusters from the single parameter cluster analysis, and the elements of the matrix represent the genes that are common between the clusters in that row and column's margins.

4.2.4. Clustering using control information

In many microarray time-course experiments, there may be a set of samples for a control treatment, in addition to other treatments applied over the time course. For example, in a pathogen experiment, comparing a mock inoculated control with a pathogen inoculation treatment. Alternatively this may be a comparison of a wildtype organism with a mutant. It may thus be useful to compare model fits from these different treatment sets.

After the models have been separately fitted to the treated and control datasets, it is possible to use the fitted models in each of the datasets to determine if there are any differences. If the cluster analysis is performed on the fitted models for the treated data, it is then possible to subdivide the clusters by taking the control fitted model into account. Two outcomes are possible - either the models will be different, in which case it is possible to further group the genes by the control model, or the model will be the same. In the latter case, it is possible to perform a regression analysis using the data from both treatment sets together to identify whether parameters are common across the treatment sets. This provides an indication of any significantly different parameters between the two fits, thus indicating where the significant differences between the expression profiles are. For example, a gene may have very similar parameters between the treatments except for a different midpoint parameter, implying that the gene reacts similarly in both the control and treated datasets, just at different times.

To determine if two treatments with the same fitted model have significantly different parameters, an iterative process was developed whereby more complex models were compared to simpler models using the change in the residual variance. This process uses a top-down approach where a simple model is compared to more complex models, in order to find the most parsimonious model. These complex models are created through the addition of a grouping factor, which makes it possible to vary each parameter while keeping the others constant. Initially, the model with the grouping factor for a particular parameter is compared to the original model using an F-test. If the reduction in residual variance is significant (p-value < 0.05), this means that the model with the grouping

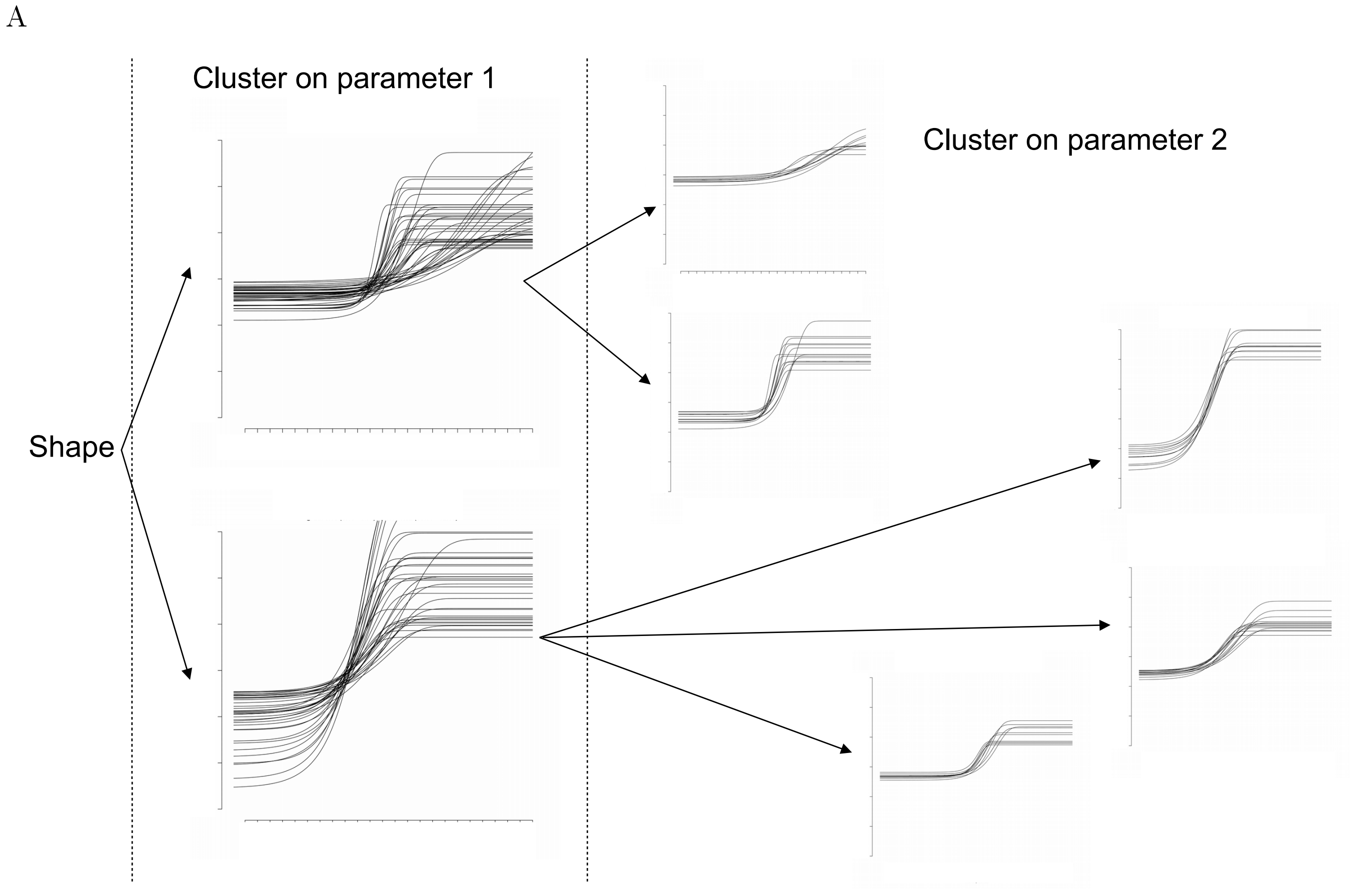


Figure 4.2: Schematic diagrams illustrating the two re-clustering approaches. (A) The sequential clustering, where genes are clustered using one parameter, and these original clusters are individually re-clustered according to the second parameter. Here the Gompertz1 model is first clustered using the *5per* parameter followed by the *grad* parameter. (B) A cross clustering (overleaf) where the genes are clustered independently using two parameters, and new clusters are identified through the intersections of pairs of clusters.

B

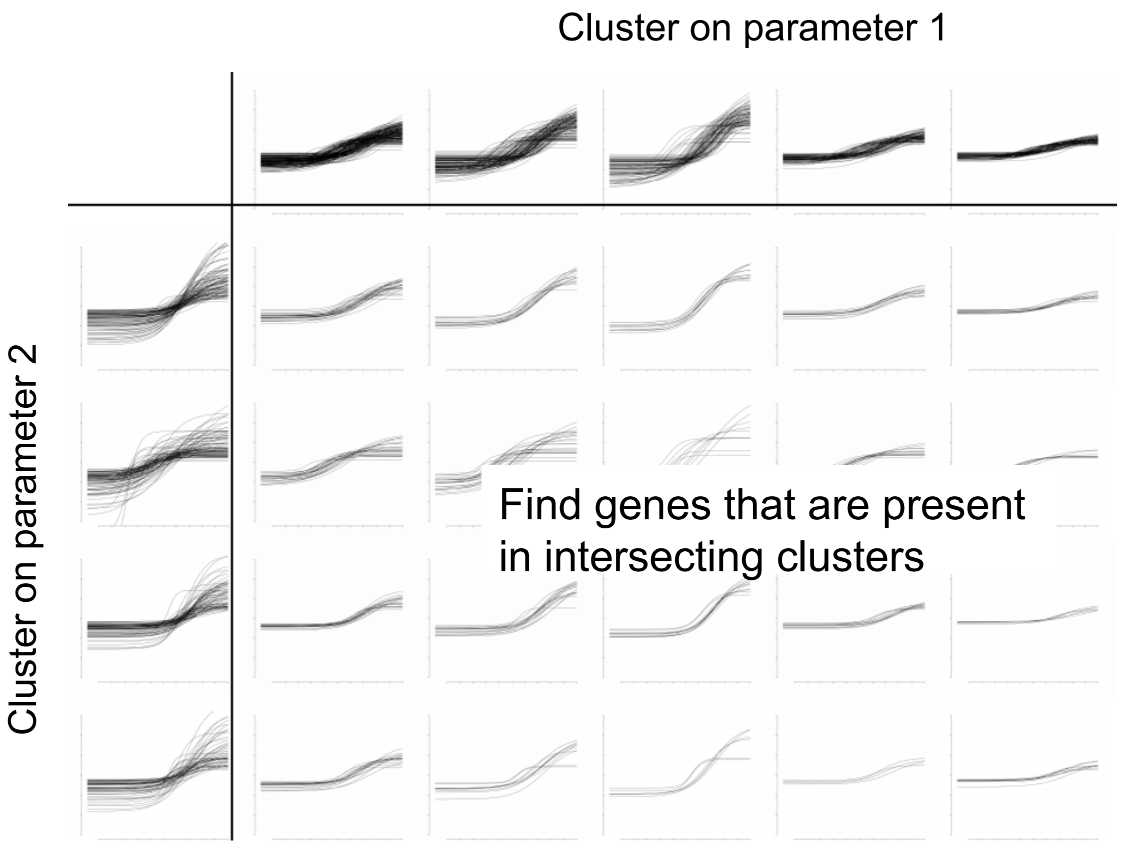


Figure 4.2: (cont.) (B) An example of cross clustering, where the genes are clustered independently using two parameters, and new clusters are identified through the intersections of pairs of clusters. Here the Gompertz1 model is clustered using the *5per* parameter (left) followed by the *grad* parameter (top).

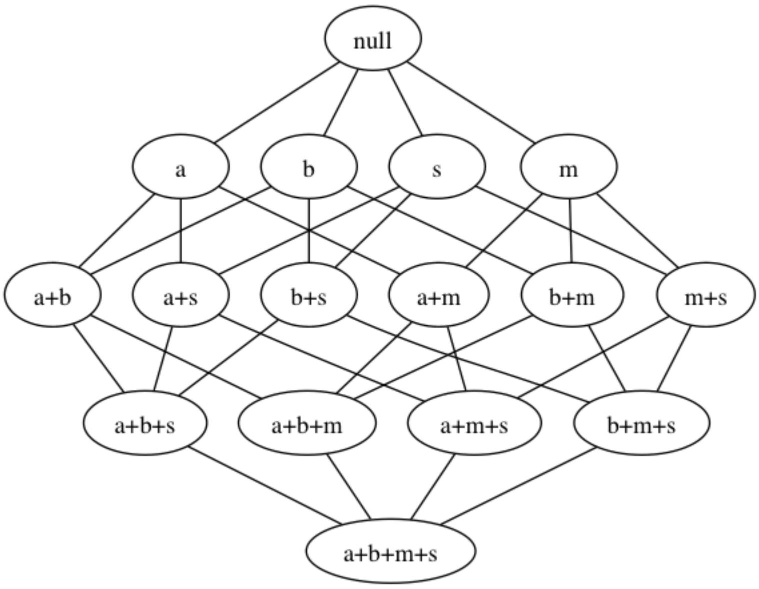


Figure 4.3: Hasse diagram showing the different combinations of parameters for the logistic curve, where each node represents the parameters that were allowed to vary. The top of the figure represents a model where none of the parameters were allowed to vary, and each level down represents an increase in complexity with the addition of a grouping factor, allowing the parameter value to change while keeping the others constant. By allowing each of the combinations of parameters to vary, more complex models may be formed. Models are compared upwards to the less complex models, and significance determined using an F-test.

factor fits the data better, and the more complex model is retained. If the addition of a grouping factor improved the fit, an additional parameter was allowed to vary, and compared to the model where only one parameter was allowed to vary. This process is continued until varying additional parameters does not reduce the residual variance.

Figure 4.3 shows a Hasse diagram providing an illustration of the different models that can be identified for the logistic curve, as well as which models pairs are compared. The top of the figure represents a model where none of the parameters were allowed to vary, and each level down represents an increase in the complexity of the model, where more parameters are allowed to vary. Thus, in the first instance, the grouping factor is incorporating within each of the four parameters (a, b, s, m), thus allowing each of the parameters to vary. Each of these models is then compared upwards to the model where no parameters were allowed to vary. If allowing a parameter to vary improved the model fit, the same procedure was performed on combinations of parameters models. For example, if varying the m parameter improved the fit, this model would be compared to models where the $m+a$, $b+m$ and $m+s$ parameters were allowed to vary. This process continues until allowing additional parameters to vary does not provide a significant improvement to the fit, or all parameters are allowed to vary.

Example results for this approach are shown in the following chapter, in Section 5.3.

4.2.5. Methods for assessing clusters

Once the clusters are formed, the biological significance of these clusters needs to be determined. The Gene Ontology (GO) is one of the most commonly used methods to annotate genes (Blüthgen *et al.*, 2005; Leonelli *et al.*, 2011; Lewin and Grieve, 2006). As an alternative to using GO to determine biological significance, it is also possible to use word lists. Thus, the words and phrases in the annotation terms were analysed. In both these cases, significance was determined using a hypergeometric test, comparing the terms in the cluster to all the genes used in the analysis (the background distribution of terms). The annotation terms were obtained from the BioConductor annotation databases, which use the TAIR (Lamesch *et al.*, 2012) and SGD (Saccharomyces Genome Database) annotations (Cherry *et al.*, 2012) for Arabidopsis and yeast, respectively. In addition, it is possible to identify over-represented motifs in the upstream promoter regions of the genes in the cluster, that may indicate a transcriptional regulator, as well as investigating the presence of genes that encode enzymes that function in the same, or similar, metabolic pathways. The details all these analysis function are described in Section 2.1.4-2.1.8.

The Biological Homogeneity Index (BHI) is a metric developed by Datta and Datta (2006), which provides an indication of the level of similarity in the annotations of the genes in the cluster. It is thought that a clustering algorithm which generates biologically meaningful clusters will group genes that have a similar biological function. The BHI value ranges between 0 and 1, where 0 indicates that there are no common annotations within genes in a cluster, and 1 indicates that all the genes in each cluster are annotated as having the same function. By using the GO terms associated with the genes in the cluster, it is possible to determine if there are a large number of genes annotated with the same GO terms, and thus, involved in the same biological function. As a result, this metric can be used to compare clusters using different parameter combinations, as well as comparing to other clustering algorithms. This algorithm is described in Section 2.1.8.

4.3. Results

4.3.1. Yeast data

It has been shown that the Arabidopsis genome is relatively poorly functionally annotated, with only approximately a quarter of the genes in the genome experimentally characterised (Quanbeck *et al.*, 2012). Thus, as a demonstration of the regression methodology described in Chapter 3, in addition to the clustering approach described above, an analysis was performed on time course data from a well annotated organism, namely *Saccharomyces cerevisiae* (baker's yeast). In addition, this data will be used to assess the performance of the cluster analysis, as well as a comparison to other clustering methods. The application of the clustering methods described above to the Arabidopsis data, shown in Chapter 3, is described in the following chapter.

As a test for the developed methods, a dataset from a yeast experiment was used

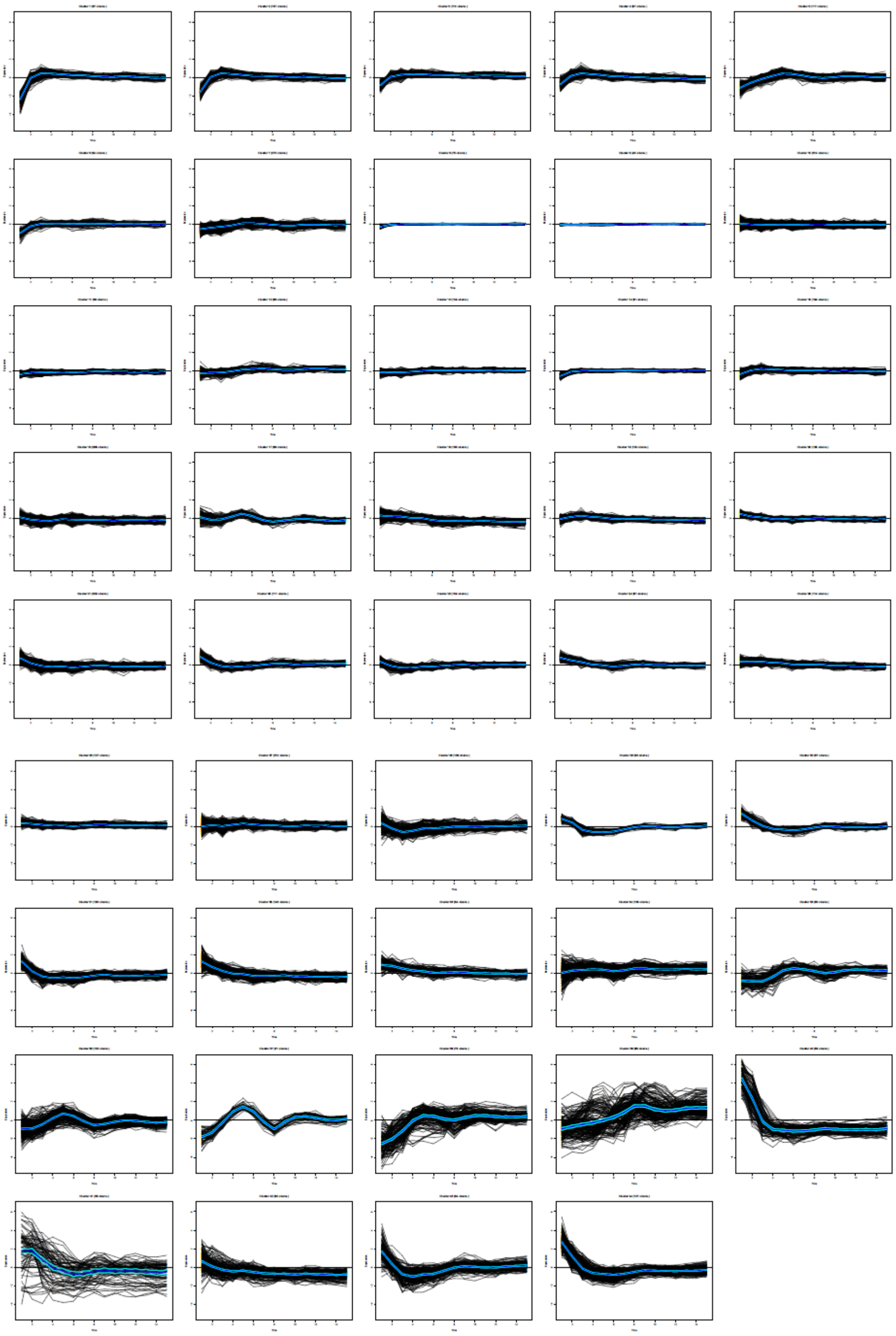


Figure 4.4: Clusters from SplineCluster on the wild-type yeast data. Each black line represents an expression profile, and the blue line represents the mean response of the cluster.

(Orlando *et al.*, 2008). A gene expression analysis was performed on budding yeast cells that were harvested at 16 minute intervals for 270 minutes (15 time points), which is equivalent to approximately two cell cycles in wild-type cells. The cells were initially synchronised using early G1 cells obtained through centrifugal elutriation. There were two biological replicates at each time point, and a total of 5661 genes in the dataset. The data was obtained from the GEO database, under the GSE8799 experiment identifier, and was normalised as described in the article. Annotations were obtained from the SGD (Saccharomyces Genome Database) (Cherry *et al.*, 2012). The dataset also contained expression data for cells where all S-phase and mitotic cyclins were mutated (*clb1,2,3,4,5,6*). However, this data was not used, as the mutation resulted in the deactivation of a large number of genes. The gene expression profiles were initially clustered using SplineCluster (Heard *et al.*, 2006) and this is shown in Figure 4.4. Despite being a time series across two cell cycles, a large number of expression profiles are not cyclic, and do not change much after the first few time points.

The regression analysis was also applied to the yeast data, and the distribution of models that fitted to the gene expression profiles is shown in Figure 4.5. The same thresholds determined in Chapter 3 were used to identify fits which fitted well, namely $R_a^2 > 0.6$, $R_{LoF}^2 > 0.6$, and $F - test < 0.05$. As can be seen, the exponential model fitted the majority of the shapes, with the second most common being the linear model. However, in this latter case, most of the fitted curves were poor fits and were identified as being flat and unchanging (mean and intercept close to zero). The Gaussian model was the second most abundant shape with good fits. The sigmoid models (logistic and the two forms of the Gompertz) did not occur as frequently as the others, but had good fits most of the time. The critical exponential, line+exponential and hyperbolic models did not occur very frequently. This roughly corresponds to the shapes that are seen when using SplineCluster (Figure 4.4).

4.3.2. Clustering results

In this section, some examples of the types of clusters that can be obtained from ShapeCluster are shown. Since ShapeCluster forms the clusters based on a specific shape and the values of specific parameters, a number of different cluster analyses were performed to obtain an overall representation of the gene expression profiles. As can be seen in Figure 4.5, the most abundant models that were found in the yeast data were the exponential, linear, Gaussian, and Gompertz2 (growth rate faster to the left of the midpoint) models, and these subsets of genes are examined in detail below. The cluster analysis was performed multiple times for each model, using different combinations of parameters. As described in Section 3.8, all fits with a $\Delta AIC < 2$ were retained as being indistinguishable from the best fit. This means that the clusters may contain genes that were also fitted by other models. In the cluster analysis, all genes that fitted the respective model were used. To assess the biological significance of the clusters, the `GOstats` package was used, which uses a hypergeometric test to determine GO term over-representation (Falcon and Gentleman, 2007, Section 2.1.4).

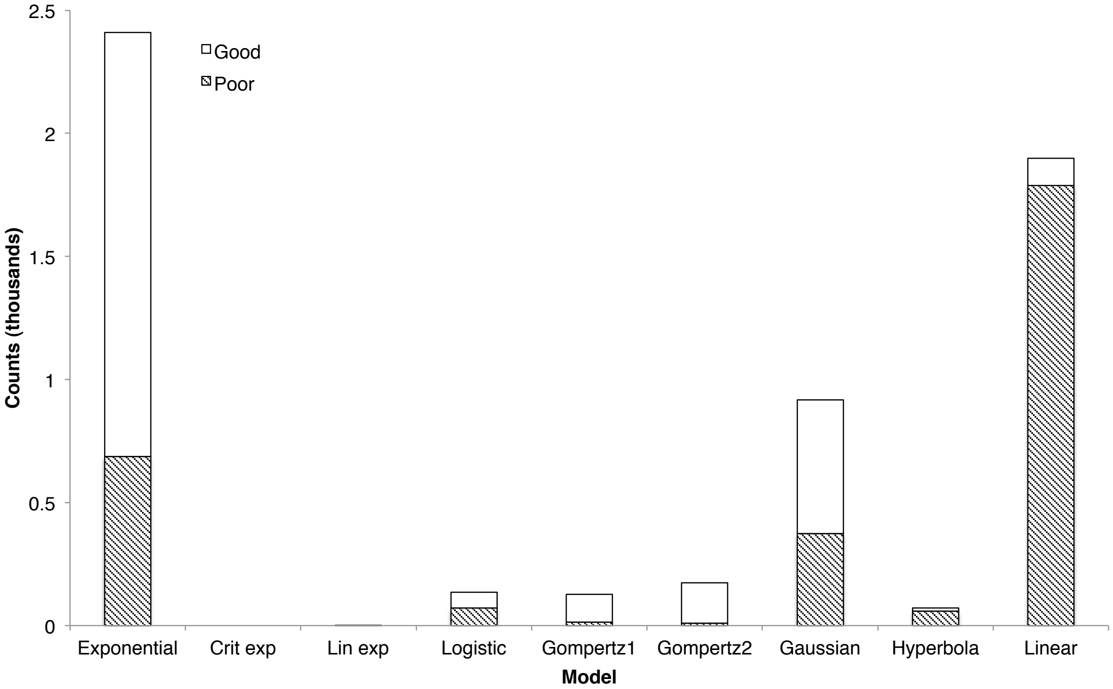


Figure 4.5: Distribution of the fitted models from the wildtype yeast data. Good fits were determined using the thresholds used were the same as the ones used for the Botrytis and senescence datasets, namely $R_a^2 > 0.6$, $R_{LoF}^2 > 0.6$, and $F{-}test < 0.05$.

Figure 4.6 shows an example of the clusters resulting from clustering the exponential shapes based on the rate of change parameter (r), and the significant GO terms are shown in Table 4.1. This parameter was selected as it would identify genes that have a similar rate of change in their gene expression, and thus may be involved in the same biological processes. Clusters with a high rate of change (r>0.06, clusters 14-21) were associated with RNA processing (transcription and translation), whereas the clusters with a slower rate of change were more associated with general metabolic processes. This would suggest that gene regulation is activated very quickly, whereas other cellular processes occur at a slower rate. This clustering also groups together genes that are both activated and repressed at the same rate. This type of cluster would not be formed with traditional clustering methods, which would cluster based on the entire expression profile. As a result, this provides an indication of genes that are being activated and repressed by some controlling set of genes.

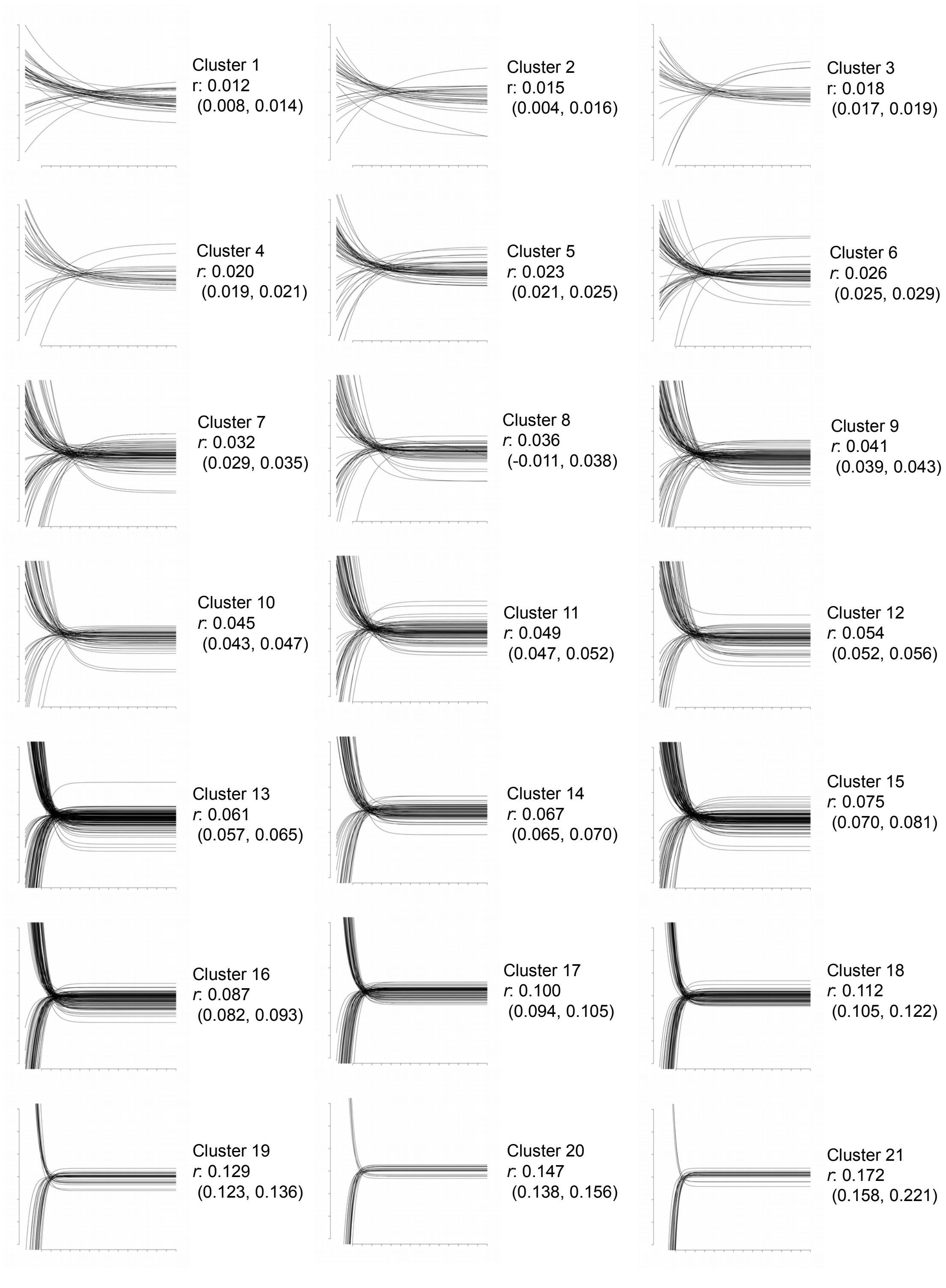


Figure 4.6: Clustering of the exponential curves based on the rate of change parameter (r). The greater the value of r, the greater the rate of change in gene expression. Thus, genes with a higher value of r, would be activated or repressed at a higher rate than those with a lower value of r. Shown are the mean parameter values, as well as the range of the parameter. The x-axis shows the time scale of the sampled data points, and the y-axis is the $\log_2$ gene expression level (from -3 to 3).

Table 4.1: Significant GO terms from the clustering of the exponential curves, based on the rate of change parameter (r).

Cluster	Mean parameter value	Cluster size	GO term	GO count	p-value
2	0.015	29	cofactor binding	6	1.30e-04
			ubiquitin-protein ligase activity	4	7.10e-04
			flavin adenine dinucleotide binding	3	8.42e-04
5	0.023	54	outer mitochondrial membrane organization	3	1.71e-04
			protein import	7	2.10e-04
			inositol phosphate dephosphorylation	3	2.68e-04
			protein import into mitochondrial matrix	3	8.40e-04
9	0.041	99	carboxylic acid metabolic process	20	1.04e-05
			cellular ketone metabolic process	20	2.98e-05
			sulfur compound metabolic process	8	2.69e-04
			glutamate catabolic process	3	9.02e-04
13	0.061	139	fatty acid beta-oxidation	4	2.12e-04
			lipid catabolic process	6	9.34e-04
14	0.067	75	cytosolic large ribosomal subunit	6	7.53e-04
15	0.075	186	cytosolic ribosome	14	6.47e-04
16	0.087	123	cytoplasmic translation	13	3.84e-05
17	0.100	85	cytoplasmic translation	14	8.10e-08
			structural constituent of ribosome	14	2.44e-06
			ribosomal subunit	14	4.21e-06
			rRNA export from nucleus	5	4.27e-05
18	0.112	18	ribosome	18	2.34e-06
			cytoplasmic translation	12	5.00e-06
			translation initiation factor activity	6	1.78e-05
			tRNA metabolic process	9	9.07e-04
19	0.129	35	cytosolic large ribosomal subunit	5	1.40e-04
			cellular macromolecule metabolic process	26	2.83e-04
20	0.147	21	cytoplasmic translation	5	2.42e-04
21	0.172	24	non-membrane-bounded organelle	17	1.91e-07
			ribosomal subunit	7	2.50e-05
			rRNA processing	7	4.50e-05
			RNA binding	9	5.29e-05

Furthermore, it is also possible to cluster on any of the other parameters from the exponential model, such as the asymptote (a) or the concavity (b) parameters. In the latter case, the b parameter determines if the exponential shape will be concave or convex, and a variety of these shapes can be seen in Figure 4.7. If $b<0$, this means that the shape is convex, indicating that the gene expression level is increasing to a plateau. Conversely, if $b>0$, this means that the shape is concave, indicating that the gene expression is decreasing to the asymptote. This parameter thus provides an indication of the dynamics of the response profile, and allows researchers to find genes with the same magnitude and type of response (up- or down-regulated). Table 4.2 shows the over-represented GO terms for these clusters, and it can be seen that the convex shapes ($b<0$, clusters 1-4), were primarily involved in primary metabolism, such as amino acid, nucleic acid and fatty acid metabolism, as well as genes that are involved in translation. For the concave shapes ($b>0$, clusters 5-8), the genes appeared to be predominately involved in secondary metabolism. This indicates that during cell division, the cells are diverting energy from the synthesis of secondary metabolites towards the activation of genes that will provide the necessary cellular components for the new daughter cells.

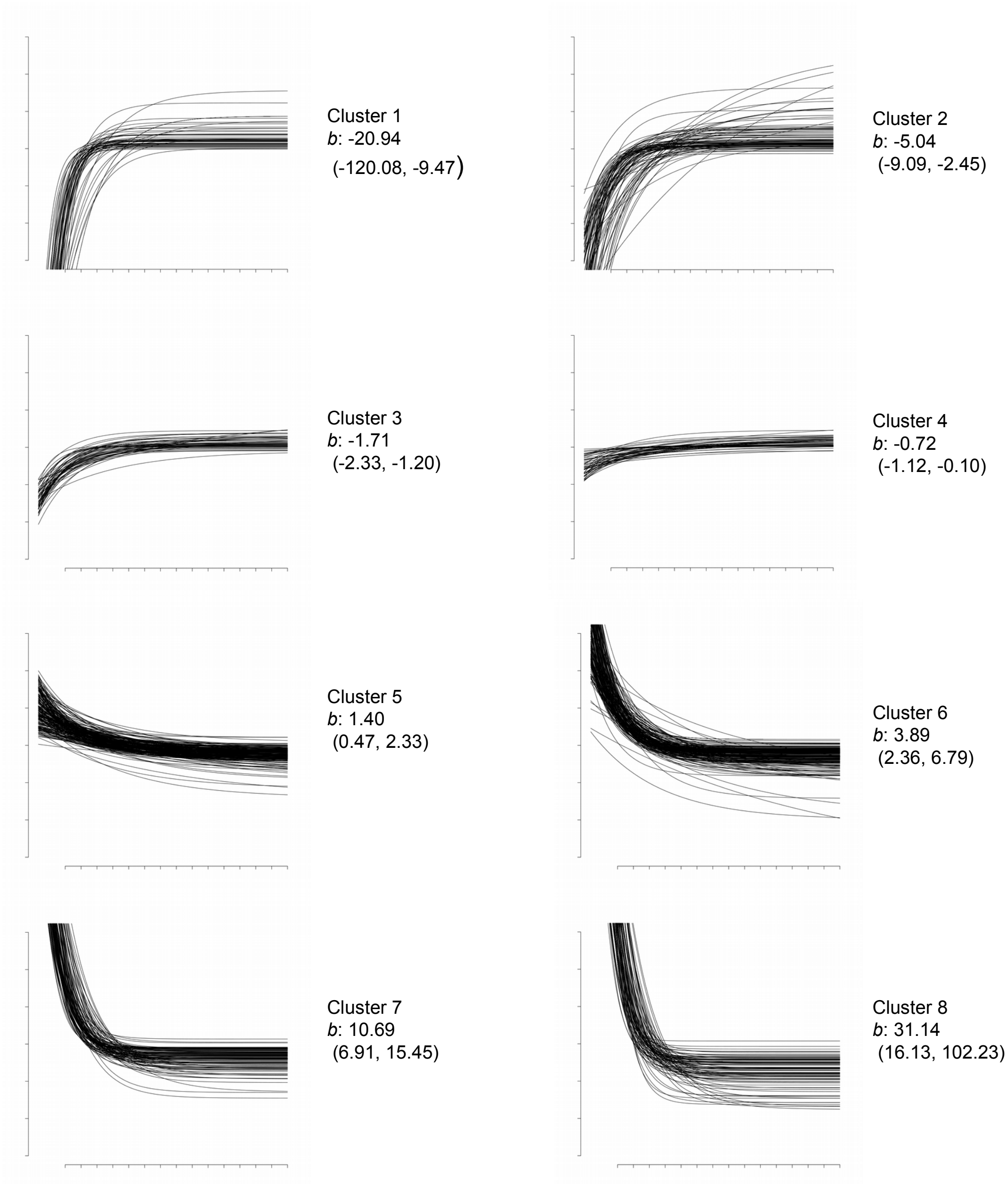


Figure 4.7: Clustering of the exponential curve shapes, based on the concavity (b) parameter. Clusters 1-4 exhibit a convex shape ($b<0$) and clusters 5-8 show a concave shape ($b>0$)

Table 4.2: Significant GO terms from the clustering of the exponential curves, based on the concavity parameter (b).

Cluster	Mean parameter value	Cluster size	GO term	GO count	p-value
1	-20.94	53	cellular nitrogen compound biosynthetic process	14	5.78e-07
			organic acid biosynthetic process	11	1.73e-06
			ribonucleoside monophosphate biosynthetic process	4	1.80e-04
			fatty acid elongase activity	2	2.58e-04
			transferase activity, transferring glycosyl groups	6	3.84e-04
2	-5.04	87	transmembrane transport	19	2.81e-05
			cellular amine metabolic process	15	4.17e-05
			protein transporter activity	6	3.78e-04
			cellular amino acid biosynthetic process	9	5.77e-04
			protein glycosylation	6	6.20e-04
3	-1.71	54	protein import into mitochondrial matrix	5	9.21e-07
			ribosomal subunit	10	3.62e-05
			structural constituent of ribosome	9	1.47e-04
			protein transporter activity	5	3.06e-04
4	-0.72	53	structural constituent of ribosome	11	2.93e-06
			macromolecular complex	34	1.13e-05
			cytoplasmic translation	9	1.43e-05
			large ribosomal subunit	8	2.23e-05
			translational elongation	5	1.23e-04
			regulation of translation	6	4.48e-04
5	1.4	222	protein autoubiquitination	3	2.30e-04
			cellular amide catabolic process	4	2.46e-04
			mRNA metabolic process	23	5.61e-04
			nuclear mRNA splicing, via spliceosome	12	7.46e-04
6	3.89	192	ligase activity	17	3.86e-04
			heterocycle biosynthetic process	15	6.47e-04
			small molecule metabolic process	52	6.73e-04
			cellular aromatic compound metabolic process	10	7.33e-04
7	10.69	113	cellular ketone metabolic process	20	2.07e-04
			carboxylic acid metabolic process	19	2.35e-04
			organic acid metabolic process	19	2.43e-04
			cellular lipid catabolic process	5	4.46e-04
			aerobic respiration	7	8.73e-04
8	31.14	94	glycogen catabolic process	3	8.62e-05
			energy reserve metabolic process	5	2.18e-04
			generation of precursor metabolites and energy	11	5.87e-04

In addition, it is possible to cluster on multiple parameters. As described in Equation (4.1), it is possible to produce a single value describing the distance between sets of genes based on two parameters (simultaneous parameter clustering - Section 4.2.2). For example, Figure 4.8 shows the exponential curves as before, except clustered simultaneously on the values of both the r and b parameters. This provides an indication of the rate of change, as well as the concavity of the curve. There were a few cases where the threshold for the cluster formation was not sufficiently stringent, and clusters with very different b parameters were merged (for example, cluster 7). This illustrates the problem with the simultaneous clustering, where combining parameters with very different scales can have a large effect on the structure of the resultant dendrogram. This issue is addressed in the meta-clustering section (Section 4.3.3). Investigating the GO terms for these clusters (Table 4.3), it can be seen that the genes are again primarily involved in various metabolic processes. However, in several cases, there were clusters with no significant GO terms, further suggesting that the simultaneous parameter clustering in this case may be suboptimal.

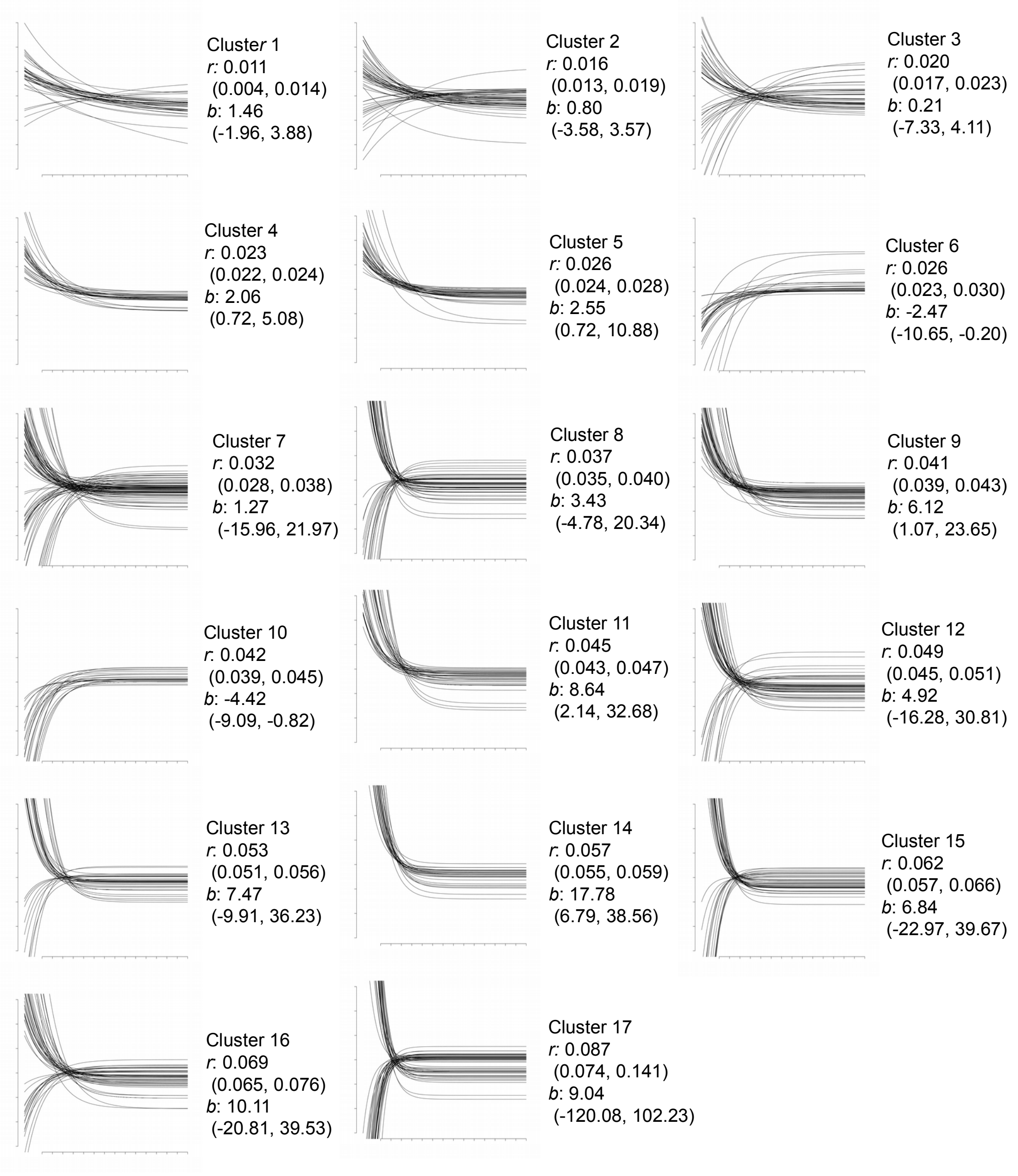


Figure 4.8: Clustering of the exponential curve shapes, based on both the rate of change (r) and concavity (b) parameters.

Table 4.3: Significant GO terms from the clustering of the exponential curves, based on both the rate of change (r) and concavity (b) parameters.

Cluster	Mean parameter value	Cluster size	GO term	GO count	p-value
2	r: 0.016, b: 0.80	53	coenzyme binding	6	4.05e-04
			ubiquitin-protein ligase activity	5	8.98e-04
3	r: 0.020, b: 0.21	48	protein import into peroxisome matrix	3	5.94e-04
			protein import into mitochondrial matrix	3	5.94e-04
4	r: 0.023, b: 2.06	30	organophosphate catabolic process	3	8.16e-05
			nucleotide diphosphatase activity	3	5.64e-04
			negative regulation of catabolic process	3	7.49e-04
6	r: 0.026, b: -2.47	29	protein transmembrane transporter activity	3	1.52e-04
			transporter activity	8	7.69e-04
9	r: 0.041, b: 6.12	55	oxidation-reduction process	14	3.84e-05
			small molecule metabolic process	22	7.89e-05
			cellular response to oxidative stress	6	1.28e-04
			carboxylic acid catabolic process	5	5.10e-04
			monocarboxylic acid metabolic process	6	8.81e-04
10	r: 0.042, b: -4.42	38	inositol phosphorylceramide metabolic process	2	1.72e-04
			protein exit from endoplasmic reticulum	2	1.72e-04
17	r: 0.087, b: 9.04	57	homocysteine S-methyltransferase activity	2	2.98e-04
			ribosome	11	4.26e-04

For the Gompertz2 models, the clustering was performed based on the gradient (*grad*) and 5% point of maximum change (*5per*) parameters (Section 3.7). The *grad* parameter describes the maximum rate of change, whereas the *5per* parameter provides an indication of the timing of the genes. That is, it identifies the time points at which genes are becoming activated or repressed. Therefore, clustering on the *5per* parameter would identify genes that are being up- or down-regulated at the same time, and thus possibly under the control of a particular transcription factor. Alternatively, clustering on the *grad* parameter would identify genes that have gene expression changing at the same rate, suggesting that these genes are being co-regulated. As before, it is possible to cluster on any of the parameters that describe the model. However, the *5per* and *grad* parameters would provide a better indication of the underlying biology, as they directly pertain to timings and rates of change in the gene expression. These parameters were clustered separately, as well as using the simultaneous parameter clustering.

Figure 4.9 shows the Gompertz2 curves clustered based on the *5per* parameter alone. By clustering on the *5per* parameter, genes that are being activated or repressed at the same time would be identified, and thus potentially being involved in the same biological process. Clusters 1-8 of this figure show curves that are increasing (genes being activated), while clusters 9-14 show curves with a decreasing response (genes being repressed). Table 4.4 shows the over-represented GO terms for these clusters. Once again, the increasing shapes were primarily involved in the biosynthesis of metabolites (cluster 8), protein transport (clusters 1, 5 and 6), as well as cell cycle related genes (cluster 7). This cell cycle related cluster had a mean *5per* value of 96 minutes, which was very similar to the cluster from the Gaussian clustering above (94 minutes, Figure 4.12, cluster 11), indicating that these genes were related and involved in the same process. The clusters with a decreasing response (clusters 9-14) again were primarily involved in secondary metabolism, such as oxidation-reduction reactions.

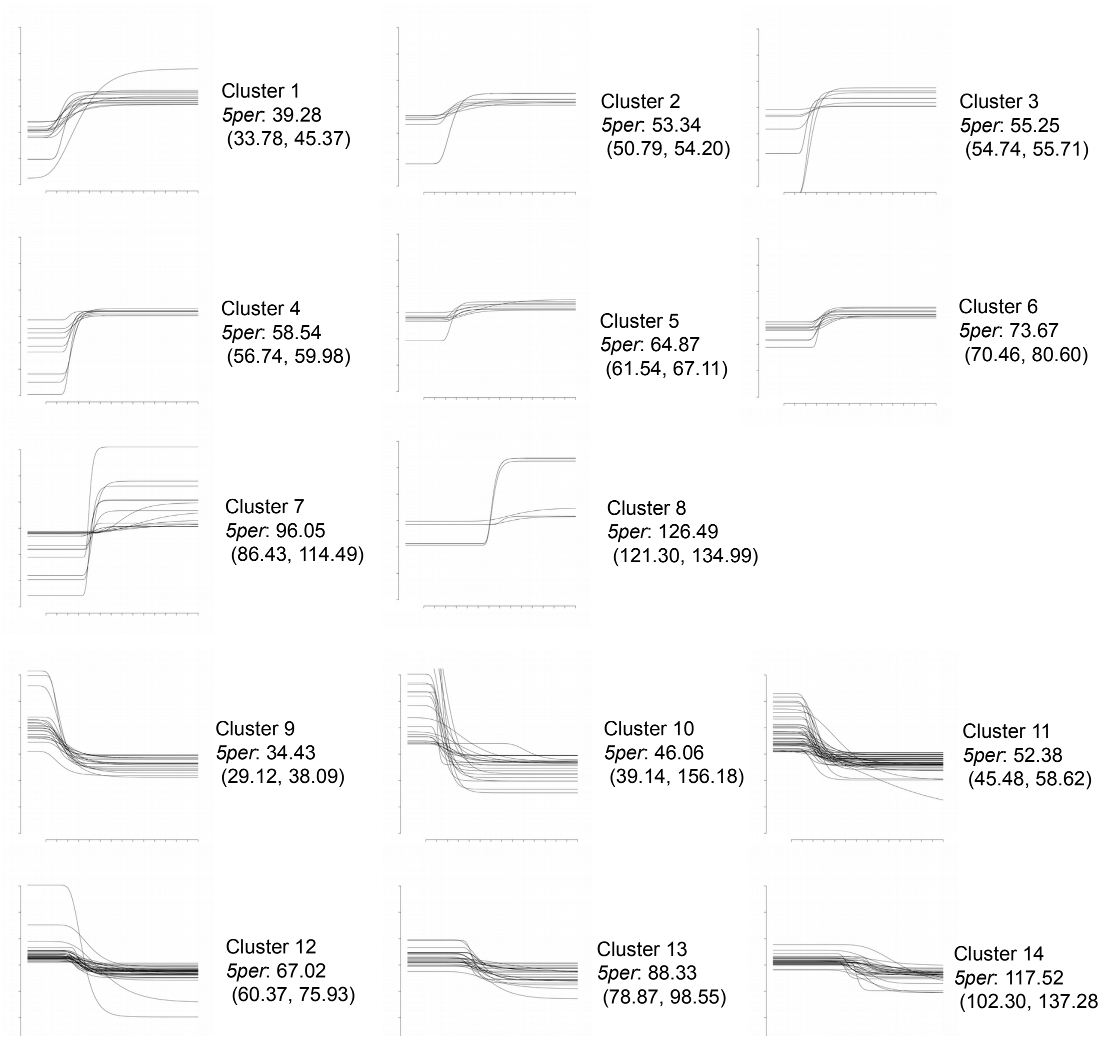


Figure 4.9: Gompertz2 shapes clustered on 5% of maximum point (*5per*). Clusters 1-8 show curves that are increasing, while clusters 9-14 possess a decreasing response.

Table 4.4: Significant GO terms from the clustering of the Gompertz2 curves, clustered on the 5% of maximum point (*5per*).

Cluster	Mean parameter value	Cluster size	GO term	GO count	p-value
1	39.28	14	negative regulation of transport	2	2.54e-04
5	64.87	8	intracellular transport	5	7.86e-04
6	73.67	11	protein N-linked glycosylation via asparagine	2	2.07e-05
			mitochondrial outer membrane translocase complex	2	9.60e-05
7	96.05	15	cellular bud	5	1.06e-04
			cytokinesis	4	1.75e-04
			cytoskeletal part	5	1.95e-04
			site of polarized growth	5	2.45e-04
			spindle pole body separation	2	4.28e-04
			cell cycle	7	7.83e-04
8	126.49	5	[acyl-carrier-protein] S-acetyltransferase activity	1	8.85e-04
9	34.43	20	antioxidant activity	3	1.45e-04
			glutathione peroxidase activity	2	2.48e-04
			pyridoxal phosphate binding	3	4.29e-04
			oxidoreductase activity	6	6.12e-04
10	46.06	23	oxidation-reduction process	8	1.75e-04
11	52.38	47	oxidoreductase activity	10	2.29e-04
14	117.52	32	mitochondrial respiratory chain	3	3.49e-04
			DNA damage checkpoint	3	4.41e-04

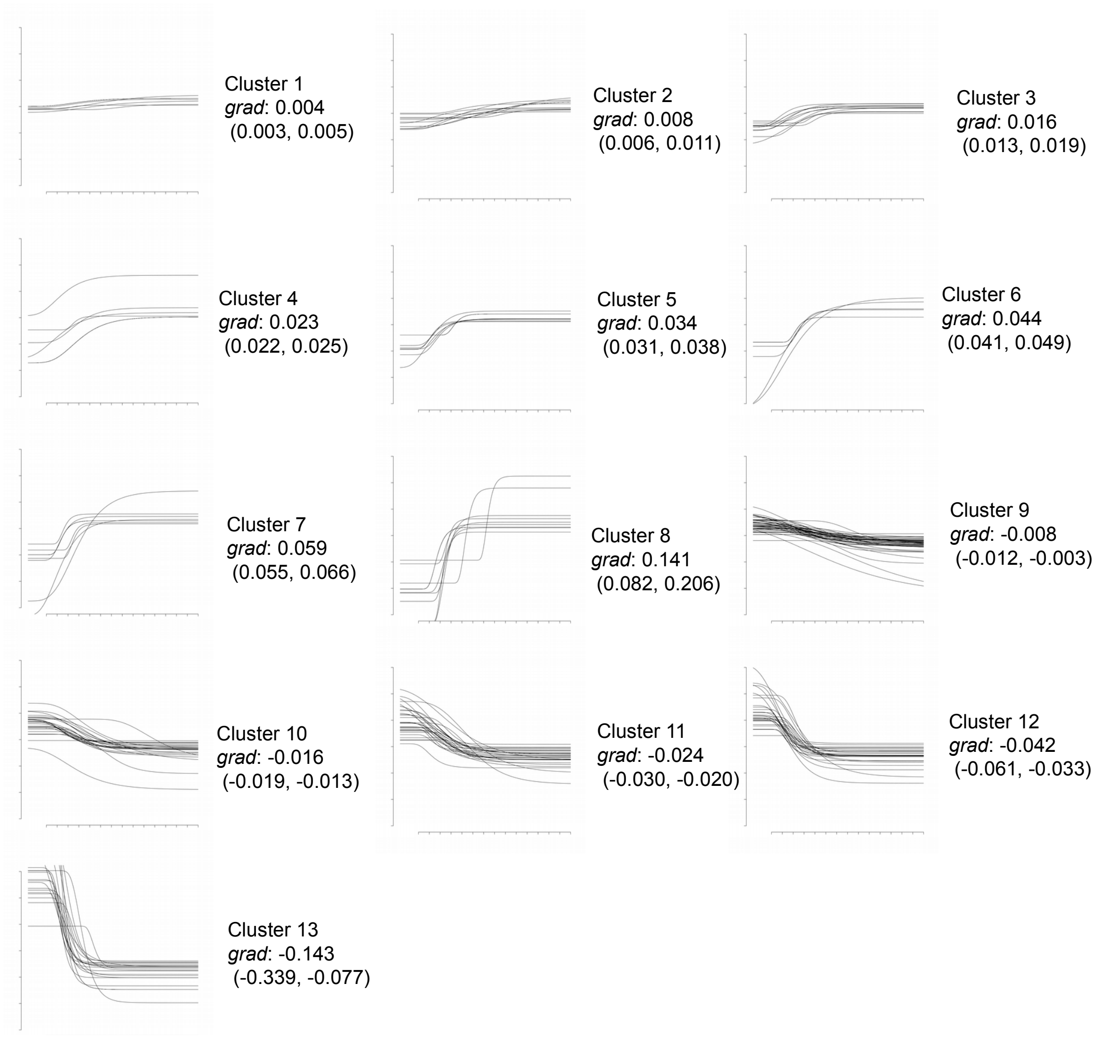


Figure 4.10: Gompertz2 shapes clustered on gradient (*grad*). Clusters 1-8 show curves that are increasing, while clusters 9-13 possess a decreasing response.

Similarly, the genes that fitted a Gompertz2 model were also clustered on the *grad* clustering (Figure 4.10). This parameter identifies genes that have similar rates of change in transcription, and are thus could possibly be co-regulated by a common regulator or transcription factor. Clusters 1-8 have an increasing response, whereas clusters 9-13 have a decreasing response. In Table 4.5, the over-represented GO terms are shown, and it can be seen that the increasing clusters are involved in macromolecule metabolism and transport, likely in preparation for cell division. In particular, genes that are being transcribed at a slower rate (smaller *grad* value), are involved in protein transport, whereas those with a larger *grad* value are involved in small molecule biosynthesis.

Table 4.5: Significant GO terms from the clustering of the Gompertz2 curves, clustered on gradient ($grad$).

Cluster	Mean parameter value	Cluster size	GO term	GO count	p-value
2	0.008	11	protein import into mitochondrial matrix	2	6.43e-04
3	0.016	11	protein channel activity	3	2.47e-06
			establishment of protein localization in mitochondrion	4	4.11e-06
			protein import into mitochondrial matrix	3	6.16e-06
			macromolecule transmembrane transporter activity	3	7.18e-06
			passive transmembrane transporter activity	3	1.96e-05
5	0.034	7	negative regulation of transport	2	5.90e-05
6	0.044	3	cellular bud site selection	2	8.81e-04
7	0.059	7	GDP-mannose biosynthetic process	2	1.32e-06
			small molecule biosynthetic process	5	1.00e-05
8	0.141	9	involved in negative regulation of transcription	2	1.23e-04
9	-0.008	47	mitochondrial respiratory chain	4	4.69e-05
			nucleotide-excision repair	5	1.15e-04
12	-0.042	27	oxidoreductase activity	8	8.19e-05
			glutathione transferase activity	2	4.56e-04
13	-0.143	20	oxidation-reduction process	9	5.77e-06
			oxidoreductase activity, acting on peroxide as acceptor	3	2.50e-05
			reduction of molecular oxygen	3	1.31e-04

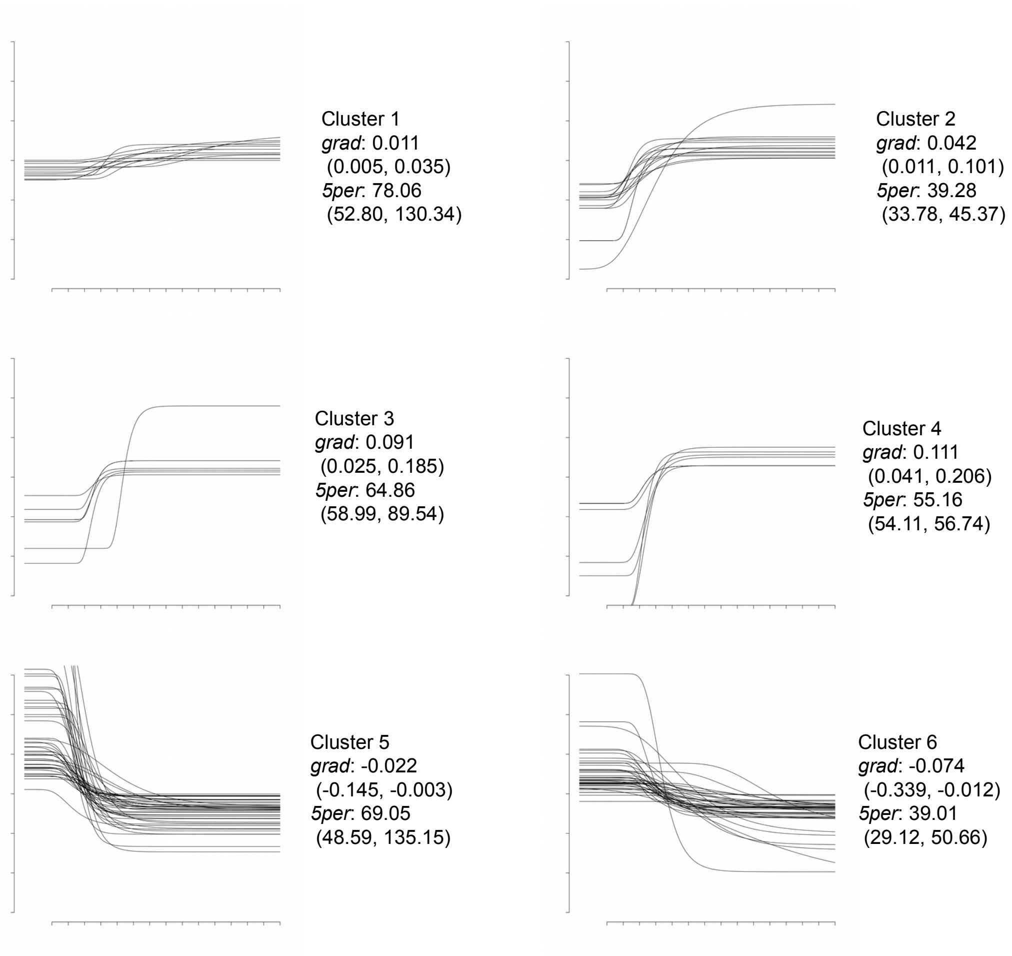


Figure 4.11: Gompertz2 shapes clustered on gradient (*grad*) and 5% of maximum point (*5per*). Clusters 1-4 show an increasing response, while clusters 5-6 show a decreasing response. Over-represented GO terms for these clusters are shown in Table 4.6.

As previously performed with the exponential model, it is possible to cluster using multiple parameters by joining the parameters using the simultaneous parameter clustering. Figure 4.11 shows the clusters from the Gompertz2 curves, clustered on both the *grad* and *5per* parameters, identifying genes that are being activated or repressed at the same time points, as well as having the same rate of change in gene expression. The GO terms from these clusters are shown in Table 4.6, and although they are less definitive than the individual cluster analyses, it can still be seen that the increasing curves are involved in protein metabolism and transport (clusters 1-4), whereas the clusters with a decreasing response (clusters 5-6) are involved in oxidative reactions.

Table 4.6: Over-represented GO terms from the clustering of the Gompertz2 curves, clustered on both the 5% of maximum point (*5per*) and gradient (*grad*) parameters.

Cluster	Mean parameter value	Cluster size	GO term	GO count	p-value
1	*grad*: 0.011, *5per*: 78.06	12	protein import into mitochondrial matrix	2	7.70e-04
2	*grad*: 0.042, *5per*: 39.28	14	small molecule biosynthetic process	5	6.90e-04
			negative regulation of transport	2	2.54e-04
3	*grad*: 0.091, *5per*: 64.86	6	involved in negative regulation of transcription	2	5.15e-05
5	*grad*: -0.022, *5per*: 69.05	41	oxidoreductase activity	13	1.96e-07
			mitochondrial respiratory chain	3	7.31e-04
			aryl-alcohol dehydrogenase (NAD+) activity	2	7.58e-04
6	*grad*: -0.074, *5per*: 39.01	45	oxidation-reduction process	15	4.79e-07
			oxidoreductase activity	13	6.60e-07
			peroxidase activity	3	2.98e-04

With the Gaussian shapes, the clusters were formed based on the mean (m) and standard deviation (s) parameters (Figure 4.12), using the simultaneous parameter clustering. The m parameter provides an indication of the time when an up-regulated gene becomes down-regulated, or *vice versa*, whereas the s parameter provides an indication of the duration of the gene expression response. Thus the *m&s* clusters would identify genes that have a maximum (or minimum) at the same time, as well as having a similar spread in the data points.

It is also possible to perform the clustering on these parameters separately (provided in Appendix D), as well as on the other parameters, namely the asymptote value (a), or the magnitude of the peak (b). The over-represented GO terms for these clusters are shown in Table 4.7, as with before, the clusters with a decreasing shape (clusters 1-9 in Figure 4.12) are primarily involved in metabolic processes, such as gluconeogenesis and nucleosome assembly. In contrast, the clusters with an increasing shape (clusters 10-13) are largely involved in ATP generation (cluster 10), and cell division (cluster 11). Investigating the parameter values of these clusters shows that the cluster involved in ATP generation has a mean m parameter of 62±4 minutes, whereas the cluster involved in the cell division is later at 94±9 minutes. This indicates that there is an increase in the production of energy in preparation for the processes necessary for cell division.

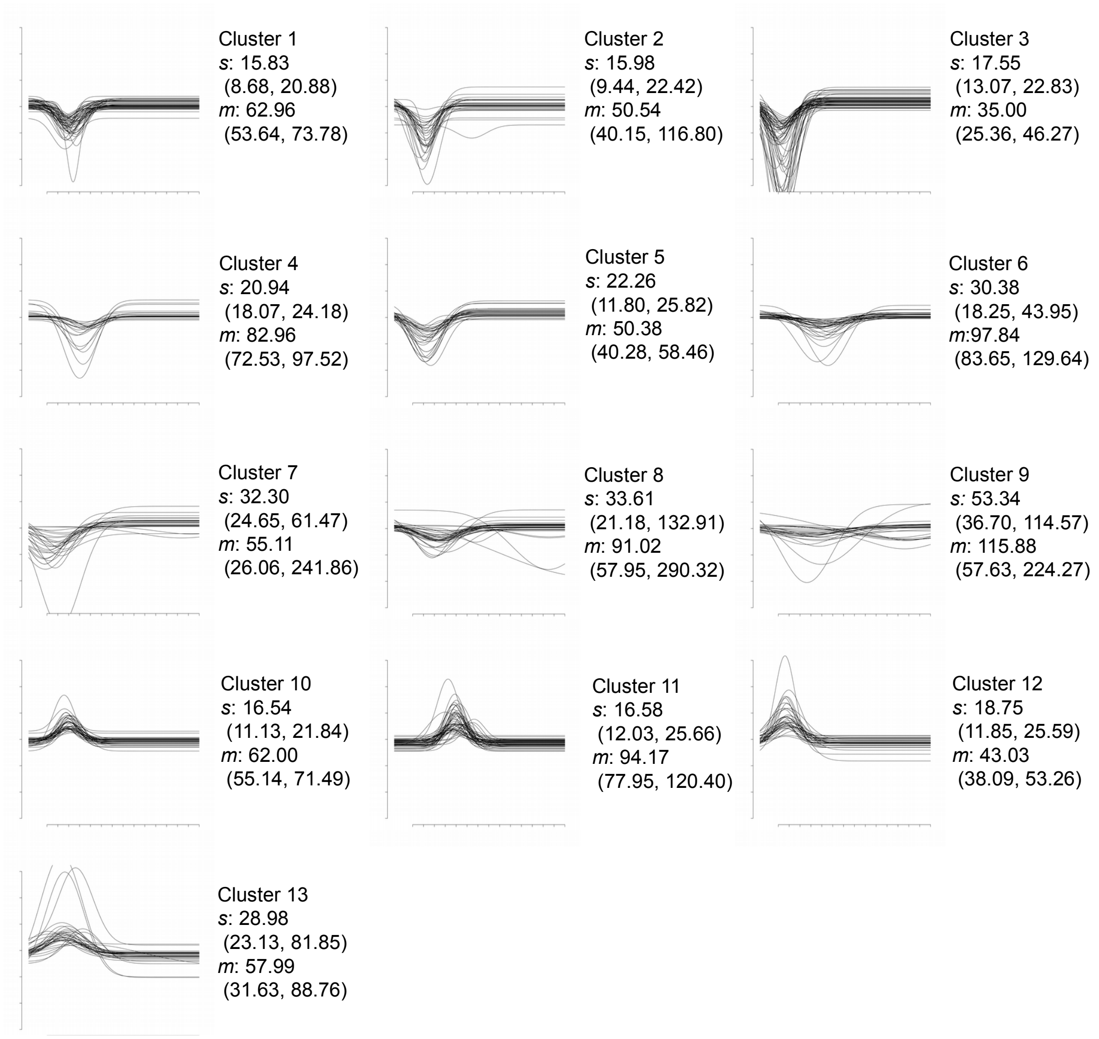


Figure 4.12: Gaussian shapes clustered on mean (m) and standard deviation (s). Clusters 1-9 are Gaussian shapes that are decreasing, whereas clusters 10-13 show Gaussian shapes that are increasing.

Table 4.7: Significant GO terms from the clustering of the Gaussian curves, based on the clustered on mean (m) and standard deviation (s) parameters.

Cluster	Mean parameter value	Cluster size	GO term	GO count	p-value
1	s: 15.83, m: 62.96	49	AMP-activated protein kinase activity	3	6.07e-06
			endoplasmic reticulum part	10	9.12e-04
3	s: 17.55, m: 35.00	61	cellular bud neck	8	1.49e-04
			cellular nitrogen compound biosynthetic process	11	3.48e-04
4	s: 20.94, m: 82.96	20	transmembrane transport	6	6.75e-04
5	s: 22.26, m:	29	nucleosome assembly	3	2.60e-04
6	50.38 s: 30.38, m: 97.84	33	small molecule biosynthetic process	9	5.47e-05
			regulation of intracellular pH	3	4.84e-04
			glycolysis	3	6.63e-04
			gluconeogenesis	3	7.31e-04
7	s: 32.3, m: 55.11	26	G-protein alpha-subunit binding	2	2.04e-05
8	s: 33.61, m: 91.02	35	mitochondrial translation	6	6.27e-05
			mitochondrial ribosome	5	1.32e-04
9	s: 53.34, m: 115.88	27	L-phenylalanine metabolic process	2	3.26e-04
10	s: 16.54, m: 62.00	39	ATP synthesis coupled proton transport	5	6.08e-08
			proton-transporting ATPase activity, rotational mechanism	5	1.85e-06
			nucleoside triphosphate biosynthetic process	5	2.74e-06
			mitochondrial envelope	13	9.91e-06
			ATPase activity, coupled to transmembrane movement of ions	5	1.92e-05
			RNA polymerase II core promoter binding transcription factor activity	5	4.21e-05
			ATP metabolic process	5	1.46e-04
			CCAAT-binding factor complex	2	4.03e-04
11	s: 16.58, m: 94.17	51	chromosome segregation	12	5.58e-09
			nuclear replication fork	6	2.18e-06
			DNA-dependent DNA replication	8	4.76e-06
			mitotic cell cycle	13	6.21e-06
			single-stranded DNA binding	5	7.67e-05
			M phase	11	3.79e-04
			microtubule-based process	6	4.86e-04
12	s: 18.75, m: 43.03	33	telomeric DNA binding	3	3.18e-04
			negative regulation of telomere maintenance	2	4.13e-04
			nuclear chromosome, telomeric region	3	6.07e-04

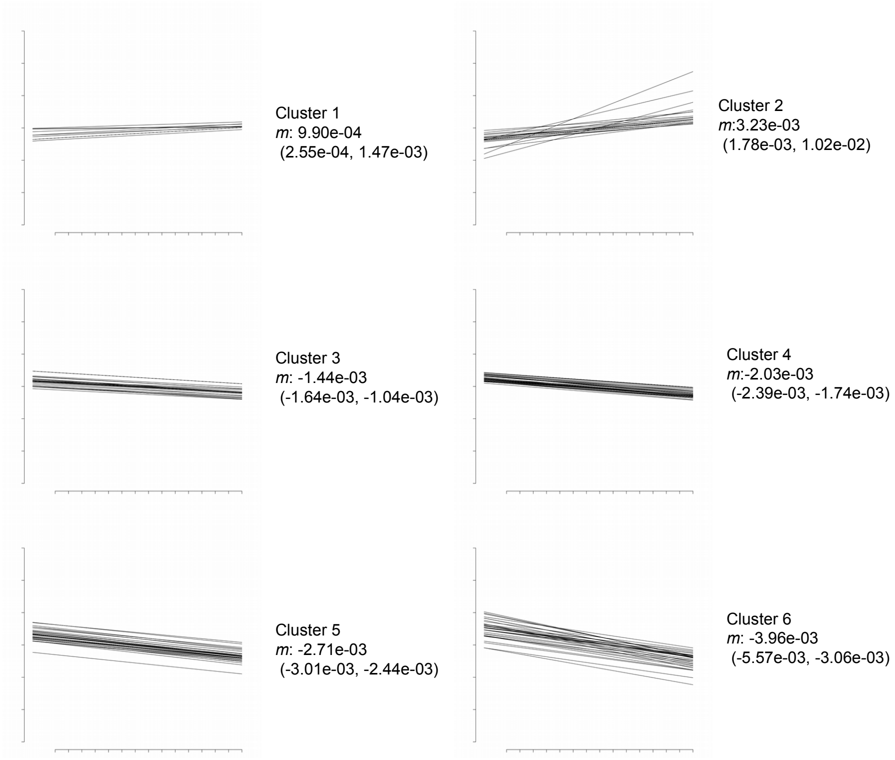


Figure 4.13: Linear shapes clustered on the gradient (m). Clusters 1-2 are clusters showing an increasing response, and clusters 3-6 show clusters with a decreasing response.

Finally, the linear shapes were analysed by clustering on the gradient parameter (m) (4.13). This clustering would find genes that fitted the linear model, and are transcribed at the same rate. As discussed above, this could identify genes that are being co-regulated. Clusters 1-2 in this figure show genes that are increasing, and clusters clusters 3-6 are decreasing. The GO terms for these clusters are shown in Table 4.8, and do not show much in terms of distinct function related to the parameter value.

Table 4.8: Significant GO terms from the clustering of the linear functions, based on the gradient (m).

Cluster	Mean parameter value	Cluster size	GO term	GO count	p-value
1	0.001	9	COPI vesicle coat	2	6.29e-05
			Golgi-associated vesicle membrane	2	4.23e-04
2	0.003	20	retrotransposon nucleocapsid	2	6.43e-04
3	-0.001	25	Rab guanyl-nucleotide exchange factor activity	2	6.65e-04
6	-0.004	35	riboflavin biosynthetic process	2	7.69e-04

4.3.3. Meta-clustering examples

In the previous section, multiple parameters were analysed in a simultaneous manner, where a single distance matrix was calculated by merging multiple parameter values into one. In this section, a variation on the clustering is performed, namely meta-clustering (described in Section 4.2.3). As in Section 4.3.2, the clustering was performed for the most abundant shapes, and using the same parameters. That is r and b for exponential, m and s for the Gaussian, and *5per* and *grad* for the Gompertz2 shapes. However, for brevity, only the results from the Gaussian meta-clustering is shown here, and the other results are contained in the Appendix D.

As described previously, two types of meta-clustering were developed. First, the sequential meta-clustering, where clusters were first formed by clustering on one parameter, followed by re-clustering these clusters using the second parameter. The second type of meta-clustering is the cross-clustering, and is discussed below. Figure 4.14 shows an example of the sequential meta-clustering for the Gaussian model, using the decreasing shapes. In this case, the shapes were first clustered using the m parameter (time of maximum expression), and then these clusters were in turn clustered using the s parameter (spread around the time of maximum response, relating to the duration of the response). As would be expected, it can be seen that the subclusters all have similar m parameters, to the original cluster. However, upon reclustering on the s parameter, the range of the s parameters in the subclusters is significantly narrower.

The over-represented GO terms for the Gaussian model are shown in Tables 4.9 and 4.10 for the increasing and decreasing shapes, respectively. Not all the subclusters had significant over-representation of GO terms, so it could indicate that these subsets of genes were not involved in any particular process, or it could mean that the clusters were too small and there was insufficient information to calculate meaningful statistics for the over-representation tests. In the increasing shapes, it can be seen that the over-represented terms progress from RNA processing, to DNA replication, to chromosome segregation, and finally cell wall organisation. This process roughly follows the processes involved during cell division. In contrast, the over-represented GO terms for the decreasing Gaussian shapes primarily pertained to cell cycle and translation. However, the values of the m parameters in these clusters are generally lower than the m values from the increasing shapes. This could mean that the genes related to the cell cycle and cell division are being down-regulated in preparation for the actual division process. Alternatively, since it is thought that the time series corresponds to two cell division cycles (Orlando *et al.*, 2008), it may be that the increasing and decreasing shapes are picking up different cycles. While the GO terms found in this clustering are similar to those identified using the simultaneous parameter clustering (Table 4.7), this clustering has provided a much finer means of identifying functions for the GO terms.

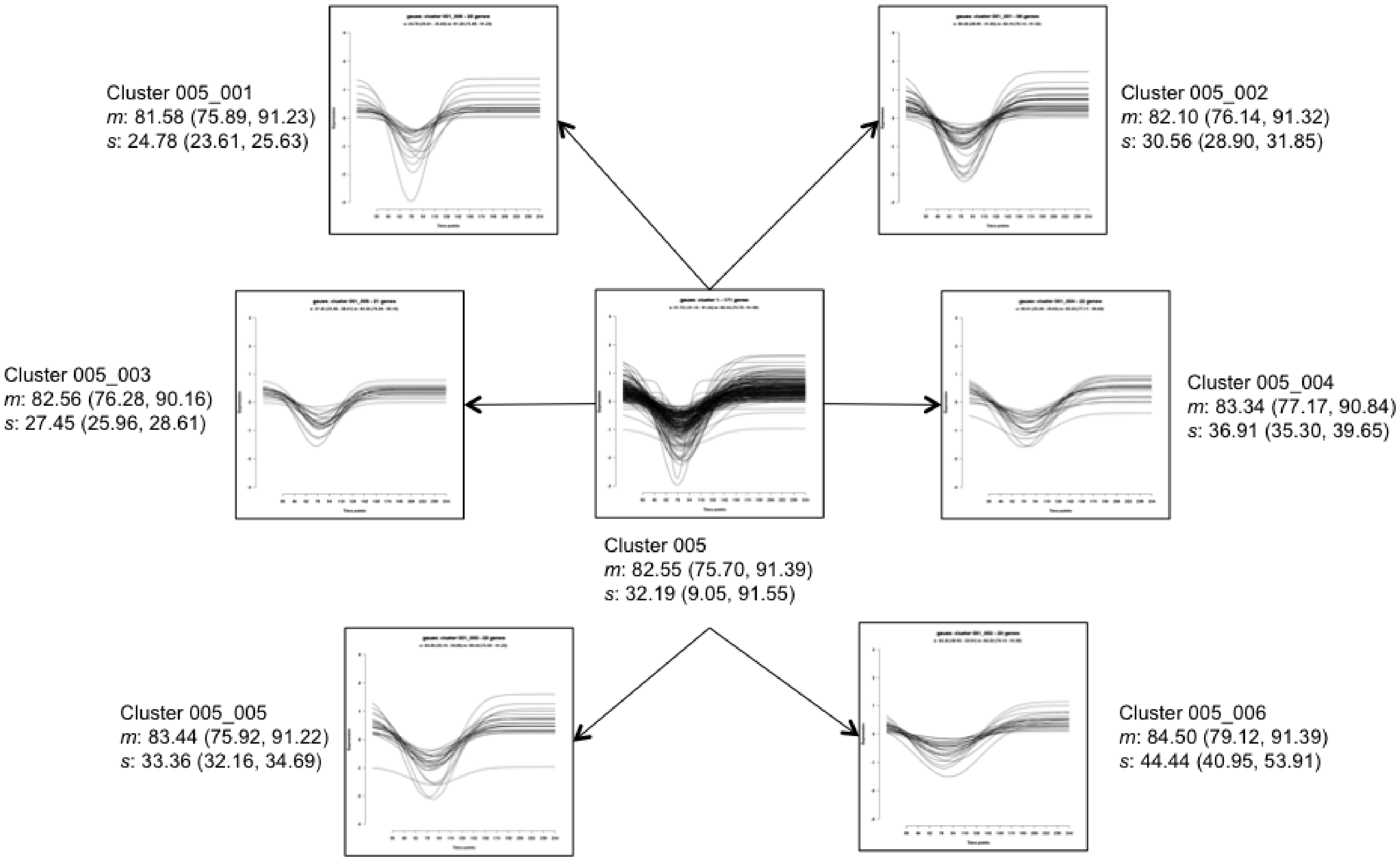


Figure 4.14: One of the clusters from the sequential meta-clustering for the Gaussian decreasing shapes. The cluster in the middle was formed from clustering all the shapes on the m parameter. The clusters surrounding it were formed by clustering this subset of genes on the s parameter. The over-represented GO terms for these clusters are shown in Table 4.9.

Table 4.9: Table of the over-represented GO terms for the increasing Gaussian shapes, clustered using the sequential meta-clustering, on the m and s parameters.

Cluster	Mean parameter value	Cluster size	GO term	GO count	p-value
001_001	m: 57.73, s: 16.44	14	purine nucleobase biosynthetic process	2	3.10e-04
002_001	m: 64.80, s: 16.97	16	ncRNA 3'-end processing	4	5.45e-06
			nuclear exosome (RNase complex)	3	8.32e-06
			rRNA metabolic process	5	5.23e-04
			mRNA metabolic process	5	7.90e-04
003_001	m: 74.15, s: 58.60	21	nucleolar part	5	4.74e-06
			ribosomal small subunit biogenesis	4	9.05e-04
004_001	m: 83.97, s: 65.31	15	nucleolar part	4	2.82e-05
			intracellular non-membrane-bounded organelle	10	1.78e-04
			transcription from RNA polymerase I promoter	3	5.73e-04
004_002	m: 84.98, s: 45.99	17	nucleolus	7	8.06e-06
			ribosomal large subunit biogenesis	4	7.72e-05
005_001	m: 91.73, s: 18.51	18	DNA replication	7	2.64e-07
			response to DNA damage stimulus	7	3.12e-05
			response to stress	9	1.59e-04
			cell cycle	8	4.78e-04
			nucleic acid metabolic process	12	5.70e-04
			sequence-specific DNA binding	5	8.08e-04
005_002	m: 92.26, s: 23.87	16	DNA metabolic process	7	1.71e-04
			DNA-dependent DNA replication	4	2.00e-04
			nuclear replication fork	3	2.30e-04
			chromosome segregation	4	6.95e-04
005_003	m: 92.26, s: 60.44	19	ncRNA processing	7	5.28e-05
006_001	m: 98.14, s: 22.51	16	recombinational repair	3	5.79e-04
006_002	m: 98.60, s: 65.29	12	ribonucleoprotein complex	7	1.06e-04
			ribosome biogenesis	5	4.33e-04
006_003	m: 98.99, s: 38.28	15	maintenance of DNA repeat elements	2	8.75e-04
007_001	m: 165.24, s: 66.08	11	attachment of spindle microtubules to chromosome	2	3.10e-04

Table 4.9 (cont.)

Cluster	Mean parameter value	Cluster size	GO term	GO count	p-value
009_001	m: 204.53, s: 53.93	11	mitochondrial translation	5	1.40e-06
			ribonucleoprotein complex	7	4.88e-05
010_001	m: 243.65, s: 36.52	17	cellular cell wall organization or biogenesis	5	5.61e-04

Table 4.10: Table of the over-represented GO terms for the decreasing Gaussian shapes, clustered using the sequential meta-clustering, on the m and s parameters.

Cluster	Mean parameter value	Cluster size	GO term	GO count	p-value
001_001	m: 33.93, s: 29.33	26	protein glycosylation	4	2.35e-04
001_003	m: 35.73, s: 15.87	15	cell cycle	7	7.67e-04
001_004	m: 35.80, s: 16.85	22	cytokinesis	4	6.85e-04
002_001	m: 43.14, s: 17.21	22	spindle pole body organization	3	2.37e-04
			M phase	7	3.83e-04
			microtubule cytoskeleton organization	4	5.30e-04
			DNA-dependent DNA replication	4	7.34e-04
			mitosis	5	8.14e-04
002_002	m: 44.34, s: 22.96	16	spindle microtubule	3	4.13e-05
			chromosome segregation	4	6.95e-04
002_004	m: 48.45, s: 16.35	19	mitotic anaphase B	2	1.07e-04
			anaphase	2	2.97e-04
			spindle pole body separation	2	8.18e-04
			nuclear outer membrane-endoplasmic reticulum membrane network	6	8.44e-04
002_005	m: 50.17, s: 28.07	17	membrane coat	3	2.25e-04
			cytoplasmic membrane-bounded vesicle	4	4.76e-04
003_001	m: 59.33, s: 32.32	18	organelle membrane	12	1.61e-05
			intracellular protein transport	6	4.85e-04

Table 4.10 (cont.)

Cluster	Mean parameter value	Cluster size	GO term	GO count	p-value
003_002	m: 60.54, s: 22.08	18	purine ribonucleoside triphosphate binding	8	7.47e-04
			ATPase activity	5	9.69e-04
004_001	m: 69.98, s: 21.86	25	Golgi apparatus	7	4.60e-05
			cell communication	8	2.03e-04
004_003	m: 70.90, s: 34.27	17	mitochondrion organization	6	2.37e-04
			cellular component organization	12	9.95e-04
005_001	m: 81.58, s: 24.78	20	biological regulation	13	7.39e-04
			membrane organization	5	7.56e-04
005_002	m: 82.10, s: 30.56	36	proteasome storage granule	5	4.71e-07
			regulation of transcription initiation from RNA polymerase II promoter	3	1.26e-04
			cellular catabolic process	13	4.27e-04
005_003	m: 82.56, s: 27.45	21	proteasome complex	4	2.91e-05
			modification-dependent protein catabolic process	5	6.33e-04
			cellular catabolic process	9	7.92e-04
005_004	m: 83.34, s: 36.91	22	proteolysis	8	9.14e-06
			mitochondrial intermembrane space	3	6.55e-04
005_005	m: 83.44, s: 33.36	23	proteolysis involved in cellular protein catabolic process	6	1.46e-04
			modification-dependent macromolecule catabolic process	6	1.71e-04
			ubiquitin-dependent protein catabolic process	5	9.66e-04
006_001	m: 233.09, s: 41.95	12	rRNA binding	3	9.24e-05
			sequence-specific transcription regulatory region DNA binding RNA polymerase II transcription factor recruiting transcription factor activity	2	1.48e-04
			cytosolic large ribosomal subunit	3	6.12e-04
006_002	m: 235.18, s: 45.35	18	cytosolic small ribosomal subunit	3	7.90e-04
006_003	m: 239.43, s: 57.45	29	cytosolic large ribosomal subunit	4	7.79e-04

As mentioned previously, the second form of meta-clustering is the cross-clustering. In this method, the parameters are clustered independently, and the intersections between the clusters are identified. Figure 4.15 shows an example of this clustering, again using the m (time of maximum expression) and s (duration of response) parameters of the increasing Gaussian shapes. The clusters in the margins are the genes that were clustered with the m (side) and s (top) parameters independently. The intersection of the genes in the different pair combinations are identified, and the GO term analysis performed on these clusters. These GO terms are shown in Tables 4.11 and 4.12, for the increasing and decreasing shapes, respectively. The GO terms identified followed a similar pattern as the sequential meta-clustering above, where the increasing GO terms showed RNA processing, DNA replication, cell cycle, and chromosome segregation. Similarly, with the decreasing shapes, the early repressed genes are involved in cell division and proteolysis. This suggests that the two types of meta-clustering do not produce significantly different clusters. Nonetheless, they still provide a different means of interrogating the data to identify new patterns.

Table 4.11: Table of the over-represented GO terms for the increasing Gaussian shapes, clustered using the cross meta-clustering, on the m and s parameters.

Cluster	Mean parameter value	Cluster size	GO term	GO count	p-value
003_007	m: 63.08, s: 87.81	3	tRNA modification	2	4.64e-04
			nuclear replication fork	3	6.13e-04
003_010	m: 64.11, s: 22.84	3	ATPase regulator activity	3	1.12e-04
003_002	m: 64.97, s: 16.67	15	ncRNA 3'-end processing	4	4.11e-06
			mRNA metabolic process	5	5.66e-04
004_008	m: 74.22, s: 30.93	4	microtubule associated complex	4	3.79e-04
004_003	m: 74.51, s: 49.44	8	astral microtubule	4	1.43e-05
004_004	m: 74.51, s: 49.44	4	RNA biosynthetic process	4	2.91e-04
			transcription from RNA polymerase I promoter	2	7.47e-04
004_001	m: 74.61, s: 56.15	5	microtubule motor activity	3	7.20e-04
005_009	m: 83.59, s: 26.56	3	cell cycle checkpoint	2	6.49e-04
005_001	m: 83.73, s: 54.50	8	transcription initiation from RNA polymerase I promoter	2	1.05e-05
			RNA polymerase I transcription factor binding	2	1.05e-05

Table 4.11 (cont.)

Cluster	Mean parameter value	Cluster size	GO term	GO count	p-value
005_006	m: 85.71, s: 42.20	3	ribosomal large subunit export from nucleus	2	5.18e-05
006_002	m: 91.73, s: 18.51	18	DNA replication	7	2.64e-07
			sequence-specific DNA binding	5	8.08e-04
006_009	m: 91.79, s: 25.69	7	recombinational repair	3	3.87e-05
006_001	m: 92.51, s: 54.24	5	cellular protein complex disassembly	5	4.97e-04
			structural constituent of cytoskeleton	5	4.97e-04
006_010	m: 92.63, s: 22.46	9	spindle	6	9.12e-04
007_005	m: 99.12, s: 37.88	4	negative regulation of transcription, DNA-dependent	3	3.05e-04
008_010	m: 106.84, s: 22.25	5	negative regulation of signal transduction	2	2.70e-04
			regulation of cell communication	2	5.32e-04
008_009	m: 109.64, s: 27.65	5	microtubule-based process	3	2.35e-04
009_006	m: 125.68, s: 43.65	7	establishment of spindle localization	2	1.19e-04
009_008	m: 127.46, s: 32.76	4	microtubule nucleation	2	7.87e-05
			structural constituent of cytoskeleton	2	2.62e-04
010_004	m: 159.55, s: 48.21	4	cellular response to heat	2	1.97e-04
011_001	m: 204.53, s: 53.93	11	mitochondrial translation	5	1.40e-06
013_006	m: 244.10, s: 42.96	8	spindle microtubule	3	9.67e-05

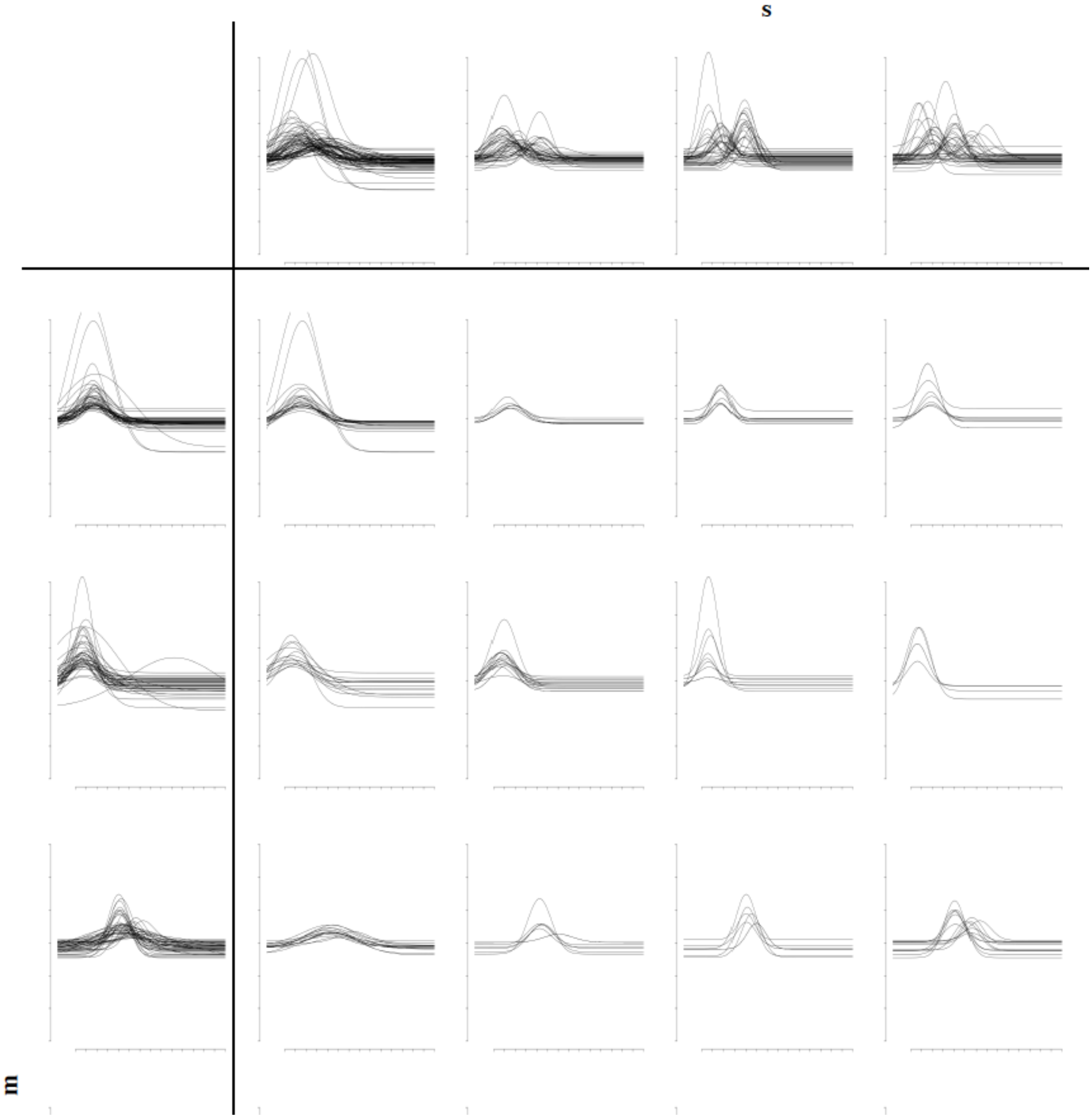


Figure 4.15: A portion of the clusters formed from the cross meta-clustering for the increasing Gaussian shapes. The clusters in the margins are the genes clustered with the m (side) and s (top) parameters. For each pair of these clusters, the common genes are identified.

Table 4.12: Table of the over-represented GO terms for the decreasing Gaussian shapes, clustered using the cross meta-clustering, on the m and s parameters.

Cluster	Mean parameter value	Cluster size	GO term	GO count	p-value
001_007	m: 34.01, s: 30.47	22	transferase activity, transferring glycosyl groups	4	5.93e-04
001_009	m: 34.84, s: 22.33	8	external encapsulating structure	4	6.38e-06
001_006	m: 36.17, s: 16.04	37	reproduction	10	5.62e-04
			cell cycle process	11	6.99e-04
002_006	m: 43.14, s: 17.21	22	spindle pole body organization	3	2.37e-04
			M phase	7	3.83e-04
			microtubule cytoskeleton organization	4	5.30e-04
			DNA-dependent DNA replication	4	7.34e-04
002_009	m: 43.77, s: 22.00	12	spindle microtubule	3	1.64e-05
			chromosome segregation	4	2.06e-04
			mitotic cell cycle	5	4.45e-04
			nucleosome assembly	2	9.34e-04
002_007	m: 46.59, s: 33.09	15	external encapsulating structure	4	1.13e-04
			glycerolipid metabolic process	4	1.22e-04
002_003	m: 47.21, s: 13.84	40	microtubule depolymerization	3	1.80e-05
			motor activity	3	4.01e-04
			mitotic anaphase B	2	4.83e-04
			spindle microtubule	3	6.80e-04
002_002	m: 50.69, s: 24.71	10	vesicle targeting, to, from or within Golgi	2	5.90e-05
			cytoplasmic vesicle part	3	1.89e-04
			ER to Golgi vesicle-mediated transport	3	4.19e-04
003_001	m: 59.33, s: 32.32	12	intracellular protein transport	6	4.85e-04
			protein targeting to mitochondrion	3	6.78e-04
			carboxylic acid transmembrane transporter activity	3	9.58e-04
003_002	m: 60.60, s: 23.71	33	AMP-activated protein kinase activity	2	3.27e-04
004_002	m: 70.32, s: 23.14	38	transport	19	8.51e-05
			cell communication	10	2.04e-04

Table 4.12 (cont.)

Cluster	Mean parameter value	Cluster size	GO term	GO count	p-value
004_001	m: 70.77, s: 31.59	29	'de novo' protein folding	3	1.95e-05
			mitochondrion organization	9	2.06e-05
			ATPase regulator activity	3	1.12e-04
005_002	m: 81.57, s: 24.44	40	ion homeostasis	6	3.42e-04
005_001	m: 82.93, s: 32.34	90	proteasome accessory complex	11	1.04e-14
			cellular catabolic process	34	3.16e-09
			positive regulation of RNA polymerase II transcriptional preinitiation complex assembly	4	1.83e-05
007_004	m: 233.93, s: 44.92	42	structural constituent of ribosome	10	2.12e-06
			gene expression	26	1.02e-05
007_005	m: 239.30, s: 62.50	72	gene expression	40	2.37e-06
			cytosolic large ribosomal subunit	7	7.82e-05

The different cluster approaches described above all produce different numbers of clusters with different sizes, and this affects the results that are obtained from the over-representation tests. For the well-fitting increasing Gaussian models, the variation in cluster size and number can be observed from Tables 4.7, 4.9 and 4.11, for the simultaneous parameter clustering, sequential meta-clustering, and cross meta-clustering, respectively. From these tables, the differences between the different clustering approaches can be identified, where the simultaneous parameter cluster approach produced a set of 5 clusters with a mean cluster size of 41 genes, the sequential meta-clustering approach produced 20 clusters with a mean size of 15 genes, and the cross meta-clustering produced 67 clusters with a mean size of 5 genes.

These results can be visually observed as a scatter plot, by using the fitted parameters of the genes as coordinates and labelling by cluster (Figure 4.16). From these figures, it can be seen that the simultaneous parameter clustering (A) has fewer but larger clusters than the others. However, some of the clusters have a broader range in one, or both, of the parameters. For example, cluster denoted by the orange crosses have a large range in the m parameter, while the other clusters have a larger range in the s parameter. The green triangles form a cluster with a large range in both parameters. The clusters formed by using the meta-clustering methodology, on the other hand, produces more clusters, which are smaller in size. The cross meta-clustering (C) in particular has multiple small clusters, as compared to the sequential meta-clustering (B).

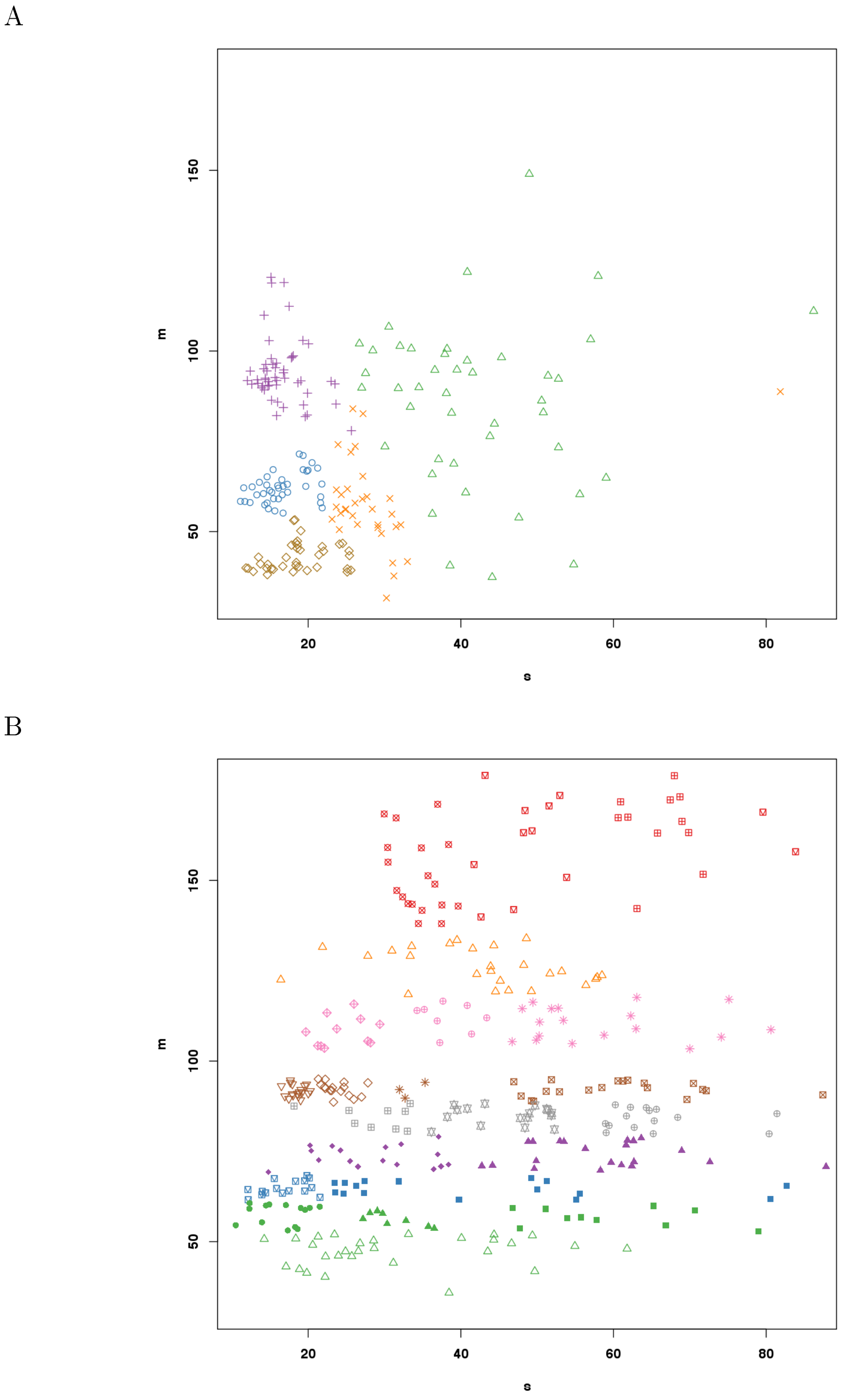


Figure 4.16: Parameter plots of the clusters showing which cluster each gene was assigned to, using the increasing Gaussian shapes. In each graph, the x-axis is the s parameter value, and the y-axis is the m parameter value. The different colours and shapes indicate different clusters. (A) Simultaneous parameter clusters; (B) Sequential meta-clustering, (C) Cross meta-clustering (overleaf). In the meta-clustering examples, the colours indicate the clusters formed by the m parameter, and the symbols indicate the clusters formed by the s parameter.

C

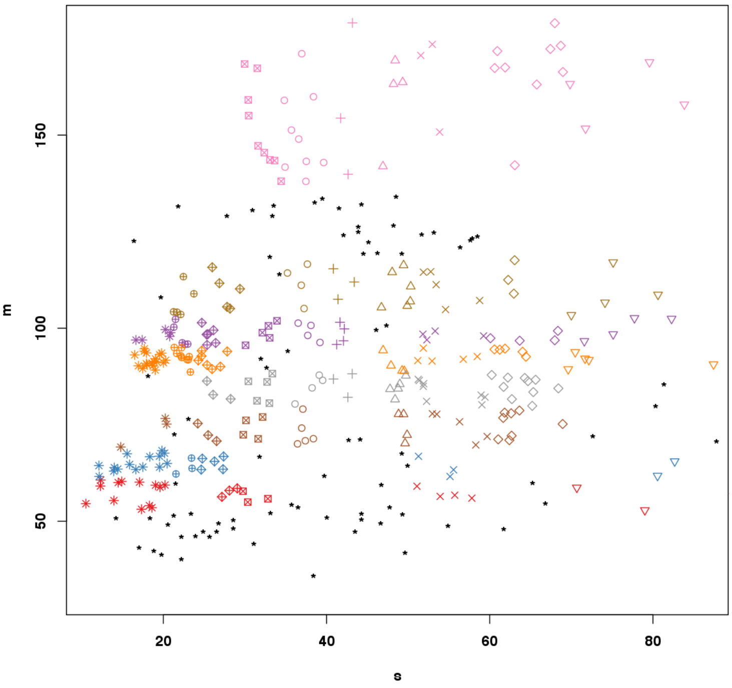


Figure 4.16: (cont.) Parameter plots of the clusters showing which cluster each gene was assigned to, using the increasing Gaussian shapes, using the cross meta-clustering approach. The x-axis represents the s parameter value, and the y-axis is the m parameter value. The colours indicate the clusters formed by the m parameter, and the symbols indicate the clusters formed by the s parameter. The clusters denoted by the black asterisks were unassigned.

4.3.4. Identification of common regulatory transcription factors

By clustering on a particular aspect of the expression profiles, such as the rate of change, it was hypothesised that it would be possible to identify genes that are influencing the expression of other downstream genes, that is, identify co-regulated genes. The genes that have similar rates of change in gene expression (similar *grad* parameters in the sigmoid curves, or m parameters in the linear models) could possibly be regulated by a common transcription factor (TF) or set of TFs. In addition, it may be possible that genes that are activated or repressed at the same time (similar *5per* parameters in the sigmoid curves, or m parameters in the Gaussian) could be activated in response to some stimulus under the control of common TFs.

To test this, a tool called YEASTRACT (YEAst Search for Transcriptional Regulators And Consensus Tracking) was used (Teixeira *et al.*, 2006). This database contains over 200 000 documented transcription regulatory associations between TFs and target genes, and was recently updated in June 2013 (Teixeira *et al.*, 2014). YEASTRACT analyses sets of genes for common regulators and identifies which TFs occur significantly based on a hypergeometric test comparing the occurrence of interactions in the clusters against all interactions in the database. YEASTRACT considers TFs with p-value $< 10^{-5}$ as highly significant. In addition, it calculates the proportion of genes in the cluster are regulated by that TF.

To determine if ShapeCluster was able to identify genes that are co-regulated, the clusters from the ShapeCluster analysis described in Section 4.3.2 were analysed. That is, the clusters from the exponential (r and b parameters) (Figures 4.6-4.8), Gompertz2 (*5per* and *grad* parameters) (Figures 4.9-4.11), Gaussian (m and s parameters) (Figure 4.12) and linear (m parameter) (Figure 4.13) were analysed to try identify any common transcriptional regulators. The results are shown in Table 4.13. From this table, it can be seen that regulatory TFs were primarily found using the exponential model with the b parameter, as well as Gompertz2 with the *grad* parameter. The Gaussian and linear clusters did not have many significant regulatory TFs, suggesting that these shapes and parameters were not optimal for identifying such relations.

Throughout the various clusters, a few common TFs could be identified such as ACE2, a TF which activates genes required for cytokinesis (Butler and Thiele, 1991), and BAS1, a MYB-related TF involved in regulating purine and histidine biosynthesis, as well as meiotic recombination (Tice-Baldwin *et al.*, 1989). These TFs are known to have many interactions (Teixeira *et al.*, 2014), so may not necessarily be relevant.

In the exponential models clustering with the r parameter, the majority of significant clusters were in clusters 10-24. These clusters had r>0.05 (Table 4.1), indicating a steeper rate of change. As discussed in Section 4.3.2, these clusters were associated with transcription and translation, whereas the clusters with a slower rate of change were associated with general metabolic processes. This reinforces the suggestion that gene regulation is activated very quickly, whereas other cellular processes occur at a slower rate. A common TF throughout these clusters was MGA2, a membrane protein that is involved in the regulation of the production of monounsaturated fatty acids, a critical component of cell membranes (Chellappa *et al.*, 2001). When clustered with the *b* parameter (Table 4.7), distinct sets between the concave and convex shapes (positive and negative values of *b* respectively) could be seen. In the clusters where $b<0$ (up-regulated genes, clusters 1-4), TFs included RAP1, an essential TF involved in transcription activation and repression (Lieb *et al.*, 2001) and SPT10, a histone acetylase with a role in transcriptional silencing. In contrast, the clusters with down-regulated genes ($b>0$, clusters 5-8) contained TFs such as MSN2 and MSN4, both transcriptional activators involved in stress responses (Martínez-Pastor *et al.*, 1996) and BAS1.

The clusters obtained by clustering the Gompertz2 profiles on the *5per* parameter did not have many clusters with significant TFs, with the exception of cluster 3 (up-regulated at approximately 55 minutes) with SNF6, a TF involved in chromatin re-modelling (Estruch and Carlson, 1990) and cluster 8 (up-regulated at approximately 126 minutes) which was significantly regulated by SFP1, a regulator of ribosomal biosynthesis and G2/M transitions during mitotic cell cycle (Xu and Norris, 1998), and MSN4. By clustering the Gompertz2 fits with the *grad* parameter, a larger number of significant regulatory TFs were found. In general, ACE2 and SFP1 were commonly found in the clusters with up-regulated genes (clusters 1-9). In cluster 4, significant TFs included STE12, a TF involved in mating and invasive growth (Roberts and Fink, 1994), and CRZ1, a stress response activated TF (Matheos *et al.*, 1997).

In addition, several sets of random genes were selected for comparison, both genes with the same shape, and sets with different shapes. The average cluster size across the various cluster analyses was calculated as 37 genes, and so 10 random sets of genes of this size were analysed for regulatory TFs. In most of these clusters, no significant regulators were identified, and in the three cases where there were, it was one of the regulators with many interactions, such as ACE2, SFP1 or BAS1. In addition, less than 30% of the cluster was regulated by the TF, as opposed to the much higher proportions in sets of genes identified by ShapeCluster. These results suggest that it is possible to identify different sets of TFs that are regulating gene expression using the clusters obtained from ShapeCluster.

Table 4.13: Top results from the YEASTRACT tool. The cluster number indicates the clusters shown in Figures 4.6-4.13. For each of these clusters, significant regulatory TFs are indicated as well as the proportion of genes in the cluster that are regulated by each TF. A dash (-) indicates that no significant regulators were found for that cluster.

Cluster number	Exponential			Gompertz2			Gaussian	Linear
	r	*b*	*r&b*	*5per*	*grad*	*5per&grad*	*m&s*	*m*
1	PDC2 (10%)	SWI5 (85%) RAC1 (49%)	-	-	SFP1 (100%) RIC1 (83%) SUM1 (67%)	-	-	GCR1 (78%)
2	-	RAP1 (74%) GAL4 (43%) THI2 (31%)	-	-	-	-	SNF7 (3%) ROX1 (50%) SUT1 (31%)	-
3	-	GCR1 (48%) PIP2 (41%) SPT10 (72%)	BAS1 (73%)	SNF6 (100%) ACE2 (100%) THO2 (29%)	-	SFP1 (100%) MBP1 (67%)	SWI6 (38%) MBP1 (36%) AZF1 (16%)	-
4	BAS1 (78%)	GCR1 (60%) SPT10 (53%) RAP1 (70%)	-	-	SFP1 (100%) STE12 (100%) CRZ1 (80%)	BAS1 (100%) ACE2 (100%) AZF1 (50%)	-	-
5	GCR2 (42%)	DAL80 (9%)	-	-	ACE2 (100%)	BAS1 (82%) RLM1 (38%) CAD1 (36%)	HPC2 (7%) HIR1 (24%)	-
6	-	BAS1 (66%) CIN5P (46%)	-	-	ACE2 (100%)	BAS1 (68%)	SUA7 (6%) SIN4 (64%)	BAS1 (74%)

Cluster number	Exponential			Gompertz2			Gaussian	Linear
	r	*b*	*r&b*	*5per*	*grad*	*5per&grad*	*m&s*	*m*
7	-	BAS1 (73%) SOK2 (59%) SFP1 (85%)	ACE2 (89%) ABF1 (61%)	ACE2 (100%) FKH2 (53%)	RDS2 (71%)		ACE2 (100%)	
8	-	MSN2 (95%) MSN4 (91%) PDR1 (63%)	MGA2 (51%) SOK2 (64%) BAS1 (71%)	SOK2 (100%) MSN4 (100%) SFP1 (100%)	ACE2 (100%) ATE12 (100%) SFP1 (100%)		HSF1 (54%)	
9	MGA2 (34%)		-	-	-		-	
10	SOK2 (57%) MGA2 (36%) SIN3 (48%)		SWI3 (75%) STP1 (42%) YAP6 (46%)	BAS1 (91%) RLM1 (52%) SRB8 (39%)	-		-	
11	MGA2 (39%) RIC1 (23%)		BAS1 (74%) HOT1 (13%) GIS1 (24%)	-	BAS1 (79%)		-	
12	MGA2 (38%) RPN4 (48%) HSF1 (44%)		MGA2 (42%) DAL81 (20%)	PHO2 (41%)	BAS1 (85%) GIS1 (29%) SPT23 (63%)		BAS1 (70%)	
13	MGA2 (38%) RPN4 (48%) HSF1 (44%)		XBP1 (35%) RPN4 (57%) SFP1 (89%)	-	MSN2 (100%) SFP1 (100%) ACE2 (100%)		-	

Cluster number	Exponential			Gompertz2			Gaussian	Linear
	r	*b*	*r&b*	*5per*	*grad*	*5per&grad*	*m&s*	*m*
14	MGA2 (34%) SFP1 (83%)		BAS1 (90%) MGA2 (53%) CIN5 (73%)	HAP4 (44%) HAP3 (25%)				
15	YRM1 (59%) MGA2 (33%)		MGA2 (53%) SPT23 (60%) SWI5 (58%)					
16	SOK2 (58%) RGM1 (20%) ACE2 (87%)		-					
17	MGA2 (46%) SFP1 (83%) ESA1 (2%)		SWI5 (77%) SOK2 (77%) SFP1 (95%)					
18	SNF2 (58%) SFP1 (88%) MGA2 (38%)							
19	-							
20	SFP1 (100%) SWI5 (83%) MGA2 (63%)							
21	STB5 (57%) SPT20 (67%)							

4.3.5. Comparison to other clustering methods

Generally when applying a clustering methodology to a set of data, the output will be one set of clusters. However, ShapeCluster constructs clusters based on a specific shape and specific parameters, so a variety of different cluster analyses can be performed to obtain an overall representation of the similarity amongst gene expression profiles.

In order to determine the quality of the clusters formed by the different approaches, the Biological Homogeneity Index (BHI) was used, described in Section 2.1.8. The BHI provides a simple metric that ranges from 0 to 1, where 1 indicates that all the genes have at least one term in common all the others. To obtain an overall representation of the gene expression, cluster analyses were performed using a range of regression models and different parameter combinations, namely exponential (a,r,b), Gaussian (a,b,m,s), Gompertz1 (a,b,*5per*,*grad*), Gompertz2 (a,b,*5per*,*grad*), logistic (a,b,*5per*,*grad*), and linear (m,c). For the sigmoid models, the *5per* and *grad* parameters were used in lieu of the m and s parameters, as an indication of timing and rate of change, respectively. The BHI scores for these different sets of clusters are shown in Table 4.14. All the fitted parameters were clustered individually, as well as the most biologically significant terms for the simultaneous and meta-clusterings. The shapes were also separated into increasing and decreasing forms prior to clustering. From this table, the highest scoring clusterings were the increasing logistic clustered on *grad* for the single parameter; decreasing Gaussian clustered on *m&s* for the simultaneous parameter clustering; and decreasing logistic clustered on *5per&grad* for both types of meta-clustering.

To compare the efficacy of ShapeCluster, the results of the clustering were compared to results from two other methods. The first is SplineCluster (Heard *et al.*, 2006), which uses a Bayesian model-based hierarchical clustering algorithm to model the gene expression profiles through the use of linear models and nonlinear basis functions. The other is Bayesian Hierarchical Clustering (BHC) (Cooke *et al.*, 2011; Savage *et al.*, 2009). This method is also a model-based hierarchical clustering, except it uses Gaussian process regression to capture the structure of the data. The clustering process uses a fast approximate interface method for a Dirichlet process mixture model, which performs agglomerative hierarchical clustering in a Bayesian framework. To obtain a fairer comparison, the full 5661 gene yeast dataset (Section 4.3.1) was filtered using the same thresholds as ShapeCluster, namely $R_a^2 > 0.6$, $R_{LoF}^2 > 0.6$, and F-test$<$0.05, and this resulted in a set of 1286 genes. The cluster analyses using SplineCluster and BHC were performed using the default parameters, and the BHI score for SplineCluster was 0.56, and BHC resulted in a score of 0.52. These clustering algorithms produce one set of clusters, thus there is a single BHI score for each method. Compared to the cluster results from ShapeCluster (Table 4.14), it can be seen that the values are often similar, if not higher, thus indicating that ShapeCluster produces clusters that contain genes with common annotations at a similar or better level compared to other clustering methods. These values depended on the shape and parameter combinations, potentially identifying the most biologically significant factors.

Table 4.14: Table of all the BHI scores for the various clusterings performed on the yeast data. *Single* refers to clustering performed on a single parameter, *Simul* is the clustering on multiple parameters, using the simultaneous parameter clustering, and *Meta* refers to the two types of meta-clustering. The meta-clustering was performed on the same parameters as the simultaneous parameter clustering. *Sequential* is the clustering of one cluster followed by another, and *Cross* is the cross-clustering where clusters were identified based on the genes found from overlapping clusterings. The highest BHI score for each model is in green, and the lowest BHI score is in red.

	Linear	Exponential	Gaussian	Gompertz1	Gompertz2	Logistic
Single	*m* Inc: 0.565 Dec: 0.532	*a* 0.547	*a* Inc: 0.478 Dec: 0.551	*a* Inc: 0.531 Dec: 0.474	*a* Inc: 0.521 Dec: 0.494	*a* Inc: 0.507 Dec: 0.478
	c Inc: 0.558 Dec: 0.510	*b* 0.509	*b* Inc: 0.472 Dec: 0.538	*b* Inc: 0.487 Dec: 0.515	*b* Inc: 0.485 Dec: 0.535	*b* Inc: 0.542 Dec: 0.472
		r 0.492	*m* Inc: 0.490 Dec: 0.507	*5per* Inc: 0.467 Dec: 0.480	*5per* Inc: 0.490 Dec: 0.463	*5per* Inc: 0.497 Dec: 0.520
			s Inc: 0.502 Dec: 0.522	*grad* Inc: 0.600 Dec: 0.418	*grad* Inc: 0.574 Dec: 0.456	*grad* Inc: 0.618 Dec: 0.380
Simul		*r,b* 0.483	*m,s* Inc: 0.504 Dec: 0.522	*5per,grad* Inc: 0.487 Dec: 0.496	*5per,grad* Inc: 0.493 Dec: 0.447	*5per,grad* Inc: 0.502 Dec: 0.471
Meta		*Cross* 0.531	*Cross* Inc: 0.584 Dec: 0.542	*Cross* Inc: 0.515 Dec: 0.671	*Cross* Inc: 0.593 Dec: 0.509	*Cross* Inc: 0.554 Dec: 0.780
		Sequential 0.533	*Sequential* Inc: 0.555 Dec: 0.556	*Sequential* Inc: 0.514 Dec: 0.580	*Sequential* Inc: 0.564 Dec: 0.518	*Sequential* Inc: 0.541 Dec: 0.587

4.4. Discussion

In this chapter, a collection of methods were developed for clustering data, whereby gene expression profiles are clustered using fitted regression models. These fitted curves represent a smoothed description of the gene expression profiles from a time series experiment, and the fitted parameters reflect aspects of the underlying biology. By performing a cluster analysis on aspects of the curves, such as the gradient or time of activation, it is possible to obtain more information regarding the biological processes that are occurring at a given point in time. For example, identify the genes which may be co-regulated by a common transcription factor, or are part of similar metabolic functions. This methodology presents a different philosophy and analysis approach for investigating gene expression profiles, where the profiles are grouped based on important aspects of the profile, instead of simply looking at the entire profile at once.

This clustering methodology was termed ShapeCluster, and operates in a two-part process. First, one of the regression models is selected, and gene expression profiles which fitted this model are used in the second step, namely cluster on one or more of the fitted model parameters. When multiple parameters are used, a number of options are available. One is the simultaneous parameter calculation, where the parameters are combined using a Euclidean distance to produce a single measure. The alternative is to use a meta-clustering approach, where the clusters are reclustered. The first type of meta-clustering is a sequential meta-cluster, where genes are clustered on the first parameter, and then these clusters are clustered based on a second parameter. The second type of meta-clustering is the cross meta-cluster, where the genes are independently clustered on each of the two parameters, and the genes in common between the two clusters are identified. The meta-clustering approaches are ideally used with two parameters, although they can be expanded to use more. However, this could lead to clusters with few members. It is possible to combine all the parameters using the simultaneous parameter method, although this would make the clustering more like "traditional" clustering approaches, where the clustering is performed over the entire expression profile, instead of investigating only a specific aspect of the profile, possibly leading to less biologically relevant genes being clustered together.

As a validation of the method, the expression data from a yeast time-series gene expression experiment was used (Orlando *et al.*, 2008). This data was selected as yeast possesses a smaller and more tractable genome than Arabidopsis, and thus, has a better level of annotation. Therefore it was expected that the clusters from ShapeCluster would contain more similar biological terms than other clustering algorithms. In general, it was found that clusters with up-regulated genes were involved in cell division and energy production, as would be expected from growing wildtype cells. On the other hand, the down-regulated genes were primarily involved in secondary metabolic processes and proteolysis. Between the different models analysed, different sets of genes were identified, indicative of the various expression patterns. For example, the exponential models represent a gene that has an exponential rate of transcription, before flattening to a plateau. These genes were primarily involved in ribosomal activity and translation. In contrast, the Gaussian models represent a gene that is activated, reaches a maximum level, and then becomes deactivated. Alternatively, the Gaussian model can also represent a gene that is repressed, reaches a minimum, and then becomes reactivated. The genes with increasing shapes were primarily involved in cell cycle processes, such as cell division and spindle pole body activity, whereas the decreasing shapes were involved in general secondary metabolic processes.

Within each of the different models, it was possible to cluster using different parameters, or sets of parameters, and each combination would reveal sets of genes with different functions. By focussing on a single parameter, and thus a single aspect of the expression profile, it was possible to investigate one aspect of the response, such as the timing of the activation of genes, or the rates that gene expression is changing. These types of clusters would not be formed with traditional clustering methods, which group genes based on the entire expression profile. As a result, this may provide an indication of genes that are being activated or repressed by some controlling set of genes despite appearing different in other aspects of their response profiles.

Clustering on the *5per* parameter (time of 5% of maximum response) provided insights into genes that were up- or down-regulated at a given time, and using this parameter made it possible to determine when specific sets of genes were activated or repressed. This in turn provided a means of identifying what biological processes were being activated or repressed in response to the stimulus, allowing the times that metabolic functions occurred to be elucidated. In the Gompertz2 case, it could be seen that transporter activity was followed by cytoskeletal growth, which in turn was followed by spindle pole body activity. Using the average parameter values from each cluster, as well as the information described above, a simple timeline of the biological processes that were occurring over time could be determined, and is shown in Figure 4.17. Thus, the ShapeCluster analysis provided a quick means to develop a timeline of biological events that were taking place at a given time, and may aid in identifying key time points for further investigation or experiments.

On the other hand, clustering on the *grad* parameter (rate of change of gene expression), showed the genes that were being up- or down-regulated with a similar rate of change, and thus could possibly be controlled by the same TFs. It also provided an indication of the genes that are changing rapidly or slowly. Again using the genes with fits to the Gompertz2 model, the slowly changing genes were involved in protein transport, while the rapidly changing genes were involved in transcription regulation.

These clusters were also analysed to determine if the clusters contained any common regulatory TFs, and it was found that the *grad* parameter is a better parameter to use to identify common transcriptional regulators than the *5per* parameter. This suggests that rate of change in gene expression is a better indicator of co-regulation than the time of activation or repression. The clusters formed using the Gompertz2 fits with the *grad* parameter were primarily regulated by TFs involved in mitosis, cytokinesis, and stress responses. In the exponential clusters, the clusters with a steeper rate of change (r>0.05, Table 4.1), were again implicated with involvement in transcription and translation, whereas the clusters with a slower rate of change were more associated with general metabolic processes. This reinforced the suggestion that gene regulation is activated very quickly, whereas other cellular processes occur at a slower rate. Interestingly, a cluster with the exponential model that contained down-regulated genes, and a cluster with the Gompertz2 model that contained up-regulated genes were both controlled by MSN4, a transcriptional activator involved in stress responses (Martínez-Pastor *et al.*, 1996). The genes in the cluster with the exponential shape were involved in secondary metabolism, while the genes in the cluster with the Gompertz2 shape were primarily involved in cell division. This could suggest a dual role for TFs, where these regulators influence multiple aspects of gene expression. These results show that it is possible to identify different sets of TFs that are regulating gene expression using the clusters obtained from ShapeCluster.

These analyses provided a biologically oriented description of individual gene expression profiles, and through careful choice of appropriate models, these methods could allow for an improved comparison of gene expression profiles, and may provide an improved understanding of common regulatory mechanisms between genes.

As mentioned previously, a number of different cluster approaches were used in the clustering process, namely single parameter, simultaneous multi-parameter, sequential meta-clustering, and cross meta-clustering. The single parameter clustering allows the researcher to investigate a single aspect of the model shape. The other methods provide the means to investigate multiple parameters, each in a different manner. The simultaneous parameter approach uses an Euclidean distance to merge the parameter values together to form a single metric. This may not ideal, particularly if the parameter values possess very different ranges, for example *5per* and *grad* in the sigmoid functions, or r and b for the exponential-type functions. In these cases, the merged value may distort the true value of each parameter, and as a result, when the dendrogram is cut to produce the clusters, the clusters may not be neatly assigned. For example, with the exponential clustering on *r&b*, it would be expected that the clusters would be formed with the same rate of change and concavity. However, since the r parameter is significantly smaller (usually $r<0.1$) than the b values, it resulted in some clusters with both concave and convex shapes. Alternative distance metrics, such as a Mahalanobis distance could be used instead.

This problem is avoided in the meta-clustering approaches, as the original parameter values are always taken into account. Compared to the simultaneous parameter clustering, the meta-clustering produced very similar over-represented GO terms, and in most cases, provided greater detail when investigating the functions of the clusters, as well as revealing results that previously had not been seen using the simultaneous parameter clustering. However, these meta-clustering approaches produce a larger number of clusters with fewer members as compared to the simultaneous parameter clustering. This can particularly be a problem with the cross meta-clustering, where if there are m clusters after clustering on parameter 1, and n clusters after clustering on parameter 2, there will be up to mxn clusters. This large number of small clusters could potentially be

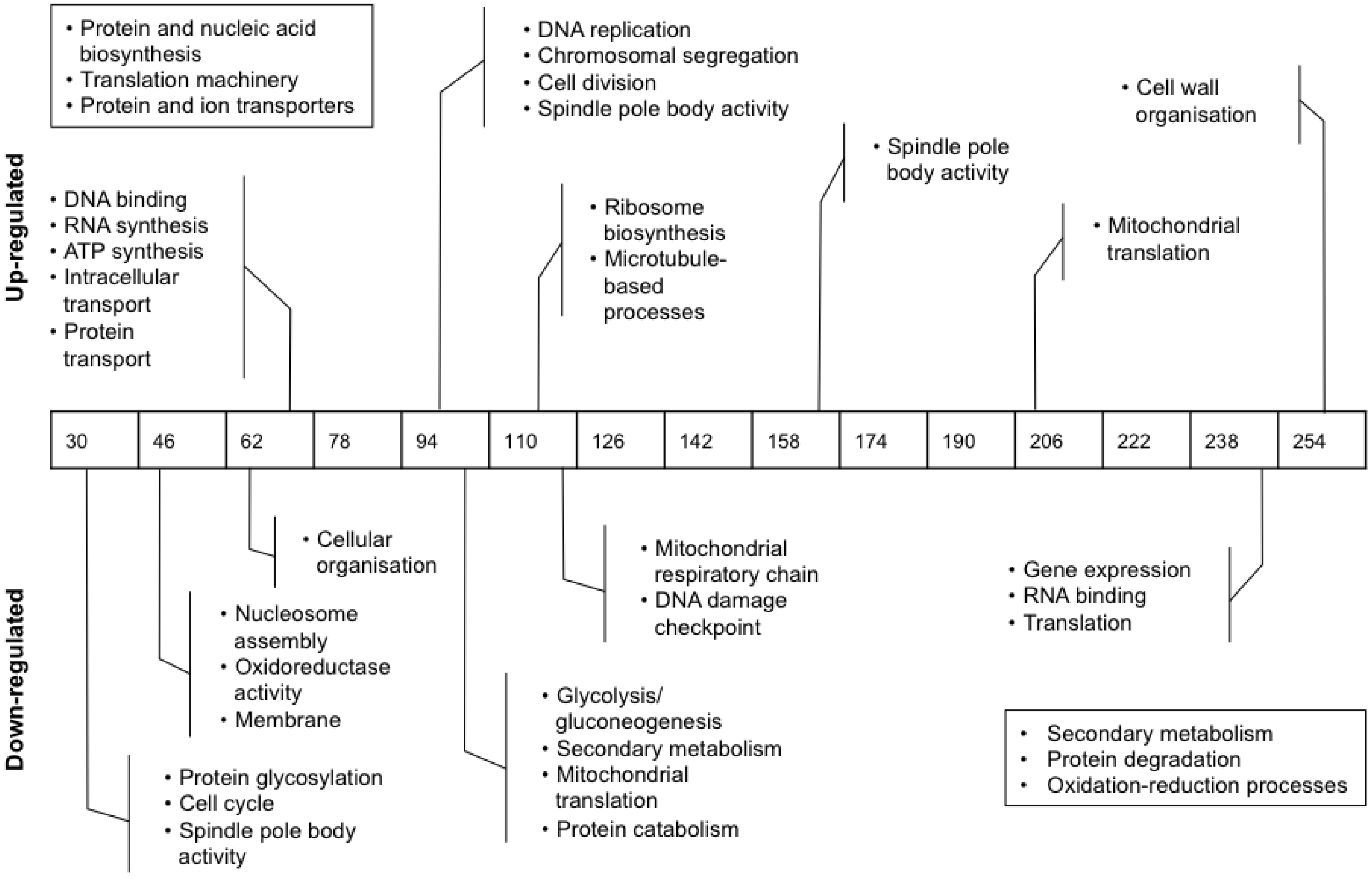


Figure 4.17: A simple timeline representation of the processes that are occurring during the yeast time course, as elucidated using the results from ShapeCluster. The bar in the middle represents the sampling time points. The up-regulated genes are shown above the time bar, and the down-regulated genes are shown below. Processes that were generally up- or down-regulated are shown in the boxes.

undesirable, particularly when looking for over-representation in the annotation terms. The sequential meta-clustering appears to be a good compromise between the simultaneous parameter clustering and the cross meta-clustering, both in terms of cluster size and cluster number.

While this dataset was analysed to demonstrate the application of the regression analysis and clustering approaches, it was possibly not the best dataset due to the time-series consisting of two cell division cycles. This means that many of the genes would show cyclic patterns. None of the functions really take this into account, and those that do would likely have a poor fit. However, investigating the expression profiles fitted from SplineCluster (Figure 4.4), there are very few genes which process this cyclic behaviour, and the shapes that are present are covered by the models that were used in the ShapeCluster. Nonetheless, it would be possible to extend the set of nonlinear functions in ShapeCluster to include appropriate shapes. Despite this, it was still possible to use the regression analysis approach to identify interesting groups of genes.

The results of the clustering were also compared to other clustering methods, namely Bayesian Hierarchical Clustering (BHC) and SplineCluster. Through the use of the Biological Homogeneity Index (BHI), it was found that ShapeCluster produced a number of clusters that were more biologically significant clusters than the other methods, despite these methods being able to take cyclic patterns into account. The BHI score was used to show the level of biological similarity between members of a cluster, based on the common GO annotations. However, this score is not a completely fair comparison, since different sets of genes were used in the different cluster analyses. That is, SplineCluster and BHC performed the cluster analysis using all the genes, whereas ShapeCluster uses the genes that fitted a particular model. Nonetheless, it does provide a simple indicator of the similarity of annotations in the different clusters.

Through the use of the BHI, it was also possible to compare the results from ShapeCluster, using different models and parameter combinations. In other clustering approaches, there is only one set of clusters generated for a given set of data. However, with ShapeCluster, there are a number of different sets of clusters possible, depending on the model and the parameters being clustered. With the sigmoid models, performing the cluster analysis on the decreasing shapes, and using the *grad* parameter on its own produced the clusters with the least homogeneous clusters. However, the increasing shapes showed relatively high BHI scores. This suggests that there is some relation in the functions of genes that are increasing at the same rate, but not necessarily decreasing, possibly indicating that the genes are being regulated by genes at the same rate, or activating each other in successive waves, such as in signal transduction cascades. It also suggests that the down-regulation of genes does not follow this type of regulatory mechanism. Clustering using the *5per* parameter alone also did not produce many functionally similar sets of genes. However, when genes were grouped using both the *5per* and *grad* parameters, more biologically homogeneous clusters were found. This indicates that both the rate of change and the time of activation are important in identifying biologically relevant sets of genes. The *a* and *b* parameters (asymptote and range parameters, respectively) also did not produce clusters with similar functions. While it may be interesting to identify genes that have the same starting, or ending, expression levels, in general these clusters may not be informative. As a result, these clusters would not be as useful in understanding the underlying biological system, particularly since normalisation processes often distort these values. The exception to this is the exponential model, where the clusters formed using the a parameter produced the highest BHI score. This possibly indicates that genes which reach similar expression levels after an exponential rate of change are functionally related. Similarly for the Gaussian results, the m and s parameters, which provide an indication of the maximum response, and duration of the response, respectively provide more biological information than the other parameters.

The best BHI scores for most of the models were generally found by performing the meta-clustering analyses. The cross meta-clustering had better scores, although this could be a result of the analysis producing numerous small clusters. On the other hand, using a Euclidean distance in the simultaneous parameter approach could distort the distance matrix, resulting in clusters that are not distinct.

The BHI score is not a perfect scoring metric, as it is dependent on the annotations available, and thus the amount of information available for a particular gene would influence the score. In addition, the BHI value is not very sensitive to changes in the degree of annotation. In particular, each pair of genes is scored 0 or 1 if there are any common annotations between them, and does not take the proportion of common annotations into account. In addition, if sets of genes do not have an exact, known function, and are only annotated with the very top level terms of GO (i.e. "cellular component", "molecular function" and "biological process"), this may artificially inflate the value of the indices. Finally, the index does not take the numbers of clusters formed, nor the cluster size into account.

In conclusion, it is important for a researcher to understand what the biological question is. While it may be tempting to cluster individually on each of the parameters separately, caution should be exercised to ensure that the most biologically relevant parameters are used. Although it may be interesting to identify genes that have the same starting, or ending, expression levels, in general these clusters may not be as informative as using the timing or rate of change parameters. As a result of using the yeast data, it was possible to obtain some general guidelines when performing the clustering on the fitted models. In general, the timing parameters (such as m or *5per*) and rate of change parameters (*grad* or r) are more informative in determining the molecular processes that are taking place. Table 4.15 shows some of the shape and parameter combinations that can be used in order to solve a specific task, such as identifying potentially co-regulated genes.

Prior to clustering, various thresholds were used to identify the profiles that had a good fit to one or more of the models. Here R^2 thresholds of 0.6 were used, although these can easily be adjusted to allow more or fewer fits into the analysis. Increasing the threshold would increase the stringency and reduce the number of expression profiles to

Table 4.15: Table of recommended shape and parameter combinations to investigate specific biological questions.

Task	Shape	Parameter
Identify co-regulated genes	Gompertz, logistic	*grad*
	Exponential	*r, b*
	Linear	*m*
Identify gene responses with the same response duration	Gaussian	*s*
Determine time of activation or repression	Gompertz, logistic	*5per, m*
	Gaussian	*m*
	Linear+exponential	*linpnt*
	Critical exponential	*turnp*

analyse, while decreasing the thresholds would permit expression profiles with poorer fits, potentially including relevant genes that were excluded due to poor fit statistics. In most cases, a threshold of 0.6 should be sufficient to provide a balance between goodness-of-fit and quantity of expression profiles to analyse.

The use of the simultaneous parameter approach is simple and useful for a general idea of the function of a group of genes. However, for more refined clusters, the use of the sequential meta-clustering provides a good balance between loss of information and number of clusters.

In the following chapter, these clustering algorithms will be applied to the Arabidopsis datasets introduced in the Section 3.9.

5. Analysis of large time series datasets

5.1. Introduction

In Chapter 3, a regression analysis approach was described where linear and nonlinear functions could be fitted to time-series gene expression data. In particular, the data used was obtained from the PRESTA long day senescence (Breeze *et al.*, 2011) and *Botrytis cinerea* infection (Windram *et al.*, 2012) time series experiments, and the details about these projects are described in Section 1.2. In brief, the senescence dataset using the morning samples consisted of 11 time points, taken every second day for 22 days, while the Botrytis dataset consisted of 24 time points, taken every 2 hours for 48 hours. In both cases, there were four biological replicates. A set of 23 802 unique probes provided a comprehensive set of genes to represent the majority of the Arabidopsis genome. The regression analysis was performed on this set of genes, and the results of the analysis are shown in Section 3.9.

In the previous chapters, parametric regression models were fitted to the gene expression profiles, and an application named ShapeCluster was developed which used these fitted parameters to group expression profiles together. In this way, a biologically oriented description of gene expression profiles could be formed and used to identify genes that are potentially functionally related in a mechanistic manner. By clustering genes on a single parameter, it was possible to focus in on a specific aspect of the expression profile, and find genes with that aspect in common. In Chapter 4, data from the model organism, *Saccharomyces cerevisiae*, which possesses a relatively small, well annotated genome was used to demonstrate the capabilities of ShapeCluster. It was shown that the clusters produced by ShapeCluster contained more biologically significant genes, as compared to other clustering algorithms. In this chapter, the models fitted from the regression analysis on the Arabidopsis data will be analysed using the clustering algorithms described in Chapter 4.

5.2. Clustering results for the Arabidopsis datasets

5.2.1. Senescence data

The ShapeCluster analysis was applied to the 11 time point senescence data, using the 23 802 genes that were representative of the Arabidopsis genome, and a broad overview of the shapes and over-represented annotations for the observed gene expression profiles are described in this section. The summary of the results from the fitting of the regression models are shown in Section 3.9. The same thresholds determined in Section 3.9 were used to identify models which fitted well, namely $R_a^2 > 0.6$, $R_{LoF}^2 > 0.6$, and

F-test<0.05. As with the yeast data, all good fits were used in the cluster analysis, meaning a gene could be fitted by multiple models, and thus could appear in cluster analyses for more than one model. For this dataset, the most abundant models were the exponential, Gaussian, Gompertz1 (faster growth rate to the right of the midpoint), and logistic models, and so these subsets of genes are presented in the subsequent analyses. It was also shown in Chapter 4 that the most biologically informative parameters were the timing and rate of change parameters, and the most relevant shape and parameter combinations are indicated in Table 4.15. That is, the *5per* and *grad* parameters in the sigmoid models, the m and s in the Gaussian model, and the r parameter for the exponential, and so these parameter combinations were used in the cluster analysis. Thus, these clusters would be able to identify genes that are being activated or repressed at the same time, and thus involved in the same biological process, as well as genes that have the same rate of gene expression, and thus potentially under the control of common transcriptional regulators. The results below were generated using the simultaneous parameter clustering (Section 4.2.2).

Biological significance was determined using the `GOstats` package (Falcon and Gentleman, 2007, Section 2.1.4) to determine over-represented GO terms, as well as detecting over-represented words in the gene annotations (Section 2.1.5). Both of these calculations use a hypergeometric test to determine significance. Since this dataset was much larger than the yeast dataset, there were a much greater number of clusters. Thus, for brevity, only a few of the clusters for each model will be discussed here, and the full results are provided in Appendix D. The x-axis of this dataset represents different times of harvests, where the first time point is 19 days after sowing (DAS), and the y-axis represents the log_2 gene expression level.

The clusters of genes fitted by exponential models are shown in Figure 5.1. These genes were clustered on the *r&b* parameters, and resulted in a total of 33 clusters. The combination of the r (rate of change) and b parameter values (shape of response) affects the overall shape of the model, and genes with similar values in both parameters were often involved in the same biological process. Analysing the annotation terms revealed that the genes in clusters 1 and 2 were enriched in terms involving transporter activity (Table 5.1). Here the combination of the parameter values resulted in shapes that were concave increasing. The genes in clusters 3-5 were enriched for ribosomes and RNA processing, and the shapes for both clusters were convex decreasing. This suggests that as the plant ages, the ribosomal and translational activities decrease. The genes in cluster 6 had lower r values than the genes in the previous clusters, indicating a more linear response, as well as a positive b value indicating a decreasing response. This response makes sense since the cluster contained genes that were involved in photosynthesis, which are known to be repressed as senescence progresses. The genes in cluster 7 had a similar rate of change to the genes in cluster 6, but had b parameter values of an opposite sign, resulting in an increasing shape. These genes were enriched for terms relating to stress responses and water deprivation, and could be involved in the activation of senescence responses.

Table 5.1: Table showing the over-represented annotation terms for the exponential model in the senescence data.

Cluster	Mean parameter value	Cluster size	Annotation term	Count	p-value
1	r: 0.350, b: -1.230	36	intracellular signal transduction	5	8.14e-06
			transport	10	2.02e-04
			metal ion transport	4	3.08e-04
			monovalent inorganic cation transport	3	8.42e-04
2	r: 0.410, b: -2.729	37	actin binding	3	1.70e-04
			transporter	8	2.21e-05
			secondary active transmembrane transporter activity	4	4.58e-04
3	r: 0.531, b: 1.928	37	translational elongation	6	2.73e-11
			structural constituent of ribosome	9	2.01e-09
4	r: 0.653, b: 2.850	46	ribosome	13	2.86e-13
			translation	14	1.00e-12
5	r: 0.887, b: 3.812	52	Ribosomal	24	2.04e-25
			structural constituent of ribosome	22	3.26e-27
			translation	22	2.03e-22
6	r: 0.118, b: 5.358	31	chloroplast	14	2.89e-05
			detection of abiotic stimulus	2	4.16e-04
7	r: 0.175, b: -4.200	42	response to desiccation	2	4.88e-04
			response to water	4	8.22e-04
			response to nutrient levels	3	8.34e-04

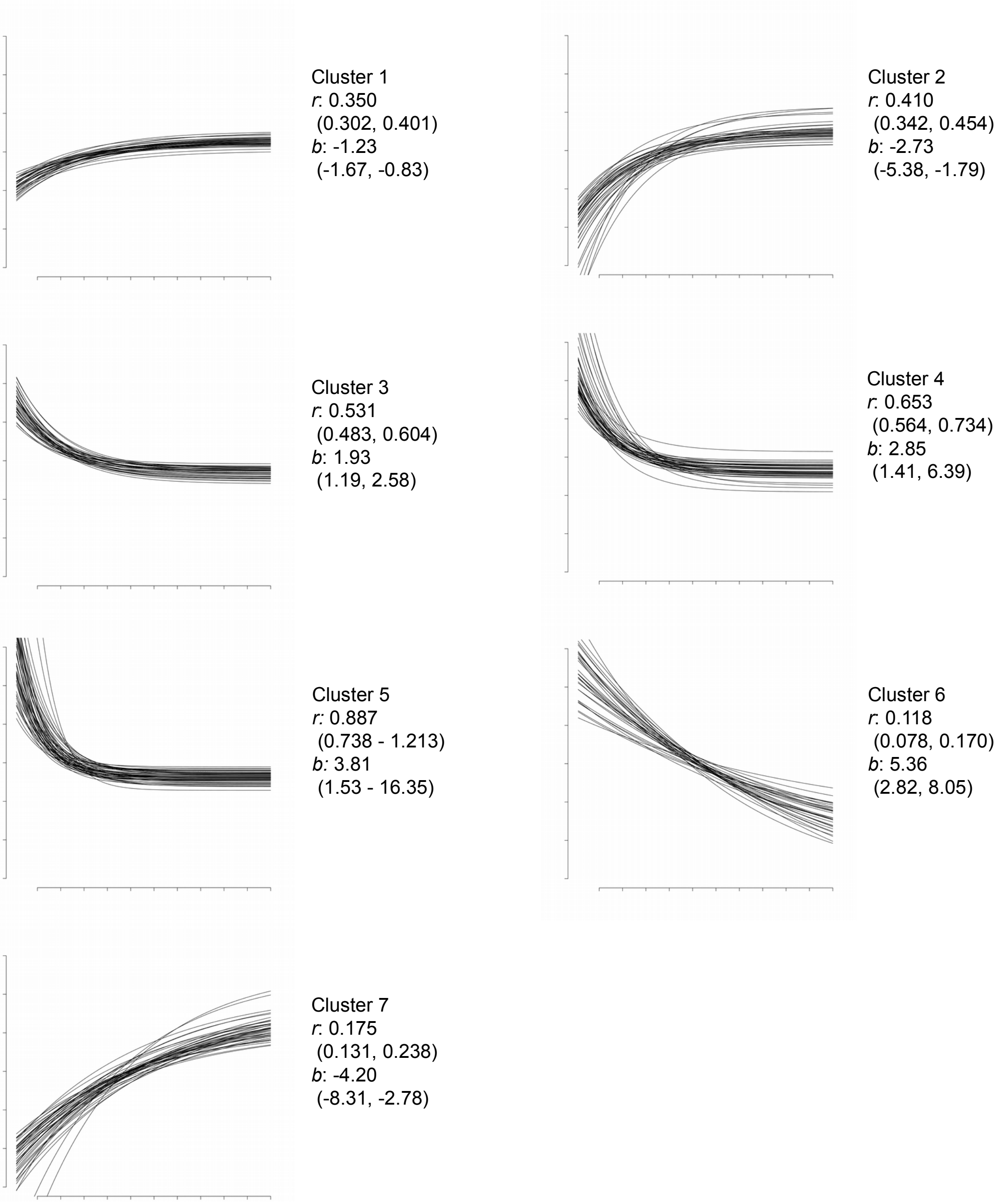


Figure 5.1: Exponential shapes from the senescence experiment, clustering on both the r (rate of change) and b (concavity) parameters. Shown are the mean parameter values, as well as the range of the parameter. The x-axis shows the sampled data points (age of plants), and the y-axis is the $\log_2$ gene expression level.

Figure 5.2 shows some examples of the genes that fitted a Gaussian model, and clustered on the *m&s* parameters, where the m parameter indicates the time of maximum response, and the s parameter describes the duration of the gene activation. Therefore clustering on these parameters would identify genes that reach their maximum or minimum at the same time, as well as having the same duration of response. There were a total of 72 clusters generated for this shape. As with the yeast data, this model represents a gene expression profile that increases to a maximum, before decreasing again. Alternatively, it may represent a gene being repressed, and then reactivated. The former shape is shown in clusters 1-6, and the latter in clusters 7-14. In the clusters being activated, genes in cluster 1 were enriched for carbohydrate metabolism, cluster 2 with transporter activity, and clusters 3 and 4 with vacuole regulation and metabolic activities (Table 5.2). This activity indicates that the plant is beginning to activate the transport processes in order to mobilise the macronutrients to other parts of the plant, such as storage organs. Notably, genes in clusters 5 and 6 were enriched for transcription factor (TF) activity. These clusters had a relatively late maximum response time of time point 7 and 8 respectively, thus indicating TFs that are activated near the end of the senescence process. Cluster 6, contained the genes *ANAC014* (*AT1G33060*) and *ANAC089* (*AT5G22290*). The NAC TF family has been shown to be involved in the senescence process (Breeze *et al.*, 2011; Hickman *et al.*, 2013), so it is possible that these genes are involved in the regulation of the senescence response. ANAC089 is involved in regulating the flowering time in Arabidopsis (Li *et al.*, 2010), and ANAC014 currently has no known biological function.

In contrast, in the genes being repressed, clusters 7-9 contained genes that were involved with RNA binding activity, and the genes in clusters 10-14 were involved with photosynthesis and chloroplasts. The former set of clusters were all down-regulated early in the time series (around time point 2-4), and were primarily involved in translation. The latter clusters were repressed at a much later stage (after time point 10) indicating that the photosynthesis genes are becoming down-regulated towards the end of the time series, as discussed above. The Gaussian model describes a change in expression to a maximum, or minimum, followed by a transition back to the starting expression level. This suggests that these responses are all transient, and are being activated or repressed in response to some biological signal.

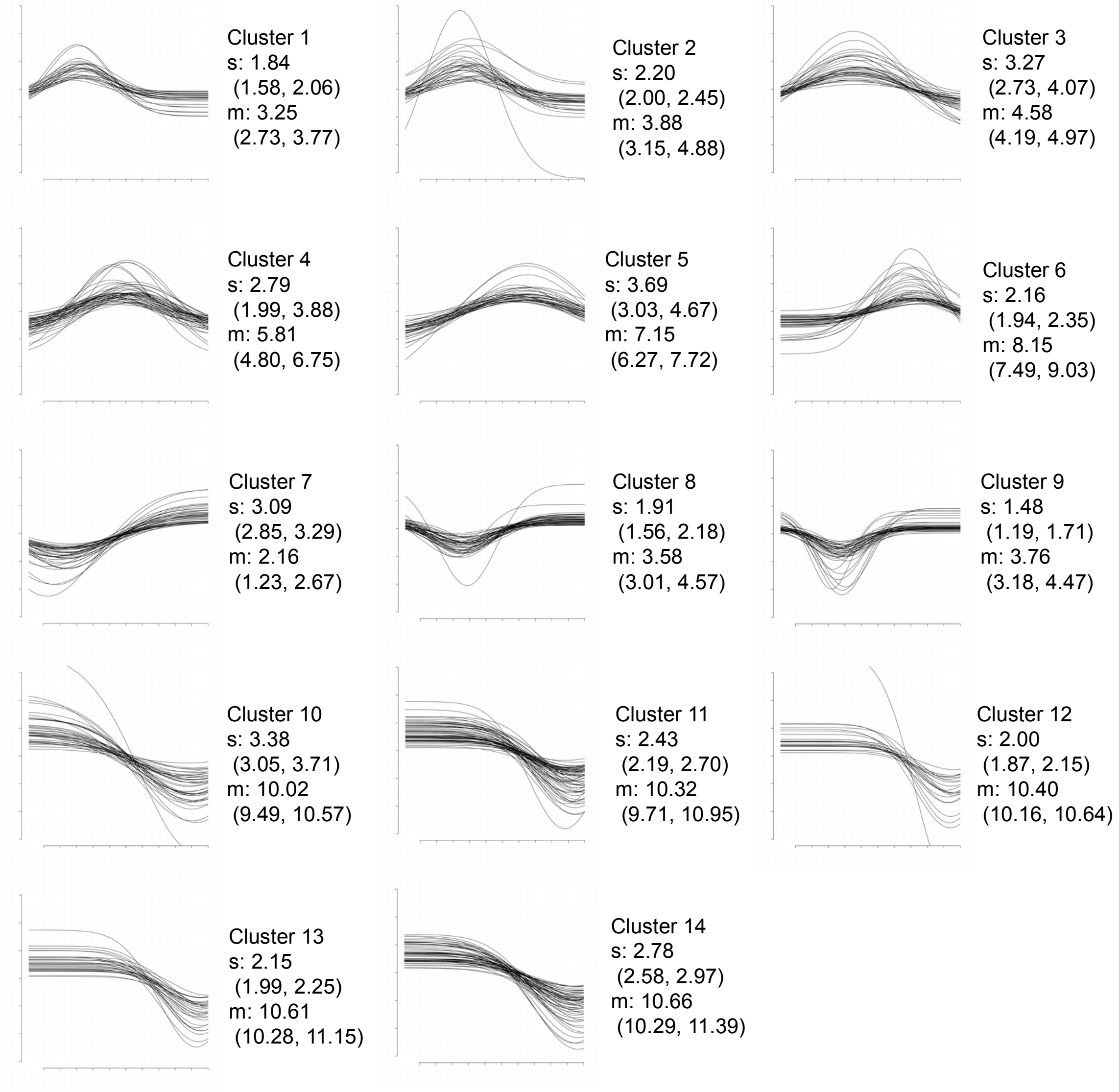


Figure 5.2: Gaussian shapes from the senescence experiment clustered on the mean (m) and standard deviation (s) parameters. Clusters 1-6 show curves that are increasing, while clusters 7-14 possess a decreasing response.

Table 5.2: Table showing the over-represented annotation terms for the Gaussian model in the senescence data.

Cluster	Mean parameter value	Cluster size	Annotation term	Count	p-value
1	s: 1.836, m: 3.253	28	cellular carbohydrate biosynthetic process	4	8.99e-05
2	s: 2.204, m: 3.876	34	inorganic phosphate transmembrane transporter activity	2	1.33e-04
3	s: 3.271, m: 4.576	36	ferric-chelate reductase activity	2	2.59e-05
			oxidoreductase activity, oxidizing metal ions	2	2.59e-05
4	s: 2.788, m: 5.808	48	sequence-specific	5	8.00e-03
			G-protein coupled photoreceptor activity	2	2.25e-05
5	s: 3.689, m: 7.150	38	sequence-specific	5	3.51e-03
			finger	6	6.22e-03
			zinc	6	6.29e-03
6	s: 2.158, m: 8.146	40	sequence-specific	7	6.82e-05
			factor	9	1.82e-03
			DNA	7	3.09e-03
			transcription	7	3.10e-03
7	s: 3.092, m: 2.164	41	nucleic acid metabolic process	13	1.58e-04
8	s: 1.909, m: 3.575	46	translation	5	6.80e-04
			RNA	6	2.65e-03
			DNA-directed RNA polymerase IV complex	2	2.31e-04
			DNA-directed RNA polymerase II, core complex	2	5.00e-04
			translation initiation factor activity	3	6.19e-04
9	s: 1.481, m: 3.756	43	GTPase activity	3	4.01e-04
			intracellular non-membrane-bounded organelle	7	6.46e-04
			ribosomal subunit	4	7.08e-04
			RNA binding	6	9.47e-04
10	s: 3.377, m: 10.015	40	chloroplast	15	5.21e-05
			photosynthesis, light reaction	3	3.66e-04
11	s: 2.432, m: 10.318	63	chloroplast	25	1.04e-07
			single-stranded RNA binding	3	1.41e-04
			plastid thylakoid	7	1.57e-04
12	s: 2.001, m: 10.398	22	chloroplast	9	1.63e-04
			cellular nitrogen compound biosynthetic process	5	3.40e-05
13	s: 2.150, m: 10.611	30	chloroplast	14	4.48e-07
			photosynthesis, light reaction	3	1.45e-04
14	s: 2.782, m: 10.662	56	chloroplast	19	1.65e-05

Some of the Gompertz1 clusters are shown in Figure 5.3, after clustering on the *5per&grad* parameters, which provided an indication of the time of first transcriptional change, and rate of change of gene expression. These clusters would identify genes that are activated or repressed at the same time point, as well as changing at the same rate. In total there were 36 clusters from this model. Clusters 1-6 contained genes with increasing shapes, and cluster 7-12 contained decreasing shapes. In the increasing responses, cluster 1 contained genes that were involved in RNA metabolism, cluster 2 with DNA binding and TF activity, cluster 3 with anthesis and ageing related genes, cluster 4 with ATP generation and clusters 5 and 6 involved in pectinesterase activity (Table 5.3). The last two clusters have a mean *5per* value of approximately 7, indicating that these genes are becoming activated in plants that are beginning to senescence (Figure 1.1). Pectinesterases are involved in breaking down the cell walls, possibly storing the breakdown products, or using them for cellular respiration (Breeze *et al.*, 2011). Clusters 2 and 3 may provide interesting sets of genes that could be involved in the activation of the senescence response and other ageing related stresses. For example, cluster 2 contained several TF related genes, including *WRKY58* (*AT3G01080*), which has been shown to act downstream of another WRKY TF that is involved in the senescence process (Miao *et al.*, 2004). In addition, there were a number of genes without a known biological function, including *HEAT SHOCK TRANSCRIPTION FACTOR B2B* (*HSFB2B*, *AT4G11660*), a member of a stress related TF family, and *AT5G28040*, a DNA-binding storekeeper protein-related transcriptional regulator. Storekeeper proteins are involved in regulating the expression of storage proteins in potatoes (Zourelidou *et al.*, 2002), and so may be involved in the transport and storage of macronutrients during senescence.

The genes with decreasing responses, as before, were primarily involved with photosynthesis. Several different sets of genes were found with different *5per* and *grad* parameter values, and so this could indicate that there are different parts of photosynthesis that are becoming down-regulated at different times and rates. For example, the genes in both clusters 7 and 12 are over-represented for terms relating to photosynthesis, although the average *5per* value for cluster 7 is 4.4, whereas the average *5per* for cluster 12 is 7.1.

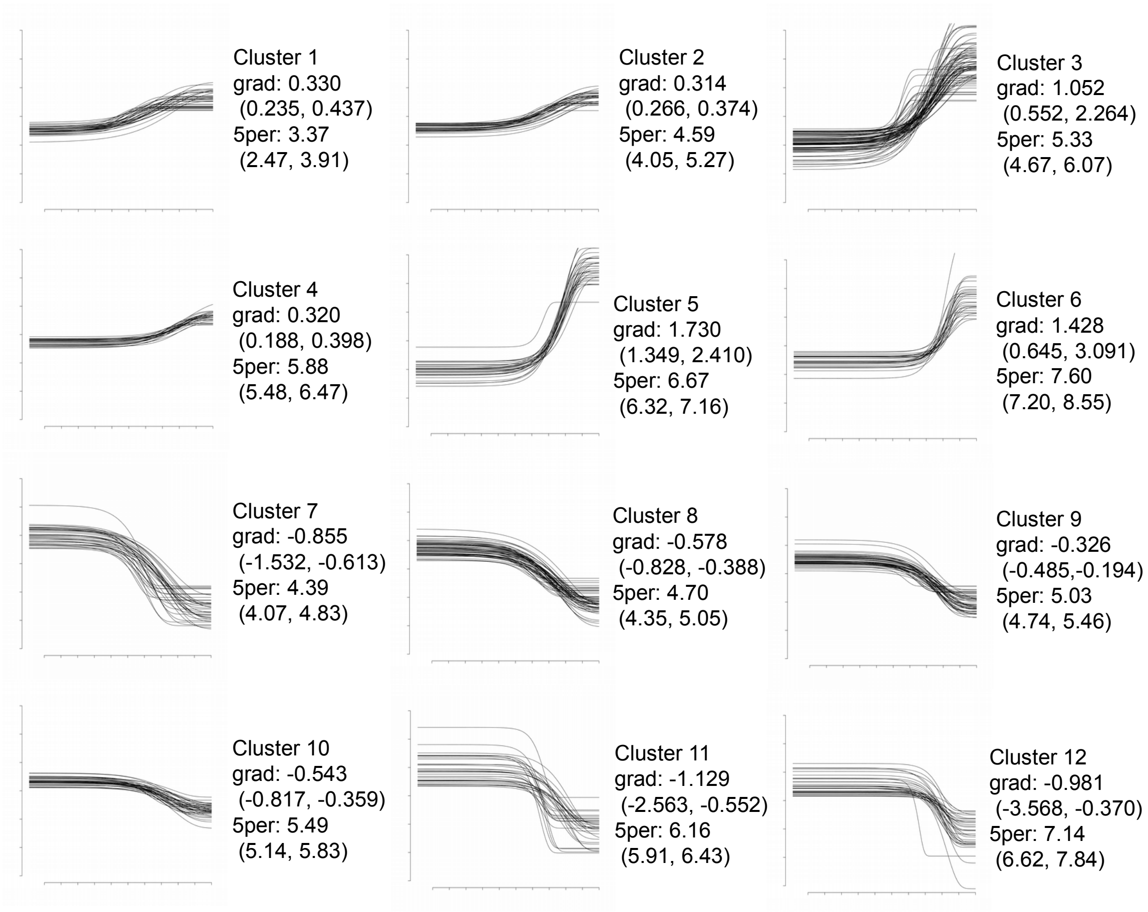


Figure 5.3: Selected clusters from the senescence data, clustered on the Gompertz1 shapes based on the 5% of maximum (*5per*) and gradient (*grad*) parameters. Clusters 1-6 show curves that are increasing, while clusters 7-12 possess a decreasing response.

Table 5.3: Table showing the over-represented annotation terms for the Gompertz1 model in the senescence data.

Cluster	Mean parameter value	Cluster size	Annotation term	Count	p-value
1	grad: 0.330, 5per: 3.373	36	nuclear mRNA splicing, via spliceosome	2	8.89e-04
2	grad: 0.314, 5per: 4.591	33	to zinc	5	1.37e-12
			DNA-binding	5	2.05e-03
3	grad: 1.052, 5per: 5.326	82	anthesis	22	1.16e-09
			differentiation	20	1.43e-07
			lignin biosynthetic process	3	6.14e-04
			aging	4	6.88e-04
4	grad: 0.320, 5per: 5.875	28	ATP	7	1.11e-03
			proton-transporting ATP synthase complex	2	1.84e-04
5	grad: 1.730, 5per: 6.672	31	anthesis	11	7.48e-07
			differentiation	11	2.07e-06
			pectinesterase activity	3	6.80e-04
6	grad: 1.428, 5per: 7.600	28	anthesis	10	8.74e-07
			indole-3-acetic acid amido synthetase activity	2	1.53e-05
			pectinesterase activity	4	1.92e-05
			auxin homeostasis	2	1.69e-04
			cell wall organization	3	9.39e-04
7	grad: -0.855, 5per: 4.394	40	chloroplast	10	8.19e-03
			thylakoid membrane	11	4.84e-12
			photosynthetic electron transport in photosystem I	5	2.57e-11
8	grad: -0.578, 5per: 4.699	59	chloroplast	26	1.86e-09
			photosynthesis	8	5.20e-09
			poly(U) RNA binding	3	9.53e-06
			photosystem I	3	1.54e-05
			photosynthesis, light harvesting	3	1.68e-05
9	grad: -0.326, 5per: 5.033	50	chloroplast	21	2.20e-08
			plastid thylakoid	8	1.25e-06
10	grad: -0.543, 5per: 5.491	41	chloroplast	16	2.64e-06
11	grad: -1.129, 5per: 6.162	34	chloroplast	14	1.07e-05
			photosynthesis, light reaction	3	2.78e-04
12	grad: -0.981, 5per: 7.142	35	photosystem I reaction center	2	7.90e-05
			photosynthesis	4	9.42e-05

Finally, selected clusters of the logistic model are shown in Figure 5.4, and as before the cluster analysis was performed on the *5per* and *grad* parameters. There were 44 clusters in total. Clusters 1-4 contained genes with an increasing response, and clusters 5-8 contained genes with a decreasing response. The genes with increasing responses showed enriched terms that were similar to those found in the Gompertz1 clusters, where genes with the early activation times were involved with stress responses and nucleic acid metabolism (clusters 1-2), and genes with later activation times were involved in the ageing response (cluster 3-4) (Table 5.4). Interestingly, in cluster 4, there were genes involved in auxin homoeostasis, such as *AUXIN UPREGULATED 3* (*AUR3*, *AT4G37390*). Auxins are a group of plant hormones, which are involved in a number of developmental processes in plants, including senescence (Osborne, 1959), so this set of genes may be involved in regulating the senescence responses. Again, the genes with decreasing responses were related to photosynthesis and chloroplasts. Like the Gompertz1 clusters above, there were several sets of genes that were found with different times of down-regulation and with different rates of change.

Table 5.4: Table showing the over-represented annotation terms for the logistic model in the senescence data.

Cluster	Mean parameter value	Cluster size	Annotation term	Count	p-value
1	grad: 0.356, 5per: 2.556	36	DNA metabolic process	5	2.34e-04
			cellular response to stress	5	3.45e-04
2	grad: 0.255, 5per: 3.51	58	domain-containing	5	4.78e-03
			RNA	6	6.67e-03
			nucleic	5	7.11e-03
3	grad: 1.595, 5per: 6.49	20	anthesis	8	6.36e-06
			expansion	7	7.76e-05
			differentiation	7	9.35e-05
4	grad: 1.011, 5per: 7.09	65	indole-3-acetic acid amido synthetase activity	3	1.88e-07
			mature	12	2.83e-07
			anthesis	15	1.82e-06
			auxin homeostasis	3	1.20e-05
5	grad: -0.368, 5per: 2.78	43	ATP	8	7.09e-03
			chloroplast	11	8.55e-03
			thylakoid lumen	3	6.31e-04
6	grad: -0.606, 5per: 3.01	51	chloroplast	20	3.46e-06
			small molecule metabolic process	11	8.88e-05
			generation of precursor metabolites and energy	5	1.34e-04
			response to temperature stimulus	6	2.79e-04
7	grad: -0.296, 5per: 5.94	43	chloroplast	12	5.85e-04
8	grad: -0.838, 5per: 7.30	43	positive regulation of catalytic activity	3	9.38e-06
			thylakoid	5	2.58e-04
			chloroplast	9	7.93e-04

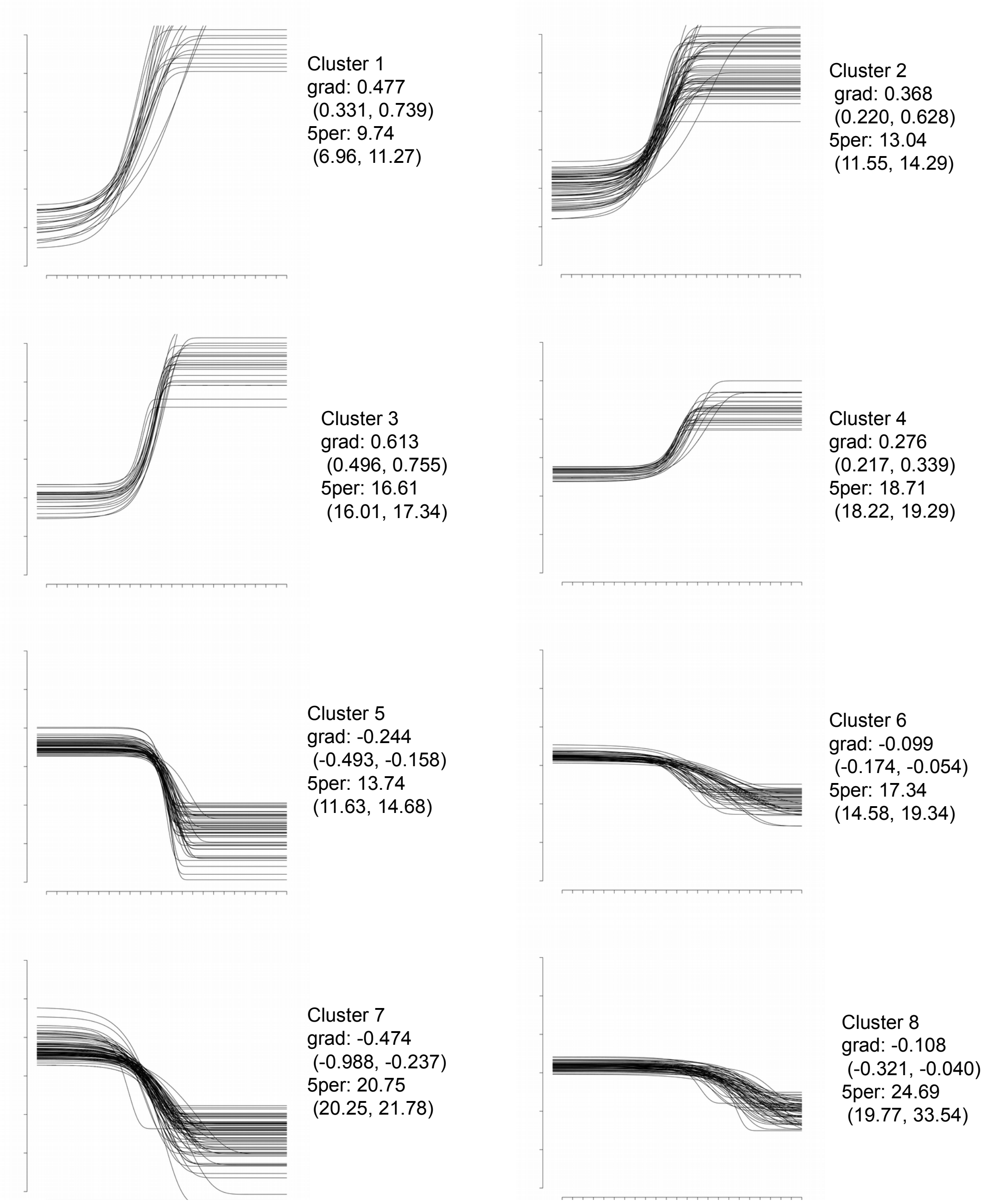


Figure 5.4: Selected clusters from the senescence data, clustering on the logistic shapes based on the 5% of maximum (*5per*) and gradient (*grad*) parameters. Clusters 1-4 show curves that are increasing, while clusters 5-8 possess a decreasing response.

In addition to the simultaneous parameter clustering, the meta-clustering approaches (Section 4.2.3) were also applied to the senescence data (Appendix D). In general, the over-represented annotation terms were the same as the simultaneous clustering, described above. A few interesting terms included those related to the response to abscisic acid (ABA) in the Gaussian models, using the sequential meta-clustering. These genes had a increasing response, reaching a maximum at around time point 10 before becoming down-regulated again. This is consistent with other findings, where it has been shown that there is an accumulation of ABA due to the up-regulation of ABA biosynthetic genes during senescence (Breeze *et al.*, 2011; Buchanan-Wollaston *et al.*, 2005; van der Graaff *et al.*, 2006). Another term that was not seen with the simultaneous parameter clustering is the presence of genes that were involved in ethylene mediated signalling pathways. These genes were again found using the Gaussian models, and sequential meta-clustering. The genes initially decreased until around time point 2 before becoming up-regulated. It has been shown that ethylene levels increase during the senescence process, due to the up-regulation of ET biosynthetic genes as the plant ages (van der Graaff *et al.*, 2006).

In the meta-clustering analyses, there were a greater number of clusters, with fewer members (approximately 70 clusters, with 20 genes in each cluster) compared to the clusters from the simultaneous clustering (approximately 40 clusters, with 35 genes in each cluster). While this smaller cluster size can help refine the significant annotation terms, if the cluster size is too small, it may become difficult to determine if an annotation is truly significant.

Figure 5.5 shows a summary of the processes that were identified at different times during the senescence process, and was primarily determined using the timing parameters from the different shapes (*5per* and m). By using these timing parameters together with the over-represented annotation terms, it was possible to determine when specific biological events were taking place. Generally up- or down-regulated genes were identified from the exponential model using the r and b parameters, and are shown in the boxes in the figure. Up-regulated genes were initially involved in metabolic processes, such as macronutrient metabolism or nucleic acid activity, before becoming more involved in stress responses. The down-regulated genes were primarily involved in chloroplast activity and photosynthesis. These results are similar to those found in the published results from the senescence time course (Breeze *et al.*, 2011). In particular, responses to water deprivation and pectinesterases were up-regulated at the same points. Down-regulated in both were genes involved in amino and nucleic acid metabolism, as well as several series of photosynthesis related genes. There were however a few differences. While both analyses identified chlorophyll related genes being down-regulated between time points 5-7, ShapeCluster did not identify photosynthesis related genes that were down-regulated at time point 3. It did however, identify photosynthesis genes that are down-regulated later, at time point 9. Other new discoveries include the identification of early up-regulation of ethylene signalling, auxin homeostasis at time point 7, and late ABA signalling.

5.2.2. Botrytis data

The same clustering process was also performed with the Botrytis data. In the Botrytis data, the x-axis shows the times that the samples were taken after infection with Botrytis, and is thus on the scale of hours (hours post-infection [hpi]). From the results in Section 3.9, it was observed that the most abundant models were the Gompertz1 (faster growth rate to the right of the midpoint), Gompertz2 (faster growth rate to the left of the midpoint), Gaussian, logistic and linear-exponential models. The cluster analysis was performed on these models and the over-represented annotation were identified, again using `GOstats` for the GO terms, and a hypergeometric test for the over-represented annotation terms (Section 2.1.4-2.1.5). Representative clusters from each model are shown here, and are presented in full in Appendix C.

Figure 5.6 shows selected clusters from the Gompertz1 model, clustered on the *5per&grad* parameters resulting in 43 clusters. In this clustering, it was found that the genes in cluster 1 were involved in hypoxia and oxygen deprivation, and cluster 2-4 were involved with stress and immune responses (Table 5.5). The hypoxia term in cluster 1 was unusual, and warranted further investigation. It has previously been shown that genes that are involved responses to hypoxia, may also be involved in defence against Botrytis (Zhao *et al.*, 2012). The hypoxia term was due to a FAD-binding Berberine family protein (*AT1G26380*), a gene encoding the cytochrome P450 enzyme CYP81F2 (*AT5G57220*), and a gene of unknown function (*AT2G23270*) which was found to be differentially expressed during hypoxia (Yang *et al.*, 2011). These genes were also found to be involved in ethylene response, which has been shown to be important in the plant's defence response to Botrytis (Windram *et al.*, 2012; Zhao *et al.*, 2012). Other genes of possible interest in this cluster were genes encoding a putative cytochrome P450, CYP71B22 (*AT3G26200*), as well as two WRKY TFs (WRKY45 and 75 - *AT3G01970* and *AT5G13080*, respectively). This family of TFs is known to be involved in a variety of biotic and abiotic stresses (Eulgem *et al.*, 2000; Pandey and Somssich, 2009), and it is possible that these genes may have a role in the plant defence response.

The decreasing responses, as with the senescence data, were primarily involved in chloroplast activity, particularly the genes contained in clusters 5 and 7. Interestingly, the mean *5per* values for these clusters are relatively far apart, having a difference of almost 10 hours. This suggests that there were two separate repressions of chloroplast activity taking place during the course of the experiment, or different aspects of photosynthesis being down-regulated at different times. Similarly, the genes in clusters 6 and 8 showed over-represented terms pertaining to ribosomes, transcription and translation. The mean *5per* values for these clusters were approximately 7 hours apart, again possibly suggesting separate sets of genes with different functions, and different points of down-regulation. However, the ranges of the *5per* parameters in these clusters are almost contiguous, possibly suggesting that the repression process may be continuous throughout the infection. Nonetheless, in all decreasing responses, it is likely that the

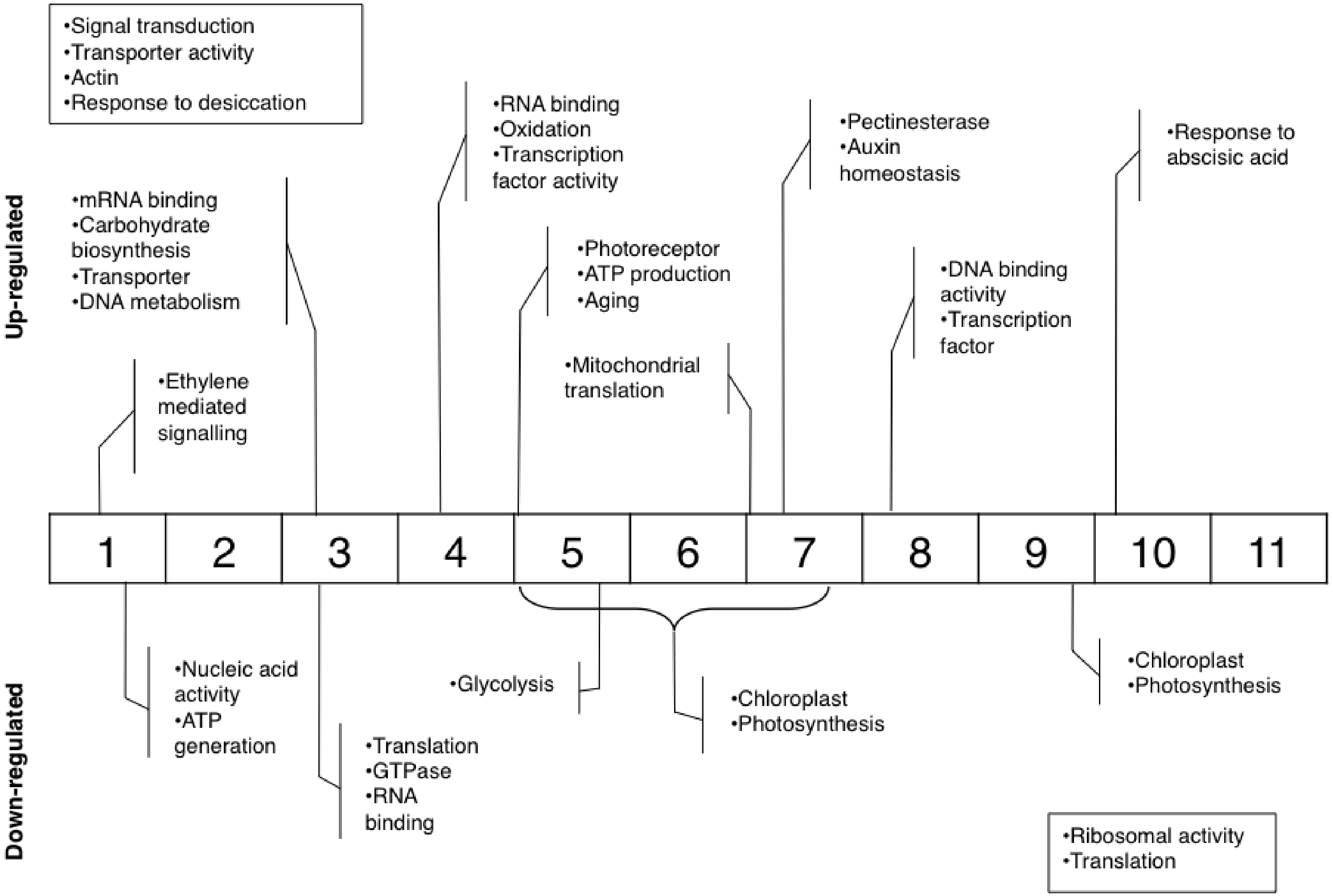


Figure 5.5: A simple timeline representation of the processes that are occurring during the senescence time course, as elucidated using the results from ShapeCluster. The bar in the middle represents the sampling time points. The up-regulated genes are shown above the time bar, and the down-regulated genes are shown below. Processes that were generally up- or down-regulated are shown in the boxes.

plants are succumbing to the Botrytis infection, which is causing the plant to divert energy from growth and maintenance processes, and focus on mounting a defence response.

Table 5.5: Table showing the over-represented annotation terms from the Gompertz1 models in the Botrytis data.

Cluster	Mean parameter value	Cluster size	Annotation term	Count	p-value
1	grad: 0.477, 5per: 9.741	22	cellular response to hypoxia	3	1.24e-06
			response to oxygen levels	3	1.46e-05
			oxygen binding	3	6.83e-04
			indole-containing compound biosynthetic process	2	7.35e-04
2	grad: 0.368, 5per: 13.040	60	Cytochrome	6	7.36e-06
			anthesis	9	1.85e-03
			response to stress	14	6.68e-04
3	grad: 0.613, 5per: 16.614	22	anthesis	5	1.34e-03
			response to stimulus	12	1.04e-04
			response to chitin	3	1.92e-04
4	grad: 0.276, 5per: 18.707	26	regulation of response to stress	3	3.43e-04
			activation of innate immune response	2	4.47e-04
			positive regulation of immune response	2	7.01e-04
5	grad: -0.244, 5per: 13.736	57	photosynthesis	5	2.98e-08
			chloroplast	25	1.06e-08
6	grad: -0.099, 5per: 17.344	50	Ribosomal	6	3.00e-05
			chromatin assembly or disassembly	3	4.39e-04
7	grad: -0.474, 5per: 20.745	81	chloroplast	38	1.16e-13
			translation	8	4.58e-04
			organelle	22	7.62e-04
			chloroplast thylakoid	7	7.95e-04
8	grad: -0.108, 5per: 24.692	64	structural constituent of ribosome	11	1.29e-09
			translation	12	1.11e-08

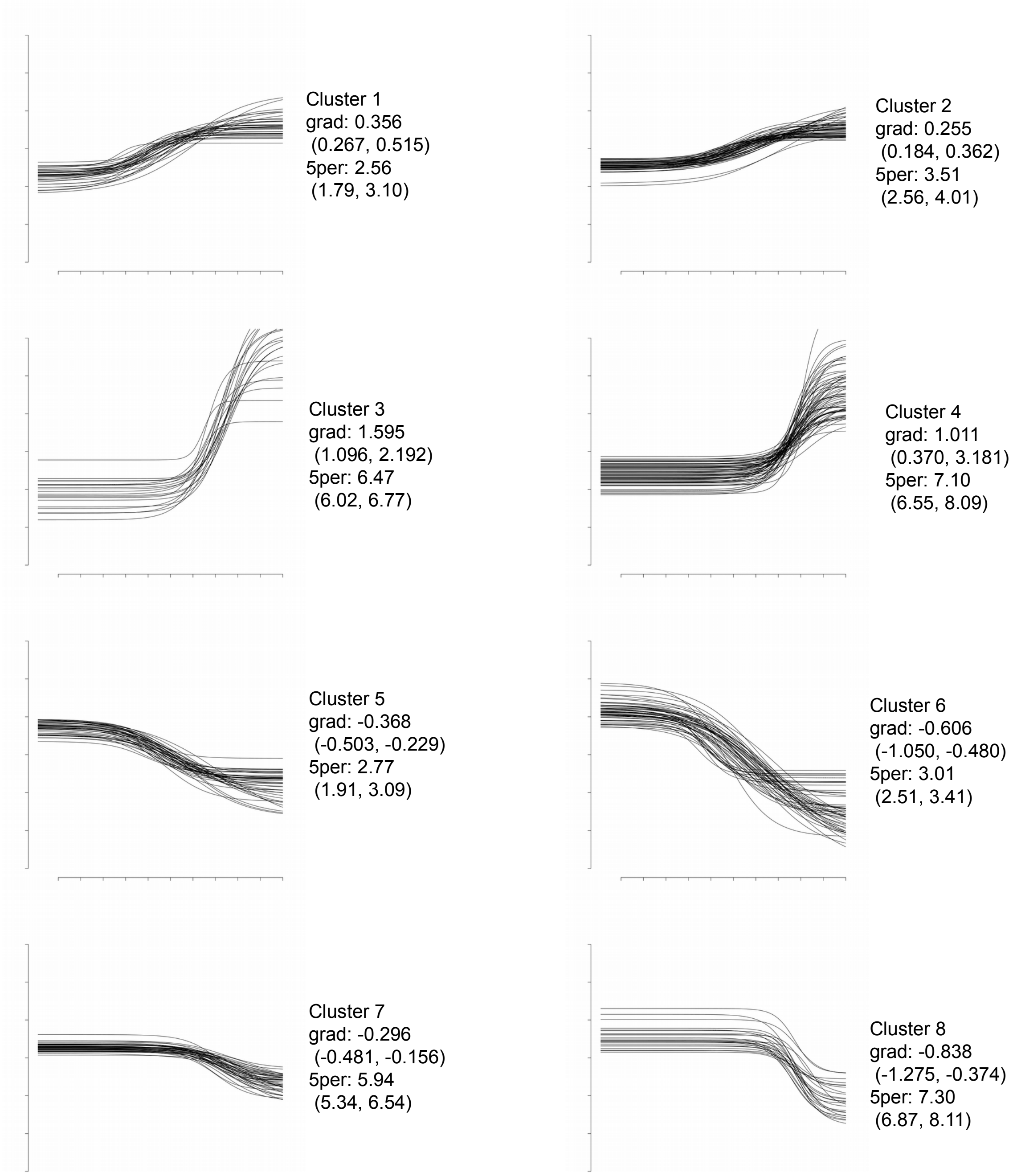


Figure 5.6: Selected clusters from the Botrytis data, clustering on the Gompertz1 shapes based on the 5% of maximum (*5per*) and gradient (*grad*) parameters. Clusters 1-4 show curves that are increasing, while clusters 5-8 possess a decreasing response. The x-axis shows the sampled data points (hours after infection with Botrytis), and the y-axis is the $\log_2$ gene expression level.

Figure 5.7 shows the clusters from the Gompertz2 models, again clustering with the *5per&grad* parameters, with a total of 66 clusters. While this model is similar to the Gompertz1 model, there were still some differences between the clusters that were formed. The genes in cluster 1 had annotation terms that were enriched for transporter activity, the genes in clusters 2 and 4 were enriched for response to chitin and stress, and the genes in cluster 3 were enriched for hormone metabolism (Table 5.6). Chitin is a characteristic component of fungal cell walls, and the detection of it by plants triggers a defence response (Windram *et al.*, 2012). The genes involved in this response become up-regulated at around 16 hpi, indicating the point at which the plant detects the pathogen. Shortly after, at around 17 hpi, a number of genes involved in indoleacetic acid metabolism are up-regulated. Indoleacetic acid is an auxin, which have been found to be important in stress responses, and have been implicated in plant defence (Llorente *et al.*, 2008; Windram *et al.*, 2012).

The decreasing shapes had similar over-represented terms to the clusters found in Gompertz2, where the genes in clusters 5 and 6 were involved in RNA processing, and the genes in clusters 7 and 8 were involved in chloroplast activity.

Table 5.6: Table showing the over-represented annotation terms from the Gompertz2 models in the Botrytis data.

Cluster	Mean parameter value	Cluster size	Annotation term	Count	p-value
1	grad: 0.075, 5per: 12.529	20	intrinsic to membrane	5	2.27e-04
2	grad: 0.433, 5per: 16.758	33	response to stress	10	5.34e-04
			response to chitin	3	6.55e-04
			multi-organism process	6	8.01e-04
3	grad: 0.044, 5per: 17.228	56	indoleacetic acid metabolic process	2	3.48e-04
			hormone metabolic process	3	9.66e-04
4	grad: 0.763, 5per: 19.910	38	anthesis	6	5.74e-03
			response to chitin	4	4.59e-05
5	grad: -0.083, 5per: 17.000	49	Ribosomal	5	4.77e-04
			RNA processing	5	6.35e-04
6	grad: -0.108, 5per: 17.735	43	membrane-enclosed lumen	6	1.79e-04
7	grad: -0.405, 5per: 18.367	46	chloroplast	18	8.28e-06
			tRNA aminoacylation for protein translation	3	2.34e-04
			amino acid activation	3	2.69e-04
			photosynthesis, light reaction	3	6.86e-04
			translation elongation factor activity	2	8.70e-04
8	grad: -0.430, 5per: 21.124	47	chloroplast thylakoid membrane	9	1.75e-08
			thylakoid membrane	9	2.88e-08

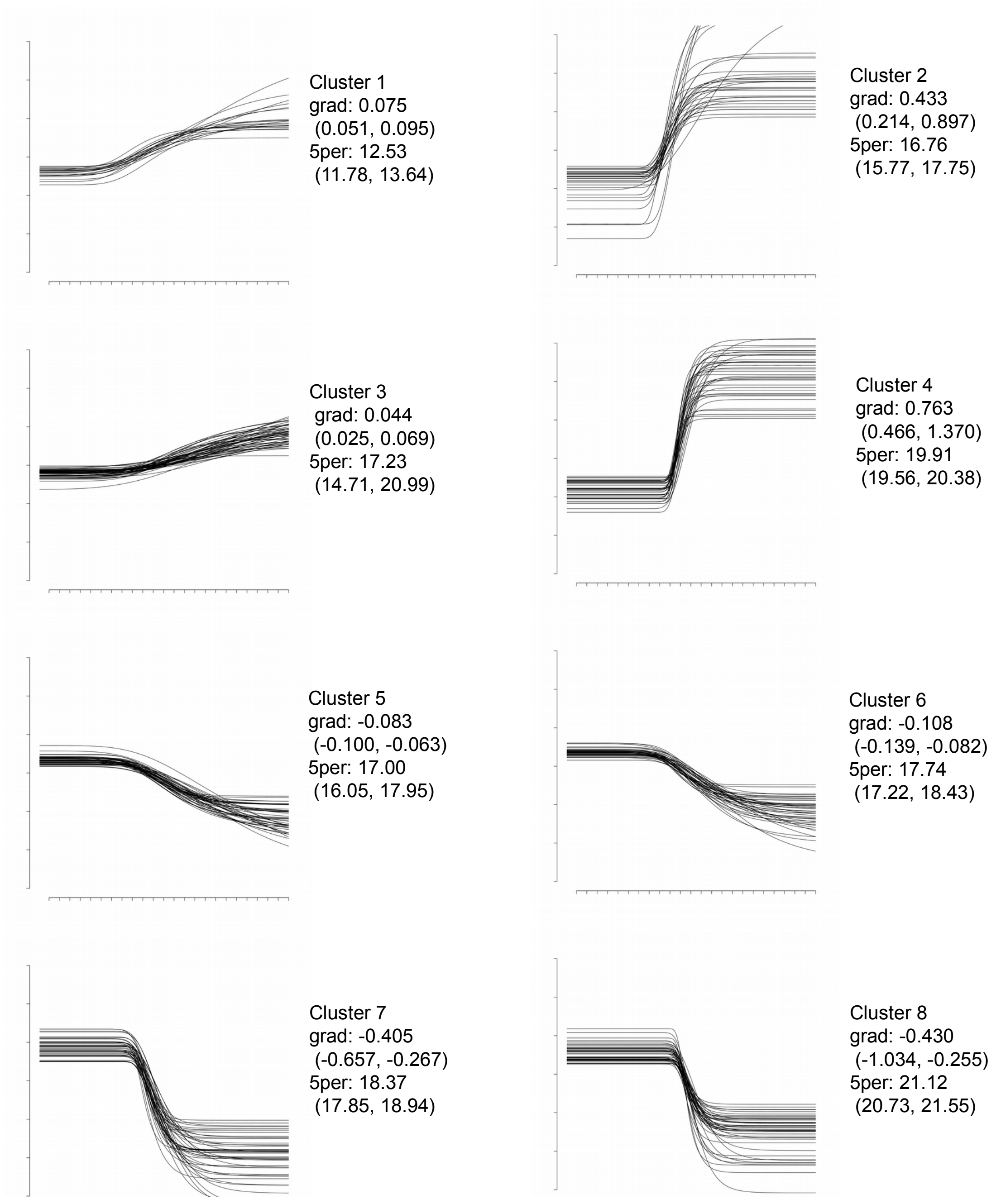


Figure 5.7: Clusters from the Botrytis data, clustering on the Gompertz2 shapes based on the 5% of maximum (*5per*) and gradient (*grad*) parameters. Clusters 1-4 show curves that are increasing, while clusters 5-8 possess a decreasing response.

Figure 5.8 shows the some of Gaussian responses, clustered on the *m&s* parameters, out of a total of 64 clusters. Increasing responses included cluster 1 with genes that were involved in chloroplast activity, cluster 2 contained a number of terms related to transcription, cluster 3 contained genes involved in jasmonic acid synthesis, cluster 4 contained genes involved in defence response and ATP generation, and cluster 5 contained genes involved in transport (Table 5.7). Jasmonic acid is known to be involved in the Botrytis defence response (Windram *et al.*, 2012), and the ATP generation could be associated with the defence response. Cluster 2 contained *CIRCADIAN CLOCK ASSOCIATED 1* (*CCA1*, *AT2G46830*), a gene that is involved in the circadian clock, and may be involved in immune responses (Zhang *et al.*, 2013), as well as *REVEILLE 1* (*RVE1*, *AT5G17300*), which encodes a MYB-like TF that is involved in both the circadian clock and auxin signalling pathways (Rawat *et al.*, 2009). These genes are normally cyclic, and the Botrytis infection resulted in s dampened oscillation (Windram *et al.*, 2012). In addition, there were a number of genes encoding TFs in the cluster with unknown function, including *AT1G71030*, a MYB-like TF, and *AT3G09320*, *AT3G60300*, *AT4G38960* and *AT1G49200*, all zinc-finger TFs. These genes may all be involved in regulating the response to pathogens, possibly with crosstalk with the circadian clock.

The decreasing shapes showed similar annotation terms as before, with genes in cluster 6 being involved with chloroplast activity, and the genes in clusters 7 and 8 involved with ribosomal activity.

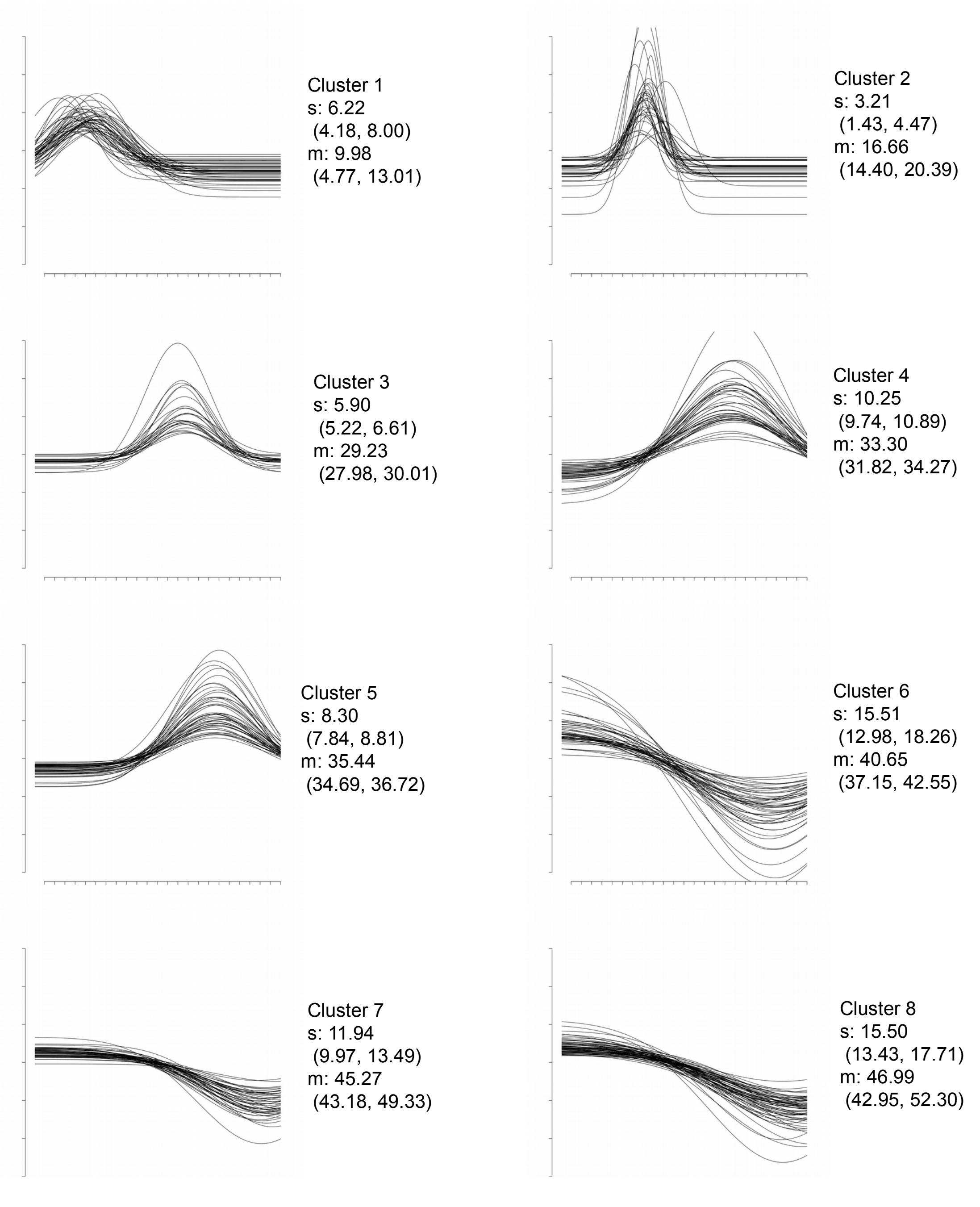


Figure 5.8: Clusters from the Botrytis data, clustering the Gaussian shapes on the mean (m) and standard deviation (s) parameters. Clusters 1-5 show curves that are increasing, while clusters 6-8 possess a decreasing response.

Table 5.7: Table showing the over-represented annotation terms from the Gaussian models from the Botrytis dataset.

Cluster	Mean parameter value	Cluster size	Annotation term	Count	p-value
1	s: 6.217, m: 9.983	49	thylakoid	6	2.87e-06
			chloroplast	18	9.41e-06
2	s: 3.205, m: 16.655	35	Homeodomain-like	6	1.20e-06
			zinc	7	5.07e-04
			DNA-binding	5	2.11e-03
			negative regulation of circadian rhythm	2	2.17e-06
			sequence-specific DNA binding transcription factor activity	11	1.55e-05
			response to ethylene stimulus	4	9.41e-05
			regulation of gene expression	10	1.83e-04
			cellular nitrogen compound metabolic process	13	5.77e-04
			response to gibberellin stimulus	3	7.39e-04
			response to auxin stimulus	4	8.03e-04
			negative regulation of transcription, DNA-dependent	3	8.32e-04
3	s: 5.902, m: 29.231	25	response to jasmonic acid stimulus	3	7.75e-04
4	s: 10.251, m: 33.304	44	DNA-binding	5	9.66e-03
			ATP biosynthetic process	3	2.07e-04
			response to fungus	4	5.50e-04
5	s: 8.296, m: 35.439	46	transporter	8	1.12e-04
			disaccharide transmembrane transporter activity	2	1.40e-04
6	s: 15.508, m: 40.653	49	chloroplast	17	2.40e-05
			cellular response to cold	2	5.51e-04
			response to cytokinin stimulus	3	5.96e-04
7	s: 11.944, m: 45.268	47	translation	6	6.85e-05
			cytosolic large ribosomal subunit	4	8.26e-05
8	s: 15.496, m: 46.993	64	biogenesis	7	3.63e-08

The clusters from the logistic model bear a great deal of similarity to the previous results from the Gompertz models (Figure 5.9, Table 5.8). Since these models have similar shapes, many of the genes in the clusters are the same (approximately 30% of the genes that fitted the logistic model also fitted a Gompertz model - Table 3.3). A total of 54 clusters were found from this cluster analysis. Much like the previous results, in the up-regulated responses, there were clusters with genes involved in hypoxia, stress, and auxin responses. With the down-regulated responses, again the genes in the clusters were primarily involved in photosynthesis and protein translation.

Table 5.8: Table showing the over-represented annotation terms from the logistic models in the Botrytis dataset.

Cluster	Mean parameter value	Cluster size	Annotation term	Count	p-value
1	grad: 0.496, 5per: 12.50	25	cellular response to hypoxia	3	1.65e-06
			oxygen binding	4	3.69e-05
			response to stress	9	2.57e-04
2	grad: 0.270, 5per: 15.838	31	peroxidase activity	4	9.71e-06
			response to stress	9	9.07e-05
			oxidoreductase activity	4	1.97e-04
			response to other organism	5	9.82e-04
3	grad: 0.629, 5per: 18.577	39	auxin:hydrogen symporter activity	3	1.31e-06
			response to chitin	4	5.73e-05
			transmembrane transport	5	8.03e-04
4	grad: 0.089, 5per: 21.333	51	dehydrogenase	6	2.06e-05
			organelle membrane	9	1.61e-04
			oxidation	6	2.51e-04
			reduction	6	3.32e-04
			plant-type cell wall organization	3	5.21e-04
5	grad: -0.088, 5per: 14.835	64	Ribosomal	5	1.43e-03
			translation	5	2.39e-03
			DNA-binding	7	4.20e-03
6	grad: -0.151, 5per: 16.096	47	dormancy	5	4.17e-06
			development	8	1.02e-05
			protein metabolic process	15	6.54e-04
7	grad: -0.466, 5per: 21.015	60	chloroplast	24	1.38e-07
			thylakoid part	9	5.60e-07
			chloroplast thylakoid membrane	8	1.50e-06
			chloroplast stroma	10	7.33e-06
8	grad: -0.125, 5per: 27.344	20	structural constituent of ribosome	7	4.99e-09
			intracellular non-membrane-bounded organelle	9	4.03e-08
			ribosomal subunit	6	6.71e-08
			translation	7	2.03e-07

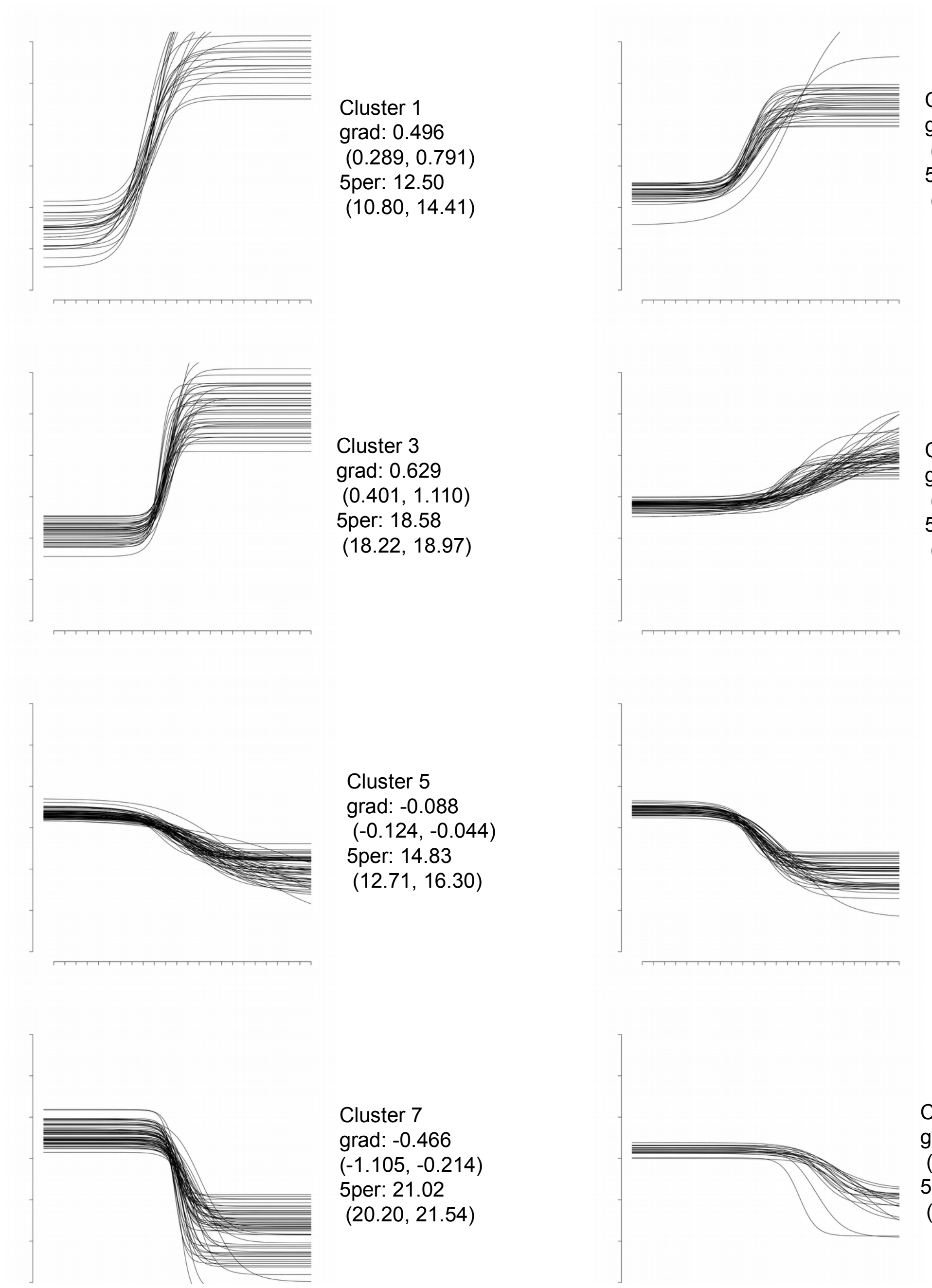


Figure 5.9: Clusters from the Botrytis data, clustering on the logistic shapes based on the 5% of maximum (*5per*) and gradient (*grad*) parameters. Clusters 1-4 show curves that are increasing, while clusters 5-8 possess a decreasing response.

Finally, Figure 5.10 shows a few clusters from the linear+exponential model, out of a total of 12 clusters. A number of different shapes are possible with this model, which are described in Section 3.2. In this case, the curves were clustered using the time point where the shape changes from exponential to linear (or *vice versa*) (the *linpnt* parameter), as well as the gradient of the linear portion (c). The *linpnt* parameter acts as a timing parameter, much like the *5per* parameter in the sigmoid models, or the m parameter for the Gaussian models. In this case, *linpnt* provides an indication of the time point at which gene expression is changing from one response shape to another, possibly indicating the activation of a regulator, or the response to some biological stimulus. The c parameter acts as a rate of change of gene expression parameter, describing the rate at which genes are responding to the previous stimulus. Cluster 1 represents an expression profile that mostly increases exponentially and decreases to a linear down-regulation response, and contains genes that are primarily involved in protein translation. Cluster 2 represents an expression profile that mostly decreases exponentially, before increasing linearly, and contains genes that are involved in DNA binding and topoisomerase, such as *DNA GYRASE A* (*GYRA*, *AT3G10690*) and *AT4G31210*. These genes encode topoisomerases, which are involved in unwinding the DNA so that it can be transcribed. The cluster also contained a number genes that encode TFs, such as *WRKY26* (*AT5G07100*), and *GATA TRANSCRIPTION FACTOR 29* (*GATA29*, *AT3G20750*). These TFs could be responsible for activating the genes necessary for the defence response, which become more active as the infection progresses. Cluster 3 represents an expression profile that increases linearly, and then decreases exponentially. These genes are involved in transporter activity, and are down-regulated relatively late in the time course (around 40 hpi). It could be that the expression of these genes is being suppressed as the plant succumbs to the Botrytis infection. Finally cluster 4 represents an expression profile where the expression decreases linearly, and then increases exponentially. The genes in this cluster are involved in the nucleolus and DNA binding, again possibly representing TF binding.

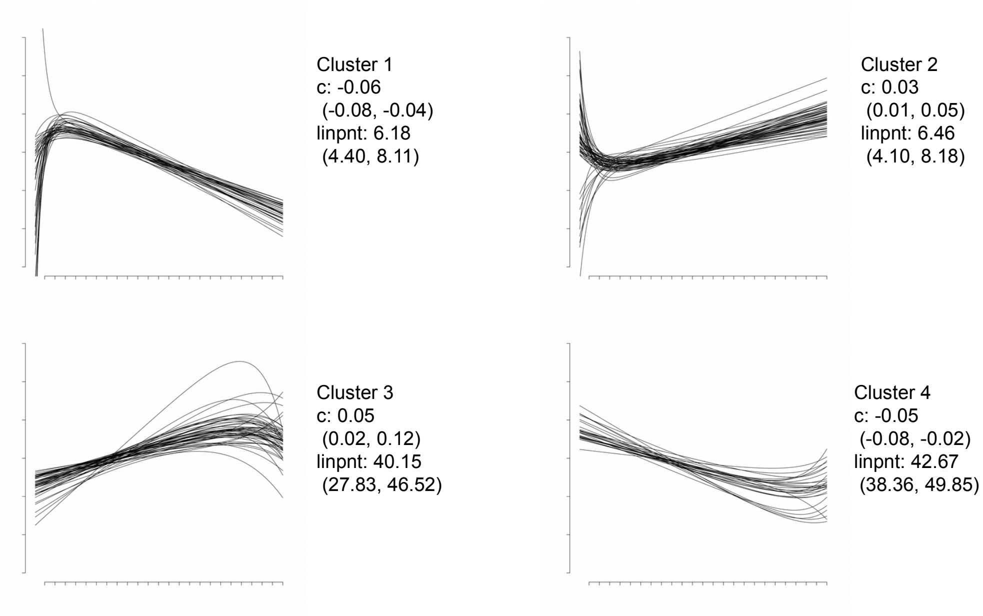


Figure 5.10: Linear+exponential models from the Botrytis experiment, clustered on the *linpnt* (x-value were the shape becomes linear) and c (gradient of the linear portion).

Table 5.9: Table showing the over-represented annotation terms from the linear+exponential models.

Cluster	Mean parameter value	Cluster size	Annotation term	Count	p-value
1	c: -0.057, linpnt: 6.18	37	translation	6	1.78e-05
			rRNA binding	2	3.47e-04
			ribosome	5	4.44e-04
2	c: 0.034, linpnt: 6.463	58	to zinc	5	4.43e-11
			activity binding	5	1.11e-05
			DNA topoisomerase activity	2	6.28e-04
3	c: 0.048, linpnt: 40.15	52	transporter	6	4.75e-03
			respiratory chain complex I	3	2.69e-04
4	c: -0.046, linpnt: 42.67	28	intracellular organelle lumen	5	5.38e-04
			nucleolus	4	6.00e-04

As with the senescence dataset, the meta-clustering on the above models revealed the same over-represented annotation terms (Appendix D). One difference that was identified in the Gaussian model using the sequential meta-clustering was a cluster that was down-regulated late in the time course (after 40 hpi). The genes in this cluster were involved in biotic stimulus and innate immune response. The genes responsible for these terms included *ABSCISIC ACID RESPONSIVE ELEMENT-BINDING FACTOR 1* (*ABF1*, *AT1G49720*), an ABA responsive element-binding factor, which is required for ABA signalling, as well as *RESISTANT TO P. SYRINGAE 5* (*RPS5*, *AT1G12220*), which encodes a disease resistance protein of the CC-NBS-LRR family. ABA induced signalling is activated during *Pseudomonas syringae* pv. tomato infection, and results in increased susceptibility (de Torres-Zabala *et al.*, 2007), and some members of the (CC-NBS-LRR class) family confer resistance to *P. syringae* (Simonich and Innes, 1995). Since these genes are down-regulated, it suggests that these genes are involved in the responses to biotrophic pathogens, which work antagonistically to necrotrophic pathogens, such as Botrytis (Section 1.1).

Figure 5.11 shows a timeline of the processes that are occurring over the course of the Botrytis time course, and determined using the timing parameters from the clusters. Over-represented terms from genes which fitted the exponential model are shown in the boxes. The majority of the up-regulated genes are activated between 10-20 hpi, and suggests that this is when the plant detects the pathogen and mounts a defence response. Of note are the activation of genes involved in hypoxia, cytochrome activity, ethylene activity, auxin binding, peroxidase activity, response to jasmonic acid and response to chitin, all of which are involved in the defence response, and are activated before 20 hpi. Genes that are up-regulated later in the time course include genes involved in water deprivation and proteolysis. The down-regulated genes are involved in ribosomes, translation and photosynthesis. As with the senescence data, there are different points at which the photosynthesis and chloroplast-related genes are down-regulated, possibly indicating different parts of the photosynthetic machinery being deactivated in phases. Again, these results are similar to those identified by Windram *et al.* (2012), where it was also shown that the majority of genes are differentially expressed between 14-18 hpi. The over-represented terms are approximately the same, except that the ShapeCluster analysis identified the early hypoxia response.

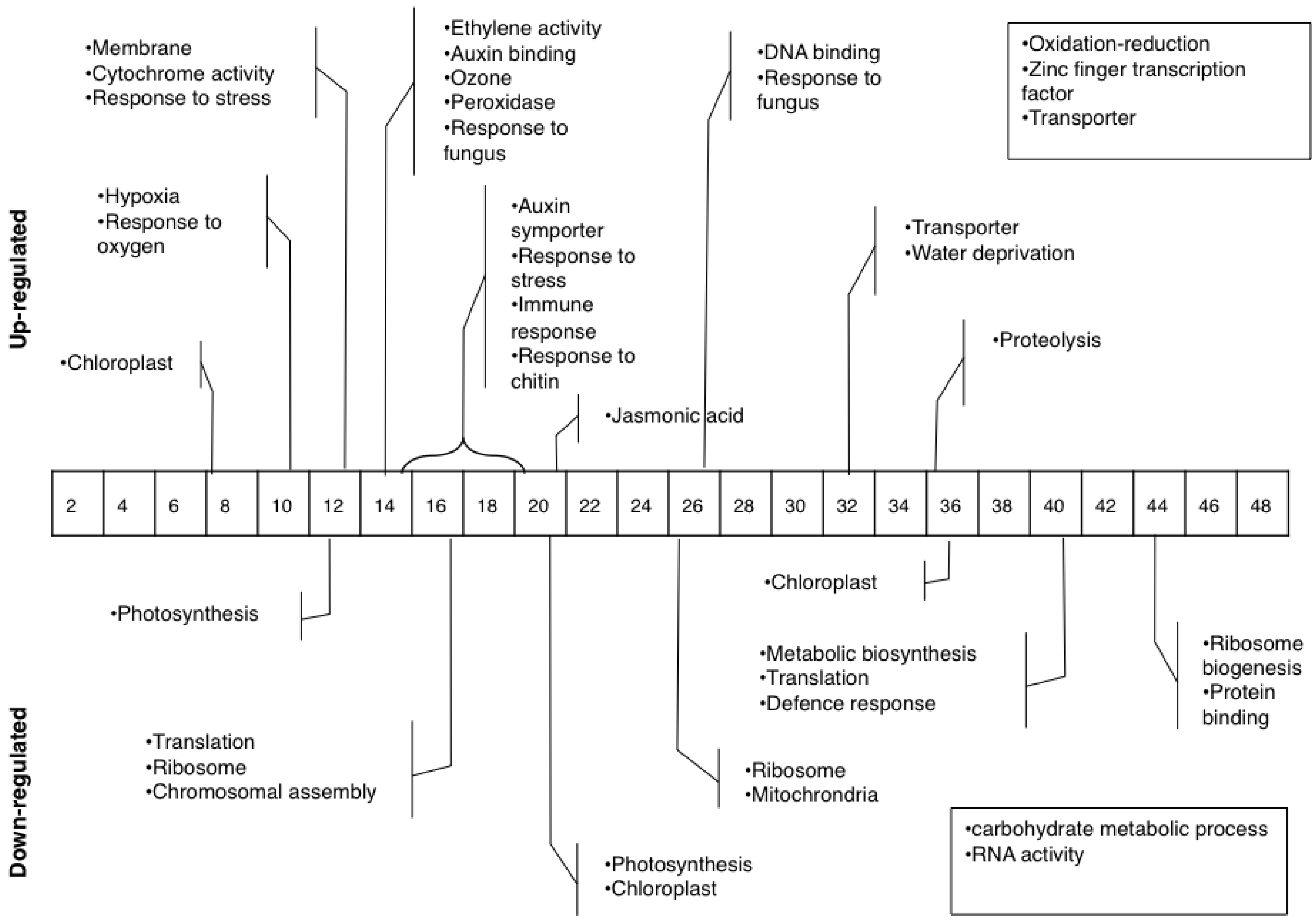


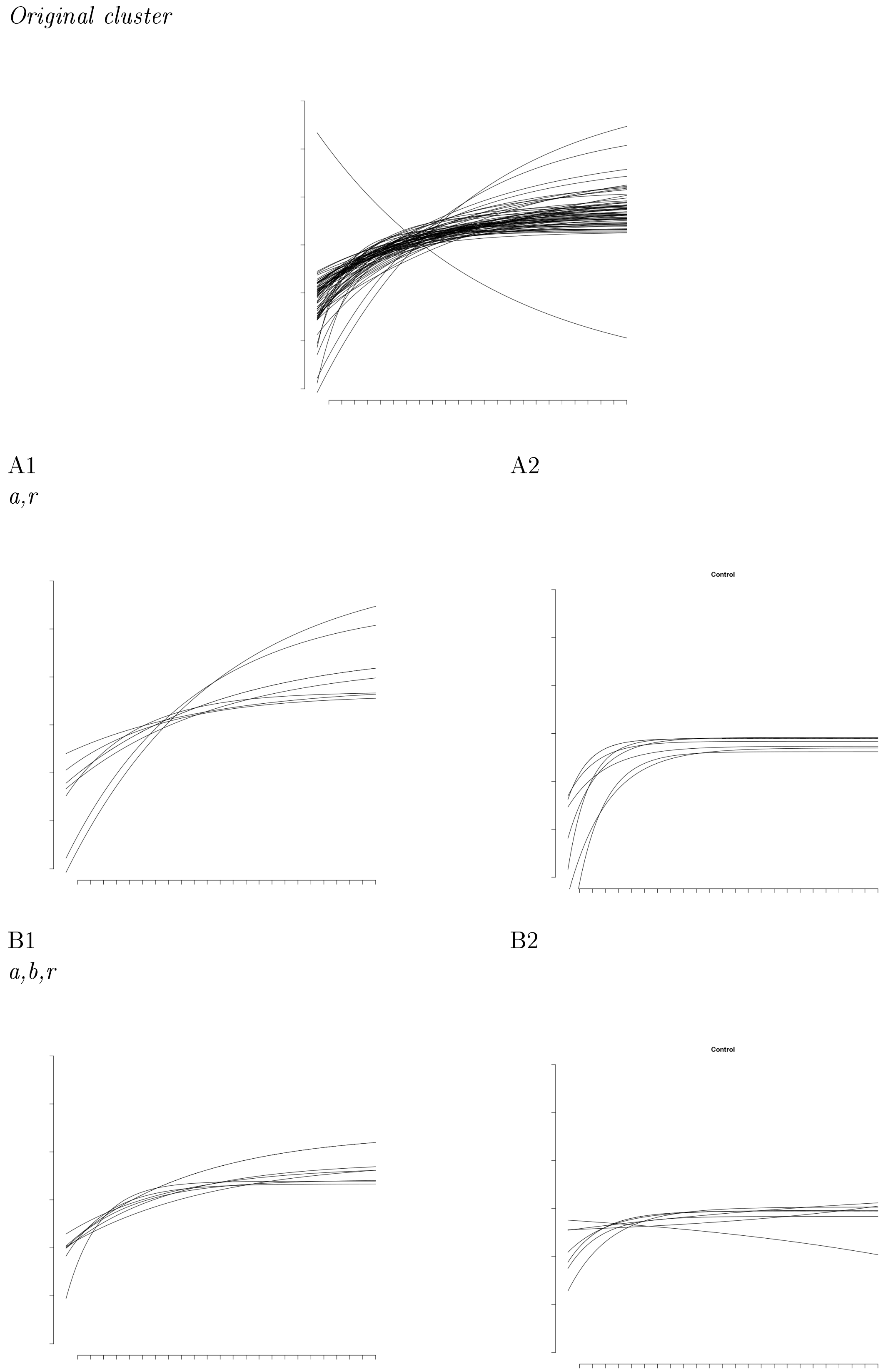
Figure 5.11: A timeline representation of the processes that are occurring during the Botrytis time course, as elucidated using the results from ShapeCluster. The bar in the middle represents the sampling time points. The up-regulated genes are shown above the time bar, and the down-regulated genes are shown below. Processes that were generally up- or down-regulated are shown in the boxes.

5.3. Cluster using control data

Since it is expected that there would be distinct differences in the expression profiles between the control and treated samples, the Botrytis data was used to illustrate a further subdivision using the control information. The hypothesis was that genes that are involved in the same biological process may have similar gene expression profiles under the control conditions, in addition to responding similarly in the treatment. Therefore, genes were sorted based on the expression profile of the control state. The methodology is described in Section 4.2.4. In brief, after clustering, the genes may be grouped based on the model that fitted the control data for the same gene. If the fitted model for the control and treated sets are the same, it is possible to identify which of the parameters are similar. Figures 5.12 and 5.13 show some results of this form of clustering. In Figure 5.12, an example from the exponential clustering is shown. Here, a cluster can be divided into subclusters (A1 and B1), based on the shape and parameters of the models fitted to the control data (A2 and B2). The control model for the genes in these subclusters is also the exponential model, so it is possible to compare the parameters between the treatment sets. In the case of A, the a and r parameters were similar in both treatment sets, meaning the starting expression level and rate of change of gene expression were similar, but the concavity differed between the treatment sets. In Figure B, all the parameters were found to be similar. The original cluster had over-represented GO terms for oxidoreductase activity and cation transmembrane transporter activity. However, using this subclustering revealed that the genes in A are involved in energy production and transmembrane transport. The genes in subcluster B were involved in defence response to fungus. Thus, using the control information to subcluster revealed new annotation terms, and thus aid in identifying genes that may be of interest.

In contrast, Figure 5.13 shows a case where the subclusters did not have any similar parameters, or were a different model fit, to the original cluster. This cluster is the same as cluster 1 in Figure 5.6, where the shapes were from the Gompertz1 model, clustered on the *5per&grad* parameters, and contained genes involved in hypoxia and oxygen deprivation. By subclustering based on the control model fit, subclusters were found where there were no similar parameters (i), the control model was the linear model (ii), and the control model was Gaussian model (iii). Subclusters (i) and (iii) were found to be involved in the jasmonic acid and ethylene signalling pathways, and related to stress and other external stimuli, respectively. Subcluster (ii) mostly contained genes that were genes of unknown function. Again, this shows that by including the control information, it is possible to find the significant annotation terms.

In addition to filtering the clusters to find the genes that are enriched for a particular annotation, it may be possible to use this methodology to identify genes that are differentially expressed. For example, if the control shape is linear and roughly unchanging, and the treatment shape has a distinct response, this may indicate that the gene is differentially expressed compared to the control.



Original cluster

A1
a,r

A2

B1
a,b,r

B2

Figure 5.12: Example of clustering using the control data from the Botrytis dataset, where the control model is the same as the treated data. The original cluster from the exponential cluster analysis is shown on the top. The treated data is shown in the figures on the left, and the control data for each cluster is shown to the right. The parameters that are similar between the treatment sets are shown above the clusters.

5.4. Comparison to SplineCluster

SplineCluster (Heard *et al.*, 2006) was also used to cluster the Arabidopsis datasets. To provide a fairer comparison, the full set of 23 802 genes was filtered using the R^2_a, R^2_{LoF}, F-test p-value, and standard error of the parameter estimate filters (Section 3.9). This resulted in a set of 8216 genes in the senescence dataset, and 5303 genes in the Botrytis dataset. After applying SplineCluster to these genes, using the default parameters, this resulted in 98 clusters for the senescence dataset and 26 clusters for the Botrytis dataset. In order to determine the quality of the clusters formed by the different approaches, the Biological Homogeneity Index (BHI) was used, described in Section 2.1.8. The clusters from the senescence dataset produced a BHI value of 0.284, while the clusters from the Botrytis dataset produced a BHI value of 0.299.

For ShapeCluster, the cluster analyses were performed using a range of regression models and different parameter combinations, namely exponential (r,b), Gaussian (m,s), Gompertz1 (*5per,grad*), Gompertz2 (*5per,grad*), and logistic (*5per,grad*). These parameters refer to aspects of the timing of gene expression responses, as well as the rate of change in gene expression, and were determined in the yeast analyses to be the most informative parameters. The BHI scores for these different sets of clusters are shown in Table 5.10 for the senescence dataset and Table 5.11 for the Botrytis dataset. The clustering with control information was also performed for the Botrytis dataset, and is included in Table 5.11 in italics.

From these tables, it can be seen that the BHI scores from ShapeCluster are slightly lower than the score from SplineCluster in most of the cluster analyses, except for the decreasing sigmoid functions. Since these clusters are generally involved in photosynthesis, it indicates that the clusters formed by ShapeCluster are grouping these genes together better than SplineCluster. It is possible that the lower BHI scores are due to the poorer annotation level in Arabidopsis, which is reflected in the lower overall scores in BHI as compared to the yeast results. For the Botrytis data, it can also be seen that by subclustering the genes based on the control shape, a higher BHI score can generally be obtained.

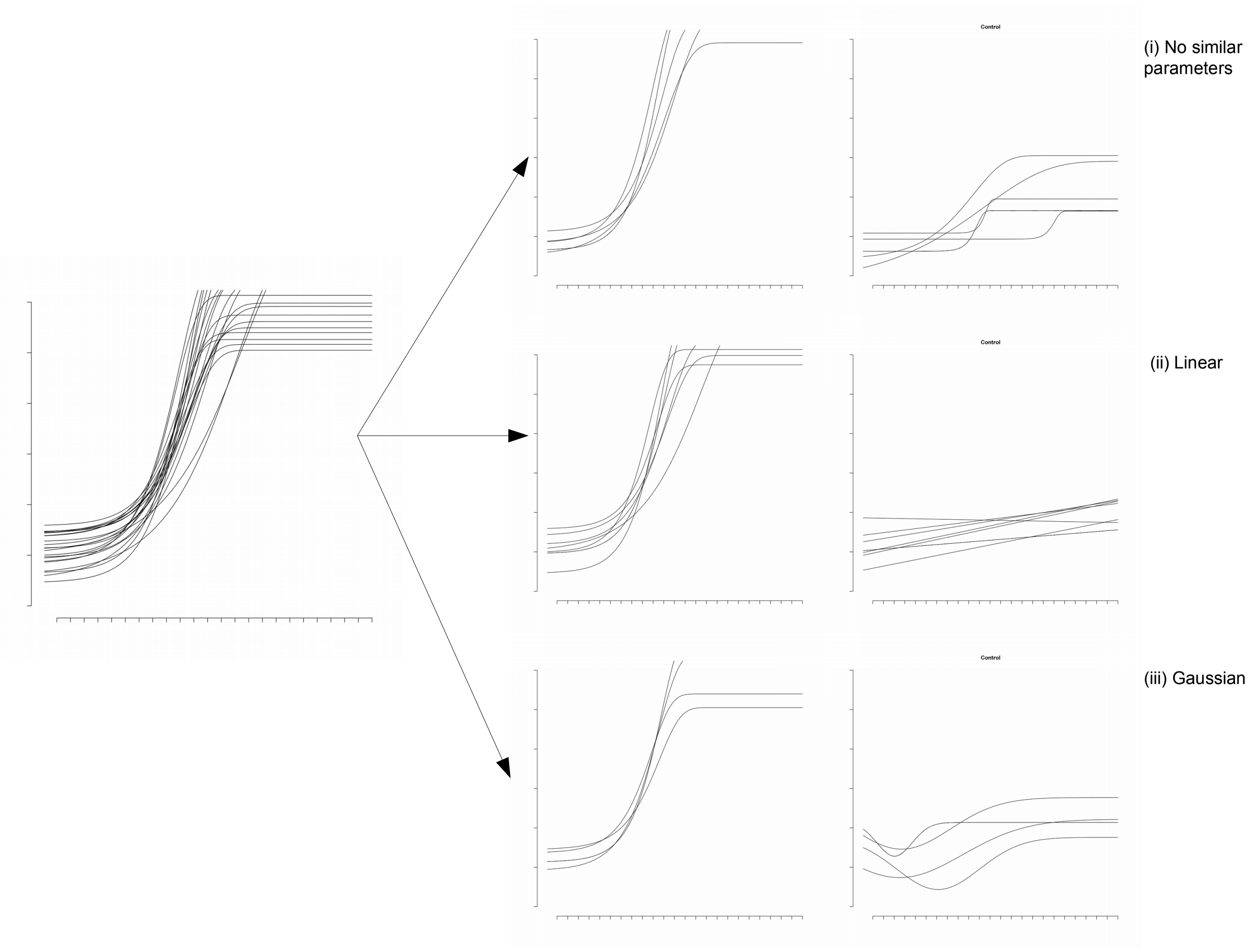


Figure 5.13: An example from the Botrytis dataset of the clustering using the control data. The figure on the left is cluster 1 in Figure 5.6 (Gompertz1), and the figures on the right show some of the different subclusters obtained. (i) shows where the control model was the same as the treatment model, but there were no similar parameters, and (ii) and (iii) show where the control models were different (linear and Gaussian, respectively).

Table 5.10: Table of all the BHI scores for the various clusterings performed on the senescence data. *Single* refers to clustering performed on a single parameter, *Simul* is the clustering on multiple parameters, using the simultaneous parameter clustering, and *Meta* refers to the two types of meta-clustering. The meta-clustering was performed on the same parameters as the simultaneous parameter clustering. *Sequential* is the clustering of one cluster followed by another, and *Cross* is the cross-clustering where clusters were identified based on the genes found from overlapping clusterings.

	Exponential	Gaussian	Gompertz1	Gompertz2	Logistic
Single	*b* 0.233	*m* Inc: 0.267 Dec: 0.263	*5per* Inc: 0.219 Dec: 0.284	*5per* Inc: 0.286 Dec: 0.282	*5per* Inc: 0.234 Dec: 0.278
	r 0.235	*s* Inc: 0.234 Dec: 0.252	*grad* Inc: 0.255 Dec: 0.287	*grad* Inc: 0.273 Dec: 0.256	*grad* Inc: 0.260 Dec: 0.286
Simul	*r,b* 0.252	*m,s* Inc: 0.254 Dec: 0.288	*5per,grad* Inc: 0.234 Dec: 0.325	*5per,grad* Inc: 0.252 Dec: 0.301	*5per,grad* Inc: 0.229 Dec: 0.312
Meta	*Cross* 0.249	*Cross* Inc: 0.279 Dec: 0.267	*Cross* Inc: 0.242 Dec: 0.338	*Cross* Inc: 0.240 Dec: 0.350	*Cross* Inc: 0.263 Dec: 0.326
	Sequential 0.242	*Sequential* Inc: 0.267 Dec: 0.278	*Sequential* Inc: 0.221 Dec: 0.370	*Sequential* Inc: 0.228 Dec: 0.323	*Sequential* Inc: 0.231 Dec: 0.320

Table 5.11: Table of all the BHI scores for the various clusterings performed on the Botrytis data. The values in italics refer to the clusters formed when taking the control shape into account. *Single* refers to clustering performed on a single parameter, *Simul* is the clustering on multiple parameters, using the simultaneous parameter clustering, and *Meta* refers to the two types of meta-clustering. The meta-clustering was performed on the same parameters as the simultaneous parameter clustering. The numbers in italics for the Single and Simul clusters indicate the use of the control shape.

	Exponential	Gaussian	Gompertz1	Gompertz2	Logistic
Single	*b* 0.242 *0.247*	*m* Inc: 0.258 *0.261* Dec: 0.245 *0.230*	*5per* Inc: 0.244 *0.237* Dec: 0.273 *0.304*	*5per* Inc: 0.239 *0.255* Dec: 0.253 *0.265*	*5per* Inc: 0.239 *0.249* Dec: 0.256 *0.265*
	r 0.265 *0.283*	*s* Inc: 0.241 *0.220* Dec: 0.243 *0.780*	*grad* Inc: 0.257 *0.281* Dec: 0.301 *0.298*	*grad* Inc: 0.250 *0.270* Dec: 0.282 *0.290*	*grad* Inc: 0.260 *0.255* Dec: 0.296 *0.302*
Simul	*r,b* 0.216 *0.204*	*m,s* Inc: 0.250 *0.239* Dec: 0.255 *0.255*	*5per,grad* Inc: 0.244 *0.270* Dec: 0.333 *0.370*	*5per,grad* Inc: 0.257 *0.301* Dec: 0.290 *0.270*	*5per,grad* Inc: 0.249 *0.252* Dec: 0.308 *0.329*
Meta	*Cross* 0.197	*Cross* Inc: 0.261 Dec: 0.274	*Cross* Inc: 0.242 Dec: 0.289	*Cross* Inc: 0.243 Dec: 0.267	*Cross* Inc: 0.263 Dec: 0.326
	Sequential 0.185	*Sequential* Inc: 0.249 Dec: 0.237	*Sequential* Inc: 0.238 Dec: 0.337	*Sequential* Inc: 0.241 Dec: 0.291	*Sequential* Inc: 0.245 Dec: 0.310

5.5. Investigation of specific genes

As a further application of the cluster analysis, the clusters were examined for selected genes of interest, to identify which genes co-cluster together. These genes were *PHOTOSYSTEM I LIGHT HARVESTING COMPLEX GENE 6* (*LHCA6*, *AT1G-19150*), *ARABIDOPSIS NAC DOMAIN CONTAINING PROTEIN 92* (*ANAC092*, *AT5G39610*) and *ARABIDOPSIS NAC DOMAIN CONTAINING PROTEIN 55* (*ANAC055*, *AT3G15500*). *LHCA6* encodes a light-harvesting complex I protein, which forms part of photosystem I (Peng and Shikanai, 2011). It is thus expected that genes that cluster with it would also form part of the photosynthetic machinery. *ANAC092* and *ANAC055* both encode TFs which contain a NAC-binding domain, and were identified as being involved in a number of stress responses (Balazadeh *et al.*, 2010; Hickman *et al.*, 2013; Ooka *et al.*, 2003).

The expression profile of *LHCA6* was fitted by the logistic model in the senescence time course, and the Gompertz2 model in Botrytis time course. These models were clustered using the *5per*, *grad*, and *5per&grad* parameters, and are shown in Figure 5.14. By using these parameters, it was possible to identify genes that were repressed at the same time as *LHCA6* (using the *5per* parameter), the genes that had the same rate of change in gene expression (*grad* parameter), and the genes that were being repressed at the same time as well as having the same rate of change (*5per&grad*). The multiple parameter clustering was performed using the simultaneous parameter clustering. In all these clusters, the significant GO terms were related to the chloroplast (e.g. thylakoid, stroma, photosynthesis). Investigating if any of the genes were involved in the same metabolic pathways, it was found that the clusters contained genes that are involved in the photosystem I and II pathways. Genes that co-clustered using the *5per* parameter included *PHOTOSYSTEM I SUBUNIT K* (*PSAK*, *AT1G30380*), indicating that other photosystem I genes are down-regulated at the same time point. In the *grad* clusters, the genes *PHOTOSYSTEM II LIGHT HARVESTING COMPLEX GENE 2.3* (*LHCB2*, *AT3G27690*) and *PHOTOSYSTEM II SUBUNIT T* (*PSBTN*, *AT3G21055*) were found. This suggests that the photosystem II genes are down-regulated at the same rate as the photosystem I genes, but at different times.

In the Botrytis clusters, in addition to the photosynthesis related genes, the *5per* parameter clustering (Figure 5.14 B1) contained annotation terms related to ribosomes, biosynthetic processes, as well as stress and pathogenesis-related (PR) proteins. These included genes encoding proteins belonging to the disease resistance protein (TIR-NBS-LRR class) family, such as *AT5G36930*, *AT5G11250*, and *AT5G39730*, in addition to *AT5G40060*, which encodes an AIG2-like (avirulence induced gene) family protein. Altogether, this indicates that disease resistance genes were becoming down-regulated at the same time as photosynthesis genes, and may be an indication of the defence response failing, and the plant succumbing to the pathogen infection.

Figure 5.15A shows the clusters containing *ANAC092* in the senescence dataset, where the best fit was to the logistic model, and the genes were clustered using the *5per*, *grad*,

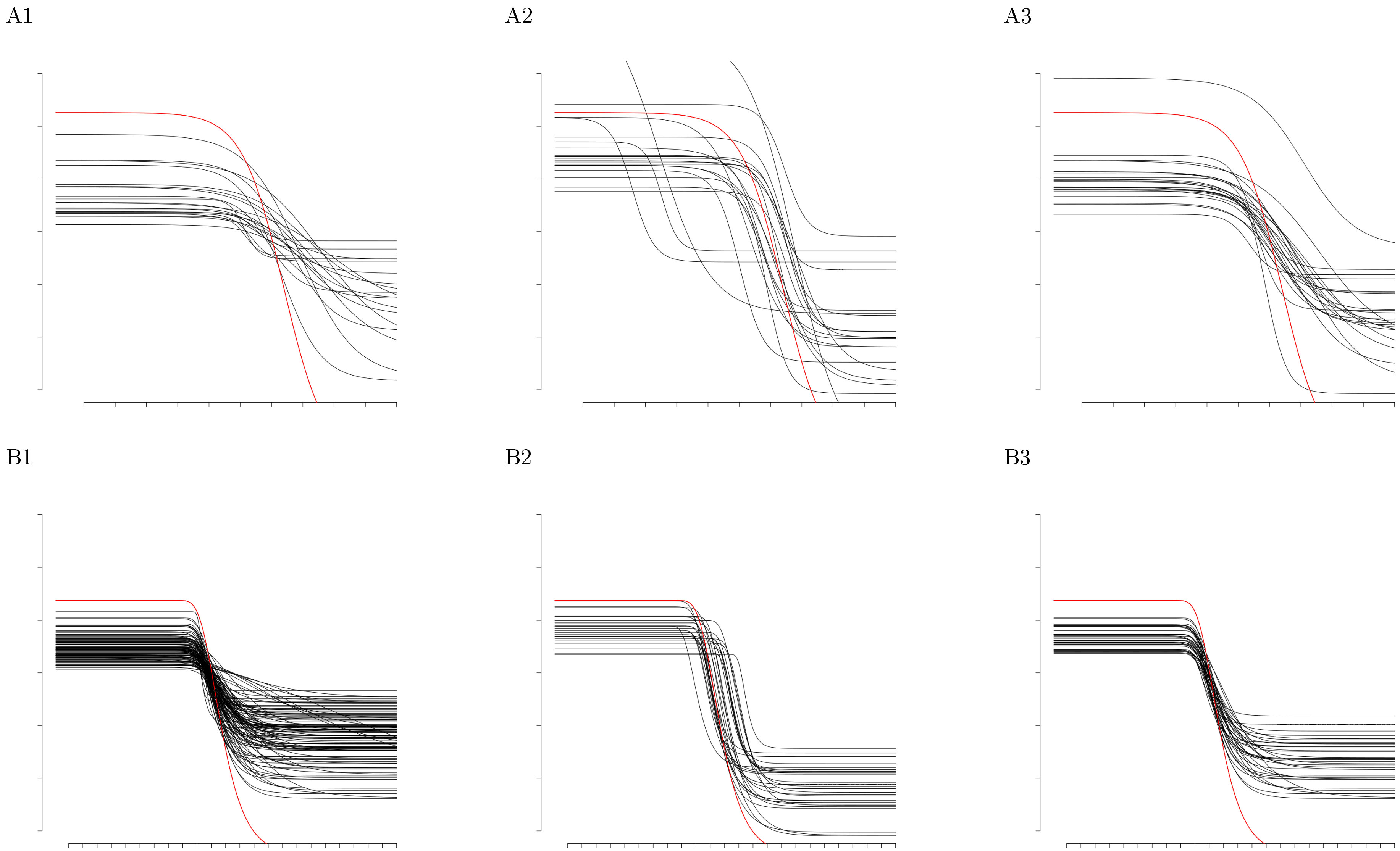


Figure 5.14: The clusters containing *LHCA6* in the senescence (row A, logistic) and Botrytis (row B, Gompertz2) datasets. *LHCA6* is shown in red. The clusters were formed by clustering on the *5per* (column 1), *grad* (column 2), and *5per&grad* parameters (column 3).

and *5per&grad* parameters. The *5per* clustering showed genes that were activated at the same time as ANAC092, and contained genes that were over-represented for nucleotide excision repair, whereas the *grad* cluster showed genes that were changing at the same rate, and contained genes involved in abiotic stress, water and salt stress, including *SENESCENCE-ASSOCIATED GENE 113* (*SAG113*, *AT5G59220*), which has been shown to be involved in ABA signalling pathway (Zhang and Gan, 2012), and a cytochrome *CYP76C* (*AT2G45570*). The combined *5per&grad* cluster primarily showed annotated terms enriched for anthesis, although the cluster contained a number of other stress related genes, including *ANAC055* (see below), *WRKY45* (AT3G01970), *WRKY65* (*AT1G29280*), and another senescence associated gene *SAG21* (*AT4G02380*). These genes have been shown to be involved in stress responses (Eulgem *et al.*, 2000; Shimono *et al.*, 2007; Weaver *et al.*, 1998). These clusters of genes were also analysed for the presence of any gene products that are involved in particular metabolic pathways, using the annotations in MAPMAN (Section 2.1.6). The *5per&grad* cluster showed a number of gene products involved in pathways involving auxin regulated hormone metabolism, MYB domain TF family regulation, and WRKY domain TF family regulation. These clusters were also compared to the results from an *anac092* mutant study to investigate downstream targets (Balazadeh *et al.*, 2010). A number of ANAC092 targets were found in the clusters formed with the *grad* parameter, and included *SAG113*, *WRKY45*, and *CYP76C*.

In the Botrytis treatment, the best fit to *ANAC092* was the Gaussian model, and clusters were formed using the *m*, *s*, and *m&s* parameters (Figure 5.15B). The *m* parameter indicated the time of maximum gene expression (and the time of gene repression), the *s* parameter indicated the duration of the response signal, and the *m&s* parameters indicated the genes that reached the maximum at the same time, as well as having the same response duration. *ANAC092* had a relatively large *s* parameter value, implying that the activation of the gene was slow, and then gradually subsided. This could indicate that the genes with this *s* parameter have a persistent signal. Performing the cluster analysis on the *m* parameter revealed annotation terms enriched in catalytic activity, and included *HIGHLY ABA-INDUCED PP2C GENE 3* (*HAI3*, *AT2G29380*), a gene which is involved in ABA signalling. The *s* cluster contained terms involved in oxireductase activity, response to stress, autophagy, ageing and secondary metabolism, and the *m&s* cluster contained terms related to abscission. Genes products that were involved in metabolic pathways were found in the cluster formed using the *s* parameter, and included involvement in auxin regulated hormone metabolism, MYB domain TF family regulation, AP2/EREBP domain TF family regulation, bZIP TF family regulation, and ubiquitin protein degradation.

Common genes between these clusters and the experimentally determined downstream genes were mostly found in the cluster formed using the *s* parameter, and included a FAD/NAD(P)-binding oxidoreductase family protein (AT4G38540), and CALCIUM EXCHANGER 7 (CAX7, AT5G17860).

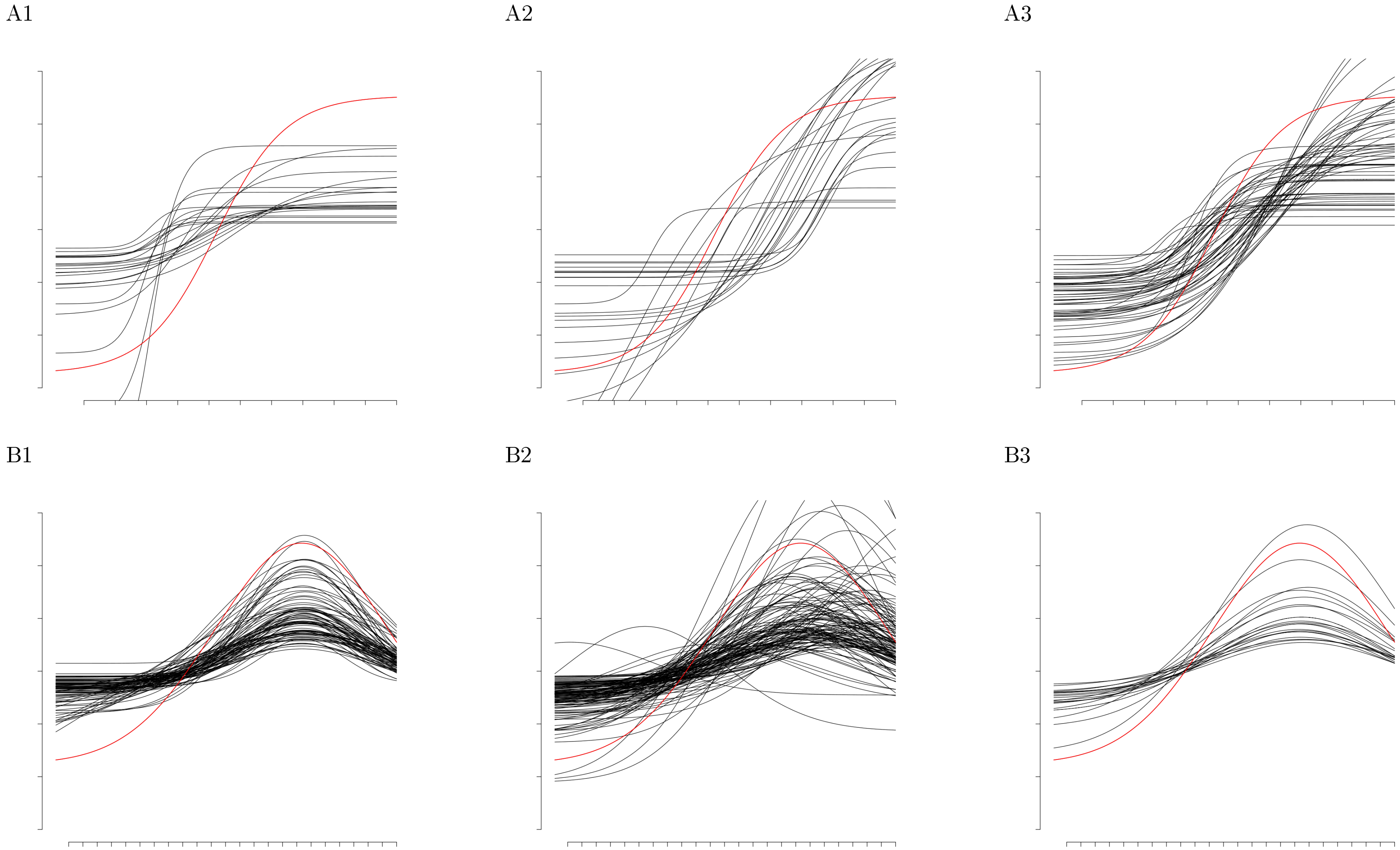


Figure 5.15: The clusters containing *ANAC092* in the senescence (row A, logistic) and Botrytis (row B, Gaussian) datasets. *ANAC092* is shown in red. The clusters were formed by clustering on the *5per* (column 1), *grad* (column 2), and *5per&grad* parameters (column 3) for senescence, and the m (column 1), s (column 2), and *m&s* parameters (column 3) for Botrytis.

Finally for the *ANAC055* clusters, in senescence the best fit was the logistic model, and the clusters were created by clustering on the *5per*, *grad*, and *5per&grad* parameters (Figure 5.16A). Since *ANAC055* co-clustered with *ANAC092*, there were several common genes, including *WRKY45*, *WRKY65*, and *SAG21*. Over-represented GO terms in the *5per* clustering contained terms related to DNA binding, as well as response to fungus and jasmonic acid. However, it is known that senescence causes the activation of pathways that are involved in the defence signalling pathways, particularly of those involved in jasmonic acid and salicylic acid, the primary stress hormones in Arabidopsis (Windram *et al.*, 2012). Clustering on the *grad* parameter revealed over-represented annotation terms involved in transporters, anthesis and secondary metabolism. In addition, there were many genes encoding putative TFs, including *AT4G17900*, *AT2G28200*, *AT1G02610*. By performing the cluster analysis on the *5per&grad* parameters a set of genes was produced with over-represented annotation terms related to ethylene, jasmonic acid and abscisic acid signalling, senescence, wounding, and water loss. In this cluster, there were also genes which encoded proteins that were involved in metabolic pathways responsible for auxin regulated hormone metabolism, MYB domain TF family regulation, WRKY domain TF family regulation, and ubiqutin protein degradation. Like the *ANAC092* results, data from an *anac055* mutant was obtained to compare experimentally determined downstream targets with these clusters (Hickman *et al.*, 2013). Common genes included *AUTOPHAGY 18A* (*AtATG18a*, *AT3G62770*), which is involved in nutrient deprivation and senescence, and two genes which encode RING/U-box superfamily proteins, *AT5G55970* and *AT1G63840*, the latter of which has been shown to be involved in ABA signalling (Xin *et al.*, 2005).

In Botrytis, the best fit to the *ANAC055* gene expression profile was the Gompertz1 model (Figure 5.16B). As with all the previous sigmoid shapes, the clusters were created by performing the cluster analysis on the *5per*, *grad*, and *5per&grad* parameters. Clustering on the *5per* only revealed over-represented annotation terms as response to stimuli. However, investigation of the cluster members revealed several genes encoding MATE (multidrug and toxic compound extrusion) efflux family proteins. These proteins are generally associated with the transport of toxic compounds out of the cytoplasm (Eckardt, 2001), but it has been suggested that they may also transport compounds to assist in the defence response (Omote *et al.*, 2006; Rowe *et al.*, 2010). These compounds may include molecules such as oxidoreductases or cytochrome P450 enzymes (Rowe *et al.*, 2010). Indeed in the cluster, there were several genes encoding cytochrome P450s. Clustering on the *grad* parameter showed annotation terms relating to water deprivation, ethylene stimulus, wounding, and hypoxia. This cluster also contained MATE family proteins, cytochrome P450s and peroxidases, as well as another NAC TF, ANAC019 (AT1G52890). When clustered on the *5per* and *grad* parameters, the clusters with *ANAC055* also contained over-represented annotation terms that are associated with Botrytis infection including ethylene stimulus, jasmonic acid stimulus, response to fungus, wounding, and stress responses (Windram *et al.*, 2012).

Common downstream genes, as compared to Hickman *et al.* (2013), included *JAS-*

MONATE ZIM DOMAIN PROTEIN 8 (*JAZ8*, *AT1G30135*), and *RESPONSIVE TO HIGH LIGHT 41* (*RHL41*, *AT5G59820*), which encodes a zinc finger TF that is involved in high light and cold acclimation (Doherty *et al.*, 2009).

In addition to the above analyses, the motifs in the upstream sequences of the genes in the above clusters were analysed, and over-represented motifs were identified using a hypergeometric test (described in Section 2.1.7). In this analysis, the presence of 350 experimentally verified plant motifs were identified in the region 500 bp upstream of transcription start site for the genes in the above clusters. After filtering for motifs that were significantly enriched (p-value$<$0.01) in at least one of the clusters, a set of 71 motifs were found, and are shown in Figures 5.17 and 5.18 for the senescence and Botrytis clusters, respectively. From these figures, it can be seen that there were significantly more over-represented motifs found in the cluster analyses when clustering on the *grad* parameter, as opposed to the *5per* parameter. This suggests that the rate of change of the gene expression levels provide a better indication of common regulators as opposed to the time of up- or down-regulation.

In the senescence clusters (Figures 5.17), most of the significant motifs were found in the cluster containing *ANAC055*, using the *grad* parameter (Figure 5.16 A2). Of note are M00660 (tcACGT), the binding site for a bZIP TF that is involved in activating a large number of genes in developing rice plants (Izawa *et al.*, 1994), and M01584 (ACGTGG), the binding site of ELONGATED HYPOCOTYL 5 (HY5, AT5G11260), a bZIP TF in Arabidopsis that is known to be a part of the response to light signalling pathway, as well as mediating ABA responses during seed germination, early seedling growth and root development (Chen and Xiong, 2008; Lee *et al.*, 2007). This binding site is similar to the G-box (CACGTG), which can be found in the promoters of many light stimulus genes, as well as response to hormones, such as ABA, ethylene and jasmonic acid (Menkens *et al.*, 1995). Another binding site for HY5 was also found in the *ANAC092* cluster with the *5per* parameter (Figure 5.15 A1), albeit from soy bean (M01186, TGACGT) (Song *et al.*, 2008).

Another significantly over-represented motif was M01136 (AAAG), which is found in the *ANAC092* cluster using the *5per&grad* parameters (Figure 5.15 A3), as well as the *ANAC055*, using the *5per&grad* parameter (Figure 5.16 A3). This motif is the binding site for the Dof1 and Dof2 TFs, and are associated with the expression of multiple genes involved in carbon metabolism in maize, as well as responses to stress and light (Yanagisawa and Sheen, 1998; Yanagisawa, 2000). Interestingly, this motif is not enriched in the previously mentioned *ANAC055* cluster using the *grad* parameter, suggesting that genes with this motif in their promoters have both the same time of up-regulation, as well as rate of change in gene expression.

In the Botrytis clusters (Figures 5.18), the majority of the enriched motifs were found when performing the cluster analysis on the *grad* parameter for the sigmoid models,

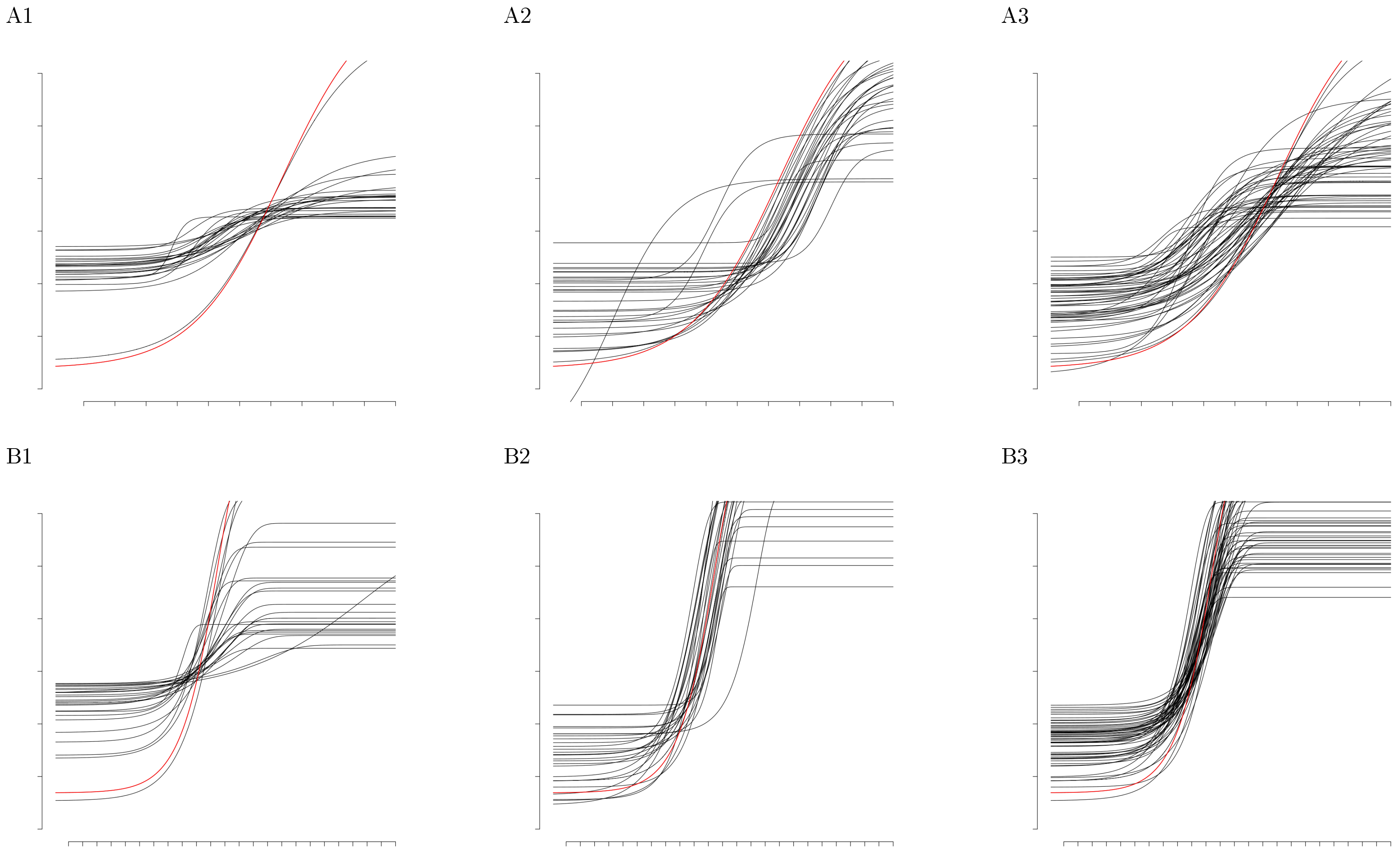


Figure 5.16: The clusters containing *ANAC055* in the senescence (row A, logistic) and Botrytis (row B, Gompertz1) datasets. *ANAC055* is shown in red. The clusters were formed by clustering on the *5per* (column 1), *grad* (column 2), and *5per&grad* parameters (column 3).

as well as the *s* parameter of the Gaussian model. As with the senescence data, the clusters that were formed using the timing parameters (*5per* for the sigmoid models, and *m* for the Gaussian model) did not have many significantly over-represented motifs. Over-represented motifs from the cluster which contained *LHCA6*, clustered on the *grad* parameter (Figure 5.14 B2) included light-responsive promoter regions, including S-000424 (GATAAGR), the I-box (Martı nez Hernández *et al.*, 2002), and M00435 (CACGTGG), the binding site for PHYTOCHROME INTERACTING FACTOR 3 (PIF3, AT1G09530), which is involved in phytochrome signalling pathways in Arabidopsis (Martı nez Garcı a *et al.*, 2000). These promoters are both involved in responses to light, and are unsurprising considering many of the genes in the cluster were photosynthesis related. In the cluster where *ANAC092* was clustered on the *s* parameter (Figure 5.15 B2), over-represented motifs included the binding site of a NAC TF in wheat, TaNAC69 (M01055, CGTaN{5}tACG), which is involved in responses to both biotic and abiotic stresses (Xue, 2005; Xue *et al.*, 2011), as well as the binding site of TGA1b in tobacco (M00946, gtgACGTgac) (Izawa *et al.*, 1993). TGA TFs are generally associated with SA signalling (Johnson *et al.*, 2003), although it has been shown that another member of the family, TGA3, is important in regulating the response to Botrytis (Windram *et al.*, 2012). The HY5 binding site (M01584, ACGTGG) was also over-represented in this cluster (Lee *et al.*, 2007). In the cluster containing *ANAC055* and clustered using the *5per&grad* parameters (Figure 5.16 B3) the WRKY TF binding site (S-000310, TTTGACY) was over-represented. As mentioned previously, many WRKY TFs have been shown to be involved in a number of stress responses (Eulgem *et al.*, 2000).

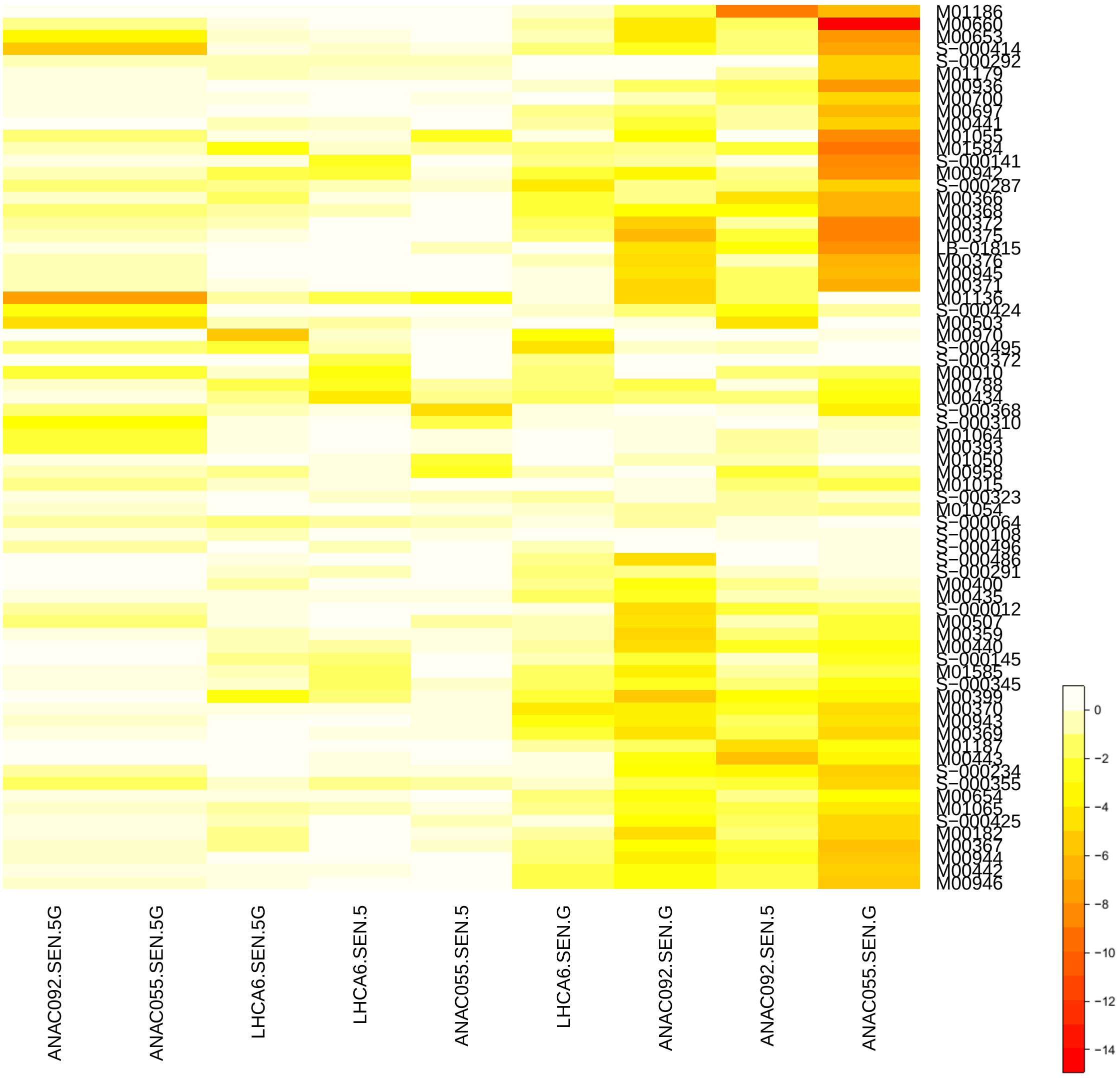


Figure 5.17: Significantly over-represented motifs that were found in the up-stream regions of the genes that are found co-clustered with *ANAC092*, *ANAC055*, and *LHCA6* using the senescence dataset. Significance was determined using a hyper-geometric test, where darker, redder colours indicate higher significance. The rows show the different motifs, and the columns show the different cluster analyses using the different parameter combinations. *G* indicates clustering on the *grad* parameter, *5* on the *5per* parameter, and *5G* is using both. Scale is given in log(p-value).

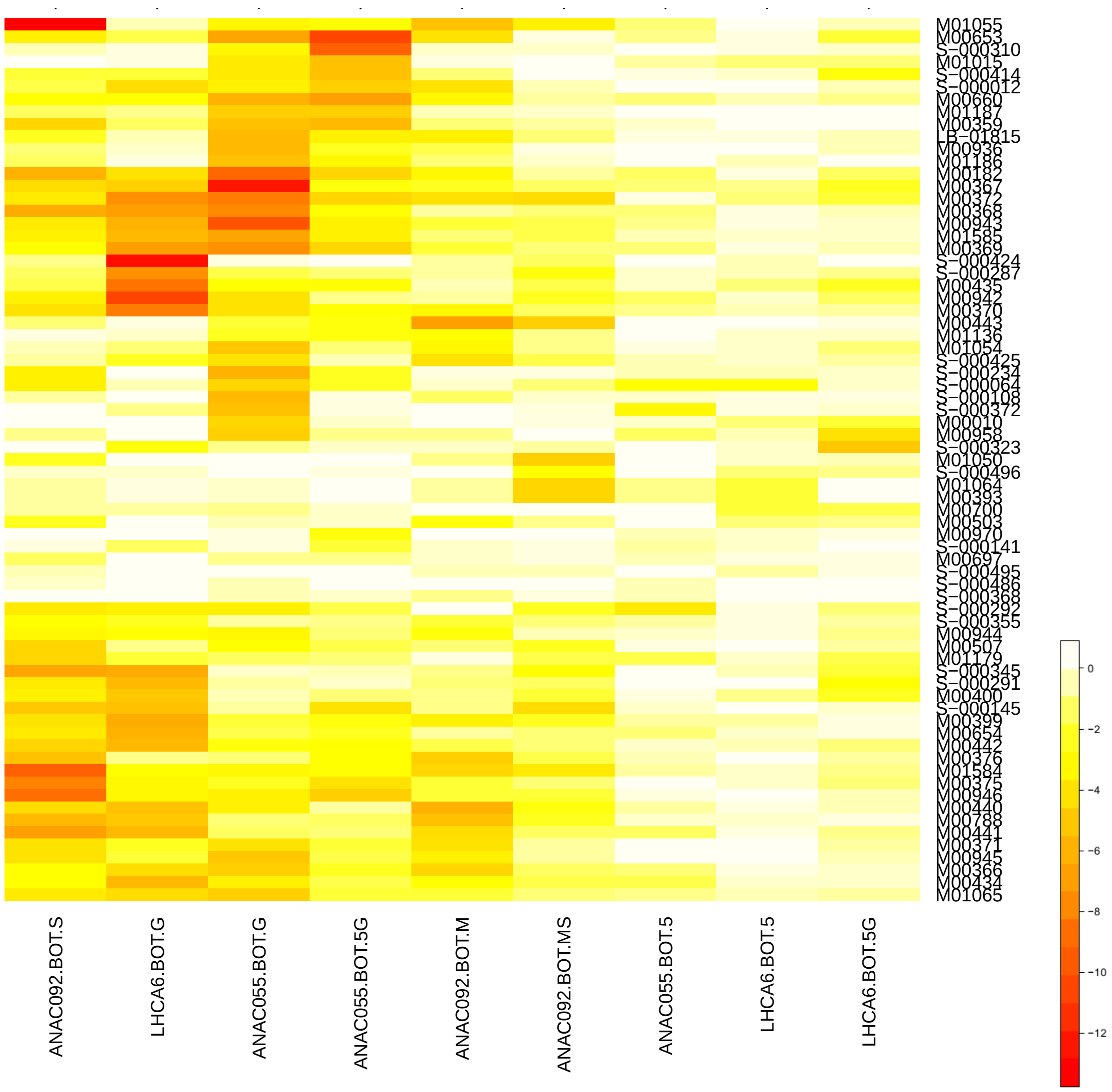


Figure 5.18: Significantly over-represented motifs that were found in the up-stream regions of the genes that are found co-clustered with *ANAC092*, *ANAC055*, and *LHCA6* using the Botrytis dataset. Significance was determined using a hypergeometric test, where darker, redder colours indicate higher significance. The rows show the different motifs, and the columns show the different cluster analyses using the different parameter combinations. *G* indicates clustering on the *grad* parameter, *5* on the *5per* parameter, and *5G* is using both. *M* indicates clustering on the *m* parameter, *S* on the *s* parameter, and *MS* is using both. Scale is given as log(p-value).

5.6. Discussion

In this chapter, the data from the Arabidopsis datasets was analysed. The datasets were created to investigate the effect of two different environmental stresses, namely the developmental stress investigating changes in gene expression as the plant ages, and in response to the necrotrophic pathogen, *Botrytis cinerea*. An application was developed using a regression approach, together with the cluster analysis, and named ShapeCluster. Using ShapeCluster, it was possible to analyse the data and determine the general function of the various sets of genes. As with the yeast data in the previous chapter, understanding what each parameter describes in terms of the biological system being studied is paramount to obtaining the best outputs. As before, a large number of clusters were produced, potentially containing a large amount of information. The analysis here focussed primarily on the most abundant models in each of the datasets.

As with the analyses in Chapter 4, the clusters were formed based on single or pairs of parameters. In doing so, this revealed different sets of genes with functions relating to a particular aspect of the expression profile, such as the timing of the activation of genes, or the rates that gene expression is changing. In the analyses of the datasets, a broad overview of the molecular functions of the organism as it undergoes the respective stress was revealed. Using the described parameters, it was found that genes that were down-regulated were generally involved in chloroplast activity and photosynthesis. This makes sense as photosynthesis is down-regulated in senescence as the cellular components of the leaf is being broken down (Breeze *et al.*, 2011), as well as in Botrytis, when the plant is attempting to mount a defence response, where it is thought that these photosynthetic genes are being repressed so that the cellular nitrogen may be used to synthesise new defence proteins, or mobilised away (Windram *et al.*, 2012). Many of the up-regulated genes were involved in the hormone signalling pathways involving the stress hormones, jasmonic acid, salicylic acid, ethylene and abscisic acid. In addition, there were numerous clusters with possible DNA binding activity, possibly indicating a presence of TFs. These could assist in the identification of key regulators in these stresses. Many of the over-represented annotation terms were repeated in the cluster analyses on different shapes, and this is likely due to the possibility of a gene expression profile fitting to multiple models (Tables 3.2 and 3.3). However, it is possible for clusters to be over-represented in the same annotation term, but have very different shapes, and this could be an avenue for further investigation. By using the fitted parameters, it was possible to infer biological information from the different shapes. In particular, by using the timing parameters a timeline of the processes that are occurring in the organism over the time course could be produced. These were compared to the published results, and found to be consistent. Additionally, a few new responses were discovered, including auxin homeostasis and a late down-regulation of photosynthesis genes in senescence, and the response to hypoxia in Botrytis.

By clustering with the control information, more information about the function of the genes was revealed. Using the model fit of the control dataset, it was possible to sort the genes into subclusters based on the fit under the control conditions. It was thought that genes with common biological functions would have similar profile shapes in both the mock and infected treatments. Thus, this process refines the clusters to uncover the significant over-represented terms that may be hidden due to the large number of terms in the cluster. In addition, this method may be useful in identifying genes that are differentially expressed compared to the control. However, as with the meta-clustering analyses, this approach did produce a large number of small clusters. Generally it was possible to obtain some sort of over-representation score for the cluster, but there were cases where the over-representation tests did not reveal anything due to the cluster sizes being too small, or a lack of annotation for the genes in the cluster.

As with the yeast analysis, the similarity of annotations in the clusters was determined using the BHI score. The BHI scores for this analysis were lower than the results obtained from the yeast data, and this reflects the level of annotation relative to the genome size. Arabidopsis has a genome of over 27 000 genes whereas yeast has just over 6500 genes, 76% of which are verified (as of October 2013, http://www.yeastgenome.org/cache /genomeSnapshot.html). It is estimated that approximately 30% of the Arabidopsis genome has been experimentally verified (Quanbeck *et al.*, 2012). The results from ShapeCluster were compared to the clusters from SplineCluster, and the values of the clusters from ShapeCluster were generally slightly lower. However, there were higher BHI scores for the decreasing sigmoid models. This is possibly due to the higher number of genes which have been identified as photosynthesis related genes. Thus, the lower scores can be attributed to the poorer annotation level. In addition, it was also shown that by using the control information, clusters that have more homogeneous annotations could be produced.

In addition, a more comprehensive analysis of the clusters was performed, where clusters containing genes with known stress responses were analysed. By performing the cluster analysis on a variety of parameters, information on these genes was obtained. By using a known photosynthesis-related gene, *LHCA6*, it was possible to determine that the results were what was anticipated. That is, it would be expected that other photosynthesis-related genes would co-cluster with *LHCA6*. When investigating the known stress genes, *ANAC092* and *ANAC055*, again some known information was found, such as the activation of pathways involving the stress-related hormones, jasmonic acid and salicylic acid, both of which are implicated in both the senescence and Botrytis stresses. In addition to this, some potentially new information was revealed, such as the presence of certain WRKY TFs, which could be important in the activation of senescence. In addition, it was found that *ANAC055* often clustered with MATE proteins, which are generally thought to be involved in toxin resistance. However, it has been suggested that this family of proteins may also be involved in defence responses. Thus, in using the regression and clustering approaches it may be possible to analyse genes with unknown function to determine what the genes may be functionally related to each other.

The promoter regions of the genes in these clusters were also investigated for motifs, which may indicate the presence of genes that are co-regulated. Some motifs that may be of interest were found, particularly in clusters formed using the rate of change parameters, which produced more enriched motifs as opposed to the initial point of activation. This suggests that these genes become up- or down-regulated at the same rate, rather than getting activated all at once. From this analysis, a large number of motifs that were over-represented were binding sites for TFs from a variety of plant species, as well as TFs that were involved in stress responses. Several light-responsive promoters were found in the clusters containing *LHCA6*, including one motif that is associated with light responsive differential expression, and another that was associated with PIF3, a TF that interacts with photoreceptors, resulting in signal transduction pathways that result in changes of up to 30% of the plant transcriptome (Shin *et al.*, 2007). In addition, in both the Botrytis and senescence clusters, the HY5 binding site was over-represented, which is also related to light responsive signalling, as well as hormone signalling. The PIF3 and HY5 signalling pathways have been shown to be related (Shin *et al.*, 2007), so this may indicate some crosstalk between these and other stress response pathways. These motifs were obtained from experimentally verified binding sites, and as a result, is a relatively small pool of results. If the aim was to find novel binding sites, tools such as MEME (Bailey *et al.*, 2006) or RSAT (Van Helden, 2003) may be used.

ShapeCluster was thus used to analyse a large set of genes, where models were fitted to the expression profiles and analysed in terms of the fitted parameters. These models were also filtered to identify the fits that were well described, thus ensuring that the most relevant information was retained. These analyses provided a biologically focussed representation of gene expression profiles, and may provide an improved understanding of the molecular mechanisms that occur in response to stimuli. In addition, this could aid in identifying previously undiscovered regulators, or assist in assigning putative functions to genes with no known function.

6. Investigation of the effect of multiple environmental stresses in plants

6.1. Introduction

Plants are constantly bombarded by a multitude of biotic and abiotic stresses, and this can lead to reduced yields in crop plants. Understanding how plants respond to more than one stress is of utmost importance. Many single stress experiments have been performed, and large sets of stress- or pathogen-responsive genes have been identified. Investigation into these sets of genes has revealed that there is a great deal of cross-talk between the various stress response pathways. Thus, investigations into the responses of plants to combinations of stresses, such as drought and pathogen attack, would be desirable. Furthermore, it would be extremely valuable to investigate whether it is possible to predict the expected response to a combination of stresses, based on the response to individual stresses.

Since senescence is related to age, the combination of this stress together with that of a pathogen response is often known as age-related resistance, where defence responses are altered by the developmental processes (Whalen, 2005). This age-related resistance is driven by a range of molecular mechanisms, and changes depending on the pathogen. Generally, it has been observed that younger plants are more susceptible to pathogens and become more resistant over time. For example, it has been shown that older Arabidopsis plants were more resistant to *Hyaloperonospora arabidopsidis* (Rusterucci *et al.*, 2005), and while 3-4 week-old Arabidopsis plants were susceptible to *Pseudomonas syringae* pv. tomato, plants that are >5 weeks old showed a 10-100 fold reduction in bacterial growth (Kus *et al.*, 2002). This effect is also true for many other plant-pathogen systems. For example, 20-25 day old *Nicotiana benthamiana* plants became infected with *Phytophthora infestans*, while mature plants were resistant to all isolates (Shibata *et al.*, 2010). In addition, there is an increased resistance with older plants between wheat and *Puccinia recondita* f.sp *tritici* (Pretorius *et al.*, 1988), soybean and *Phytophthora megasperma* var *sojae* (Ward *et al.*, 1981), cotton and *Rhizoctonia solani* (Hunter *et al.*, 1978), and cowpea and cow-pea rust fungus (Heath, 1994). The mechanisms involved in the age-related resistance appear to differ widely, and in many cases, the exact molecular processes are unknown.

However, with some pathogens, there is increased susceptibility with an increase in age, such as with onion and *Alternaria porri* (Miller, 1983). It is likely to be the case with Botrytis, as it boosts senescence through the production of ethylene (Cristescu *et al.*, 2002), which promotes leaf senescence (Grbić and Bleecker, 1995; Wang *et al.*, 2013), as well as fruit ripening and flower senescence (Klee and Clark, 2010). In addition to ethylene, some isolates of Botrytis also produce abscisic acid (Sharon *et al.*, 2007), which is a known inducer of senescence (Lim *et al.*, 2007).

When a plant is exposed to combined environmental stresses, it is thought that the response from a single stress would be modulated to form a combined response (Atkinson and Urwin, 2012; Mittler and Blumwald, 2010). Rasmussen *et al.* (2013) suggest five different predicted behaviours that may occur - combinatorial, cancelled, prioritised, independent, and similar (Figure 6.1). A *combinatorial* response occurs where the responses are similar in the single stress situations, but produces a different response when combined. The *independent* response is where one of the individual stresses does not respond to the stress, and the response to the combined stress is the same as the other stress response. A *cancelled* response is where the individual responses are different, but return to control levels in the combined response. A *prioritised* response is where the responses are different, but is the same level as one of the individual stresses. This is similar to the independent response, expect that there is a response to both stresses. Finally, the *similar* response is if both individual responses are the same and the combined response is similar to the individual response. Thus it can be seen that there a wide range of possible outcomes in response to a combination to two different environmental stresses. However, the difficulty in predicting these combined responses from the single stress data would be identifying the overriding factor that influences the response (Rasmussen *et al.*, 2013).

These predicted responses from combinations of environmental stresses provide a useful starting place to determine the effect of simultaneous stresses. However, these predictions were developed with only a single time point in mind, and only investigate three states - up-regulated, down-regulated, and unchanged. Thus by using the fitted regression models from the time series expression profiles described in Chapter 3, it may be possible to investigate the gene expression profile for the combined stress treatment using these models. This chapter describes some of the possible outcomes of a combined stress response using these models, as well as biological experiments to validate the predictions.

6.2. Combined stress predictions

Through the use of the fitted regression models from the senescence and Botrytis datasets, it was possible to make predictions as to the effect of a combined stress. Given two expression profiles that are on the same time scale, either of them could be the more influential in determining the outcome of the combined stress (Figure 6.2A-B). However, in this case, the Botrytis infection takes place on a much shorter time scale as compared to the senescence (48 hours as opposed to 22 days). Thus, it seems reasonable to assume that the shorter response would take precedence when combined.

Two simple models are proposed: an additive model, and a replacement model. The additive model is simply the cumulative expression level between the two stress responses. That is, the imposition of the second stress simply adds to the response of

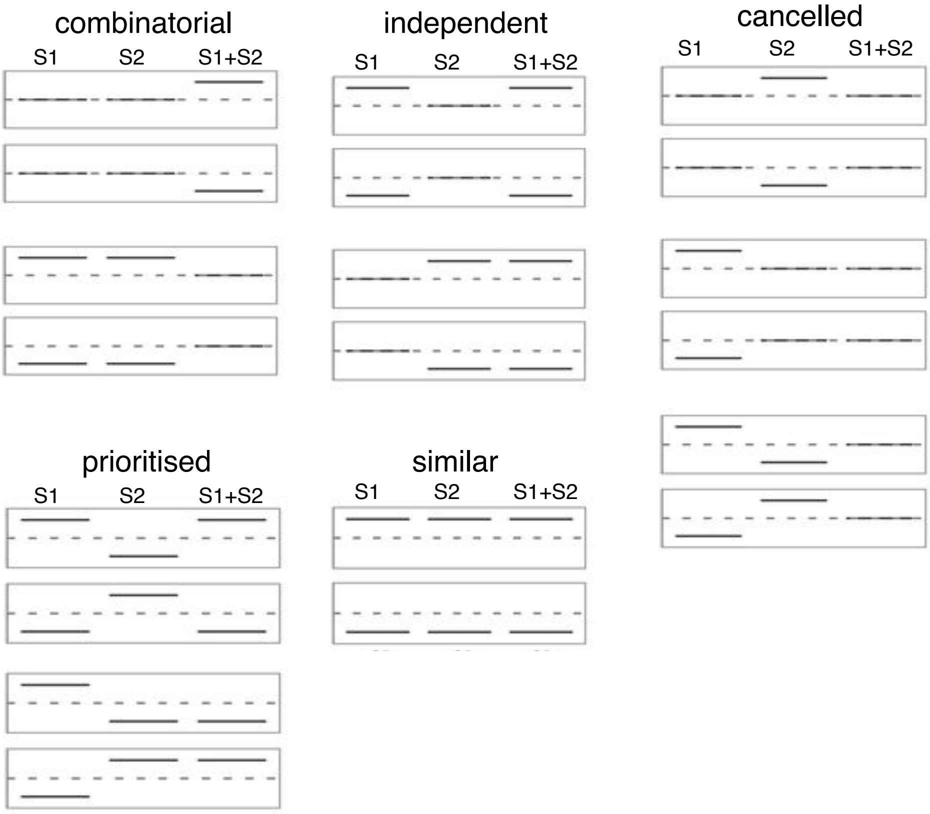


Figure 6.1: Schematic representation of possible results that may be obtained as a result of a combined stress. The two columns on the left (S1 and S2) are the responses from two individual stresses, and the column on the right shows the response from the combined stress (S1+S2). The dotted lines represent no change in gene expression from the control, and above and below the line represent up- and down-regulation, respectively. Figure adapted from Rasmussen *et al.* (2013).

the first stress. This is illustrated in Figure 6.2C, where the infection is predicted to result in the basal expression plus the expression as a result of the infection process. In the replacement model, where the expression profile from one of the stress treatments displaces the other gene expression response. This is shown in Figure 6.2D, where the Botrytis stress replaces the underlying senescence stress, but only reaches the same maximum as it would have in the single stress.

Another option, which is not addressed here, is where the responses are more complicated than simply the two combined and is instead a result of other interactions, both upstream and downstream of the gene of interest. In these cases, there may be a synergistic interaction where the combination results in a different pattern of response, or antagonistic where the gene expression response for the combined stressed is reduced below the level for either stress alone. As a result, it would be extremely difficult to predict the effect of a combined stress simply from the expression profiles from the single stress data.

6.3. Phenotype screen

To determine if there was a relationship between senescence and *Botrytis cinerea* infection, a phenotype screen was performed where different aged leaves were infected and the size of the lesion was measured. Plants were grown as described in Section 2.2.1.1. Seeds were sown at different times to obtain different aged plants, and time points of 28, 31, and 35 days after sowing (DAS) were selected based on the data from the original PRESTA long day senescence screen (Breeze *et al.*, 2011). These time points were chosen to represent a mature leaf, a leaf beginning to show visible signs of senescence, and a time point approximately in-between these points. Leaf 7 of the Arabidopsis plants was tagged with cotton string on emergence. This leaf was harvested from the 28, 31 and 35 DAS plants and placed on 0.8% w/v plant agar (Duchefa Biochemie) in propagator trays. Ten replicates at each plant age were obtained. The Botrytis inoculum was created as described in Section 2.2.1.2. A single 0.6 µl droplet of the inoculum was placed on each leaf, and the trays were covered with lids. At 48, 60 and 72 hours post-infection (hpi), photographs were taken. Using the ImageJ image analysis software (Abramoff *et al.*, 2004), the sizes of the lesions were quantified, and the results are shown in Figure 6.3.

As expected, there was a mean-variance relationship in the lesion sizes, that is, the variance became larger as the size of the lesion became larger. Thus, the data was log-transformed. It can be seen that there is a slight increase in lesion size between the 28 and 31 day old plants, and much larger lesions in the 35 day old plants as compared to the younger plants.

To obtain statistical validation of the results, a two-way ANOVA was performed using the age of the plants, and the time of infection as factors. The model used was

$$\log(lesion) \sim Harvest * Time + Error(rep/leaf)$$

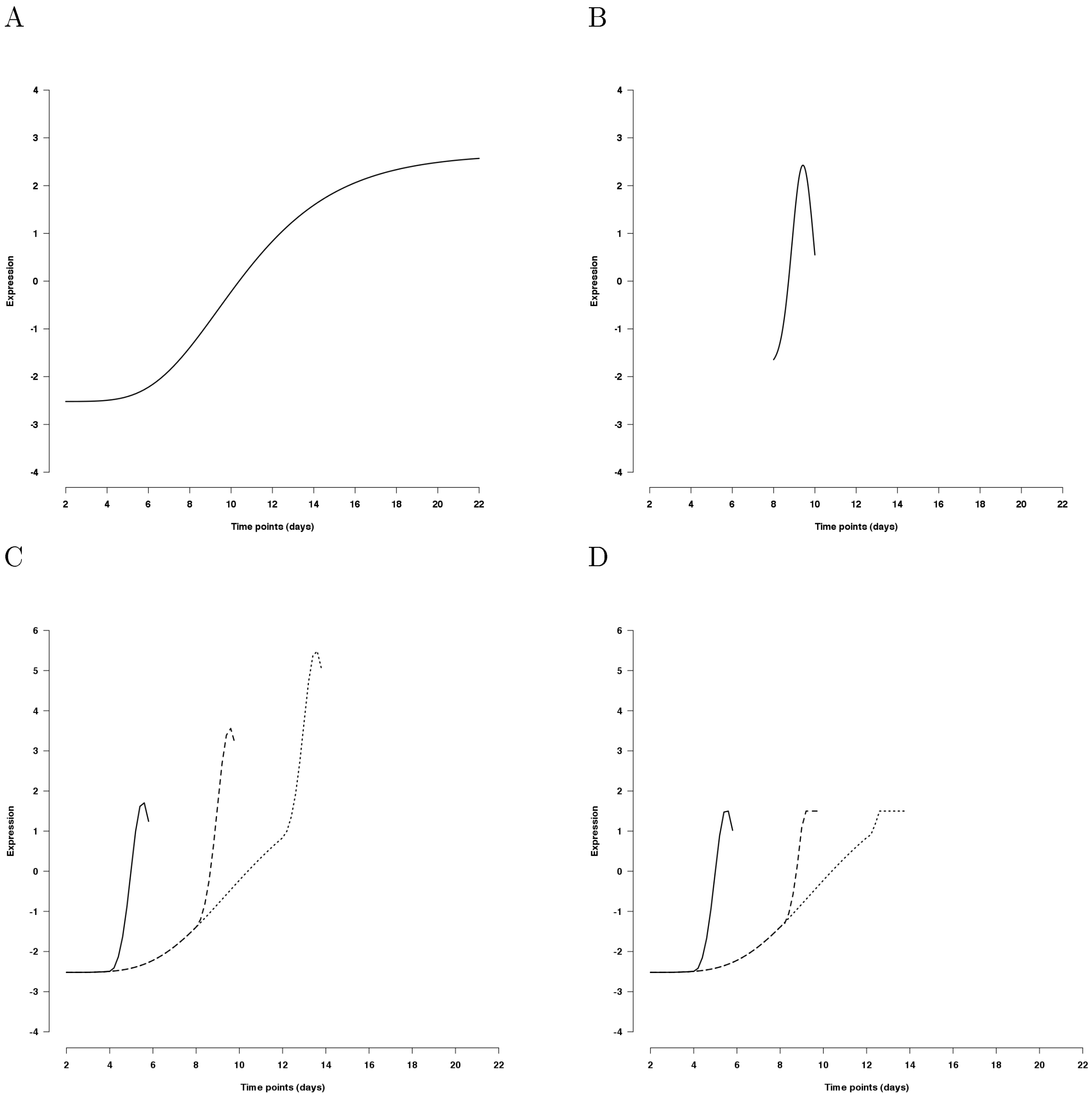


Figure 6.2: Illustration of the predicted multi-stress models using *ANAC092* as an example. The senescence (A) and Botrytis infection (B) stresses are shown on the same time scale. The additive model (C) assumes the combined model is simply the addition of the two single stress models. The replacement model (D) describes the case where one of the models replaces the gene expression, maintaining the same change in gene expression. In both models, the solid line indicates infection at time point 4, the dashed line indicates infection at time point 8, and the dotted line indicates infection at time point 12.

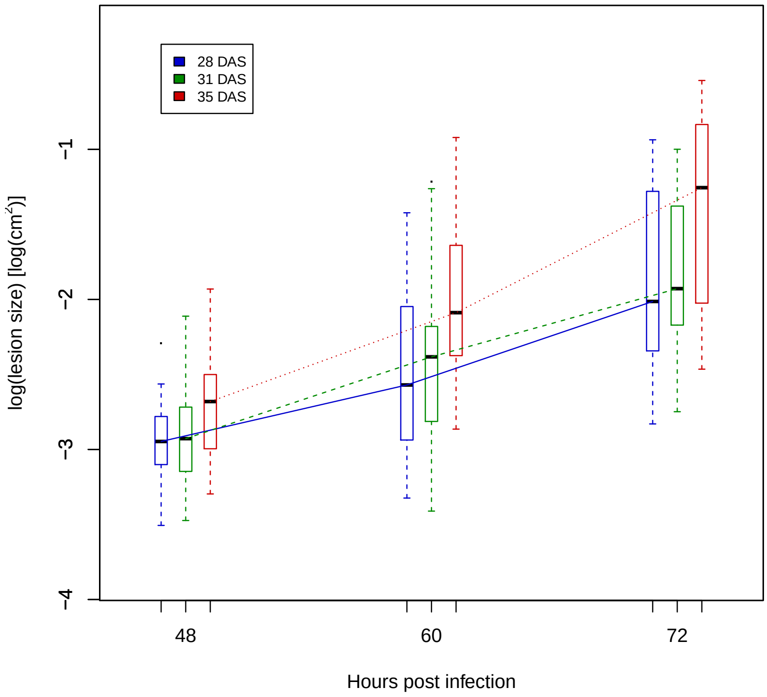


Figure 6.3: Figure showing the lesion sizes on the leaves from the phenotype screen. Shown are the three different harvest times (28, 31, and 35 DAS), and the lesion sizes at the time points post-infection (48, 60 and 72 hpi).

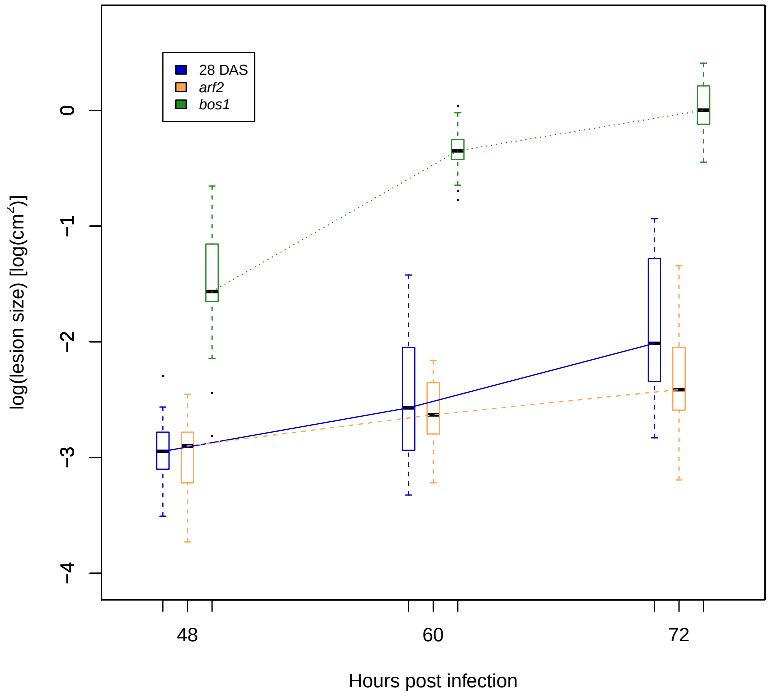


Figure 6.4: Control data for phenotype screen, where leaves from *bos1* (green) and *arf2* (dark yellow) mutants were also used as positive and negative controls to ensure the Botrytis inoculum was functioning correctly. Shown are the lesion sizes from these leaves as well as the 28 DAS wildtype samples (blue) as before.

where *lesion* is the size of the lesion (in cm^2), *Harvest* is the different ages of the leaf (28, 31 and 35 DAS plants), *Time* is the time after infection (48, 60 and 72 hpi), and the *Error* term indicates the inclusion of an error model made up of the sample leaf number nested within the replicate number. The results from this showed that the infection time was extremely significant, as was the age of the plant (both p-value$<$0.001). This indicated that the plant's susceptibility to Botrytis increases with age. The interaction between these terms was not significant.

As a confirmation that this result was true, a set of controls were also performed using a *bos1* over-expresser, and an *arf2* knockout (Vert *et al.*, 2008). The *bos1* mutant was shown to be more susceptible to Botrytis infection, and the *arf2* mutant is known to be resistant, thus acting as a positive and negative control, respectively. Leaves from both of these mutants were harvested at 28 DAS, and infected with Botrytis. The results are shown in Figure 6.4. As expected, it can be seen that the *arf2* mutant has slightly smaller lesions as compared to the wildtype, and the *bos1* mutant has significantly larger lesions.

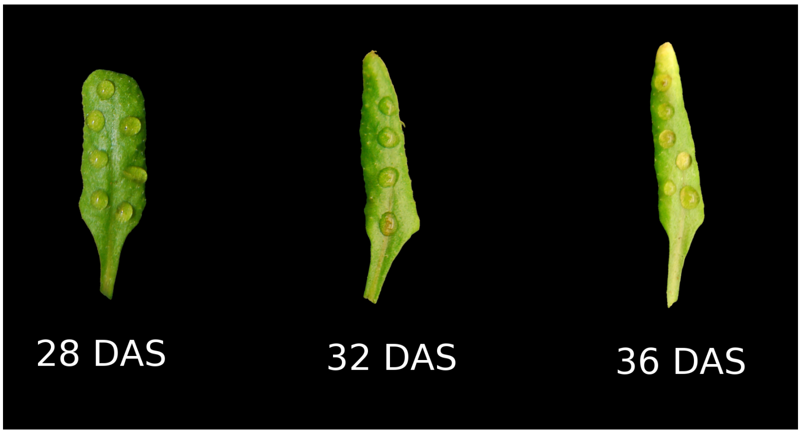


Figure 6.5: Photographs of the sampled leaves showing the different levels of senescence. Also shown are the droplets of the Botrytis inoculum.

6.4. Gene expression analysis in response to multiple stresses

With the confirmation that there is indeed an effect of leaf age on Botrytis infection, a larger experiment was performed to investigate the gene expression for a selection of genes at the various harvest times. Since each replicate sample comes from a different plant, the experiment was designed to take a variety of experimental factors into account. These included the replicate sample number, the treatment that the leaf would undergo, as well as the distance from the air intake vents. This last factor was added as it was hypothesised that the plants closest to the air vents would dry faster than those away from the vents, and thus develop differently.

This experiment was designed as a randomised complete block design with a nested factorial treatment structure, and the layout of the plants in the controlled environment room is illustrated in Figure 6.6. The plants were grown in identical growth conditions as before (Section 2.2.1.1). In the above phenotype experiment, the leaves in the final harvest time (35 DAS) were not showing visible signs of senescence, so this experiment was altered to provide slightly older leaves. Thus, leaves were harvested from plants 28, 32, and 36 DAS, and examples of these leaves are shown in Figure 6.5 to illustrate the different levels of senescence.

From the published Botrytis time series experiment (Windram *et al.*, 2012), it was identified that that majority of genes produced a maximum response between 20 and 28 hpi, and thus these time points were used to investigate the effect of Botrytis infection. Thus, there were five different treatments performed for each harvest time, namely initial time point (T0), 20 hpi mock, 20 hpi infected, 28 hpi mock, and 28 hpi infected. The Botrytis inoculum was created as before (Section 2.2.1.2), and a mock inoculum was created using only half-strength grape juice. Each inoculation was performed by applying multiple 0.6µl droplets of the respective inoculum, ensuring coverage of the leaf (between 3-6 spots, depending on the size of the leaf). Each set of treatments was replicated four times, and the position assigned such that each replicate was a different distance from the air vents. The location of treatments within each harvest replicate was randomly determined. Each treatment consisted of plants that were grown in sets of 12 plants, and the 10 most similar leaves were selected. Thus, each treatment had 10 replicate samples. At the appropriate time point, the whole leaf samples were snap frozen in liquid nitrogen and stored at -80ºC.

For each treatment in each of the harvests, a RNA extraction was performed. Six of the ten replicate leaves from each treatment were ground together using a pestle and mortar that was chilled with liquid nitrogen. A portion of this ground sample (approximately 1g) was used to extract the RNA. The RNA was extracted, and from this, cDNA was synthesised, as described in Sections 2.2.1.3-2.2.1.5.

In order to perform a quantitative PCR (qPCR) analysis, it is necessary to identify a gene of interest. A number of genes were selected either as they were predicted from the pre-existing PRESTA data (Section 1.2) to be highly connected to other stress response genes in both Botrytis or senescence, or possessing an interesting expression profile. Genes were selected based on the strength and shape of the response to both senescence and Botrytis infection. *ANAC092* (*AT5G39610*) and *ANAC055* (*AT3G15500*) were selected as they have been shown to be involved in a number of stress responses, as well as being co-expressed with other stress response genes (see Section 5.5). *LHCA6* (*AT1G19150*) and *OXI1* (*AT3G25250*) were selected they both showed strong responses to the Botrytis and senescence stresses, and *NFYA7* (*AT1G30500*) was selected as it showed opposite shape of response between the two stresses. The fitted models from the regression analysis (Chapter 3) for these genes in the two single stress datasets are shown in Figure 6.7. *PUX1* (*AT3G27310*) was used as the reference gene, as it was found to be unchanging in response to both Botrytis and senescence (Appendix A). Thus, by using this gene, it is possible to obtain a baseline level of gene expression. The primer sequences for these genes are listed in Table B.1.

A set of qPCR experiments were performed using samples from these different treatments. Using *PUX1* as a reference gene, it was possible to obtain $-\Delta Ct$ values as a measure of gene expression.

$$-\Delta Ct = -(Ct_{gene} - Ct_{PUX1})$$

where the Ct values are the cycle number required to pass a detection threshold and obtained using the `qpcR` package in R (Section 2.2.1.6). All $-\Delta Ct$ values are provided in Appendix D. This value provides an indication of how up- or down-regulated a gene expression level is relative to a baseline reference level. By plotting the $-\Delta Ct$ values for the various genes, the gene expression changes between the various infection time points and the different plant ages could be visualised. These figures are all shown below.

To provide statistical support, and based on the experimental design illustrated in Figure 6.6, the data was modelled using the following formula:

$$-\Delta Ct \sim Harvest * (Tzero/(Time * Inoc)) + BioRep. \tag{6.1}$$

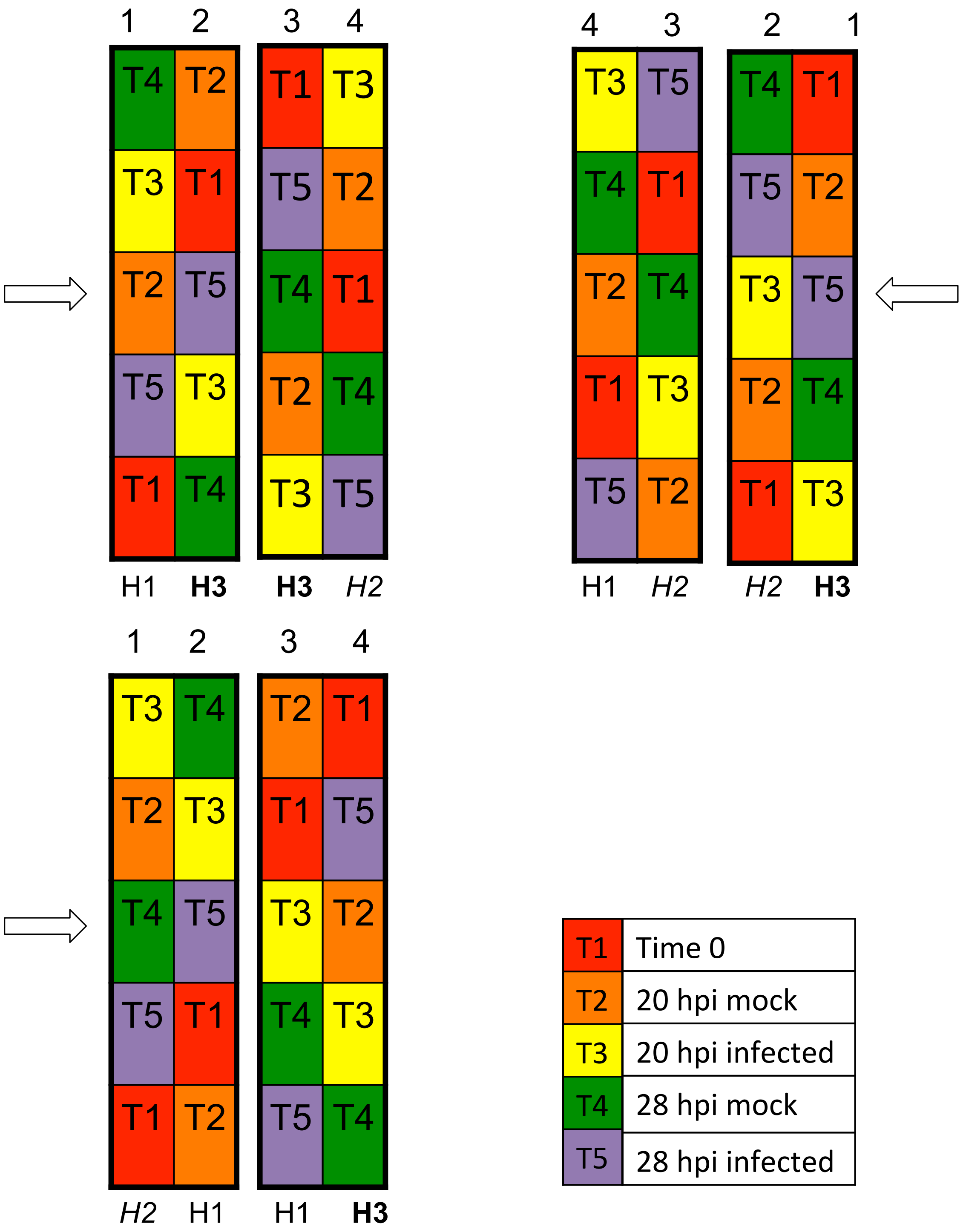


Figure 6.6: Illustration showing the randomised complete block design with a nested factorial treatment structure. Each coloured block represents a set of 12 plants with a specific treatment. The arrows on the side represent the location of the air vents, the numbers above each block is the biological replicate number, and the number below is the harvest number (H1: 28 DAS, H2: 32 DAS, H3: 36 DAS).

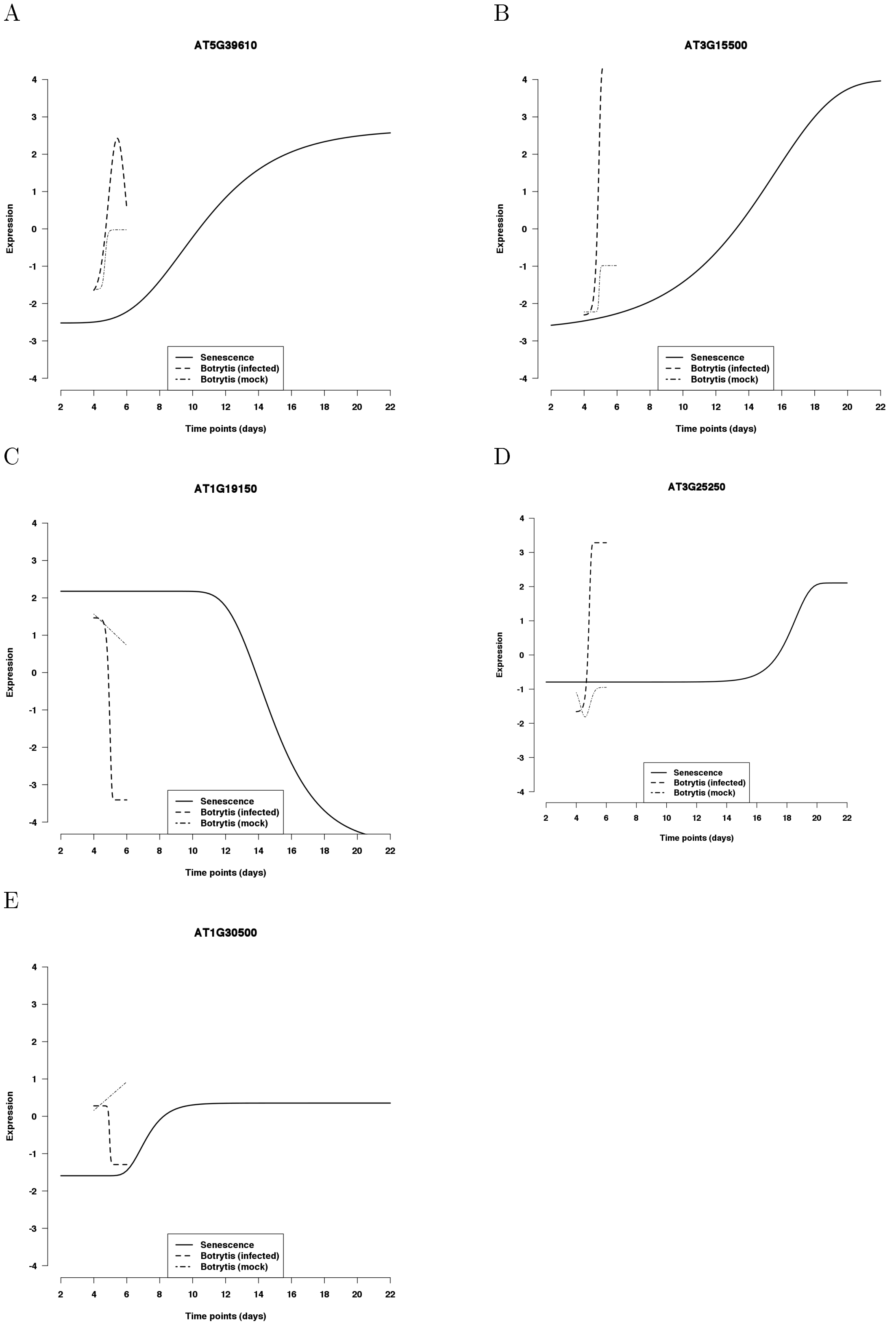


Figure 6.7: Plots of the fitted models from the microarray experiments, plotted on the same time scale. Shown is the data from the senescence experiment (solid line), as well as the Botrytis (dashed line) and Botrytis mock (dotted line). (A) *ANAC092*, (B) *ANAC055*, (C) *LHCA6*, (D) *OXI1*, (E) *NFYA7*

This meant that the $-\Delta Ct$ values were modelled on all the interactions between age of the plants (*Harvest*), and the effect of the time after infection (*Time*) and whether it is a mock or Botrytis infection (*Inoc*) nested within the first time point (*Tzero*). The biological replicate number (*Biorep*) was a block effect included to determine if there were any significant differences between the biological replicates, for example, the effect of being grown on different shelves.

In the *ANAC092* expression analysis (Figure 6.8A), it can be seen that there are expression differences between the various harvest times, as well as an effect following Botrytis infection. There are distinct differences between the different aged leaves at the first time point of infection, indicating that the expression of *ANAC092* increases with the age of the plant. This confirms the result seen in the original time course (Figure 6.7A). Botrytis infection also appears to increase expression of *ANAC092*, again confirming results from the original time course experiment.

Using the model described in Equation (6.1), an ANOVA table was generated to describe the significance of each of the treatment effects, and is shown in Table 6.1. From this table, it could be determined that the differences in expression levels between the first time point (T0) and all others were extremely significant (p-value$<$0.001). In addition, it also indicated that there were significant differences between the harvests (p-value$<$0.01), and a significant interaction between the different inocula used in the treatment and time after inoculation (p-value$<$0.01).

This table of ANOVA values could also be used to provide an indication of the variability between the technical replicates, through the use of the standard error of difference between two means (SED). This was calculated as

$$SED = \sqrt{\frac{2 \cdot RMS}{r}}$$

where *RMS* is the residual mean square from the ANOVA, and r is the number of biological replicates (in this case, four).

By extracting the effects the significant terms from the ANOVA analysis, it was possible to construct a predicted model. The magnitude of each effect was determined using the `model.tables` function in R, and together with the grand mean, a predicted model could be produced to reconstruct the expression profiles. This model removes the noise, and thus provides a clearer illustration of the gene expression profiles.

The model is illustrated for *ANAC092* in Figure 6.8B. From this figure, it can be seen that there are significant differences between the different harvests, where the 28 DAS plants have the lowest expression levels, and the 36 DAS have the highest expression levels, implying that the expression level increases as the plants age. In all the different aged plants, the expression in both mock and Botrytis inocula increased at the 20 hpi time point, and at the 28 hpi time point, the expression level decreased in the mock treated samples, but increased in the Botrytis treated samples.

A similar situation could be seen in the *ANAC055* gene expression profile (Figure 6.9A). Again, gene expression starts at different levels, depending on the age of the plants, and upon infection, the gene expression level increases to a relatively similar

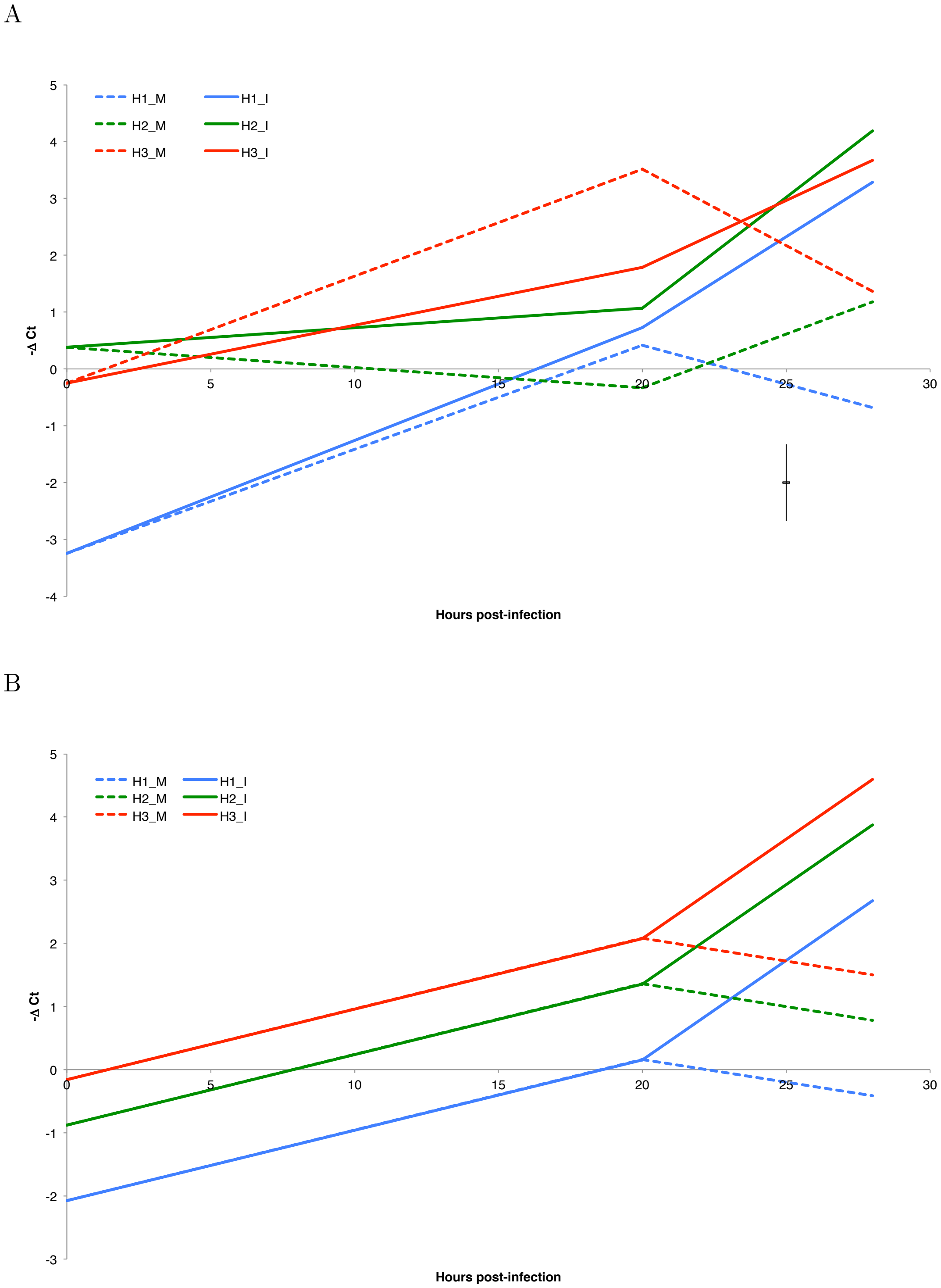


Figure 6.8: (A) *ANAC092* expression analysis, showing the gene expression for this gene at 0, 20 and 28 hours after Botrytis infection. The blue lines are the 28 DAS plants (H1), the green lines are 32 DAS plants (H2), and the red lines are the 36 DAS plants (H3). The solid lines represent the infected data, and the dotted lines are the mock data. The black bar indicates the standard error of differences between two means (SED=1.35). (B) The Predicted model for *ANAC092* analysis, using the significant interaction terms in Table 6.1.

Table 6.1: Full ANOVA table for *ANAC092* showing all the interaction terms and p-values for the terms. *BioRep* is the biological replicate number, *Harvest* is the age of the plant, *Tzero* is an indication if the sample is from the first time point, and *Inoc* indicates whether the treatment was with the mock or Botrytis inocula. A colon (:) between terms indicates an interaction between those terms.

	df	SS	MS	F value	P(>F)
BioRep	3	2.37	0.79	0.217	0.88403
Harvest	2	37.58	18.79	5.163	0.00989
Tzero	1	71.08	71.08	19.528	6.85E-005
Tzero:Time	1	11.32	11.32	3.11	0.08509
Tzero:Inoc	1	28.63	28.63	7.865	0.0076
Harvest:Tzero	2	14.85	7.42	2.04	0.14273
Tzero:Time:Inoc	1	28.78	28.78	7.908	0.00745
Harvest:Tzero:Time	2	12.37	6.18	1.699	0.19524
Harvest:Tzero:Inoc	2	9.44	4.72	1.296	0.2843
Harvest:Tzero:Time:Inoc	2	3.41	1.71	0.469	0.629
Residuals	42	152.87	3.64		

level. The mock profiles were relatively unchanging across the time points.

The significant interaction terms from the ANOVA model for *ANAC055* and all subsequent genes are shown in Table 6.2. For *ANAC055*, the significant terms were again harvest time, and T0 vs. the later time points (both p-value$<$0.001). The expression values were significantly different between the times after infection (20 vs. 28 hpi, p-value$<$0.05), as was the different between the different inocula treatments (mock vs. infected, p-value$<$0.001). The interaction between the different inocula treatments and time after infection was also significant (p-value$<$0.001), indicating that there were significant differences between the mock and infected treatments, and the infection time (20 vs. 28 hpi).

The predicted model for *ANAC055* using these significant terms is shown in Figure 6.9B. As would be expected from the significant terms described above, the predicted model has different starting expression levels for each of the different harvests (although the 32 and 36 day old plants are relatively similar). In addition, the mock and Botrytis inocula treatments are significantly different from each other, as are the expression values for the 20 and 28 hpi time points.

In the *LHCA6* analysis (Figure 6.10A), the gene expression decreased in the mock infected samples in a similar pattern to both of the original time series experiments. The 36 day old plants had a lower expression level than the other plants (28 and 32 DAS), although these younger plants appeared to have a larger change in expression level as compared to the 36 DAS samples. As expected, the Botrytis infection resulted in a further down-regulation of the gene expression levels.

From the ANOVA, there are significant differences in the expression levels between the T0 points and subsequent time points (p-value$<$0.001), as well as first time point compared to the mock and Botrytis inoculated samples (p-value$<$0.01) (Table 6.2).

From these terms, it can be determined that the main effect is the effect of the inoculation treatment, and the plant age is not significant. This is shown in the predicted

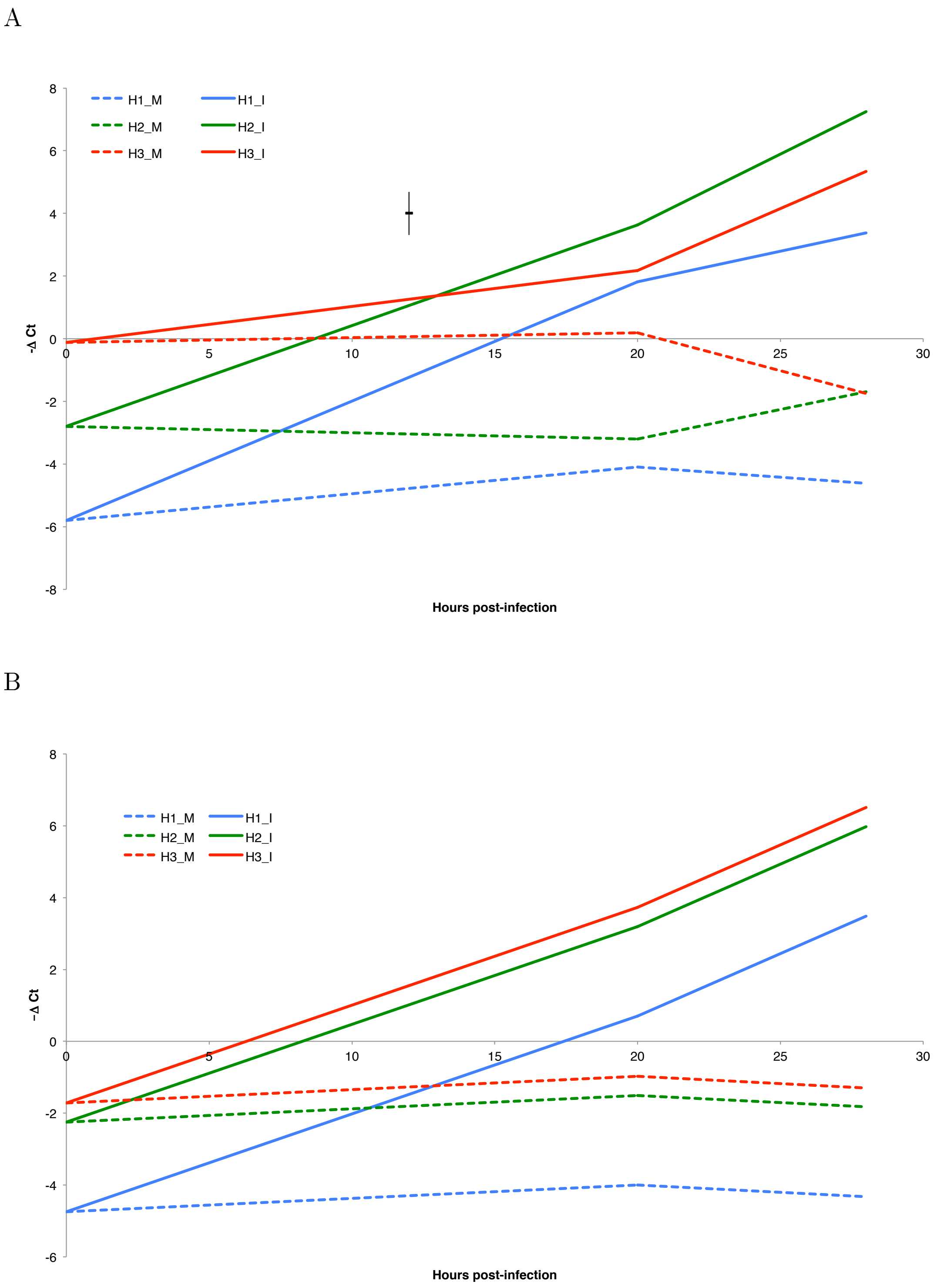


Figure 6.9: (A) *ANAC055* expression analysis, showing the gene expression for this gene at 0, 20 and 28 hours after Botrytis infection. The blue lines are the 28 DAS plants, the green lines are 32 DAS plants, and the red lines are the 36 DAS plants. The solid lines represent the infected data, and the dotted lines are the mock data. The black bar indicates the standard error of differences between two means (SED=1.38). (B) The predicted model for *ANAC055* analysis, using the significant interaction terms in Table 6.2.

Table 6.2: Table of all significant interactions for each gene.

Gene	Term	F value	P (>F)
ANAC055	Harvest	13.865	2.38E-005
	Tzero	32.997	9.26E-007
	Tzero:Time	4.797	0.03412
	Tzero:Inoc	132.45	1.44E-014
	Tzero:Time:Inoc	7.635	0.00846
LHCA6	Tzero	17.143	0.000163
	Tzero:Inoc	11.901	0.00129
OXI1	Harvest	23.154	2.09E-007
	Tzero	98.203	2.51E-012
	Tzero:Time	5.86	0.0201
	Tzero:Inoc	610.689	<2.00E-016
	Harvest:Tzero	3.247	0.0494
	Harvest:Tzero:Inoc	4.553	0.0165
NFYA7	BioRep	4.308	0.00976
	Tzero:Inoc	4.377	0.04252

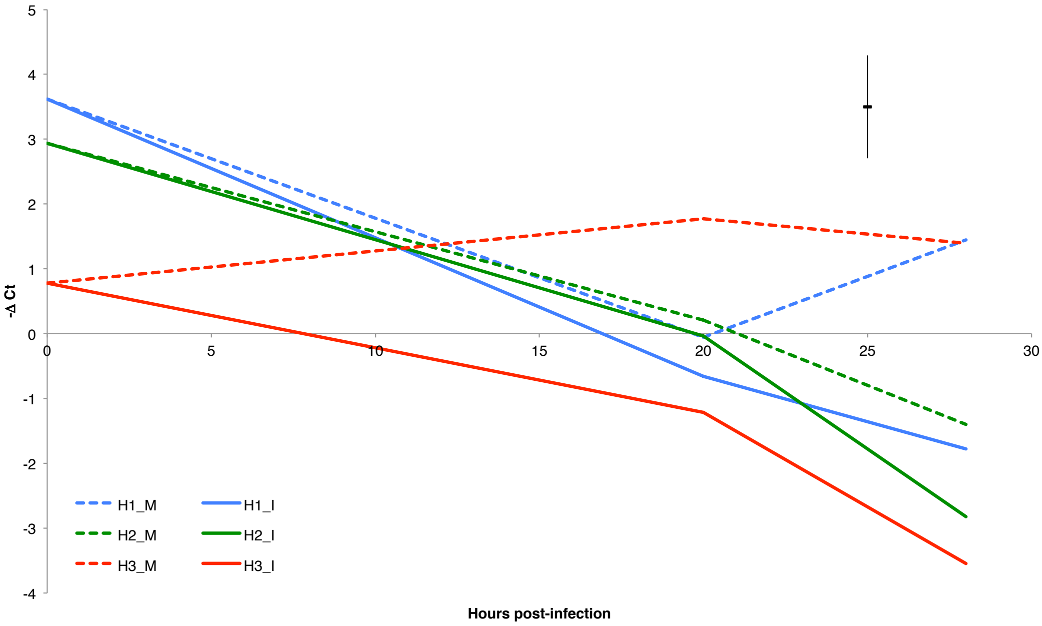
model (Figure 6.10B). In both the mock and Botrytis infection cases, the gene expression decreases, although the Botrytis infection results in a greater decrease in gene expression. However, after 20 hour post-infection, the expression level remains constant.

Similarly, with the *OXI1* data (Figure 6.11A), it could be seen that the initial time points had relatively different expression levels between the different harvests, but all approached the same expression in the Botrytis infection. The expression levels from the mock samples remained relatively constant.

Using the ANOVA analysis (Table 6.2), there were a large number of significant interaction terms. The most significant terms were the differences between the expression levels of the different aged plants, the differences between the T0 levels and later levels, and differences between the inocula treatment (all p-value<0.001). Other significant terms were the difference between the times after infection, and an interaction of the different harvests, the initial time point and the inoculation treatment (both p-value<0.05). These effects can also be seen in the predicted model (Figure 6.11B), where there are difference between the different aged plants, the mock and Botrytis infection, and difference between the times after infection.

In contrast to the previous genes, the *NYFA7* expression profiles had opposite directions of response in the different stresses, namely increasing in senescence and decreasing in Botrytis. In the combined stress experiment (Figure 6.12A), it can be seen that the data is very different compared to the previous genes, with the expression profiles seemingly cancelling each other out, resulting in a flat profile. In addition, the range of response is much smaller in comparison to the previous genes. In the mock samples, the expression profiles seemed to follow the increasing pattern with an increase in age. In the infected samples, there was a slight decrease in expression, although not as pronounced as in the single stress experiment. The SED for this data was large relative to the range of response (SED=1.74), and this could possibly be distorting the results.

A



B

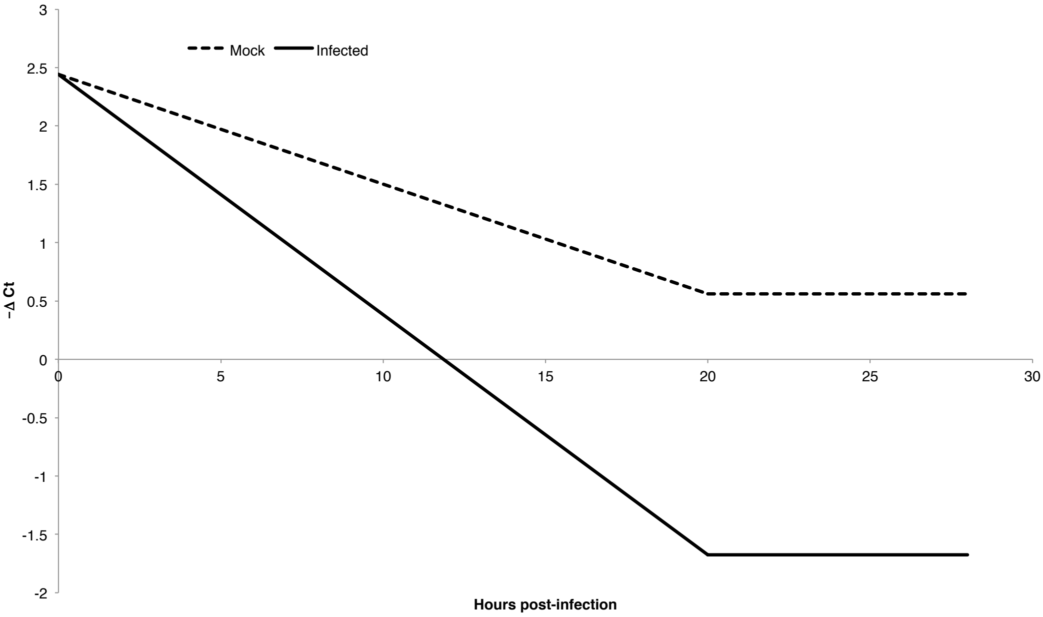


Figure 6.10: (A) *LHCA6* expression analysis, showing the gene expression for this gene at 0, 20 and 28 hours after Botrytis infection. The blue lines are the 28 DAS plants, the green lines are 32 DAS plants, and the red lines are the 36 DAS plants. The solid lines represent the infected data, and the dotted lines are the mock data. The black bar indicates the standard error of differences between two means (SED=1.59). (B) The predicted model for *LHCA6* analysis, using the significant interaction terms in Table 6.2. The solid line is the infected samples, and the dotted line is the mock samples.

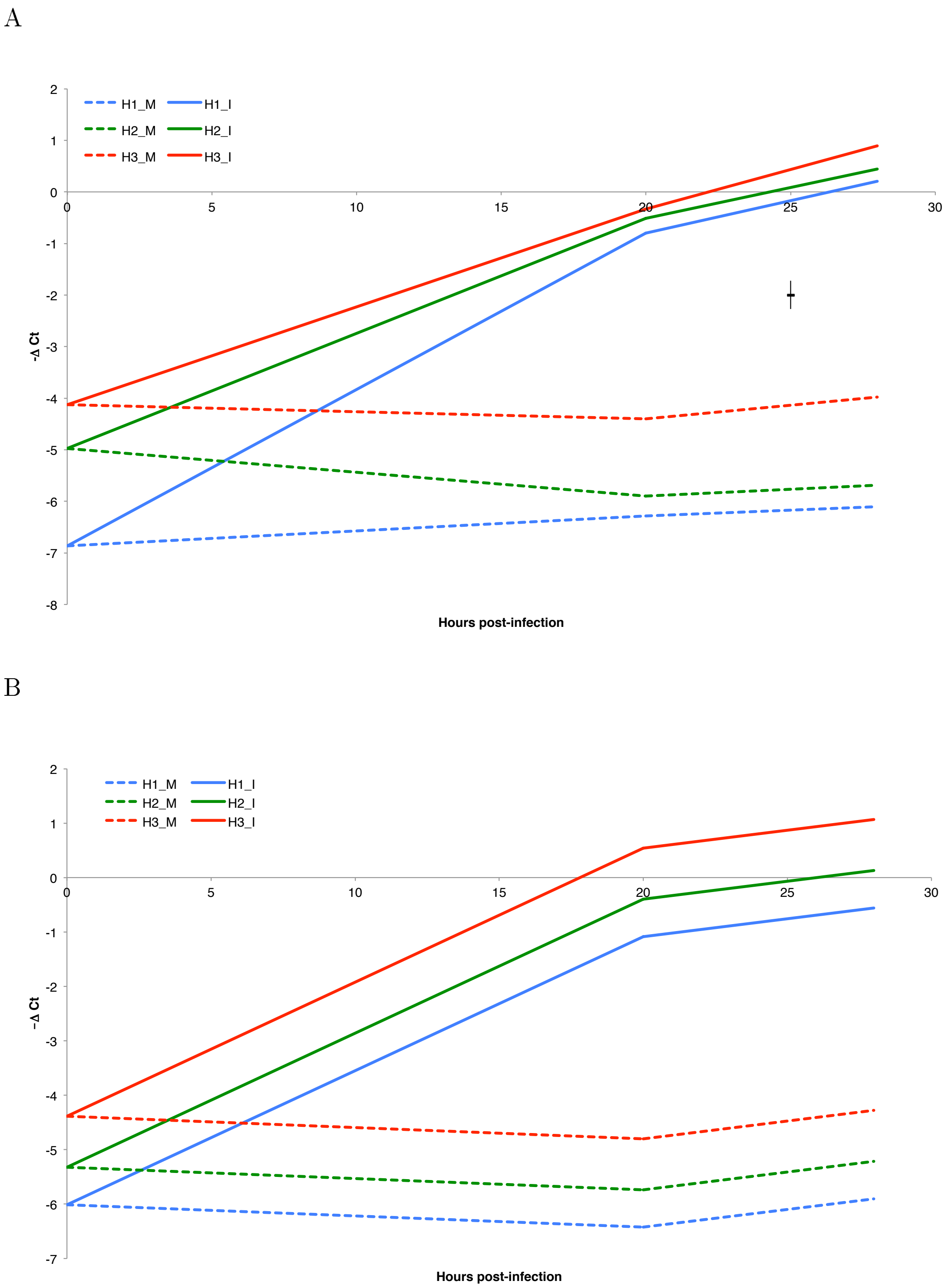


Figure 6.11: (A) *OXI1* expression analysis, showing the gene expression for this gene at 0, 20 and 28 hours after Botrytis infection. The blue lines are the 28 DAS plants, the green lines are 32 DAS plants, and the red lines are the 36 DAS plants. The solid lines represent the infected data, and the dotted lines are the mock data. The black bar indicates the standard error of differences between two means (SED=0.54). (B) The predicted model for *OXI1* analysis, using the significant interaction terms in Table 6.2.

The ANOVA analysis (Table 6.2) showed that there was a slight interaction between the inoculation treatments (p-value<0.05), and interestingly a significant difference between biological replicates (p-value<0.01). As a result, the predicted model (Figure 6.12B) has different models depending on the biological replicate. The predicted model also makes clearer the effect of the Botrytis treatment, where the mock treated samples showed an increase in gene expression, and the Botrytis infected samples resulted in an unchanging expression profile.

6.5. Real-time investigation of gene expression

In addition to the qPCR experiment, luciferase transcriptional reporters were used to identify real-time changes in the transcription of the genes of interest. The full intergenic region was used for *NFYA7* (819 bp), but only a portion of the *ANAC092* promoter (1050 bp) was used. The full length of the *ANAC092* intergenic region was over 4000 bp, but a fragment was used as there were difficulties cloning the entire promoter into the vector. Nonetheless, it was assumed that the majority of promoter regions were primarily found in the first 1000 bp upstream of the transcription start site. The promoters were cloned using the Gateway cloning system, and described in Section 2.2.2. This process is summarised in Figure 6.13. The promoter fragments were amplified from genomic DNA (using the primers in Table B.2), and cloned into the donor vector, pDONR/Zeo. The promoter fragment was then cloned into the destination vector, pBGWL7, which contained the luciferase cloning sequence and a Basta (glufosinate herbicide) resistance gene. This vector was used to transform *Agrobacterium tumefaciens* cells, which were used to transform Arabidopsis plants. Successful transformants (T1 plants) were selected from seedlings that were able to grow in the presence of Basta. Several seedlings were grown to seed, and these seeds were again selecting using Basta, to identify heterozygous transformants (sets of seeds that segregated in the presence of Basta with a 3:1 ratio of growth:death). Resistant T2 plants were transplanted to soil and grown until the correct age for the luciferase experiment.

An additional reporter was obtained for *OXI1*, which contained a 1500 bp promoter fragment fused to the luciferase coding sequence (created by Robert Ingle, University of Cape Town).

This experiment allowed for the observation of the genes of interest over a period of time, and did not require the samples to be destroyed at each time point. Seeds were sown to obtain leaves 28, 32 and 36 DAS, and the plants were grown in the same conditions as before (Section 6.3). The leaves were treated with the Botrytis and mock inocula as previously described (Section 2.2.1.2). A liquid nitrogen cooled CCD camera was used to capture the fluorescence intensity every two hours for 40 hours after infection, and these images were analysed with ImageJ (Section 2.2.2.7). The fluorescence was quantified as the mean grey intensity of the leaves (Figure 6.14).

The data for the *OXI1*:luc expression is shown in Figure 6.15. It can be seen that there are some differences between the different aged plants, most notably that the oldest plants (36 DAS) show a greater amount of activity. The 28 and 32 DAS plants

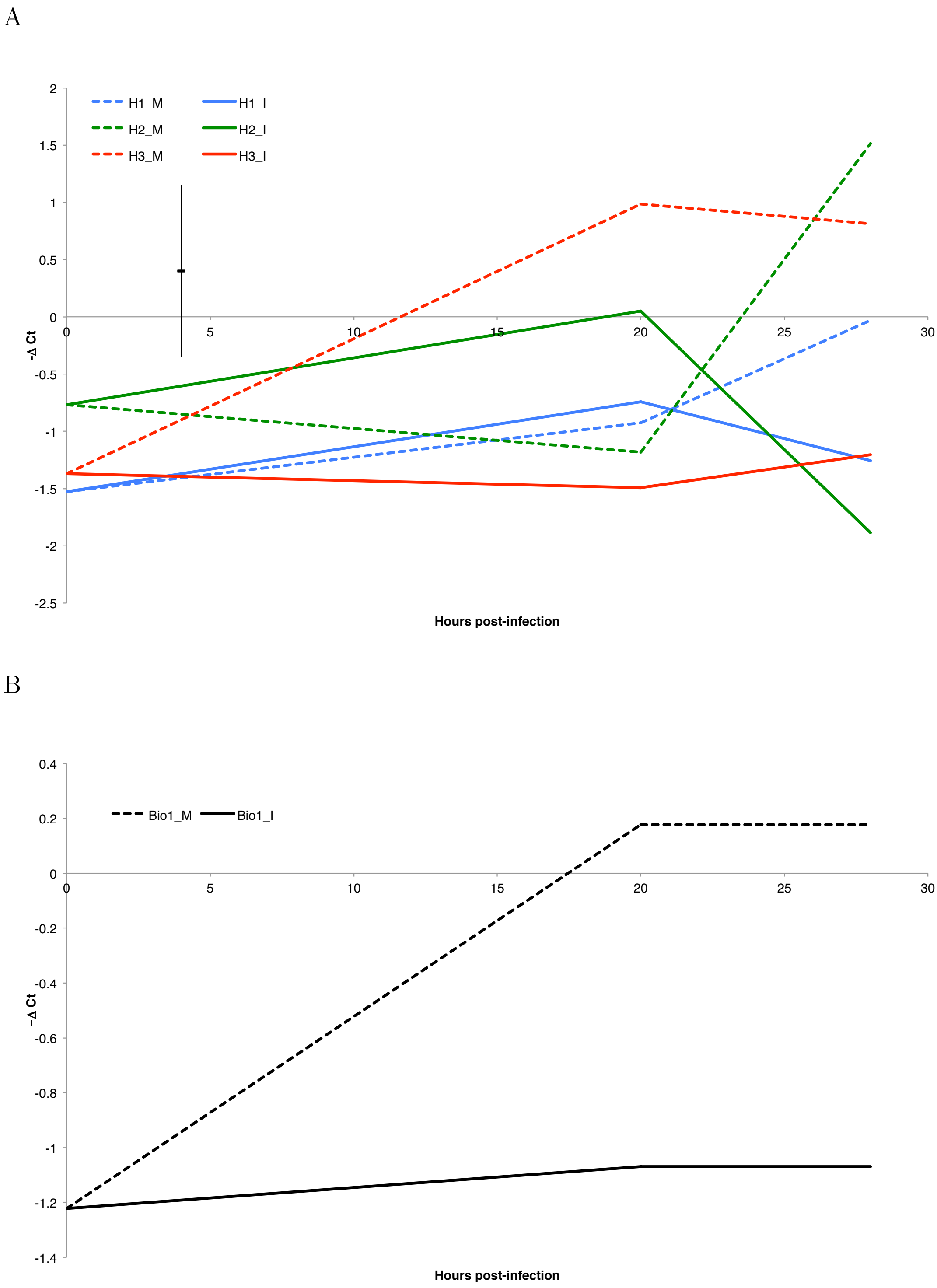


Figure 6.12: (A) *NFYA7* expression analysis, showing the gene expression for this gene at 0, 20 and 28 hours after Botrytis infection. The blue lines are the 28 DAS plants, the green lines are 32 DAS plants, and the red lines are the 36 DAS plants. The solid lines represent the infected data, and the dotted lines are the mock data. The black bar indicates the standard error of differences between two means (SED=1.50). The predicted model for *NFYA7* analysis, using the significant interaction terms in Table 6.2. The main effects were the biological replicate (not shown) and inoculation treatment. Each biological replicate would change the baseline expression level, so the results of only one biological replicate is shown. The solid lines represent the infected data (I), and the dotted lines represent the mock data (M).

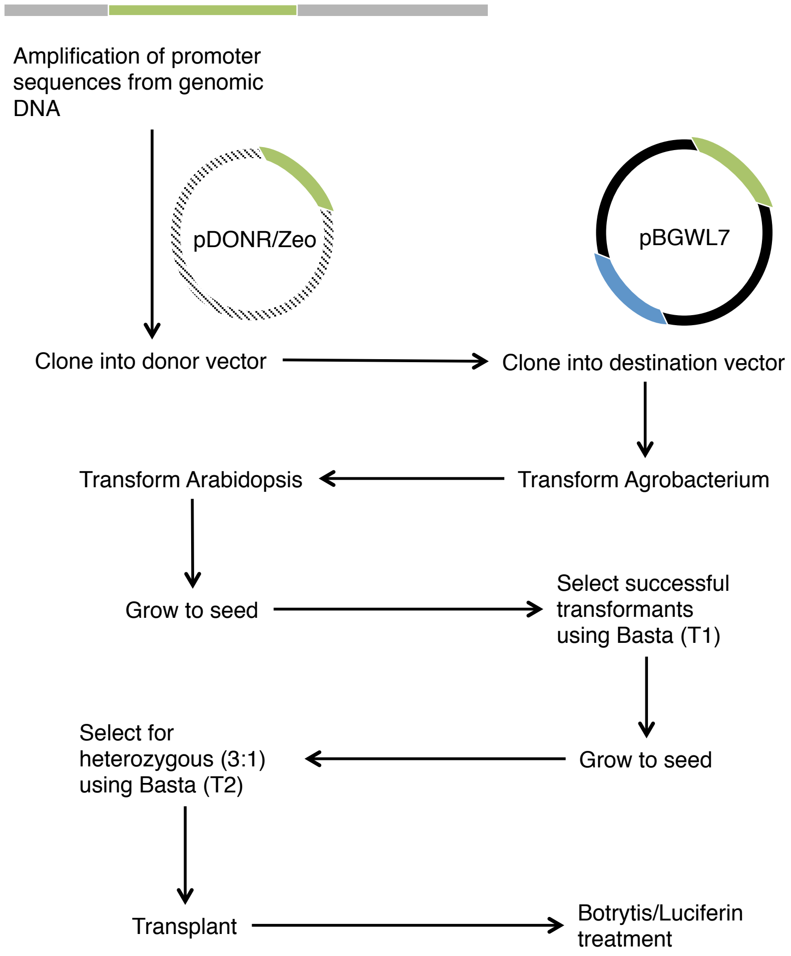

Amplification of promoter sequences from genomic DNA
pDONR/Zeo
pBGWL7
Clone into donor vector
Clone into destination vector
Transform Arabidopsis
Transform Agrobacterium
Grow to seed
Select successful transformants using Basta (T1)
Select for heterozygous (3:1) using Basta (T2)
Grow to seed
Transplant
Botrytis/Luciferin treatment

Figure 6.13: Diagram describing the transformation procedure, where the promoter fragment (green) is amplified, and cloned into a vector containing the luciferase coding sequence (blue). These vectors were integrated into Arabidopsis plants using an Agrobacterium transformation. Successful transformants were used in the luciferase experiment.

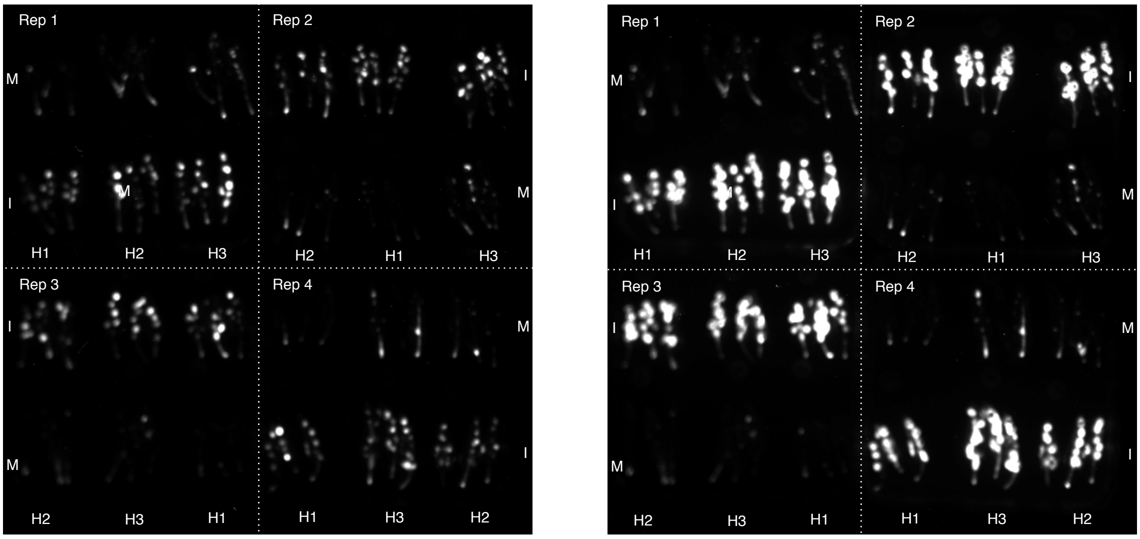


Figure 6.14: Images from the *OXI1*:luciferase construct at 20 (left) and 28 (right) hpi. *H1*, *H2* and *H3* refer to plants that were 28, 32 and 36 days old, *M* refers to treatment with the mock inoculum, and *I* refers to treatment with the Botrytis inoculum. Each treatment was performed using three biological replicates, and replicated using four technical replicates.

had a similar fluorescence profile over the time series. Similar to the qPCR data (Figure 6.11), the mock treatments exhibited a relatively constant expression level.

A tailing-off effect can be observed after approximately 30 hpi, and this is due to the Botrytis infection resulting in large lesions in the leaf where the plant tissue had died. The measurements were taken as an average over the area the leaf, and thus the "holes" in the fluorescence resulted in a lower average fluorescence value.

An ANOVA was performed at time points 20, 28 and 38 hpi to compare the results. The model used was

$$Luc \sim Harvest * Time * Inoc + Error(Rep/Group/Leaf/Time)$$

where *Luc* is the mean grey intensity, *Harvest* is the different age of leaf (28, 32 and 36 DAS plants), *Time* is the time after infection (20, 28 and 38 hpi), *Inoc* indicates whether the leaf was treated with the mock or Botrytis inocula, and the *Error* model made up of the infection time nested within the sample leaf number (*Leaf*), which is nested within position of the group of leaves (*Group*), which is nested within the replicate number (*Rep*). The results indicated that there were significant differences between the times after infection, between the mock and infected treatments, and the interaction between these two terms (all p-value<0.001). The age of the leaf was also significant, although at a lower significance level (p-value<0.05).

The results for the *ANAC092*:luc and *NFYA7*:luc expression is shown in Figures 6.16 and 6.17, respectively. In these cases, the amount of fluorescence was relatively low. In the *ANAC092*:luc results (Figure 6.16), a similar trend to the previous qPCR result can be seen, where the gene expression is up-regulated in both the mock and Botrytis

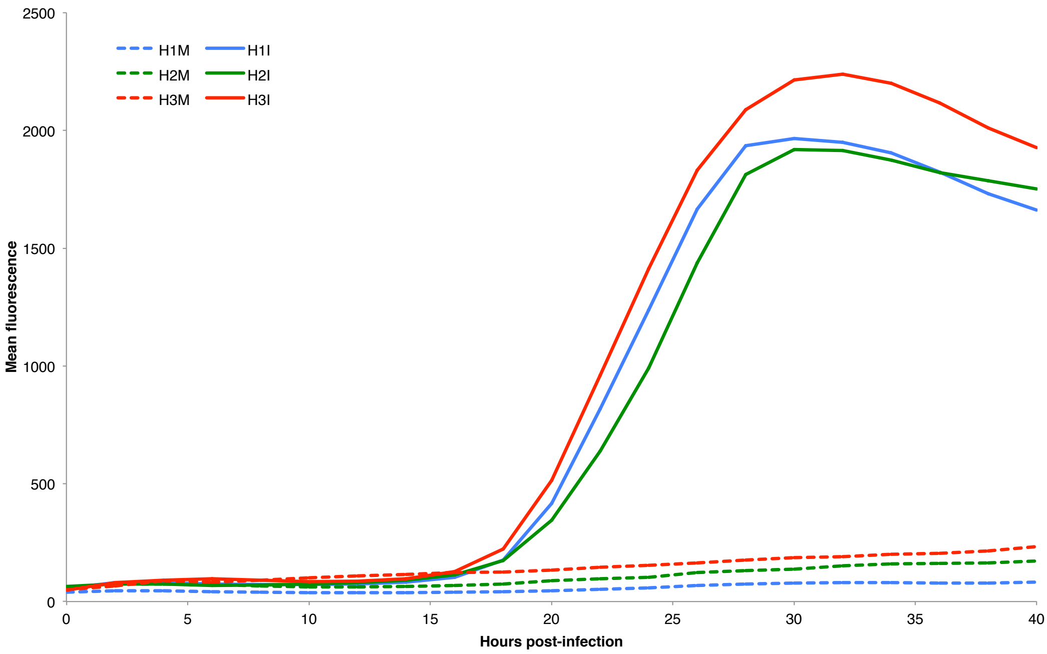


Figure 6.15: Luciferase results from the *OXI1*:luciferase reporter construct. Shown is the mean fluorescence (mean grey pixel intensity) for the different treatments. The blue lines are the 28 DAS plants, the green lines are 32 DAS plants, and the red lines are the 36 DAS plants. The solid lines represent the infected data, and the dotted lines are the mock data.

infection treatments until approximately 20 hpi. Following this, the expression profiles diverge slightly. In the 28 DAS plants, the infected sample had a higher expression level compared to the mock, which is the same as the qPCR and microarray results. However, the opposite is true for the 32 and 36 DAS plants.

In the *NFYA7*:luc results (Figure 6.17), the expression profiles across all the treatments were relatively noisy, and in general did not follow the expression profile expected from the qPCR and microarray data (that is, up-regulated in the mock treatment, and down-regulated in the infected treatment). The average fluorescence was lower than the *ANAC092* data, so this could be attributed to noise.

6.6. Discussion

In this chapter, the gene expression changes in response to a combined stress were investigated. With the use of the parametric models discussed in Chapter 3, new predictive models could be identified through the manipulation of the single stress model parameters. To this end, an experiment was designed whereby gene expression changes in responses to combined stresses could be investigated. Several different genes were selected based on shape and response to determine the effect of the multiple stresses.

The techniques used to determine the effect of multiple stresses of gene expression levels included quantitative PCR (qPCR), and a luciferase transcription reporter. qPCR experiments are relatively quick and simple to perform, especially once the RNA is extracted and reverse transcribed to cDNA. However, due to the destructive nature of

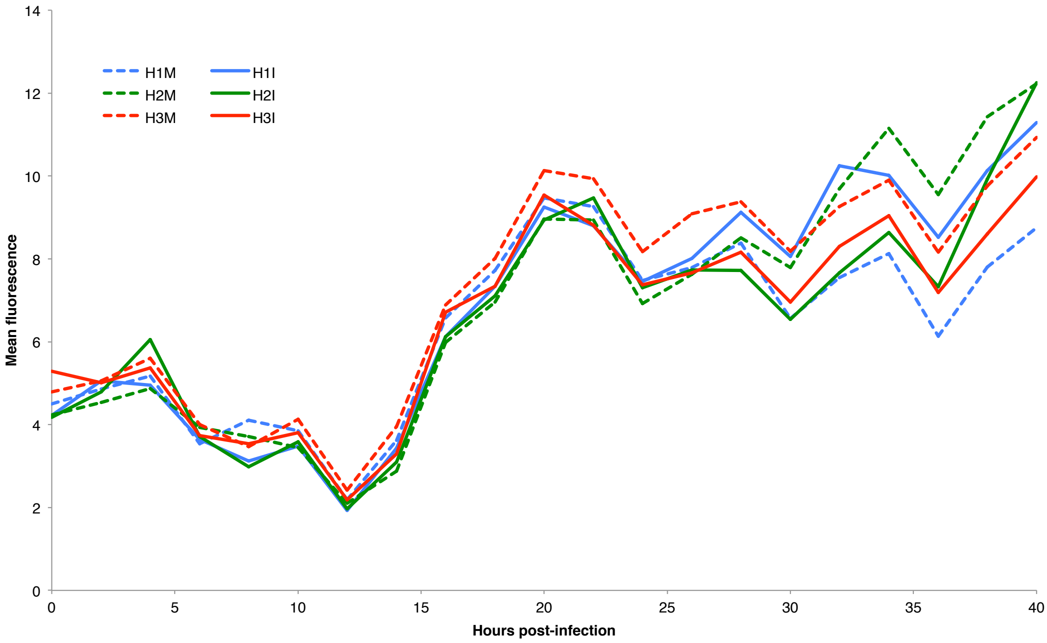


Figure 6.16: Luciferase results from the *ANAC092*:luciferase reporter construct. Shown is the mean fluorescence (mean grey pixel intensity) for the different treatments. The blue lines are the 28 DAS plants, the green lines are 32 DAS plants, and the red lines are the 36 DAS plants. The solid lines represent the infected data, and the dotted lines are the mock data.

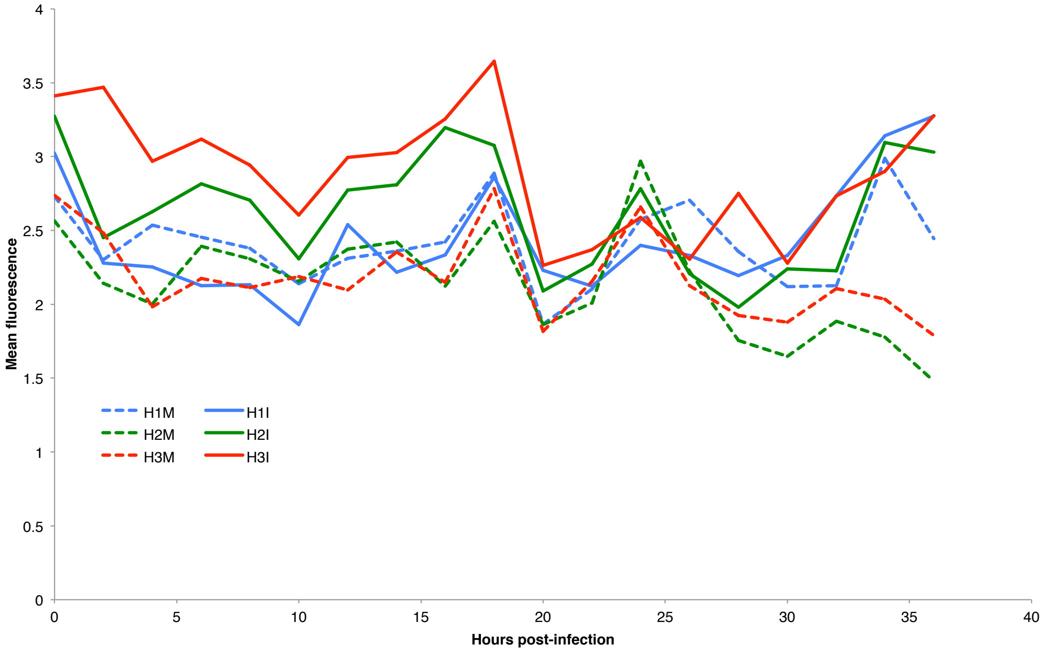


Figure 6.17: Luciferase results from the *NFYA7*:luciferase reporter construct. Shown is the mean fluorescence (mean grey pixel intensity) for the different treatments. The blue lines are the 28 DAS plants, the green lines are 32 DAS plants, and the red lines are the 36 DAS plants. The solid lines represent the infected data, and the dotted lines are the mock data.

DNA sampling in plants, the data was limited to the time points sampled. In addition, qPCR experiments require the use of a fluorescent binding agent, which is relatively expensive. On the other hand, the luciferase reporter is extremely laborious to construct, taking several months to produce the transformed plant lines. However, once the construct has been obtained, the experiment is much simpler than the qPCR experiment, and allows for multiple time points to be observed from each leaf sample, since the replicate samples are not destroyed in the sampling process. A possible caveat is that the reporter relies on the transcription and translation of the luciferase enzyme, and may not be completely related to the true transcription rate.

While qPCR experiments are useful for determining the expression profile for a number of genes, there is some disagreement about the subjectiveness surrounding qPCR data, particularly due to effects such as transcript quality, random priming, issues with normalisation, and the choice of reference gene (Bustin *et al.*, 2005; Nolan *et al.*, 2006). Typically genes such as actin or GAPDH are used as a reference gene, although there is some contention as to their value as a universal reference gene. It is generally preferable to identify a reference gene depending on the conditions used in the experiment (Bustin *et al.*, 2005), and as such, *PUX1* was used in this experiment, as it was found to be unchanging in both the senescence and Botrytis time series (Appendix A).

By performing an ANOVA analysis on the qPCR expression profiles, it was possible to identify the most significant terms in the model. Generally, the most common significant terms were the *Harvest* and *Tzero:Inoc* terms, indicating that there were significant differences between the different aged plants, and differences between the mock and infected treated samples. The exceptions to this were the *LHCA6* and *NFYA7*. *LHCA6* did not have a significant *Harvest* term, indicating that there were no differences between the different aged plants. Since this gene encodes a photosynthesis protein, it is possible that the gene is constantly down-regulated after a certain point during the senescence process where the production of photosynthesis proteins are no longer needed. *LHCA6* did have a *Tzero:Inoc* term, indicating that Botrytis infection further influences the expression of this gene. *NFYA7* on the other hand had a significant *BioRep* term. This was interesting as it indicated that there was a significant effect caused by the different biological replicates, and was also related to the position of the plant in the growth cabinet. One possible explanation is differences in the range of responses in the expression level of the genes, where the differences was 1 in *NFYA7*, and 7-10 in the other genes. This possibly indicates that the larger signal in the other genes masks the biological variability in the replicates. In addition, across all the genes in the qPCR expression analyses, the SED values were all similar (approximately 1.5), so due to the lower range of response in the *NFYA7* data, this meant there was a lower signal-to-noise ratio for the *NFYA7* as compared to the other genes.

By using the significant effects from the ANOVA tables, it was possible to obtain a predicted model for the gene expression in the combined stress. This predicted model removed the noise from the expression profiles, providing a clearer indication of the underlying expression patterns. In the *ANAC092*, *ANAC055* and *OXI1*, similar effects were found, where there were differences between harvests, infection times, and treatments. From Figures 6.8B, 6.9B and 6.11B, it can be seen that the difference between the mock and infected treatments are the same in the different aged plants. This suggests that the Botrytis infection results in the addition of an expression response to the existing response due to the senescence process. This pattern can also be seen to some extent in the *NYFA7* data (Figure 6.12B), where the up-regulation in the senescence data was seemingly cancelled out by the down-regulation of the Botrytis treatment. This indicates that the up- and down-regulation were added together, thus counteracting each other.

In addition to the qPCR experiment, a luciferase reporter was fused to the promoter of *ANAC092* and *NFYA7* to provide real time information as to the expression of these genes. The advantage in using the luciferase reporter methodology is that it provides additional time points and allows for real-time observation of the gene expression levels. The full intergenic region was used for *NFYA7* (819bp), but only a portion of the *ANAC092* promoter (1050bp) was used. In addition, a luciferase reporter was obtained for *OXI1*.

The results for the *ANAC092* and *NFYA7* luciferase reporters were not very clear, with minimal fluorescence occurring. Due to time constraints, the T2 selected plants were transplanted directly from the media containing BASTA to soil, and then used in the experiments. This meant that the plants were already stressed from the herbicide in the beginning, which may have affected the development. In addition, since a large number of plants were required for the luciferase experiment, plants from different T1 parents were used, so this could add additional variability due to different transformants. Ideally the T2 plants would have been allowed to grow to seed, and the experiment would have used T3, homozygous plants grown on soil from the start. Thus, the low signal in the *ANAC092* and *NFYA7* luciferase reporters could possibly be due to the use of the heterozygous mutants, resulting in a lower signal, or may not have been correctly transformed. Alternatively, it is possibly due to these genes encoding transcription factors, which are generally low abundance proteins. The luciferase reporter relies on observing the activity of the luciferase enzyme, as opposed to the gene expression. That is, the fluorescence is a result of the luciferase coding sequence being transcribed, translated, and reacting with the luciferin, as opposed to merely detecting the transcript level. Thus there could be some differences in the gene expression level, and the translated protein level for these genes. *OXI1* on the other hand encodes a protein kinase, and the results from the luciferase assay were much clearer. Here the expression profiles obtained roughly matched up to the equivalent data from the qPCR experiment. Thus, the use of the luciferase assay can be useful in observing the real-time effects of a stress treatment.

Using these approaches to investigate and predict the effect of multiple stresses on gene expression may aid in determining the best times for sampling. If it is possible to predict the effect of a combined stress, then sampling times and rates can be identified where the response is changing the most. If sampling points are set at random, and are too far apart, the signal may be missed due to mRNA degradation. However, if the sampling points are too close together, these points may be redundant, adding unnecessary expenses to the experiment. Thus, estimating the correct time that processes occur may assist in designing future experiments, to maximise the information obtained, and minimising the cost required (Bar-Joseph, 2004; Rosa *et al.*, 2012).

For more accurate predictions of multiple stresses, it would probably be desirable to investigate the gene regulatory networks, particularly to determine the effect of the result of a combined stress response. That is, investigate the effects of perturbations of upstream genes on downstream genes, to determine which cause the greatest effects. In addition, by investigating the responses of genes, that are up- and down-stream of a gene of interest, it may be able to better predict what the gene's response to a stress, or combination of stresses, would be. In the PRESTA project, a combined stress model is being developed to incorporate all the gene expression data from the various stresses, and identify a core set of genes that are involved in multiple stress responses.

7. Determining the uncertainty of biological replicates

7.1. Introduction

In the previous chapters, an analysis approach was developed for describing a gene expression profile. By fitting these models, it was possible to explain the majority of the variability in the data. In most analyses, including the regression analysis described in the preceding chapters, it is assumed that all between-observation variation is due to measurement error, and assigning this variability to the measured data can result in a model that is less accurate (Aach and Church, 2001; Liu *et al.*, 2010). One source of variability that is generally not considered is the variability due to the "biological age" of the organism at the time it was sampled (Liu *et al.*, 2010). That is, despite limiting the variation from external influences, each individual may be slightly different due to differing developmental rates. This is further confounded in multicellular samples where biological process occur in multiple cells, and each of these processes may occur at different rates and times. Thus this biological variation not only affects the variability in the observed expression values, but also the uncertainty in the time axis.

The issue of biological age is particularly a problem in cross-sectional time course experiments, such as in plant based experiments, where samples at each time point have to be collected destructively, and thus need to be obtained from different individuals. In most experiments, multiple samples are taken at each time point, and each of these repeated samples are called replicates. This is done to improve the confidence in the obtained results, in case of erroneous measurements. However, in cross-sectional studies each replicate is from a different plant. Therefore, it may not be appropriate to simply group together the replicates that were taken at the same sampling point, as these replicates may be of different biological ages.

An illustration of this effect is shown in Figure 7.1. These leaves were sampled from the PRESTA long day senescence screen (Breeze *et al.*, 2011). The top row shows the leaves sampled 19 days after sowing (DAS), and the bottom row shows leaves sampled 33 DAS. Besides differences in the size and shape of the leaves, in the 33 DAS samples, differences between the leaves in their senescence response can be seen, where some leaves exhibit more yellowing than others. These phenotypic differences could have an effect on the underlying genetic responses, and would thus add variability to the data.

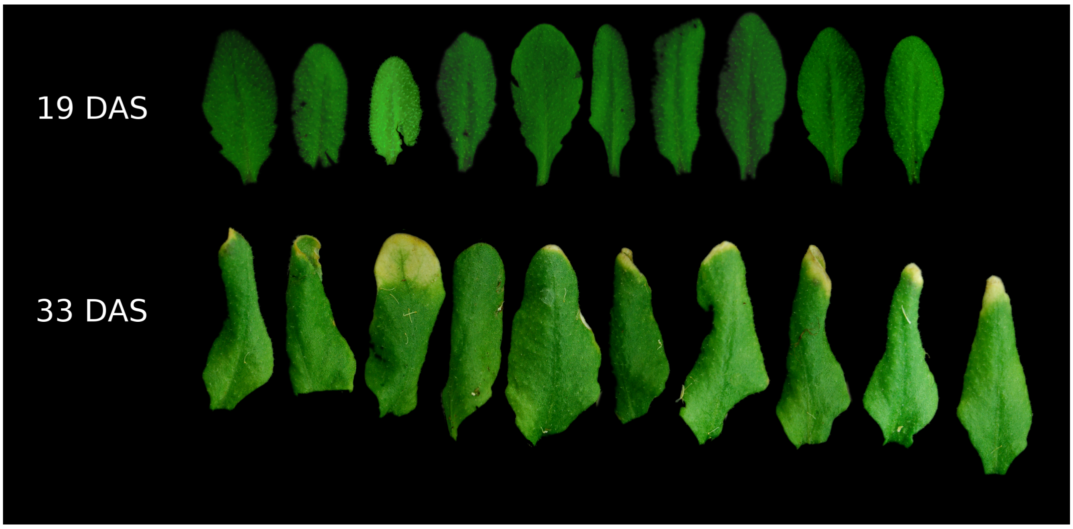


Figure 7.1: Illustration of the variability between different leaves of the same age. Each of the leaves was obtained from a plant of the indicated age (19 days after sowing, top and 33 days after sowing, bottom). Unpublished data (V. Buchanan-Wollaston).

7.1.1. Algorithms for estimating biological age

In order to estimate the effect of biological age, a number of approaches have been proposed. One such approach is described by Sankoff and Kruskal (1983) and Aach and Church (2001), and utilises a method named "dynamic time warping". This algorithm uses principles based in sequence alignment and dynamic programming. This process requires data from two different time series, with k common genes and n time points, although the datasets do not necessarily have to be the same length. The expression levels of the k genes form a k-dimensional vector of length n, and by using a global alignment algorithm, analogous to the Needleman-Wunsch sequence alignment algorithm (Needleman and Wunsch, 1970), the data from the two gene expression datasets are compared. By expanding and compressing the time series (analogous to insertions and deletions in sequence alignments), the optimal time alignment of the two datasets can be found (Aach and Church, 2001). This algorithm warps all of the genes uniformly, so a modified approach would be to individually align each gene and cluster the alignments (Smith *et al.*, 2009).

An alternate approach is the use of Gaussian process regression along with a probabilistic model to determine the uncertainty about the biological age of each replicate, which provides a means to infer the relative degree of development in each replicate sample (Liu *et al.*, 2010). By applying the algorithm to microarray data, it was possible to predict the profile shapes and biological times for each replicate. The predicted shifts were correlated to the independently obtained morphological estimates of the development. The approach was thus able to reduce the prediction error of the sample data, thus significantly reducing the mean squared error in a cross-validation study.

One approach that has not been used much in biological data are total least squares (TLS) and errors in variables (EIV) models, which have a wide range of applications in system theory and signal processing (reviewed in Markovsky and Van Huffel, 2007 and

Van Huffel *et al.*, 2007). In least-squares methods, the response variable is corrected but not the explanatory, and the sum of squared vertical distances from the data to the fitted line is minimised. In TLS models, both the response and explanatory variables can be corrected, and the sum of the squared orthogonal distances from the data points to fitted line are minimised through the use of singular value decompositions. EIV models are similar to the TLS methods, but possess less restrictive assumptions on the error distributions.

7.2. Time shifting

In this analysis, an algorithm was developed that is similar to the process described by Liu *et al.* (2010), except using splines, and termed *time shifting.* Through the use of an iterative, cross-validation, multivariate regression approach (using non-parametric smoothing spline models), the true "biological age" of the samples was estimated. Splines were used in this case as merely a fit for each gene was required, as there was no need for interpretation of the fitted curve, or a comparison of the shapes. Splines are constructed from polynomial pieces that are joined at certain x-positions, called the knots. Specifically, in this analysis, penalised splines were used, which possess a B-spline basis with equally spaced knots, and a difference penalty on the estimated regression coefficients of adjacent B-splines to adjust the smoothness of the curve (Eilers and Marx, 1996, 2010).

The methodology to calculate the time shifts is shown in Figure 7.2. For each gene in a given set, a replicate point was removed from each time point, and a spline was fitted to the remaining data, in a process described in Section 2.1.9. The removed replicates were then replaced into the data, and the time estimated by finding the closest x-value on the spline that corresponds to the y-value of the removed replicate. If the difference in time points between the original value and new, predicted value (termed the *temporal displacement*) is within some constraints, the displacement is retained (Figure 7.3). These constraints were that the temporal displacement was greater than 0.1 time points and less than 1.5 time points, and were used to ensure that the replicates did not unnecessarily wobble (shift only very slightly), or moving large, unreasonable distances. That is to say, even though there may be some uncertainty in the true biological age of a sample, it is unrealistic that a sample would be greater or less than 1.5 time points, given the other experimental constraints, such as sowing time, and controlled environmental conditions. An addition constraint was added for the first and last time points, where the maximum displacement allowed was 0.5 time points outside of the observed data points. This was to prevent excessive extrapolation into regions where there was no information. These data points were still allowed to move 1.5 time points "inwards". These constraints also served to act as stopping conditions for the time shifting process.

The displacement was determined for all the genes, and an average displacement was calculated for all retained displacements. It was assumed that despite the different expression profile shapes, even if the displacement for one gene is towards the left of the

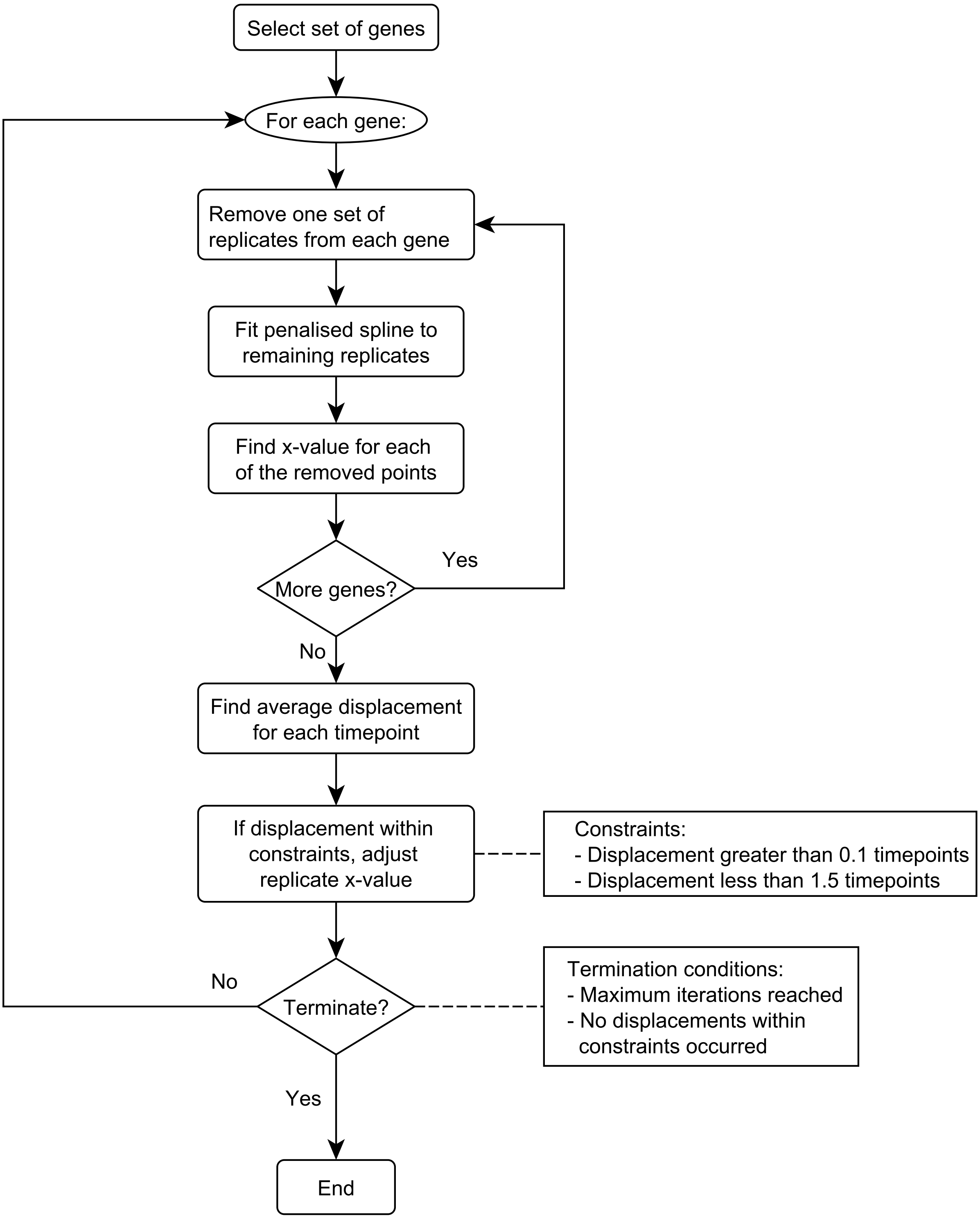

Select set of genes
For each gene:
Remove one set of replicates from each gene
Fit penalised spline to remaining replicates
Find x-value for each of the removed points
More genes?
Yes
No
Find average displacement for each timepoint
If displacement within constraints, adjust replicate x-value
Constraints:
- Displacement greater than 0.1 timepoints
- Displacement less than 1.5 timepoints
Terminate?
No
Termination conditions:
- Maximum iterations reached
- No displacements within constraints occurred
Yes
End

Figure 7.2: Schematic of time shifting process.

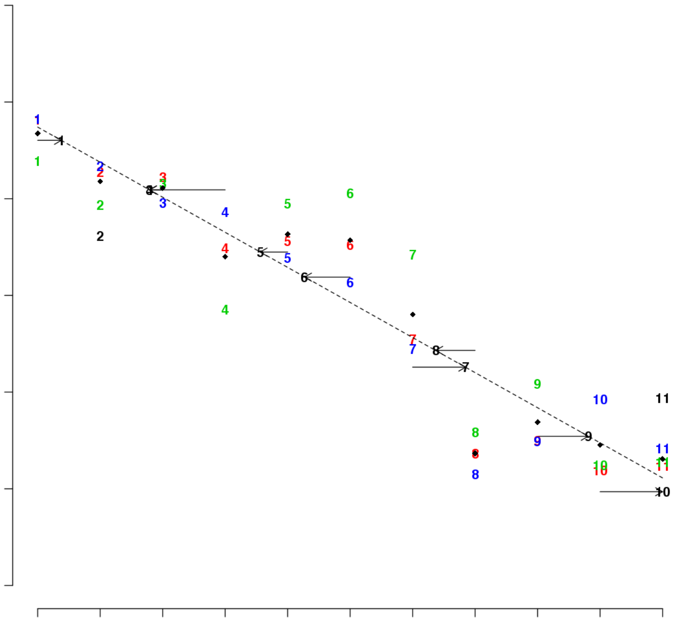


Figure 7.3: Figure describing the calculation of temporal displacements. One set of replicate points are removed from the data (black numbers), and a spline fitted to the means of the remaining data (dotted line fitted to black dots). The removed points are shifted to where they would occur on the fitted spline (indicated by the arrows), and the distance moved is the displacement. If the displacement is greater than 1.5 time points, the point is not moved (time points 2 and 11).

graph, and another towards the right, the net effect would be negligible. The time points of the removed replicate set were adjusted according to this average displacement. This process was then performed with the adjusted replicates, and the next set of replicates are removed. This procedure is performed until no displacements meet the criteria (i.e. no replicates are moved), or a maximum number of iterations is performed. This is currently set at 100 iterations. This process is reminiscent of the approach taken with the spike detection (Section 3.5), except the x-value is being predicted, instead of the location of the y-value.

Once the overall time shift was calculated, the regression models (Chapter 3) were used to fit the shifted data, and determine if the shifting process improved the model fit. This was primarily done by using the residual mean square (RMS) from the model fit, which provides an indication of the amount of variation that is unexplained by the model (Section 3.4). Thus the lower the value, the better the fit.

7.3. Results

7.3.1. Effect of temporal marker gene set size

In determining the amount to shift the time points, a subset of the total set of genes was used, as it was assumed that it would not be necessary to use the full dataset for the time shifting procedure. With the full set of genes, there may be too many conflicting displacements, which may constrain the shifting process. It was thought that a much smaller set of genes would be sufficient in estimating the necessary time shifts. Thus sets of *temporal marker genes*, were used in the time shifting procedure to provide an estimate of the overall shifting patterns that are taking place. These genes were randomly selected sets of genes, so as not to bias the analysis towards a specific molecular process. A variety of different gene set sizes were used to identify the minimum number of genes required to improve the fits.

The data from the senescence time course was used as a test of the time shifting process. The temporal marker genes were selected from genes that were previously identified as having a good fit, as determined using the techniques described in Section 3.8. That is, the model goodness-of-fit and standard error of estimation of the parameters were used to find genes that possessed a good model fit. Using the senescence dataset, this produced a list of 8216 genes that could be used as a estimation of time.

In order to determine the ideal number of genes to use for the time shifting process, a variety of different set sizes (between 5 and 4000 genes) were used, and 10 random gene sets were obtained for each sample size. In the interests of time, only 10 sets of genes were analysed for each size. For each set of genes, the time shifting and refitting process takes approximately 3 hours, and requires a large amount of computation power, meaning only a few can be performed simultaneously. Ideally, at least 100 gene sets at each size would be analysed.

For each of the gene sets, the time shifting procedure described above was performed. Using the shifting information, the time points in the data were adjusted, and the regression models were refitted to the 23 802 genes in the senescence dataset. The average residual mean square (RMS) was calculated after the refitting procedure and compared to the average RMS value from the original, unshifted data. The results for this are shown in Figure 7.4. From this it can be seen that the very small gene set sizes (less than 30 genes) resulted in the average fit being worse than the unshifted data. Larger gene set sizes produced better fits, and also required fewer iterations of the time shifting procedure. It can also be seen that the RMS value does not decrease as much when using more than approximately 100 genes. The greatest reduction in RMS was found using a set of 100 genes, and the temporal displacements calculated using these genes were used in the rest of the analyses. This set of genes is shown in Figure 7.5A, and the gene list is provided in Appendix C. To contrast, a similar sized set of genes which did not result in an overall improvement in goodness-of-fit is shown in Figure 7.5B. There are no distinct differences between the two sets of genes, where both sets show a variety of shapes. Some characteristics that are found in Figure 7.5A, but not

Figure 7.5B include a large number of expression profiles which change (become up- or down-regulated) between time points 4-8, there is a more equal distribution of increasing and decreasing shapes, and there are more genes with a larger range of response. A GO over-representation analyses of these sets of genes did not reveal anything in particular. Genes in both sets were involved in general metabolic processes, such as photosynthesis or amino acid biosynthesis, although there were more over-represented terms found in the genes in Figure 7.5A. Both sets of genes had roughly the same composition of models, which indicated that it was not necessarily a particular model that was responsible for the reduction in RMS.

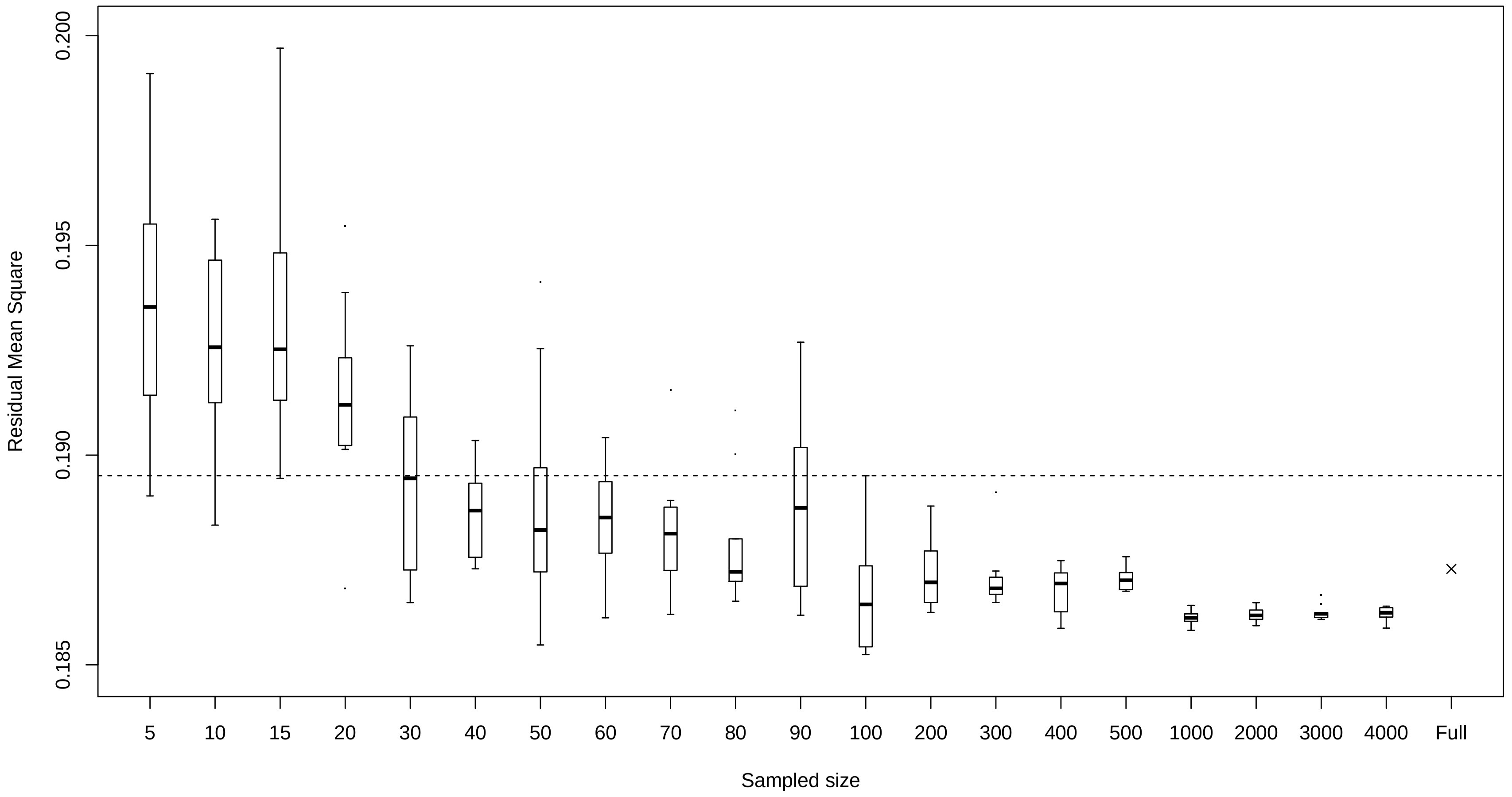


Figure 7.4: Average Residual Mean Square (RMS) values from a variety of temporal marker genes. At each set size, 10 different random sets of genes were used to perform the time shifting, and the regression models were refitted. The RMS after using the full set of genes in the shifting is also shown. The horizontal dotted line is the RMS value for the unshifted data.

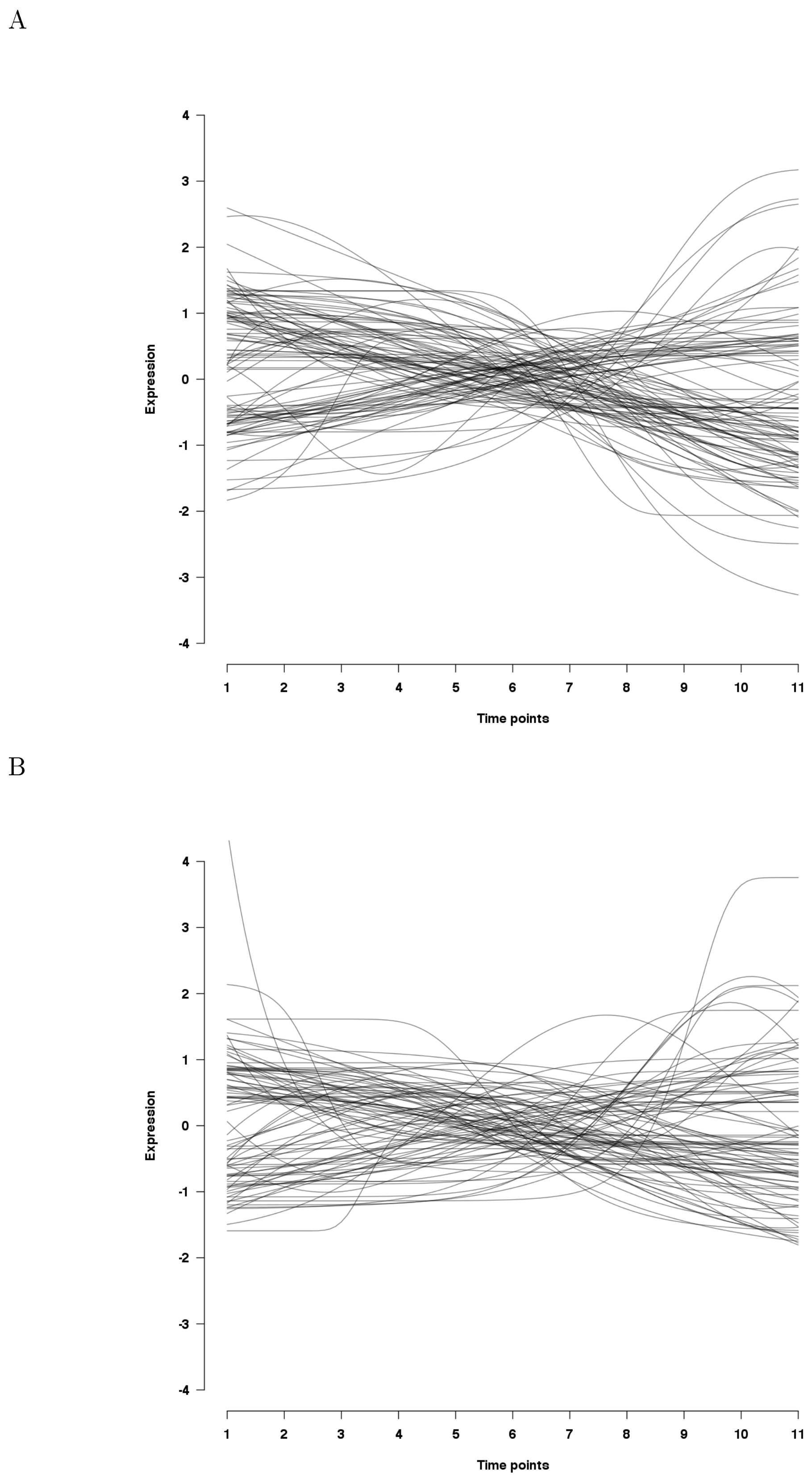


Figure 7.5: The fitted curves of the temporal marker genes. (A)The genes that resulted in the greatest reduction of RMS. (B) To contrast, a set of genes which did not result in an improvement in fit.

7.3.2. Gene displacement

Figure 7.6 shows the magnitude of the displacements across a number of different iterations in the time shifting process. As would be expected, the displacements required are large at first, but rapidly decrease in size, and smaller adjustments are required in the later iterations. While these boxplots included all calculated displacements, not all were used in the shifting process. The shifts were constrained to move no more than 1.5 time points, or 0.5 outwards for the first and last (time points 1 and 11, respectively), to prevent excessive and unrealistic movements.

The overall displacement for each time point is shown in Figure 7.7. From this figure it can be seen that the first and last time points did not have any displacements, likely due to them requiring a time shift greater than that allowed by the constraints. The displacements generally moved towards the middle time points, in particular towards time point 5. Time points adjacent to time point 5 exhibited smaller displacements than those of time points that were further away.

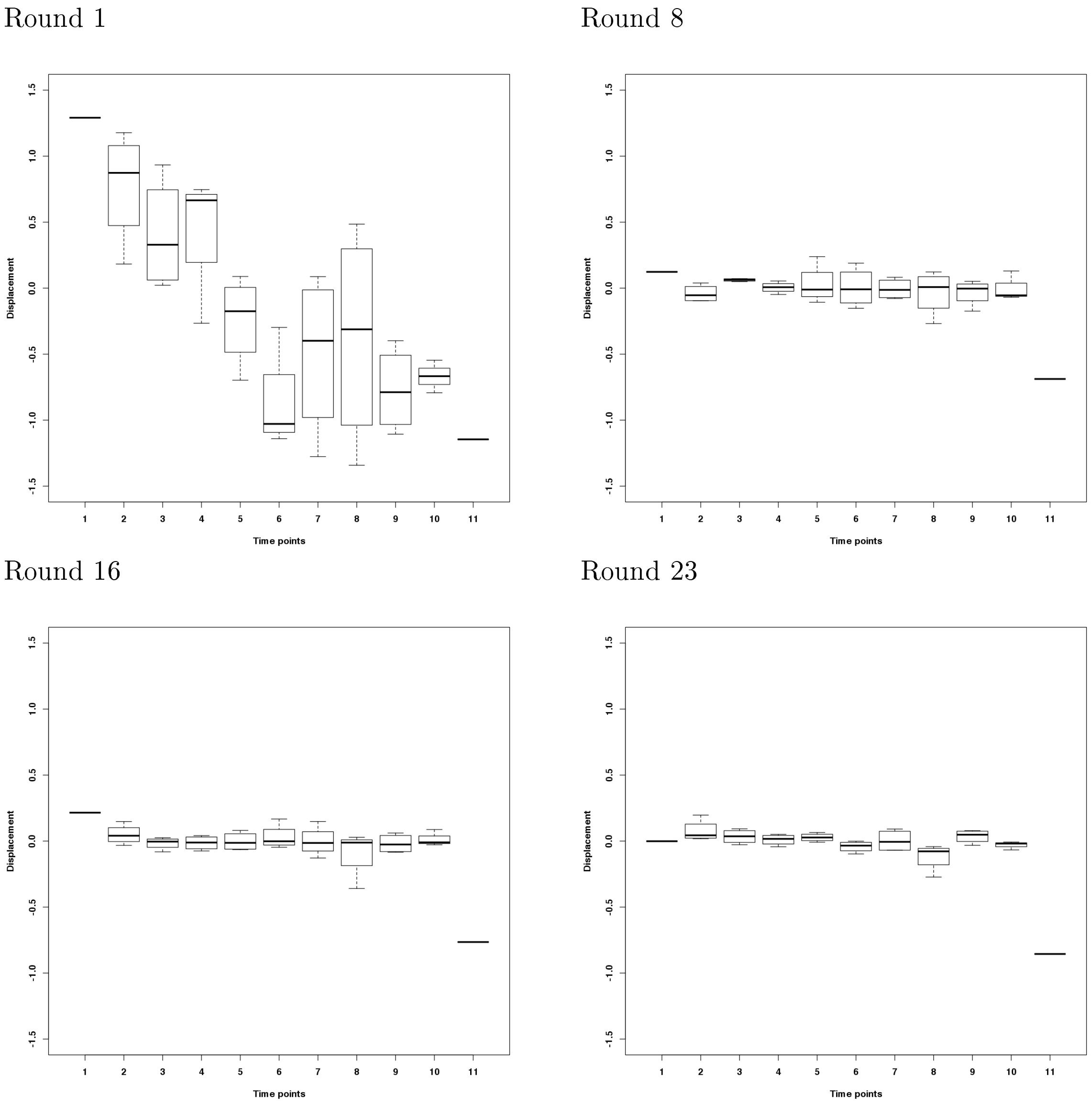
7.3.3. Refitted data

After the time shifting process was completed, the displacements were applied to the original 23 802 gene expression profiles in the senescence dataset. The shifted data was fitted using the regression models in the same process as in Section 3.8. However, in this case, the observed time points were not constrained to the reported time, but the time shifted values calculated above. Following this process there were 38 time points in the dataset, where all the replicates, except for those at the first and last time points, moved to a different time point. Using this data, the regression models were fitted, and the best model determined.

The genes of interest that were investigated in Chapter 6 (*ANAC092*, *ANAC055*, *NFYA7*, *LHCA6*, and *OXI1*) were also analysed to determine if there was an effect on the fitted models. All the genes exhibited an improvement in fit, although *ANAC092*, *OXI1* and *LHCA6* did not change much. This is unsurprising since all of these model fits were already good fits (in all cases the unshifted $R^2 > 0.9$). The two remaining genes, *ANAC055* and *NYFA7* are discussed in greater detail below (Figure 7.8).

Of these five genes, *ANAC055* showed the greatest improvement in fits between the unshifted and shifted results, where the RMS decreased from 0.90 to 0.54 (Figure 7.8A). As a result of the time shifts, the fitted model changed from the Gompertz1 model to the Gaussian model. The time points were moved towards the centre, and as a result it was possible to better detect the drop in gene expression at time point 10.

The RMS value for *NFYA7* remained approximately 0.3 in the fits to both the shifted and unshifted data. The fitted models for both the unshifted and shifted datasets are shown together in Figure 7.8B. In the unshifted data, the fitted model was Gompertz2 had a m parameter (time point of maximum growth) of 3.4, while with the shifted data, the model changed to logistic with a m parameter of 4.2. This indicates that the process is predicted to occur almost one time point later than what was indicated by the original data, and this could have implications for the predicting timing of the gene



Round 1

Round 8

Round 16

Round 23

Figure 7.6: The distribution of the temporal displacements for the time point, after a selections of iterations of replicate removals, namely rounds 1, 12, 33, 46. Positive displacements indicate that the replicate was moved to a later time point, and *vice versa* for negative displacements.

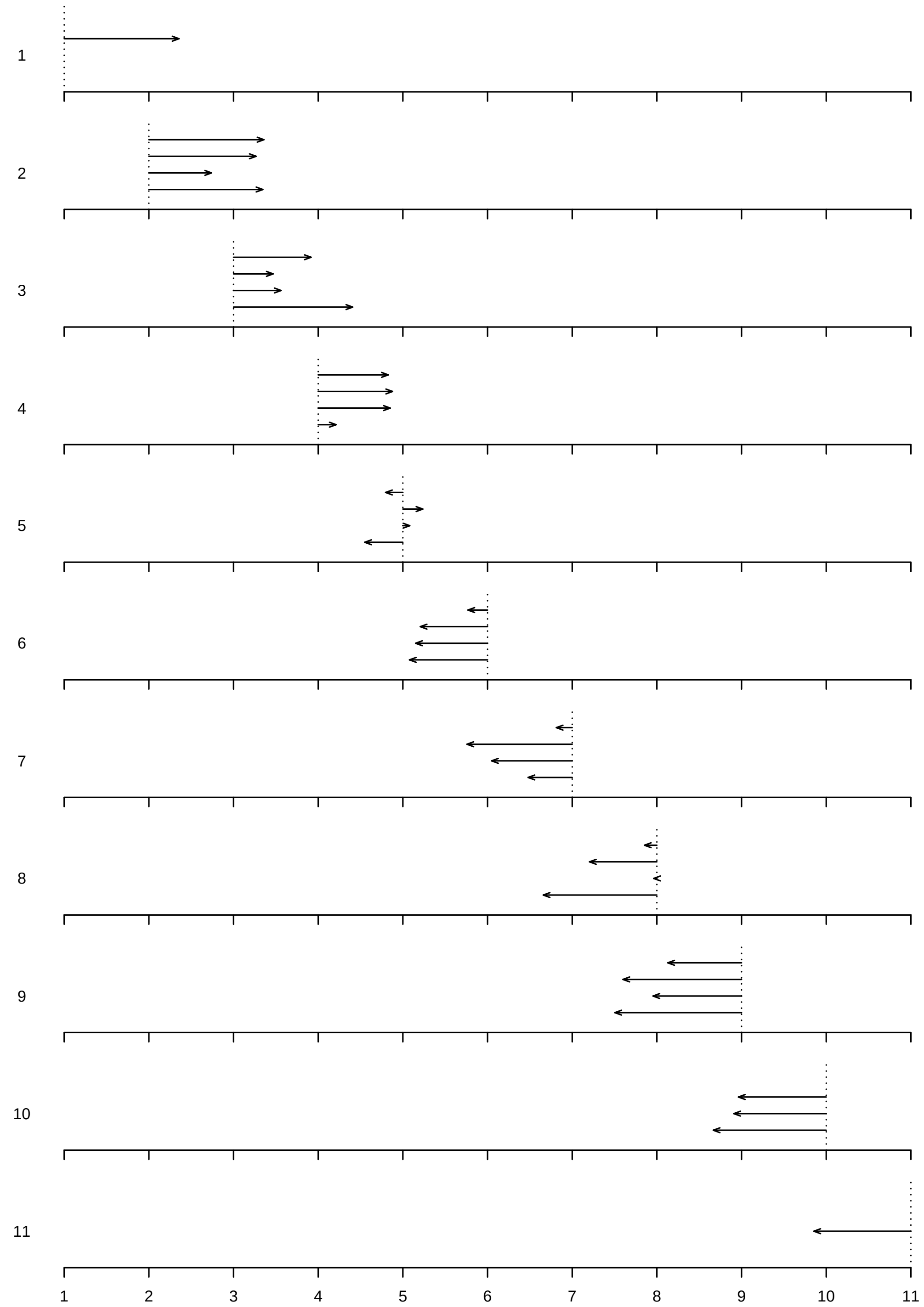


Figure 7.7: Diagram showing the overall displacements for each replicate sample. Missing arrows indicate that that replicates did not move due to the constraints.

expression response.

To investigate the overall effect of the time shifting on the model fits, the RMS values of the shifted (illustrated in blue) and unshifted (illustrated in red) are shown in Figure 7.9. From this figure, it can be seen that there is a slight improvement in the overall RMS values, with a large number of genes showing a reduction in RMS, meaning that a greater proportion of the variability was explained with the addition of the time shifting. The majority of these improvements occurred towards the centre of the graph, indicating that the fits in the tails did not change much. That is, neither the genes with already good fits (those in the left tail) nor those with poor fits (right tail) can improve much with the addition of time shifting. For the models with a large RMS value, this signifies that more significant changes are required, such as more complex models. Of the 23 802 genes that were shifted, approximately half (11 845) showed a reduction in RMS.

These fits were further analysed to determine if there was an overall improvement in the model fits. This was determined by comparing the number of "good" and "bad" fits in the shifted and unshifted datasets, where a good fit was arbitrarily defined as a model with $R_a^2 \geq 0.6$, and $R_a^2 < 0.6$ for a poor fit. If a fitted model had a poor fit in both the shifted and unshifted datasets, this implies that the model used was not the appropriate to describe the data. Thus, the time shifting was not able to provide any additional information. Similarly for a model fit that is a good fit in both datasets, the time shifting is not able to provide more information, although in this situation, it is because the fitted model was appropriate. For models that were poor in the unshifted dataset but good in the shifted data set, this indicates that the time shifting was able to improve the model fit, and *vice versa* for the models that were good in the unshifted and poor in the shifted datasets. Using these categories, it was possible to determine the change in the quality of fits when performing the time shifting, and are shown in Table 7.1. From this table, it can be seen that there were a large number of genes that had a poor fit in the unshifted data and remained poor fits in the shifted data, as well as those that had good fits in both datasets. There were over 500 genes where the time shifting improved the model fit, and for the majority, the difference in R^2 value was minor (less than 0.1). However, for those where the difference was larger, it was found that many of these genes had relatively small ranges of response, and were generally fitted by a linear function. Thus with the shifted data, the models were better able to estimate the parameters. On the other hand, there were just under 400 where the model fit was worsened with the time shifting. In these cases, again the range of response was relatively low, and application of the time shifting resulted in the expression profile "loosing" its shape. This was particularly the case for the exponential-type curves, where the r (rate of change) parameter is sensitive to the shape of the expression profiles. Nonetheless there was an overall improvement in the fits by performing the time shifting.

Some examples of these fits are shown in Figure 7.10. The original data and model fits to the data are shown in column 1, and the shifted data is shown in column 2. Row A represents a gene (*AT5G19120*) where the model fit was worse after performing the

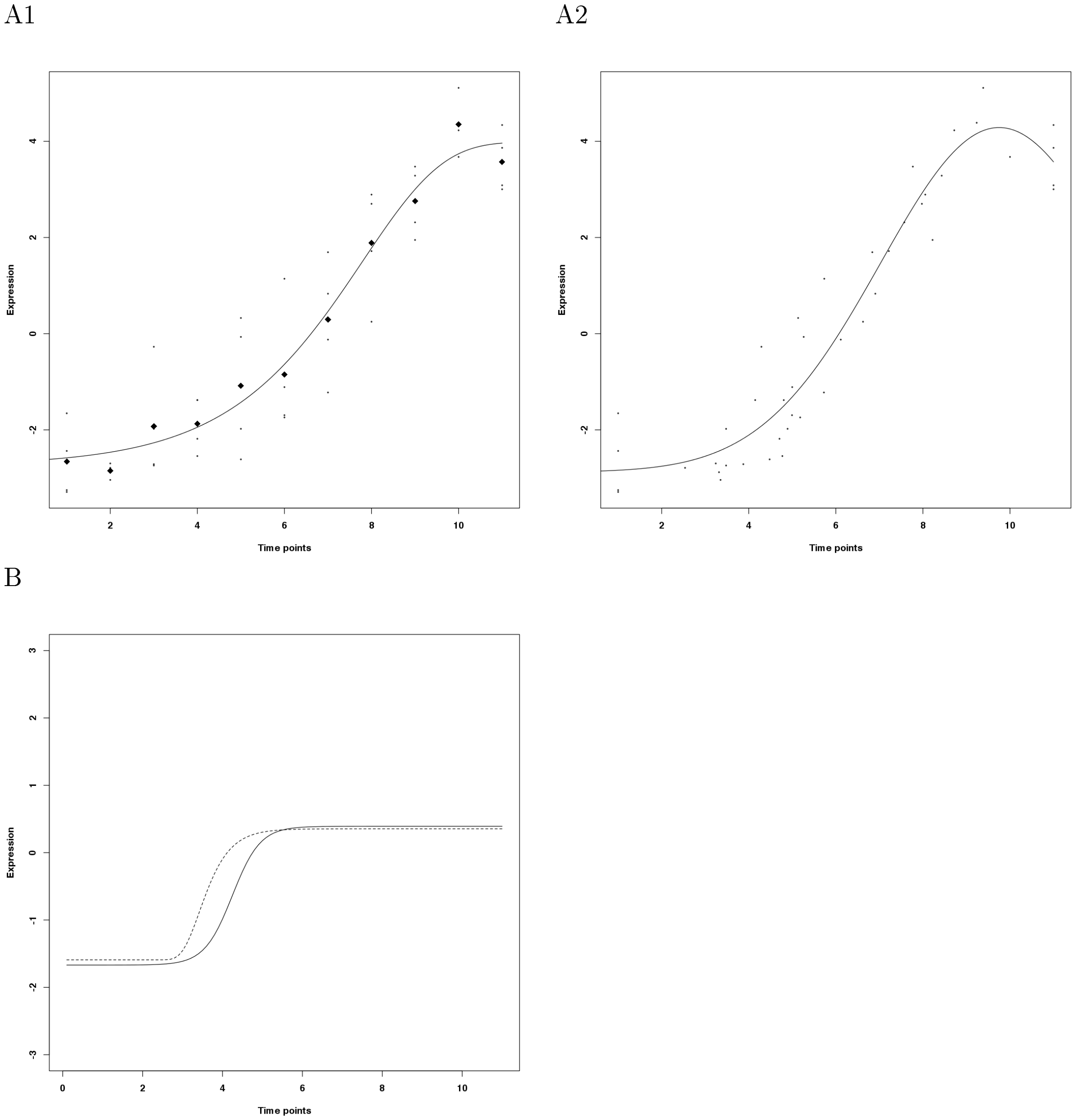


Figure 7.8: (A) *ANAC055* showing a change of model from the unshifted data (A1) to the shifted data (A2), which decreased the RMS and improved fit. The dots indicate the replicate points, and the black diamonds indicate the means of the replicates. (B) The *NFYA7* expression profile remained sigmoid, except the change in expression was shifted later to a later time point. The shifted data is represented by the solid line, and the unshifted data is indicated by the dotted line.

Table 7.1: Table comparing the number of genes with good fits in the shifted and unshifted datasets. *Good* indicated genes that had $R_a^2 \geq 0.6$, and *Poor* indicated genes had $R_a^2 < 0.6$.

		Shifted	
		Poor	*Good*
Unshifted	*Poor*	17130	523
	Good	394	5755

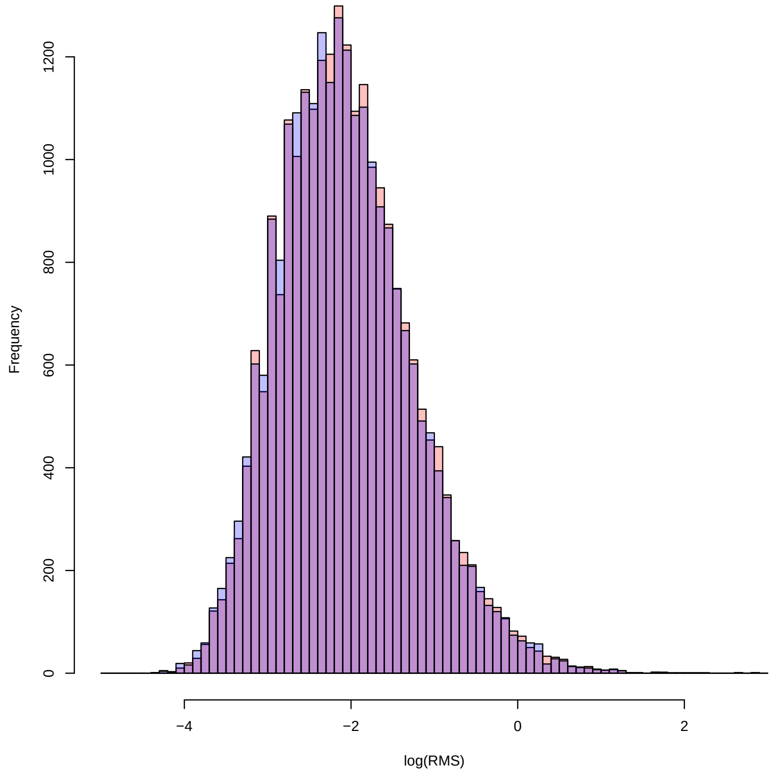


Figure 7.9: Distribution of the RMS values between the shifted (blue) and unshifted (red) data. The overlaps are shown in purple.

time shifting. The unshifted data fitted a critical-exponential model (A1), with a RMS of 0.345 and a R_a^2 value of 0.659. However, upon shifting the data, the best fit became the linear model (A2), with a RMS of 0.887 and a R_a^2 value of 0.123. Here, the original model fitted well, and the time shifting process altered the data points such that the original model was no longer able to fit the data.

In contrast, row B represents a gene (*AT5G66360*) where the time shifting improved the model fit. The unshifted data fitted a Gompertz1 model (B1), with a RMS of 0.076 and a R_a^2 value of 0.568. However, upon shifting the data, the best fit became a Gaussian model (B2), with a RMS of 0.074 and a R_a^2 value of 0.684. Thus, while the overall variability was not reduced by a large amount, there was an improvement in the fit of model to the data.

The clustering algorithms from Chapter 4 were applied to the shifted data, and the BHI scores from these clusters are shown in Table 7.2. These scores were compared to the results from the unshifted data, that was presented in Table 5.10. By applying the time shifting, it was possible to improve the annotation homogeneity of the exponential, Gompertz1 increasing, and Gompertz2 increasing shapes. A similar pattern was shown when using the meta-clustering approaches. This indicates that the shifted data was able to improve some of the cluster analyses, possibly where some parameter values changed or the fitted model in the unshifted dataset changed to another model when using the shifted data.

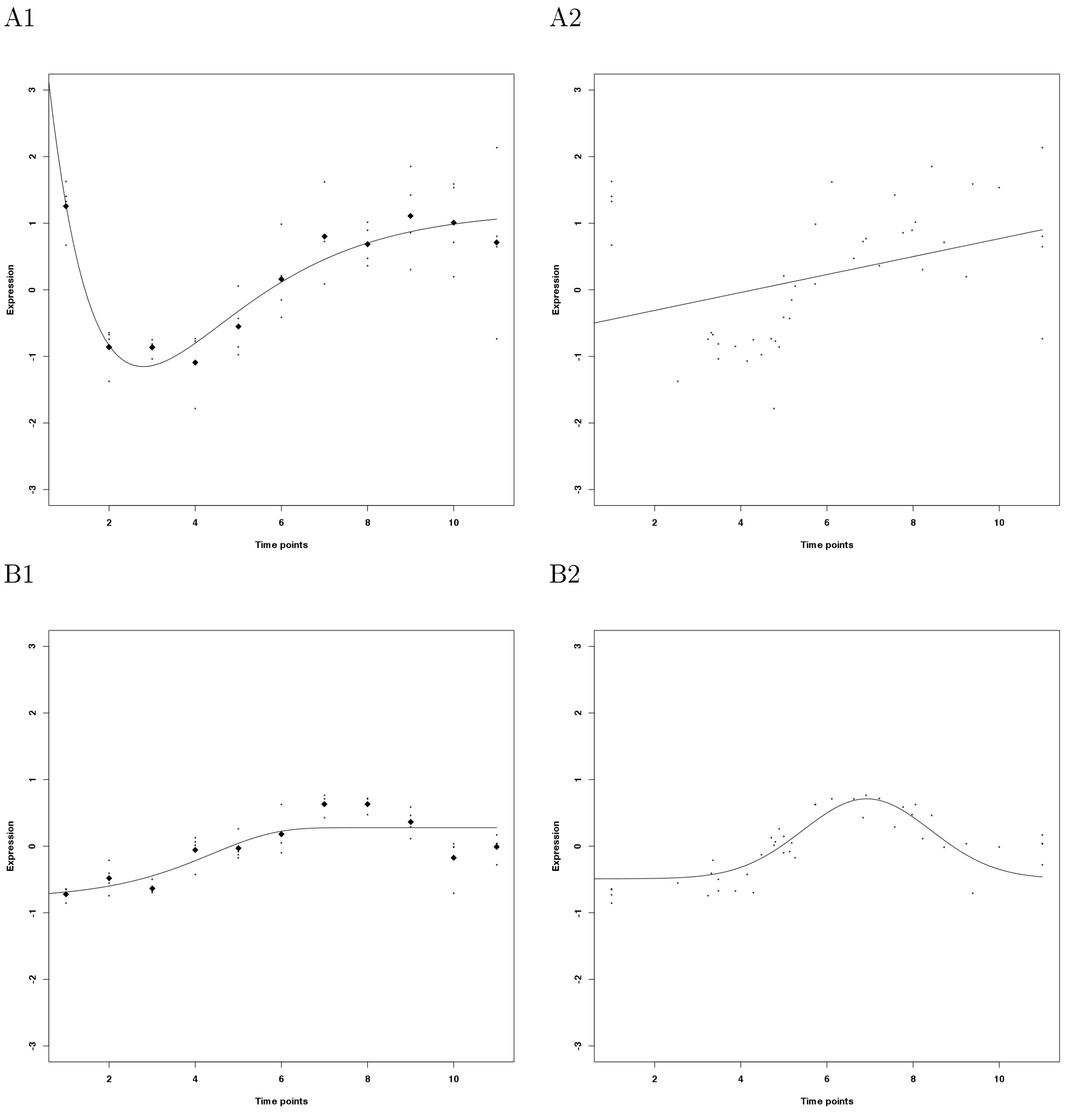


Figure 7.10: A comparison of the fits obtained with the unshifted (column 1) and shifted (column 2) data. Row (A) represents a gene where the model fit was worse after performing the time shifting, and row (B) represents a model where the model fit was improved. Each dot represents a biological replicate, and the diamonds represent the mean of replicates at the same time point.

Table 7.2: BHI scores for the shifted and unshifted data for a variety of shapes, using the simultaneous parameter cluster analysis. For each model, the shapes were separated into increasing (*Inc*) and decreasing (*D*ec) forms, except for the exponential where all the fits were used.

Model	Parameters	Shifted		Unshifted	
		Inc	*Dec*	*Inc*	*Dec*
Exponential	*r,b*	0.267		0.252	
Gaussian	*m,s*	0.244	0.273	0.254	0.288
Gompertz1	*5per,grad*	0.237	0.281	0.234	0.325
Gompertz2	*5per,grad*	0.255	0.273	0.252	0.301
Logistic	*5per,grad*	0.251	0.286	0.299	0.312

7.4. Discussion

Many analysis algorithms assume that the data is obtained from a longitudinal time series, where each replicate sample was obtained from the same organism, such as in human or animal studies. In contrast, most plant studies are from a cross-sectional time series, meaning the replicate samples were taken from different plants at each time point. This is due to the destructive nature of the sampling, i.e. destroying the leaf in order to extract the genetic material for analysis. Even though these samples were taken at the same point in time, there can still be a relatively large amount of variation due to different developmental rates, genetic variations and environmental conditions (Bar-Joseph, 2004; Liu *et al.*, 2010).

In this chapter, an iterative, cross-validation analysis was developed to investigate the variability in the replicate samples due to biological age. This process aimed to refine the underlying signal, by reducing the variability and any random fluctuations that may be present due to the slight developmental differences in the replicate samples. At each iteration, a set of replicates was removed, and a penalised spline was fitted to the remaining data. The removed replicates were replaced, and the spline was used to identify an estimate of the time point where the replicate value should occur. This process was performed for each of the replicates in a set of genes, and an average displacement across these genes was used to determine the required time shift.

Since the temporal marker genes were randomly selected, the expression profiles in the selected set exhibited a variety of different shapes. When used in the time shifting analysis, this would result in displacements in different directions, with some displacements moving towards earlier time points, and some towards later. Thus using the average provides an overall displacement size and direction. However, this is possibly not the best statistic to use, since it is prone to outliers. Thus, if there is one displacement that is large compared to the others, the overall displacement will be biased towards that large displacement. A better statistic may be to use a weighted mean, where displacements that are closer to the sampled time point are preferential. Alternatives include using a median or truncated mean.

When assessing the model fits, the R_a^2 value was used, indicates how much of the variance is explained by the model. Conversely, the RMS describes the amount of unexplained variance. Thus, a good fit would have a R^2 value close to 1, and a RMS value close to 0. In previous chapters (Chapters 3-5), the R^2_{LoF} was also used. However, it could not be used for the shifted data, as this metric requires replicate data points in order to calculate the pure error (i.e. the between-replicate variation). The R^2 value is possibly not the most appropriate statistic to use for comparing the models, as the shifted and unshifted datasets have different x-values, and are thus not exactly the same. Despite this, it still gives a rough estimate of the goodness-of-fit of the models.

To perform the time shifting, a small set of genes were used to determine the biological time, named the temporal marker genes. Fewer genes were used since it was thought larger sets would overly constrain the shifting process. When the full set of 8216 genes were used, it was found that the time shifting improved the RMS, but not as much as the smaller set sizes. To determine the number of genes that were required, a series of simulations were performed, and the average RMS was calculated after applying the time shifts. It was found that approximately 100 genes sufficiently reduces the overall RMS value for all the genes. The addition of more genes was unnecessary as there was no significant reduction in RMS with an increase in the number of genes used to determine biological time. Furthermore, how genes are selected can also influence the shifting process. Here, the genes were randomly selected, but an improved set could be found using some systematic process, for example identifying genes that exhibit a large range of response, or behaving in a particular manner. There was no obvious difference between the gene sets that were used for the time shifting. In all the cases, a variety of different models were used to determine time. The few possible differences were that sets which improved the overall fit had more genes that changed (became up- or down-regulated) between time points 4 to 6, had a large range of response. This could also explain the migration of the displacements towards time point 5. The sets of genes which resulted in poorer fits overall had a more even distribution of when the genes changed, and this possibly added more noise when trying to determine the size and direction of the displacements.

After the time shifting process, it was found that the majority of the model fits could be improved solely by the use of the time shifting, either because the model fit was already a good fit, or the model fit was poor due to the model being inappropriate for the data. However, it was possible to explain more of the variability for several hundred of the model fits. In these cases, the original model fit could not sufficiently describe the expression profiles, but using the time shifting, it allowed the fitting algorithms to better estimate the parameters. For some of the model fits, particularly the exponential-type functions (exponential, critical exponential, linear+exponential), the application of the time shifting resulting in the model no longer able to fit to the data. This is due to the r parameter being sensitive to the data. Nonetheless, it was possible to reduce the RMS for a number of genes, thus explaining more of the variability. This indicates that the differences in biological age between replicate samples results in a small but describable amount of the total variability.

The cluster analysis from the previous chapters was applied to the shifted data, and it was found that the BHI score improved for some of the shapes. It is possible that some the fits that were in the clusters using the unshifted data were changed to different shapes, and this resulting in clusters with different members. Alternatively, the value of some of the parameters could have changed and this resulted in moving to a different cluster.

The advantage of using this approach was that is possible to analyse a large (genome-wide) scale set of genes, in a simple manner, and using a few genes as a representation of time in the different samples. Ultimately, this analysis provided an estimation of the true biological age of the replicate samples, which could provide a higher resolution time series data for use in further analyses, such as gene network modelling.

8. General discussion

With high-throughput technologies becoming ubiquitous, it is possible to perform large, genome-scale analyses to investigate a variety of research areas, including diagnosis of diseases, transcription factor binding interactions, and effects on the transcriptome, proteome or metabolome after a treatment. While this large amount of data is vital to uncovering new discoveries, the vast amounts of data that is generated can often obscure the underlying signal. As a result, many analysis pipelines have been developed to aid in the examination of the underlying processes. However, the danger of these pipelines is that researchers often view them as black box analyses, where the analysis method is obfuscated. Thus, the analyses presented here emphasise the biological interpretability of the results.

In the biological sciences, the underlying processes are complex and generally not well understood, and so are modelled using a function that relates the relationship as closely as possible. Nonparametric models, such as splines, are often used to model expression profiles, as they are flexible, and make no assumptions about the underlying response. However, this is also its disadvantage, as prior information about the expression profiles does exist, namely the shape of the response. Therefore, using this information provides a better means to utilise all the available information.

Parametric models have previously been used to describe gene expression data. Eastwood *et al.* (2008) presented a methodology where the gene expression profiles were fitted with a critical-exponential model, whose parameters could be used to interpret the time of maximum gene expression. Chechik and Koller (2009) developed a flexible parametric model which was product of two sigmoid functions, and was used to capture the responses of genes to environmental perturbations. However, the models parameters were not directly connected to a mechanism. Jenkins *et al.* (2013) presented a model which used the "switch" times of gene expression, that is, the times where genes become up- or down-regulated. Using these timing points, it was possible to interpret the data in terms of transcription and degradation rates.

In this thesis, a high-throughput parametric regression analysis approach was developed to describe the gene expression profiles from a number of time course experiments. The regression models were used to describe the shape of the response, and may potentially provide a more mechanistic description of the underlying biological processes. A variety of different models were used to represent a range of different shapes, such as the exponential and sigmoid models, and each of these models reflect a different mode of action in the gene expression profile. Thus, each model provides information regarding the shape of the response. For example, the increasing sigmoid model shapes represent an expression profile similar to a growth response, where the initial lag phase indicates the start of the up-regulation of a gene, the exponential growth phase is where the gene is being rapidly transcribed, the linear phase indicates the maximal growth rate, before finally decreasing to an asymptote where the transcript production and degradation rates are equal. Since the model parameters identify the features of the responses, they are thus able to inform about when and how the expression profiles are changing. For example, in the sigmoid curves, the m parameter influences the time point of maximal growth rate. From these parameters, it is also possible to derive other features of the response, such as the *5per* parameter. For increasing sigmoid shapes, this is defined as the time at which 5% of the maximum response was attained, indicating the time that the gene is being activated. Similarly it is possible to identify other times of expression levels, for example the time of 90% of the maximum response, indicating the time where the gene is becoming repressed and the rate of gene expression is slowing down. While this analysis does provide a greater level of flexibility to the analyses, it is also possible that there may be too much flexibility, that is, there are too many different ways to analyse the data. As a result, a number of recommendations were provided in Table 4.15 to aid in the analysis depending on the biological question to be answered. Finally, most analysis methods average out replicate samples taken at the same time point and only model the relationships between the means. In contrast, the regression analysis was able to assess the variability between the replicates and use it to test the lack of model fit.

A weakness to the regression analysis is that the models need to be specified before they can be used. While the models used in this analysis covered a range of response shapes, it is not an exhaustive set. Nonetheless, it is relatively simple to extend the regression analysis to include more models, although more complex shapes would require more complex models with more parameters, and these can become difficult to interpret. Alternatively, it is also possible to modify a pre-existing model to include more information. One example of this was the use of spikes. The spikes were described in Chapter 3, and were identified in genes where the expression profile followed one of the models, except for a sharp increase or decrease of expression at a single time point. This phenomenon was thought to occur due to a rapid and sudden activation or repression of a gene. By removing the expression values at this time point, it was possible to improve the model fit. This information is valuable as it describes an underlying process that is interrupted at a single time point. Identifying expression profiles which have spikes at the same time, or are consecutive to each other could provide a better indication of the processes that are taking place. Detecting the spikes required additional considerations such as determining if the spike is a significant feature, when the spike occurs, and how many spikes are present. The spikes add an additional set of parameters that are not estimated in the other models, but by analysing the other model parameters and interpreting the spikes separately, it is possible to integrate these models. Another type of model that was not described by the current models were the expression profiles which also showed oscillatory behaviour. A possible solution would be to impose a sine function on top of the regression models.

Typically when analysing gene expression data, expression profiles are clustered together to find genes that are potentially involved in the same biological processes (reviewed in Bar-Joseph, 2004). In time series gene expression, the expression profiles are fitted across the entire time series. While this has identified genes that are co-expressed, there is no biological reason why genes that are involved in related biological processes should have exactly the same expression profile across the time course. As described above, the fitted model parameters identify features of the response, such as time of gene activation. Thus a novel approach to the clustering of genes, based on using particular features (parameters) of the gene expression profiles, was developed. This cluster analysis, called ShapeCluster, first requires the selection of one of the regression models, and then a set of parameters with which to cluster upon. In this way, the biological interpretation is at the forefront. By identifying a model to use, the response shape is specified. That is, the analysis specifically investigates the expression profiles that follow, for example, a sigmoid or exponential response pattern. Since the model parameters identify features of the response, clustering on different combinations of parameters would result in different sets of genes. For example, when trying to identify genes that are co-regulated, it might be beneficial to investigate the timings of gene expression changes, since genes that are activated at early time points, may in turn be activating genes at later time points. Alternately, investigating genes that are changing at the same rate may be indicative of TFs which are driving the activation of genes in waves. A possible extension would be to perform the cluster analysis on multiple models which have parameters that describe similar features.

The regression and cluster analyses were applied to three time course gene expression experiments, namely a yeast dataset as a validation of the clustering methodology, and two larger Arabidopsis datasets - a dataset investigating the molecular responses during the senescence process, and a dataset investigating the response to the necrotrophic pathogen, *Botrytis cinerea*. Assessment of these showed that clusters were produced with informative biological function, and by using the timing parameters, it was possible to construct a timeline of the various molecular responses, which was consistent with the published results. Investigation of the clusters showed that there were multiple transcription factors that co-clustered with the genes encoding NAC domain proteins, and these could provide new lines of research, particularly in identifying the genes that are co-regulated in stress responses. NAC proteins have been identified in a variety of plant species, and have been found to be involved in a number of stress responses (Balazadeh *et al.*, 2010; Fujita *et al.*, 2004; Hickman *et al.*, 2013; Meng *et al.*, 2009; Ooka *et al.*, 2003; Xue *et al.*, 2011). In the Botrytis dataset, a large number of genes were found that were related to light responses, and this could indicate a relation between Botrytis infection and the underlying circadian rhythms. Some research has shown that there is a relationship between the circadian clock and the plant immune response (Roden and Ingle, 2009; Wang *et al.*, 2011).

The quality of clusters was evaluated through the use of GO annotation terms and the Biological Homogeneity Index (BHI), which determines the similarity of annotations of the genes within a clusters. While the BHI score was useful in determining which cluster method produced better annotation terms, it is not ideal. It does not take cluster size into account, nor the number of common annotations between the genes in the clusters. Thus, even if a pair of genes had a generic “biological process” term in common, this would score as well as a pair of genes that had a large proportion of descriptive terms in common. In addition the score itself can be difficult to interpret, particularly when comparing scores between two different cluster analyses. With metrics such as the R^2, it is simpler to interpret the difference between two R^2 values. For example, for a given dataset, it is clear that a model with a R^2 value of 0.9 explains 10% more of the variance than a model with a R^2 of 0.8. It is not as straightforward with the BHI. Therefore, further work needs to be done to identify an easy to interpret method to assess the biological quality of a cluster analysis.

Using the fitted models for the single stresses, the effect of the combined response was predicted, and experiments were performed to validate the predictions. Through the use of properly designed experiments, it was possible to fit a factorial model, and by identifying the significant model effects, a predicted model could be identified. Had the experiment not been properly designed it would not have been as simple (if at all possible) to extract the significant effects. When fitting models, researchers generally plot the means of the data. However, there can be a large amount of noise in the data, obscuring the underlying signal. By calculating and plotting the predicted model, the noise is removed and this gives a clearer indication of the underlying mechanisms in the system. From the predicted and regression models, it was possible to identify that the multiple stresses are generally additive, where the effect of a later stress is added to the current expression level. However, this did depend on the gene, and is possibly influenced by other factors, such as upstream regulators of the gene of interest. For example, the predicted model for *ANAC055* and *ANAC092* were very similar, and the different ages of plant clearly had an additive effect on the overall gene expression response. It was shown in the cluster analysis that these two gene co-cluster together, and so genes that have the same significant models effects could also be functionally related. This would act as an extension of the cluster analysis, except instead of grouping the expression profiles by the fitted model parameters, the profiles would be grouped by the significant model effects.

qPCR is often used as a validation of results from microarrays, as it is thought that qPCR provides a more accurate determination of gene expression (Morey *et al.*, 2006). However, qPCR experiments have their own biases, such as amplification biases during the PCR process with different primer sets. As a result, the data can be equally noisy. Nonetheless, both technologies are equally valuable. It may be interesting to perform the predicted model analysis using microarrays, and thus obtain the combined stress expression profiles for a greater number of genes.

In this analysis, only two stresses were combined, although it may be possible to extend the combined stress response to predict more complicated combinations. In many ways, the combined stress analyses are similar to algorithms which integrate multiple datasets from different experiments. For example Kirk *et al.* (2012) described a Bayesian correlated clustering approach to combine the datasets from different sources, including gene expression, chromatin immunoprecipitation–chip and protein–protein interaction data. Similar approaches could thus be used to predict the effect of combined stresses from single stress data.

In plant-based, cross-sectional time course experiments, samples at each time point are collected destructively, and this introduces an additional source of variability into the observed response - that of variation in the true biological age of each plant at the time of sampling. All current analyses assume that all between-observation variation is associated with variability in the level of gene expression, but measurement error associated with the true age of the plant at sampling could contribute massively to this variability. Thus a cross-validation multivariate regression approach was developed to estimate the biological time and reduce this variability, thereby providing better estimates of the shape of the gene expression profiles, and as a consequence provide higher resolution time-series data (more time-points, fewer biological replicates) for use in identifying gene networks. In the temporal marker genes that were used in the time shifting analyses, most of the expression profiles were becoming up- or down-regulated between time points 4 and 6, and it is likely this is where many of the gene changes occur. Thus it is possible that the replicate points were moving towards the time points where there is the greatest amount of change, and this is what allowed for the improvement of the model fits. It would thus be interesting to perform the analysis on other datasets where the expression profiles are changing at different time points.

There has not been much research performed into estimating the biological time in the models. The most similar method to the time shifting is the Gaussian process regression approach by Liu *et al.* (2010). This approach uses a probabilistic model to account for the uncertainty regarding the biological age of each replicate, and both are similar in that they both shift replicate points along the x-axis. Regression models which allow both the response and explanatory variables to be corrected, such total least squares (TLS) and error in variable (EIV) models, would be an interesting extension of the regression analysis (discussed in Markovsky and Van Huffel, 2007), particularly since these approaches have not currently been applied to gene expression data.

Current methods to infer gene regulatory networks generally use a correlation type approach where the expression profiles are matched between subsequent time points. Examples include a penalised spline based, semi-parametric Bayesian model, which infers the time-invariant network structure from longitudinal data (Morrissey *et al.*, 2010, 2011), and a hierarchical, non-parametric Bayesian approach which leverages data from multiple time series (Penfold and Wild, 2011; Penfold *et al.*, 2012). By using the fitted parameters from the regression, a potentially mechanistic approach could be used to better determine which genes are related by using specific aspects of their expression profiles, such as the rate of change, or the timings of the responses. In this way, the network model that relates the mechanisms behind the expression profile between genes may provide a better indication of which genes are regulating each other.

The development of these new tools has provided a detailed means of investigating plant responses to environmental stress. While the analyses were primarily performed using data from Arabidopsis, it may be possible to apply these findings to other closely related brassica crop species such oilseed rape, cauliflower, and broccoli. Thus these tools have made it possible to perform a detailed assessment of plant stress responses, which will assist in future breeding of brassica (and other crop) species with improved yields in a changing climate.

Although only the long day senescence and Botrytis infection datasets were used, there are several other datasets from the PRESTA project that could be exploited to investigate responses to stress, such as the Pseudomonas dataset to investigate the differences between the different types of pathogens. In addition, related datasets, such as the short and long day senescence response could be compared. In a broader context, while the analyses in this thesis were performed using data obtained from microarrays, it is easily extensible for use with data from other time series gene expression analyses, such as RNA-Seq (Oh *et al.*, 2013). Moreover, this approach is applicable to any time series data, both biological and nonbiological, such as the quantitative analysis of economic phenomena.

In this thesis, a set of novel analysis approaches were presented, where a large number of gene expression profiles could be investigated using a simple regression approach, and were modelled using a small number of parameters. These parameters were biologically interpretable, and allowed the investigation of multiple aspects of the expression profiles. The analyses provided the user with greater flexibility, while also providing information that would be more difficult to extract using other analysis methodologies. Thus by investigating the same data in a different way it was possible to obtain more and better information. Through the use of these analyses, a more biological paradigm was produced, whereby a simple and direct means of interrogating the data was provided. Using this approach potentially revealed more about the biology in a mechanistic manner, where more information regarding the underlying biology was uncovered.

A. *PUX1* gene expression profile

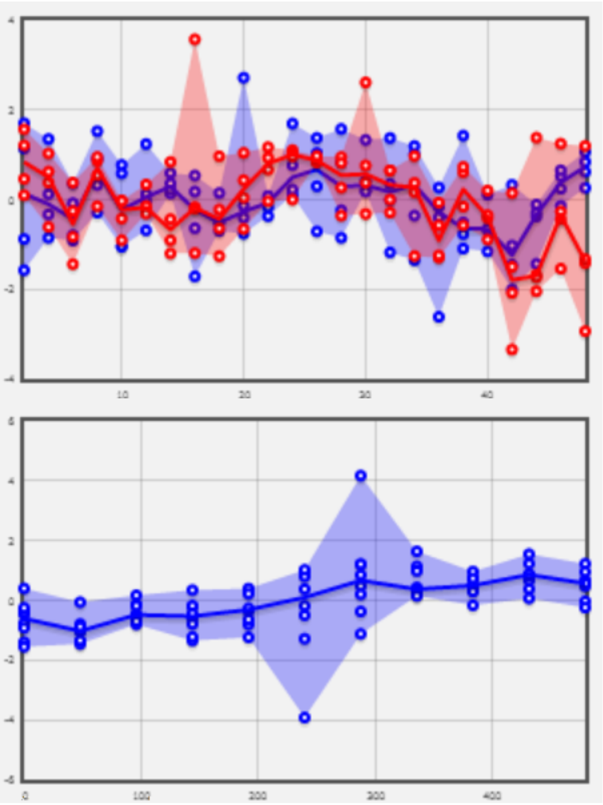


Figure A.1: Gene expression profiles for *PUX1*, from the Botrytis (top) and senescence (bottom) experiments. In the Botrytis experiment, the blue line is the mock inoculation, and the red line is the infected inoculation.

B. Primers

Table B.1: Table of primer sequences used for the qPCR. All sequences are 5' -> 3'.

Sequence	Function
TGCCGATGGTACAAAGGTTC	*ANAC092* qPCR forward primer
TTTCTTGGTCGGAGAAGCAG	*ANAC092* qPCR reverse primer
AACCGGGTTTCAGGGTTTAG	*ANAC055* qPCR forward primer
ATCCCCTCAGTTTGTTGCTG	*ANAC055* qPCR reverse primer
GCAACAAGGCGTTCCTTTAC	*NF-YA7* qPCR forward primer
CTTGATTGTCTGCGCCTTAG	*NF-YA7* qPCR reverse primer
TTAAACCGGGGTCTGTTGAC	*LHCA6* qPCR forward primer
ATCGAACCACAAACCTCCAG	*LHCA6* qPCR reverse primer
TTAGAGGCCATGACCAAAGC	*PUX1* qPCR forward primer
TCTGTGGATTTACGCTCGTG	*PUX1* qPCR reverse primer

Table B.2: Table of primer sequences in the creation of a luciferase reporter. All sequences are 5' -> 3'.

Sequence	Function
AAAAAAGCAGGCTTCATACATTGTTTTCACGAGATGGATAACATTTG	*ANAC092* promoter forward primer
CAAGAAAGCTGGGTCTTTATCCTAATAGGGTTTCTAAAAATGATC	*ANAC092* promoter reverse primer
AAAAAAGCAGGCTTCGTAACAGACAGAACCTGAGCTTC	*NFYA7* promoter forward primer
CAAGAAAGCTGGGTCTTTATCCTAATAGGGTTTCTAAGATC	*NFYA7* promoter reverse primer
GGGGACAAGTTTGTACAAAAAAGCAGGCT	Gateway sequence forward primer
GGGGACCACTTTGTACAAGAAAGCTGGGT	Gateway sequence reverse prime
GTAAAACGACGGCCAG	Colony PCR forward primer (M13 forward)
CAGGAAACAGCTATGAC	Colony PCR reverse primer (M13 reverse)
CTTCAACGTTGCGGTTCTG	pBGWL7 sequencing primer

C. Time shifting genes

Table C.1: List of the 100 genes that were used as temporal marker genes, and which resulted in the greatest reduction of residual mean square.

AT3G10270	AT1G51550	AT2G42260	AT1G07170
AT3G13730	AT1G27620	AT1G78310	AT5G45350
AT3G23710	AT3G61600	AT1G53250	AT3G17840
AT1G22280	AT3G06420	AT2G45550	AT3G60530
AT5G55160	AT5G39720	AT1G03687	AT4G32800
AT5G07030	AT3G09850	AT2G19500	AT2G07676
AT5G17210	AT2G35980	AT3G12080	AT1G30210
AT5G49740	AT1G76100	AT1G80560	AT1G23870
AT2G34430	AT1G20850	AT5G01450	AT2G43030
AT1G23060	AT3G20800	AT4G30360	AT3G55400
AT1G23960	AT4G38670	AT1G49850	AT5G17920
AT1G48450	AT2G37300	AT2G23450	AT2G22190
AT1G72710	AT5G01410	AT3G58610	AT1G18170
AT5G05740	AT3G08880	AT1G26150	AT2G39470
AT3G16000	AT1G76720	AT4G03390	AT1G77370
AT1G51090	AT1G54350	AT4G20430	AT4G40030
AT1G31190	AT5G45680	AT5G40380	AT2G38400
AT5G08330	AT2G38695	AT3G61310	AT5G52100
AT3G28200	AT1G73290	AT1G26650	AT4G32920
AT3G07460	AT1G20950	AT2G36990	AT4G36650
AT5G45020	AT2G18950	AT4G30020	AT4G25260
AT5G23060	AT4G38160	AT2G25830	AT3G01670
AT4G01050	AT4G38040	AT2G17630	AT5G05750
AT2G32520	AT3G57550	AT5G01630	AT2G39360
AT1G70100	AT5G16000	AT4G15560	AT2G40435

D. Additional results

Additional results are provided on the CD, and include:

Cluster results

Yeast

- Gaussian - *m,s*

Senescence and Botrytis

- Exponential - *r,b*
- Gaussian - *m,s*
- Gompertz1 - *5per, grad*
- Gompertz2 - *5per, grad*
- Logistic - *5per, grad*
- Linear - *m*

Figures of the clusters, the gene lists, and over-represented GO terms for the simultaneous parameter clustering, cross meta-clustering, and sequential meta-clustering results are provided.

In addition, some of the code that was written and used in the analyses is given.

qPCR results

Spreadsheet of the $-\Delta Ct$ values for all the primer pairs are provided.

Bibliography

Aach, J. and Church, G. M. (2001) Aligning gene expression time series with time warping algorithms *Bioinformatics* **17**, 6, 495–508.

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

Abraham, C., Cornillon, P. A., Matzner-Løber, E. and Molinari, N. (2003) Unsupervised Curve Clustering using B-Splines *Scand J Stat* **30**, 581–595.

Abramoff, M. D., Magelhaes, P. J. and Ram, S. J. (2004) Image Processing with ImageJ *Biophotonics International* **11**, 7, 36–42.

Anderson, J. P., Badruzsaufari, E., Schenk, P. M., Manners, J. M., Desmond, O. J., Ehlert, C., Maclean, D. J., Ebert, P. R. and Kazan, K. (2004) Antagonistic interaction between abscisic acid and jasmonate-ethylene signaling pathways modulates defense gene expression and disease resistance in Arabidopsis. *Plant Cell* **16**, 12, 3460–3479 ISSN 1040-4651 (Print).

Andre, C. M., Schafleitner, R., Guignard, C. d., Oufir, M., Aliaga, C. A. A., Nomberto, G., Hoffmann, L., Hausman, J.-F. o., Evers, D. l. and Larondelle, Y. (2008) Modification of the health-promoting value of potato tubers field grown under drought stress: emphasis on dietary antioxidant and glycoalkaloid contents in five native Andean cultivars (Solanum tuberosum L.) *J Agr Food Chem* **57**, 2, 599–609.

Androulakis, I. P., Yang, E. and Almon, R. R. (2007) Analysis of time-series gene expression data: methods, challenges, and opportunities *Annu Rev Biomed Eng* **9**, 205–228.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium *Nat Genet* **25**, 1, 25–29.

Asselbergh, B., De Vleesschauwer, D. and Höfte, M. (2008) Global switches and fine-tuning-ABA modulates plant pathogen defense *Mol Plant Microbe Int* **21**, 6, 709–719.

Atkinson, N. J. and Urwin, P. E. (2012) The interaction of plant biotic and abiotic stresses: from genes to the field *J Exp Bot* **63**, 10, 3523–3543.

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

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

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

Bar-Joseph, Z. (2004) Analyzing time series gene expression data *Bioinformatics* **20**, 16, 2493–2503.

Bari, R. and Jones, J. D. (2009) Role of plant hormones in plant defence responses *Plant Mol Biol* **69**, 4, 473–488.

Bates, D. and Watts, D. (1988) *Nonlinear regression analysis and its applications* Wiley series in probability and mathematical statistics: Applied probability and statistics Wiley.

Baxter, L., Jironkin, A., Hickman, R., Moore, J., Barrington, C., Krusche, P., Dyer, N. P., Buchanan-Wollaston, V., Tiskin, A., Beynon, J., Denby, K. and Ott, S. (2012) Conserved noncoding sequences highlight shared components of regulatory networks in dicotyledonous plants. *Plant Cell* **24**, 10, 3949–3965.

Bechtold, U., Albihlal, W. S., Lawson, T., Fryer, M. J., Sparrow, P. A., Richard, F., Persad, R., Bowden, L., Hickman, R., Martin, C. *et al.* (2013) Arabidopsis HEAT SHOCK TRANSCRIPTION FACTORA1b overexpression enhances water productivity, resistance to drought, and infection *J Exp Bot* **64**, 11, 3467–3481.

Berardini, T. Z., Mundodi, S., Reiser, L., Huala, E., Garcia-Hernandez, M., Zhang, P., Mueller, L. A., Yoon, J., Doyle, A., Lander, G., Moseyko, N., Yoo, D., Xu, I., Zoeckler, B., Montoya, M., Miller, N., Weems, D. and Rhee, S. Y. (2004) Functional annotation of the Arabidopsis genome using controlled vocabularies *Plant Physiol* **135**, 2, 745–755.

Bijlsma, R. and Loeschcke, V. (2005) Environmental stress, adaptation and evolution: an overview *J Evolutionary Biol* **18**, 4, 744–749.

Bird, S., Loper, E. and Klein, E. (2009) *Natural Language Processing with Python* O'Reilly Media Inc.

Blaschke, C., Oliveros, J. C. and Valencia, A. (2001) Mining functional information associated with expression arrays *Funct Integr Genomics* **1**, 4, 256–268.

Blüthgen, N., Brand, K., Cajavec, B., Swat, M., Herzel, H. and Beule, D. (2005) Biological profiling of gene groups utilizing Gene Ontology. *Genome Inform* **16**, 1, 106–115.

Brasier, A., Tate, J. and Habener, J. (1988) Optimized use of the firefly luciferase assay as a reporter gene in mammalian cell lines. *Biotechniques* **7**, 10, 1116–1122.

Breeze, E., Harrison, E., McHattie, S., Hughes, L., Hickman, R., Hill, C., Kiddle, S., Kim, Y.-S., Penfold, C. A., Jenkins, D., Zhang, C., Morris, K., Jenner, C., Jackson, S., Thomas, B., Tabrett, A., Legaie, R., Moore, J. D., Wild, D. L., Ott, S., Rand, D., Beynon, J., Denby, K., Mead, A. and Buchanan-Wollaston, V. (2011) High-resolution temporal profiling of transcripts during Arabidopsis leaf senescence reveals a distinct chronology of processes and regulation *Plant Cell* **23**, 3, 873–894.

Brock, G., Pihur, V., Datta, S. and Datta, S. (2008) clValid, an R package for cluster validation *J Stat Softw* **25**, i4.

Brook, R. and Arnold, G. (1985) *Applied Regression Analysis and Experimental Designs* Statistics: a Series of Textbooks and Monographs Marcel Dekker, Incorporated.

Brown, P., Baxter, L., Hickman, R., Beynon, J., Moore, J. D. and Ott, S. (2013) MEME-LaB: motif analysis in clusters *Bioinformatics* **29**, 13, 1696–1697.

Buchanan-Wollaston, V. (1997) The molecular biology of leaf senescence *J Exp Bot* **48**, 2, 181–199.

Buchanan-Wollaston, V., Earl, S., Harrison, E., Mathas, E., Navabpour, S., Page, T. and Pink, D. (2003) The molecular analysis of leaf senescence–a genomics approach. *Plant Biotechnol J* **1**, 1, 3–22.

Buchanan-Wollaston, V., Page, T., Harrison, E., Breeze, E., Lim, P. O., Nam, H. G., Lin, J.-F., Wu, S.-H., Swidzinski, J., Ishizaki, K. *et al.* (2005) Comparative transcriptome analysis reveals significant differences in gene expression and signalling pathways between developmental and dark/starvation-induced senescence in Arabidopsis *Plant J* **42**, 4, 567–585.

Burnham, K. P. and Anderson, D. R. (2002) *Model selection and multimodel inference: a practical information-theoretic approach* Springer, New York 2nd edition.

Bustin, S. A., Benes, V., Nolan, T. and Pfaffl, M. W. (2005) Quantitative real-time RT-PCR–a perspective. *J Mol Endocrinol* **34**, 3, 597–601.

Butler, G. and Thiele, D. J. (1991) ACE2, an activator of yeast metallothionein expression which is homologous to SWI5. *Mol Cell Biol* **11**, 1, 476–485.

Caldana, C., Scheible, W.-R., Mueller-Roeber, B. and Ruzicic, S. (2007) A quantitative RT-PCR platform for high-throughput expression profiling of 2500 rice transcription factors *Plant Methods* **3**, 1, 7.

Carlson, M. (2012*a*) *org.At.tair.db: Genome wide annotation for Arabidopsis* r package version 2.8.0.

Carlson, M. (2012*b*) *org.Sc.sgd.db: Genome wide annotation for Yeast* r package version 2.8.0.

Causton, H. C., Quackenbush, J. and Brazma, A. (2003) *Microarray gene expressions data analysis: a beginner's guide* Blackwell Pub., Malden, MA.

Chaussabel, D. and Sher, A. (2002) Mining microarray expression data by literature profiling *Genome Biol* **3**, 10, 1–16.

Chechik, G. and Koller, D. (2009) Timing of gene expression responses to environmental changes. *J Comput Biol* **16**, 2, 279–290.

Chellappa, R., Kandasamy, P., Oh, C. S., Jiang, Y., Vemula, M. and Martin, C. E. (2001) The membrane proteins, Spt23p and Mga2p, play distinct roles in the activation of Saccharomyces cerevisiae OLE1 gene expression. Fatty acid-mediated regulation of Mga2p activity is independent of its proteolytic processing into a soluble transcription activator. *J Biol Chem* **276**, 47, 43548–43556.

Chen, H. and Xiong, L. (2008) Role of HY5 in abscisic acid response in seeds and seedlings *Plant Signal Behav* **3**, 11, 986–988.

Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hitz, B. C., Karra, K., Krieger, C. J., Miyasato, S. R., Nash, R. S., Park, J., Skrzypek, M. S., Simison, M., Weng, S. and Wong, E. D. (2012) Saccharomyces Genome Database: the genomics resource of budding yeast *Nucleic Acids Res* **40**, D1, D700–D705.

Chini, A., Boter, M. and Solano, R. (2009) Plant oxylipins: COI1/JAZs/MYC2 as the core jasmonic acid-signalling module. *FEBS J* **276**, 17, 4682–4692.

Chini, A., Fonseca, S., Fernández, G., Adie, B., Chico, J. M., Lorenzo, O., García-Casado, G., López-Vidriero, I., Lozano, F. M., Ponce, M. R., Micol, J. L. and Solano, R. (2007) The JAZ family of repressors is the missing link in jasmonate signalling. *Nature* **448**, 7154, 666–671.

Chudova, D., Ihler, A., Lin, K. K., Andersen, B. and Smyth, P. (2009) Bayesian detection of non-sinusoidal periodic patterns in circadian expression data *Bioinformatics* **25**, 23, 3114–3120.

Clark, N. R. and Ma'ayan, A. (2011) Introduction to Statistical Methods for Analyzing Large Data Sets: Gene-Set Enrichment Analysis *Sci Signal* **4**, 190, tr4.

Clough, S. J. and Bent, A. F. (1998) Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana *Plant J* **16**, 6, 735–743.

Cook, R. D. (1982) *Residuals and influence in regression* Chapman and Hall.

Cooke, E. J., Savage, R. S., Kirk, P. D. W., Darkins, R. and Wild, D. L. (2011) Bayesian hierarchical clustering for microarray time series data with replicates and outlier measurements *BMC Bioinformatics* **12**, 399.

Cristescu, S. M., De Martinis, D., Te Lintel Hekkert, S., Parker, D. H. and Harren, F. J. M. (2002) Ethylene production by Botrytis cinerea in vitro and in tomatoes. *Appl Environ Microbiol* **68**, 11, 5342–5350.

Cutler, S. R., Rodriguez, P. L., Finkelstein, R. R. and Abrams, S. R. (2010) Abscisic acid: emergence of a core signaling network *Annu Rev Plant Biol* **61**, 651–679.

Czechowski, T., Bari, R. P., Stitt, M., Scheible, W.-R. and Udvardi, M. K. (2004) Real-time RT-PCR profiling of over 1400 Arabidopsis transcription factors: unprecedented sensitivity reveals novel root-and shoot-specific genes *Plant J* **38**, 2, 366–379.

Dalbiès-Dulout, A. and Doré, T. (2001) Management of inflorescence and viable seed production of blackgrass (Alopecurus myosuroides) on set-aside in France *Crop Prot* **20**, 3, 221–227.

Datta, S. and Datta, S. (2006) Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes *BMC Bioinformatics* **7**, 397.

de Torres-Zabala, M., Truman, W., Bennett, M. H., Lafforgue, G., Mansfield, J. W., Egea, P. R., Bögre, L. and Grant, M. (2007) Pseudomonas syringae pv. tomato hijacks the Arabidopsis abscisic acid signalling pathway to cause disease *EMBO J* **26**, 5, 1434–1443.

Dean, R., Van Kan, J. A., Pretorius, Z. A., Hammond-Kosack, K. E., Di Pietro, A., Spanu, P. D., Rudd, J. J., Dickman, M., Kahmann, R., Ellis, J. *et al.* (2012) The Top 10 fungal pathogens in molecular plant pathology *Mol Plant Pathol* **13**, 4, 414–430.

Déjean, S., Martin, P. G. P., Baccini, A. and Besse, P. (2007) Clustering time-series gene expression data using smoothing spline derivatives *EURASIP J Bioinform Syst Biol* **2007**, 70561.

Denby, K. J., Kumar, P. and Kliebenstein, D. J. (2004) Identification of Botrytis cinerea susceptibility loci in Arabidopsis thaliana *Plant J* **38**, 3, 473–486.

D'haeseleer, P. (2005) How does gene expression clustering work? *Nat Biotechnol* **23**, 12, 1499–1501.

Doherty, C. J., Van Buskirk, H. A., Myers, S. J. and Thomashow, M. F. (2009) Roles for Arabidopsis CAMTA transcription factors in cold-regulated gene expression and freezing tolerance *Plant Cell* **21**, 3, 972–984.

Dolnicar, S. (2003) Using cluster analysis for market segmentation - typical misconceptions, established methodological weaknesses and some recommendations for improvement *Australasian Journal of Market Research* **11**, 5–12.

Dong, J., Chen, C. and Chen, Z. (2003) Expression profiles of the Arabidopsis WRKY gene superfamily during plant defense response *Plant Mol Biol* **51**, 1, 21–37.

Dong, X. (2004) NPR1, all things considered *Curr Opin Plant Biol* **7**, 5, 547–552.

Draper, N. R. and Smith, H. (1981) *Applied Regression Analysis (Wiley Series in Probability and Statistics)* John Wiley & Sons Inc.

Eastwood, D. C., Mead, A., Sergeant, M. J. and Burton, K. S. (2008) Statistical modelling of transcript profiles of differentially regulated genes *BMC Mol Biol* **9**, 66.

Eckardt, N. A. (2001) Move it on out with MATEs *Plant Cell* **13**, 7, 1477–1480.

Eilers, P. H. and Marx, B. D. (1996) Flexible smoothing with B-splines and penalties *Stat Sci* **11**, 89–121.

Eilers, P. H. and Marx, B. D. (2010) Splines, knots, and penalties *Wiley Interdisciplinary Reviews: Computational Statistics* **2**, 6, 637–653.

Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns *Proc Natl Acad Sci USA* **95**, 25, 14863–14868.

Ernst, J., Nau, G. J. and Bar-Joseph, Z. (2005) Clustering short time series gene expression data *Bioinformatics* **21 Suppl 1**, i159–68.

Estruch, F. and Carlson, M. (1990) SNF6 encodes a nuclear protein that is required for expression of many genes in Saccharomyces cerevisiae. *Mol Cell Biol* **10**, 6, 2544–2553.

Eulgem, T., Rushton, P. J., Robatzek, S. and Somssich, I. E. (2000) The WRKY superfamily of plant transcription factors *Trends Plant Sci* **5**, 5, 199–206.

Everitt, B. and Bullmore, E. (1999) Mixture model mapping of brain activation in functional magnetic resonance images *Hum Brain Mapp* **7**, 1, 1–14.

Everitt, B. and Hothorn, T. (2010) *A Handbook of Statistical Analyses Using R* Statistics (CRC Press) Chapman & Hall/CRC.

Falcon, S. and Gentleman, R. (2007) Using GOstats to test gene lists for GO term association *Bioinformatics* **23**, 2, 257–258.

FAO (2012) *FAO Statistical Yearbook* Food and Agriculture Organization of the United Nations.

Ferrari, S., Plotnikova, J. M., De Lorenzo, G. and Ausubel, F. M. (2003) 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.

Flors, V., Ton, J., Van Doorn, R., Jakab, G., García-Agustín, P. and Mauch-Mani, B. (2008) Interplay between JA, SA and ABA signalling during basal and induced resistance against Pseudomonas syringae and Alternaria brassicicola *Plant J* **54**, 1, 81–92.

Fox, J. (2002) *An R and S-Plus Companion to Applied Regression* London.

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models* SAGE Publications.

Fujita, M., Fujita, Y., Maruyama, K., Seki, M., Hiratsu, K., Ohme-Takagi, M., Tran, L.-S. P., Yamaguchi-Shinozaki, K. and Shinozaki, K. (2004) A dehydration-induced NAC protein, RD26, is involved in a novel ABA-dependent stress-signaling pathway. *Plant J* **39**, 6, 863–876 ISSN 0960-7412 (Print).

Garcia-Hernandez, M., Berardini, T. Z., Chen, G., Crist, D., Doyle, A., Huala, E., Knee, E., Lambrecht, M., Miller, N., Mueller, L. A., Mundodi, S., Reiser, L., Rhee, S. Y., Scholl, R., Tacklind, J., Weems, D. C., Wu, Y., Xu, I., Yoo, D., Yoon, J. and Zhang, P. (2002) TAIR: a resource for integrated Arabidopsis data *Funct Integr Genomics* **2**, 6, 239–253.

Geiss, G. K., Bumgarner, R. E., Birditt, B., Dahl, T., Dowidar, N., Dunaway, D. L., Fell, H. P., Ferree, S., George, R. D., Grogan, T. *et al.* (2008) Direct multiplexed measurement of gene expression with color-coded probe pairs *Nat Biotechnol* **26**, 3, 317–325.

Gepstein, S. and Thimann, K. V. (1980) Changes in the abscisic acid content of oat leaves during senescence *Proc Natl Acad Sci USA* **77**, 4, 2050–2053.

Glazebrook, J. (2005) Contrasting mechanisms of defense against biotrophic and necrotrophic pathogens. *Annu Rev Phytopathol* **43**, 205–227 ISSN 0066-4286 (Print).

Glenisson, P., Coessens, B., Van Vooren, S., Mathys, J., Moreau, Y. and De Moor, B. (2004) TXTGate: profiling gene groups with text-based information *Genome Biol* **5**, 6, R43.

Govrin, E. M. and Levine, A. (2000) The hypersensitive response facilitates plant infection by the necrotrophic pathogen Botrytis cinerea *Curr Biol* **10**, 13, 751–757.

Grbić, V. and Bleecker, A. B. (1995) Ethylene regulates the timing of leaf senescence in Arabidopsis *Plant J* **8**, 4, 595–602.

Hartley, J. L., Temple, G. F. and Brasch, M. A. (2000) DNA cloning using in vitro site-specific recombination. *Genome Res* **10**, 11, 1788–1795.

Hassani-Pak, K., Legaie, R., Canevet, C., van den Berg, H. A., Moore, J. D. and Rawlings, C. J. (2010) Enhancing data integration with text analysis to find proteins implicated in plant stress response *J*

Integr Bioinformatics **7**, 3, 121.

Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Staudt, L., Chan, W., Botstein, D. and Brown, P. (2000) 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns *Genome Biol* **1**, 2, research0003.1–research0003.21.

Heard, N., Holmes, C. and Stephens, D. (2006) A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An application of Bayesian hierarchical clustering of curves *J Am Stat Assoc* **101**, 473, 18–29.

Heath, M. C. (1994) Genetics and cytology of age-related resistance in North American cultivars of cowpea (Vigna unguiculata) to the cowpea rust fungus (Uromyces vignae) *Can J Botany* **72**, 5, 575–581.

Hickman, R., Hill, C., Penfold, C. A., Breeze, E., Bowden, L., Moore, J. D., Zhang, P., Jackson, A., Cooke, E., Bewicke-Copley, F., Mead, A., Beynon, J., Wild, D. L., Denby, K. J., Ott, S. and Buchanan-Wollaston, V. (2013) A local regulatory network around three NAC transcription factors in stress responses and senescence in Arabidopsis leaves *Plant J* **75**, 1, 26–39.

Higo, K., Ugawa, Y., Iwamoto, M. and Korenaga, T. (1999) Plant cis-acting regulatory DNA elements (PLACE) database: 1999 *Nucleic Acids Res* **27**, 1, 297–300.

Huang, D. W., Sherman, B. T. and Lempicki, R. A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists *Nucleic Acids Res* **37**, 1, 1–13.

Hunter, R., Halloin, J., Veech, J. and Carter, W. (1978) Terpenoid accumulation in hypocotyls of cotton seedlings during aging and after infection by Rhizoctonia solani *Phytopathology* **68**, 347–350.

Huynh, L. N., VanToai, T., Streeter, J. and Banowetz, G. (2005) Regulation of flooding tolerance of SAG12: ipt Arabidopsis plants by cytokinin *J Exp Bot* **56**, 415, 1397–1407.

Izawa, T., Foster, R. and Chua, N.-H. (1993) Plant bZIP protein DNA binding specificity *J Mol Biol* **230**, 4, 1131–1144.

Izawa, T., Foster, R., Nakajima, M., Shimamoto, K. and Chua, N.-H. (1994) The rice bZIP transcriptional activator RITA-1 is highly expressed during seed development. *Plant Cell* **6**, 9, 1277–1287.

Izenman, A. J. (2008) *Modern multivariate statistical techniques: regression, classification, and manifold learning* chapter Cluster Analysis Springer.

Jain, A. K. and Dubes, R. C. (1988) *Algorithms for clustering data* Prentice-Hall, Inc.

Jain, A. K., Murty, M. N. and Flynn, P. J. (1999) Data clustering: a review *ACM computing surveys (CSUR)* **31**, 3, 264–323.

Jansen, M. A., Hectors, K., O Brien, N. M., Guisez, Y. and Potters, G. (2008) Plant stress and human health: Do human consumers benefit from UV-B acclimated crops? *Plant Sci* **175**, 4, 449–458.

Jenkins, D. J., Finkenstädt, B. and Rand, D. A. (2013) A temporal switch model for estimating transcriptional activity in gene expression. *Bioinformatics* **29**, 9, 1158–1165.

Johnson, C., Boden, E. and Arias, J. (2003) Salicylic acid and NPR1 induce the recruitment of trans-activating TGA factors to a defense gene promoter in Arabidopsis *Plant Cell* **15**, 8, 1846–1858.

Johnson, R. A. and Bhattacharyya, G. K. (2001) *Statistics: principles and methods* John Wiley & Sons.

Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes *Nucleic Acids Res* **28**, 1, 27–30.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M. and Hirakawa, M. (2006) From genomics to chemical genomics: new developments in KEGG *Nucleic Acids Res* **34**, suppl 1, D354–D357.

Karimi, M., De Meyer, B. and Hilson, P. (2005) Modular cloning in plant cells *Trends Plant Sci* **10**, 3, 103–105.

Katsir, L., Chung, H. S., Koo, A. J. and Howe, G. A. (2008) Jasmonate signaling: a conserved mechanism of hormone sensing *Curr Opin Plant Biol* **11**, 4, 428–435.

Kaufman, L. and Rousseeuw, P. J. (2009) *Finding groups in data: an introduction to cluster analysis* volume 344 Wiley.

Kiddle, S. J., Windram, O. P. F., McHattie, S., Mead, A., Beynon, J., Buchanan-Wollaston, V., Denby, K. J. and Mukherjee, S. (2010) Temporal clustering by affinity propagation reveals transcriptional modules in Arabidopsis thaliana *Bioinformatics* **26**, 3, 355–362.

Kirk, P., Griffin, J. E., Savage, R. S., Ghahramani, Z. and Wild, D. L. (2012) Bayesian correlated clustering to integrate multiple datasets *Bioinformatics* **28**, 24, 3290–3297.

Klee, H. J. and Clark, D. G. (2010) *Ethylene signal transduction in fruits and flowers* Springer.

Kogenaru, S., Yan, Q., Guo, Y. and Wang, N. (2012) RNA-seq and microarray complement each other in transcriptome profiling *BMC Genomics* **13**, 1, 629.

Kulkarni, M. M. (2011) Digital multiplexed gene expression analysis using the NanoString nCounter system *Current Protocols in Molecular Biology* **25**, 25B–10.

Kus, J. V., Zaton, K., Sarkar, R. and Cameron, R. K. (2002) Age-related resistance in Arabidopsis is a developmentally regulated defense response to Pseudomonas syringae *Plant Cell* **14**, 2, 479–490.

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., Karthikeyan, A. S., Lee, C. H., Nelson, W. D., Ploetz, L., Singh, S., Wensel, A. and Huala, E. (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools *Nucleic Acids Res* **40**, Database issue, D1202–10.

Langfelder, P., Zhang, B. and Horvath, S. (2008) Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R *Bioinformatics* **24**, 5, 719–20.

Lee, J., He, K., Stolc, V., Lee, H., Figueroa, P., Gao, Y., Tongprasit, W., Zhao, H., Lee, I. and Deng, X. W. (2007) Analysis of transcription factor HY5 genomic binding sites revealed its hierarchical role in light regulation of development *Plant Cell* **19**, 3, 731–749.

Lehmann, J., Atzorn, R., Brückner, C., Reinbothe, S., Leopold, J., Wasternack, C. and Parthier, B. (1995) Accumulation of jasmonate, abscisic acid, specific transcripts and proteins in osmotically stressed barley leaf segments *Planta* **197**, 1, 156–162.

Leonelli, S., Diehl, A. D., Christie, K. R., Harris, M. A. and Lomax, J. (2011) How the Gene Ontology Evolves *BMC Bioinformatics* **12**, 1, 325.

Leong, H. S. and Kipling, D. (2009) Text-based over-representation analysis of microarray gene lists with annotation bias *Nucleic Acids Res* **37**, 11, e79.

Lewin, A. and Grieve, I. C. (2006) Grouping Gene Ontology terms to improve the assessment of gene set enrichment in microarray data *BMC Bioinformatics* **7**, 426.

Li, J., Brader, G. and Palva, E. T. (2004) The WRKY70 transcription factor: a node of convergence for jasmonate-mediated and salicylate-mediated signals in plant defense *Plant Cell* **16**, 2, 319–331.

Li, J., Zhang, J., Wang, X. and Chen, J. (2010) A membrane-tethered transcription factor ANAC089 negatively regulates floral initiation in Arabidopsis thaliana *Science China Life Sciences* **53**, 11, 1299–1306.

Lieb, J. D., Liu, X., Botstein, D. and Brown, P. O. (2001) Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat Genet* **28**, 4, 327–334.

Lim, P. O., Kim, H. J. and Nam, H. G. (2007) Leaf senescence. *Annu Rev Plant Biol* **58**, 115–136.

Lin, T.-h., Kaminski, N. and Bar-Joseph, Z. (2008) Alignment and classification of time series gene expression in clinical studies. *Bioinformatics* **24**, 13, i147–i155.

Liu, Q., Lin, K. K., Andersen, B., Smyth, P. and Ihler, A. (2010) Estimating replicate time shifts using Gaussian process regression *Bioinformatics* **26**, 6, 770–776.

Llorente, F., Muskett, P., Sánchez-Vallet, A., López, G., Ramos, B., Sánchez-Rodríguez, C., Jordá, L., Parker, J. and Molina, A. (2008) Repression of the auxin response pathway increases Arabidopsis susceptibility to necrotrophic fungi *Mol Plant* **1**, 3, 496–509.

Lobell, D. B., Ortiz-Monasterio, J. I., Asner, G. P., Naylor, R. L. and Falcon, W. P. (2005) Combining Field Surveys, Remote Sensing, and Regression Trees to Understand Yield Variations in an Irrigated Wheat Landscape *Agron J* **97**, 241–249.

Lorenzo, O., Piqueras, R., Sánchez-Serrano, J. J. and Solano, R. (2003) ETHYLENE RESPONSE FACTOR1 integrates signals from ethylene and jasmonate pathways in plant defense *Plant Cell* **15**, 1, 165–178.

Lu, C., Kulkarni, K., Souret, F. F., MuthuValliappan, R., Tej, S. S., Poethig, R. S., Henderson, I. R., Jacobsen, S. E., Wang, W., Green, P. J. *et al.* (2006) MicroRNAs and other small RNAs enriched in the Arabidopsis RNA-dependent RNA polymerase-2 mutant *Genome Res* **16**, 10, 1276–1288.

Luu-The, V., Paquet, N., Calvo, E. and Cumps, J. (2005) Improved real-time RT-PCR method for high-throughput measurements using second derivative calculation and double correction *Biotechniques* **38**, 2, 287–293.

Ma, P., Castillo-Davis, C. I., Zhong, W. and Liu, J. S. (2006) A data-driven clustering method for time course gene expression data *Nucleic Acids Res* **34**, 4, 1261–1269.

Malkov, V. A., Serikawa, K. A., Balantac, N., Watters, J., Geiss, G., Mashadi-Hossein, A. and Fare,

T. (2009) Multiplexed measurements of gene signatures in different analytes using the Nanostring nCounter Assay System *BMC Research Notes* **2**, 1, 80.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. and Gilad, Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays *Genome Res* **18**, 9, 1509–1517.

Markovsky, I. and Van Huffel, S. (2007) Overview of total least-squares methods *Signal Processing* **87**, 10, 2283–2302.

Martı nez Garcı a, J. F., Huq, E. and Quail, P. H. (2000) Direct targeting of light signals to a promoter element-bound transcription factor *Science* **288**, 5467, 859–863.

Martı nez Hernández, A. d., López-Ochoa, L., Argüello-Astorga, G. and Herrera-Estrella, L. (2002) Functional Properties and Regulatory Complexity of a Minimal RBCS Light-Responsive Unit Activated by Phytochrome, Cryptochrome, and Plastid Signals *Plant Physiol* **128**, 4, 1223–1233.

Martínez-Pastor, M. T., Marchler, G., Schüller, C., Marchler-Bauer, A., Ruis, H. and Estruch, F. (1996) The Saccharomyces cerevisiae zinc finger proteins Msn2p and Msn4p are required for transcriptional induction through the stress response element (STRE). *EMBO J* **15**, 9, 2227–2235.

Matheos, D. P., Kingsbury, T. J., Ahsan, U. S. and Cunningham, K. W. (1997) Tcn1p/Crz1p, a calcineurin-dependent transcription factor that differentially regulates gene expression in Saccharomyces cerevisiae. *Genes Dev* **11**, 24, 3445–3458.

Matys, V., Fricke, E., Geffers, R., Gössling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., Kloos, D.-U., Land, S., Lewicki-Potapov, B., Michael, H., Münch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S. and Wingender, E. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31**, 1, 374–378.

Melotto, M., Underwood, W., Koczan, J., Nomura, K. and He, S. Y. (2006) Plant stomata function in innate immunity against bacterial invasion *Cell* **126**, 5, 969–980.

Meng, C., Cai, C., Zhang, T. and Guo, W. (2009) Characterization of six novel NAC genes and their responses to abiotic stresses in Gossypium hirsutum L. *Plant Sci* **176**, 352–359.

Menkens, A. E., Schindler, U. and Cashmore, A. R. (1995) The G-box: a ubiquitous regulatory DNA element in plants bound by the GBF family of bZIP proteins *Trends Biochem Sci* **20**, 12, 506–510.

Mentzen, W. and Wurtele, E. (2008) Regulon organization of Arabidopsis *BMC Plant Biol* **8**, 1, 99.

Miao, Y., Laun, T., Zimmermann, P. and Zentgraf, U. (2004) Targets of the WRKY53 transcription factor and its role during leaf senescence in Arabidopsis *Plant Mol Biol* **55**, 6, 853–867.

Millar, A. J., Carre, I. A., Strayer, C. A., Chua, N.-H. and Kay, S. A. (1995) Circadian clock mutants in Arabidopsis identified by luciferase imaging *Science* **267**, 1161–1163.

Millar, A. J., Short, S. R., Hiratsuka, K., Chua, N.-H. and Kay, S. A. (1992) Firefly luciferase as a reporter of regulated gene expression in higher plants *Plant Molecular Biology Reporter* **10**, 4, 324–337.

Miller, M. E. (1983) Relationships between onion leaf age and susceptibility to Alternaria porri. *Plant Disease* **67**, 3, 284–286.

Mittler, R. and Blumwald, E. (2010) Genetic engineering for modern agriculture: challenges and perspectives *Annu Rev Plant Biol* **61**, 443–462.

Mohr, P. G. and Cahill, D. M. (2007) Suppression by ABA of salicylic acid and lignin accumulation and the expression of multiple genes, in Arabidopsis infected with Pseudomonas syringae pv. tomato *Functional & integrative genomics* **7**, 3, 181–191.

Mooney, H. A., Winner, W. E. and Pell, E. J., editors (1991) *Response of plants to multiple stresses* Academic Press.

Morey, J. S., Ryan, J. C. and Van Dolah, F. M. (2006) Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR *Biological procedures online* **8**, 1, 175–193.

Morgan, P. W. and Drew, M. C. (1997) Ethylene and plant responses to stress *Physiol Plant* **100**, 3, 620–630.

Morris, K., Mackerness, S. A.-H., Page, T., John, C. F., Murphy, A. M., Carr, J. P. and Buchanan-Wollaston, V. (2000) Salicylic acid has a role in regulating gene expression during leaf senescence *Plant J* **23**, 5, 677–685.

Morrissey, E. R., Juárez, M. A., Denby, K. J. and Burroughs, N. J. (2010) On reverse engineering of gene interaction networks using time course data with repeated measurements. *Bioinformatics* **26**, 18, 2305–2312.

Morrissey, E. R., Juárez, M. A., Denby, K. J. and Burroughs, N. J. (2011) Inferring the time-invariant topology of a nonlinear sparse gene regulatory network using fully Bayesian spline autoregression *Biostatistics* **12**, 4, 682–694.

Motulsky, H. and Christopoulos, A. (2004) *Fitting models to biological data using linear and nonlinear regression: a practical guide to curve fitting* Oxford University Press, Oxford.

Motulsky, H. J. and Ransnas, L. A. (1987) Fitting curves to data using nonlinear regression: a practical and nonmathematical review *FASEB J* **1**, 5, 365–374.

Myers, R. H. (1990) *Classical and modern regression with applications* volume 2 Duxbury Press Belmont, CA.

Naika, M., Shameer, K., Mathew, O. K., Gowda, R. and Sowdhamini, R. (2013) STIFDB2: an updated version of plant Stress-responsive TranscrIption Factor DataBase with additional stress signals, stress-responsive transcription factor binding sites and stress-responsive genes in Arabidopsis and rice *Plant Cell Physiol* **54**, 2, e8–e8.

Ndamukong, I., Abdallat, A. A., Thurow, C., Fode, B., Zander, M., Weigel, R. and Gatz, C. (2007) SA-inducible Arabidopsis glutaredoxin interacts with TGA factors and suppresses JA-responsive PDF1. 2 transcription *Plant J* **50**, 1, 128–139.

Needleman, S. B. and Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins *J Mol Biol* **48**, 3, 443–453.

Nelder, J. A. and Mead, R. (1965) A simplex method for function minimization *The Computer Journal* **7**, 4, 308–313.

Nolan, T., Hands, R. E. and Bustin, S. A. (2006) Quantification of mRNA using real-time RT-PCR *Nat Protocols* **1**, 3, 1559–1582.

Oerke, E.-C. (2006) Crop losses to pests *J Agr Sci* **144**, 1, 31.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes *Nucleic Acids Res* **27**, 1, 29–34.

Oh, S., Song, S., Grabowski, G., Zhao, H. and Noonan, J. P. (2013) Time Series Expression Analyses Using RNA-seq: A Statistical Approach. *Biomed Res Int* **2013**, 203681.

Oh, S. A., Park, J.-H., Lee, G. I., Paek, K. H., Park, S. K. and Nam, H. G. (1997) Identification of three genetic loci controlling leaf senescence in Arabidopsis thaliana *Plant J* **12**, 3, 527–535.

Okamuro, J. K., Caster, B., Villarroel, R., Van Montagu, M. and Jofuku, K. D. (1997) The AP2 domain of APETALA2 defines a large new family of DNA binding proteins in Arabidopsis *Proc Natl Acad Sci USA* **94**, 13, 7076–7081.

Omote, H., Hiasa, M., Matsumoto, T., Otsuka, M. and Moriyama, Y. (2006) The MATE proteins as fundamental transporters of metabolic and xenobiotic organic cations *Trends Pharmacol Sci* **27**, 11, 587–593.

Ooka, H., Satoh, K., Doi, K., Nagata, T., Otomo, Y., Murakami, K., Matsubara, K., Osato, N., Kawai, J., Carninci, P. *et al.* (2003) Comprehensive analysis of NAC family genes in Oryza sativa and Arabidopsis thaliana *DNA Res* **10**, 6, 239–247.

Orlando, D. A., Lin, C. Y., Bernard, A., Wang, J. Y., Socolar, J. E. S., Iversen, E. S., Hartemink, A. J. and Haase, S. B. (2008) Global control of cell-cycle transcription by coupled CDK and network oscillators *Nature* **453**, 7197, 944–947.

Osborne, D. J. (1959) Control of leaf senescence by auxins *Nature* **183**, 1459–1460.

Pajerowska-Mukhtar, K. M., Emerine, D. K. and Mukhtar, M. S. (2013) Tell me more: roles of NPRs in plant immunity. *Trends Plant Sci* **18**, 7, 402–411.

Pandey, S. P. and Somssich, I. E. (2009) The role of WRKY transcription factors in plant immunity *Plant Physiol* **150**, 4, 1648–1655.

Paterson, A. H., Lan, T.-h., Amasino, R., Osborn, T. C. and Quiros, C. (2001) Brassica genomics: a complement to, and early beneficiary of, the Arabidopsis sequence *Genome Biol* **2**, 3, 1339–1347.

Pauwels, L., Barbero, G. F., Geerinck, J., Tilleman, S., Grunewald, W., Pérez, A. C., Chico, J. M., Bossche, R. V., Sewell, J., Gil, E., García-Casado, G., Witters, E., Inzé, D., Long, J. A., De Jaeger, G., Solano, R. and Goossens, A. (2010) NINJA connects the co-repressor TOPLESS to jasmonate signalling. *Nature* **464**, 7289, 788–791.

Pauwels, L., Morreel, K., De Witte, E., Lammertyn, F., Van Montagu, M., Boerjan, W., Inzé, D. and Goossens, A. (2008) Mapping methyl jasmonate-mediated transcriptional reprogramming of metabolism and cell cycle progression in cultured Arabidopsis cells *Proc Natl Acad Sci USA* **105**, 4, 1380–1385.

Penfold, C. A., Buchanan-Wollaston, V., Denby, K. J. and Wild, D. L. (2012) Nonparametric Bayesian inference for perturbed and orthologous gene regulatory networks *Bioinformatics* **28**, 12, i233–i241.

Penfold, C. A. and Wild, D. L. (2011) How to infer gene networks from expression profiles, revisited *Interface Focus* **1**, 6, 857–870.

Peng, L. and Shikanai, T. (2011) Supercomplex formation with photosystem I is required for the stabilization of the chloroplast NADH dehydrogenase-like complex in Arabidopsis *Plant Physiol* **155**, 4, 1629–1639.

Penninckx, I., Eggermont, K., Terras, F., Thomma, B., De Samblanx, G. W., Buchala, A., Métraux, J.-P., Manners, J. M. and Broekaert, W. F. (1996) Pathogen-induced systemic activation of a plant defensin gene in Arabidopsis follows a salicylic acid-independent pathway. *Plant Cell* **8**, 12, 2309–2323.

Pieterse, C. M., Leon-Reyes, A., Van der Ent, S. and Van Wees, S. C. (2009) Networking by small-molecule hormones in plant immunity *Nat Chem Biol* **5**, 5, 308–316.

Prasch, C. M. and Sonnewald, U. (2013) Simultaneous application of heat, drought and virus to Arabidopsis thaliana plants reveals significant shifts in signaling networks *Plant Physiol* **162**, 1849–1866.

Pré, M., Atallah, M., Champion, A., De Vos, M., Pieterse, C. M. and Memelink, J. (2008) The AP2/ERF domain transcription factor ORA59 integrates jasmonic acid and ethylene signals in plant defense *Plant Physiol* **147**, 3, 1347–1357.

Pretorius, Z., Rijkenberg, F. and Wilcoxson, R. (1988) Effects of growth stage, leaf position and temperature on adult-plant resistance of wheat infected by Puccinia recondita f. sp. tritici *Plant pathology* **37**, 1, 36–44.

Quanbeck, S. M., Brachova, L., Campbell, A. A., Guan, X., Perera, A., He, K., Rhee, S. Y., Bais, P., Dickerson, J. A., Dixon, P., Wohlgemuth, G., Fiehn, O., Barkan, L., Lange, I., Lange, B. M., Lee, I., Cortes, D., Salazar, C., Shuman, J., Shulaev, V., Huhman, D. V., Sumner, L. W., Roth, M. R., Welti, R., Ilarslan, H., Wurtele, E. S. and Nikolau, B. J. (2012) Metabolomics as a Hypothesis-Generating Functional Genomics Tool for the Annotation of Arabidopsis thaliana Genes of "Unknown Function" *Front Plant Sci* **3**, 15.

R Development Core Team (2011) *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing Vienna, Austria.

Rao, K. M., Raghavendra, A. and Reddy, K. J. (2006) *Physiology and molecular biology of stress tolerance in plants* Springer.

Rasmussen, S., Barah, P., Suarez-Rodriguez, M. C., Bressendorff, S., Friis, P., Costantino, P., Bones, A. M., Nielsen, H. B. and Mundy, J. (2013) Transcriptome responses to combinations of stresses in Arabidopsis. *Plant Physiol* **161**, 4, 1783–1794.

Ratkowsky, D. A. (1990) *Handbook of nonlinear regression models* volume v. 107 M. Dekker, New York.

Rawat, R., Schwartz, J., Jones, M. A., Sairanen, I., Cheng, Y., Andersson, C. R., Zhao, Y., Ljung, K. and Harmer, S. L. (2009) REVEILLE1, a Myb-like transcription factor, integrates the circadian clock and auxin pathways *Proc Natl Acad Sci USA* **106**, 39, 16883–16888.

Rawlings, J. O., Pantula, S. G. and Dickey, D. A. (1998) *Applied Regression Analysis: A Research Tool (Springer Texts in Statistics)* Springer.

Ritz, C. and Spiess, A.-N. (2008) qpcR: an R package for sigmoidal model selection in quantitative real-time polymerase chain reaction analysis *Bioinformatics* **24**, 13, 1549–1551.

Ritz, C. and Streibig, J. C. (2005) Bioassay Analysis using R *J Stat Softw* **12**, 5, i5.

Ritz, C. and Streibig, J. C. (2008) *Nonlinear regression with R* Springer, New York.

Rivas-San Vicente, M. and Plasencia, J. (2011) Salicylic acid beyond defence: its role in plant growth and development *J Exp Bot* **62**, 10, 3321–3338.

Rivero, R. M., Kojima, M., Gepstein, A., Sakakibara, H., Mittler, R., Gepstein, S. and Blumwald, E. (2007) Delayed leaf senescence induces extreme drought tolerance in a flowering plant. *Proc Natl Acad Sci USA* **104**, 49, 19631–19636.

Robert-Seilaniantz, A., Grant, M. and Jones, J. D. G. (2011) Hormone crosstalk in plant disease and defense: more than just jasmonate-salicylate antagonism. *Annu Rev Phytopathol* **49**, 317–343.

Roberts, R. L. and Fink, G. R. (1994) Elements of a single MAP kinase cascade in Saccharomyces cerevisiae mediate two developmental programs in the same cell type: mating and invasive growth.

Genes Dev **8**, 24, 2974–2985.

Roden, L. C. and Ingle, R. A. (2009) Lights, rhythms, infection: the role of light and the circadian clock in determining the outcome of plant–pathogen interactions *Plant Cell* **21**, 9, 2546–2552.

Rosa, B. A., Zhang, J., Major, I. T., Qin, W. and Chen, J. (2012) Optimal timepoint sampling in high-throughput gene expression experiments *Bioinformatics* **28**, 21, 2773–2781.

Roulet, E., Busso, S., Camargo, A. A., Simpson, A. J., Mermod, N. and Bucher, P. (2002) High-throughput SELEX–SAGE method for quantitative modeling of transcription-factor binding sites *Nat Biotechnol* **20**, 8, 831–835.

Rowe, H. C., Walley, J. W., Corwin, J., Chan, E. K.-F., Dehesh, K. and Kliebenstein, D. J. (2010) Deficiencies in jasmonate-mediated plant defense reveal quantitative variation in Botrytis cinerea pathogenesis *PLoS Pathog* **6**, 4, e1000861.

Rushton, P. J., Somssich, I. E., Ringler, P. and Shen, Q. J. (2010) WRKY transcription factors *Trends Plant Sci* **15**, 5, 247–258.

Rusterucci, C., Zhao, Z., Haines, K., Mellersh, D., Neumann, M. and Cameron, R. (2005) Age-related resistance to Pseudomonas syringae pv. tomato is associated with the transition to flowering in Arabidopsis and is effective against Peronospora parasitica *Physiol Mol Plant Path* **66**, 6, 222–231.

Sankoff, D. and Kruskal, J. B. (1983) *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison* volume 1 Addison-Wesley Publication.

Sasaki-Sekimoto, Y., Taki, N., Obayashi, T., Aono, M., Matsumoto, F., Sakurai, N., Suzuki, H., Hirai, M. Y., Noji, M., Saito, K. *et al.* (2005) Coordinated activation of metabolic pathways for antioxidants and defence compounds by jasmonates and their roles in stress tolerance in Arabidopsis *Plant J* **44**, 4, 653–668.

Savage, R. S., Heller, K., Xu, Y., Ghahramani, Z., Truman, W. M., Grant, M., Denby, K. J. and Wild, D. L. (2009) R/BHC: fast Bayesian hierarchical clustering for microarray data *BMC Bioinformatics* **10**, 242.

Schenk, P. M., Kazan, K., Wilson, I., Anderson, J. P., Richmond, T., Somerville, S. C. and Manners, J. M. (2000) Coordinated plant defense responses in Arabidopsis revealed by microarray analysis *Proc Natl Acad Sci USA* **97**, 21, 11655–11660.

Schliep, A., Schönhuth, A. and Steinhoff, C. (2003) Using hidden Markov models to analyze gene expression time course data *Bioinformatics* **19 Suppl 1**, i255–63.

Schoolfield, R., Sharpe, P. and Magnuson, C. (1981) Non-linear regression of biological temperature-dependent rate models based on absolute reaction-rate theory *J Theor Biol* **88**, 4, 719–731.

Sclep, G., Allemeersch, J., Liechti, R., De Meyer, B., Beynon, J., Bhalerao, R., Moreau, Y., Nietfeld, W., Renou, J.-P., Reymond, P., Kuiper, M. T. and Hilson, P. (2007) CATMA, a comprehensive genome-scale resource for silencing and transcript profiling of Arabidopsis genes. *BMC Bioinformatics* **8**, 400.

Seber, G. A. F. and Wild, C. J. (1989) *Nonlinear regression* Wiley, New York.

Sharon, A., Elad, Y., Barakat, R. and Tudzynski, P. (2007) Phytohormones in Botrytis-plant interactions in *Botrytis: biology, pathology and Control* 163–179 Springer.

Shibata, Y., Kawakita, K. and Takemoto, D. (2010) Age-related resistance of Nicotiana benthamiana against hemibiotrophic pathogen Phytophthora infestans requires both ethylene-and salicylic acid-mediated signaling pathways *Mol Plant Microbe Int* **23**, 9, 1130–1142.

Shimono, M., Sugano, S., Nakayama, A., Jiang, C.-J., Ono, K., Toki, S. and Takatsuji, H. (2007) Rice WRKY45 plays a crucial role in benzothiadiazole-inducible blast resistance *Plant Cell* **19**, 6, 2064–2076.

Shin, J., Park, E. and Choi, G. (2007) PIF3 regulates anthocyanin biosynthesis in an HY5-dependent manner with both factors directly binding anthocyanin biosynthetic gene promoters in Arabidopsis *Plant J* **49**, 6, 981–994.

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kawai, J., Carninci, P. and Hayashizaki, Y. (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage *Proc Natl Acad Sci USA* **100**, 26, 15776–15781.

Simonich, M. T. and Innes, R. (1995) A disease resistance gene in Arabidopsis with specificity for the avrPph3 gene of Pseudomonas syringae pv. phaseolicola. *Mol Plant Microbe Int* **8**, 4, 637.

Singh, K. B., Foley, R. C. and Oñate-Sánchez, L. (2002) Transcription factors in plant defense and stress responses *Curr Opin Plant Biol* **5**, 5, 430–436.

Smith, A. A., Vollrath, A., Bradfield, C. A. and Craven, M. (2009) Clustered alignments of gene-expression time series data *Bioinformatics* **25**, 12, i119–27.

Solano, R., Stepanova, A., Chao, Q. and Ecker, J. R. (1998) Nuclear events in ethylene signaling: a transcriptional cascade mediated by ETHYLENE-INSENSITIVE3 and ETHYLENE-RESPONSE-FACTOR1 *Genes & Development* **12**, 23, 3703–3714.

Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K. B., Tignor, M. and Miller, H. L., editors (2007) *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* Cambridge University Press.

Song, Y. H., Yoo, C. M., Hong, A. P., Kim, S. H., Jeong, H. J., Shin, S. Y., Kim, H. J., Yun, D.-J., Lim, C. O., Bahk, J. D. *et al.* (2008) DNA-binding study identifies C-box and hybrid C/G-box or C/A-box motifs as high-affinity binding sites for STF1 and LONG HYPOCOTYL5 proteins *Plant Physiol* **146**, 4, 1862–1877.

Spoel, S. H., Mou, Z., Tada, Y., Spivey, N. W., Genschik, P. and Dong, X. (2009) Proteasome-mediated turnover of the transcription coactivator NPR1 plays dual roles in regulating plant immunity *Cell* **137**, 5, 860–872.

St Laurent, R. T. and Cook, R. D. (1992) Leverage and superleverage in nonlinear-regression *J Am Stat Assoc* **87**, 420, 985–990.

St Laurent, R. T. and Cook, R. D. (1993) Leverage, local influence and curvature in nonlinear-regression *Biometrika* **80**, 1, 99–106.

Stegle, O., Denby, K. J., Cooke, E. J., Wild, D. L., Ghahramani, Z. and Borgwardt, K. M. (2010) A robust Bayesian two-sample test for detecting intervals of differential gene expression in microarray time series. *J Comput Biol* **17**, 3, 355–367.

Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G. and Davis, R. W. (2005) Significance analysis of time course microarray experiments *Proc Natl Acad Sci USA* **102**, 36, 12837–12842.

Straume, M. (2004) DNA microarray time series analysis: automated statistical assessment of circadian rhythms in gene expression patterning *Methods Enzymol* **383**, 149–166.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. and Mesirov, J. P. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles *Proc Natl Acad Sci USA* **102**, 43, 15545–15550.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. and Golub, T. R. (1999) Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation *Proc Natl Acad Sci USA* **96**, 6, 2907–2912.

Tanaka, Y., Sano, T., Tamaoki, M., Nakajima, N., Kondo, N. and Hasezawa, S. (2005) Ethylene inhibits abscisic acid-induced stomatal closure in Arabidopsis. *Plant Physiol* **138**, 4, 2337–2343.

Teixeira, M. C., Monteiro, P., Jain, P., Tenreiro, S., Fernandes, A. R., Mira, N. P., Alenquer, M., Freitas, A. T., Oliveira, A. L. and Sá-Correia, I. (2006) The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. *Nucleic Acids Res* **34**, Database issue, D446–D451.

Teixeira, M. C., Monteiro, P. T., Guerreiro, J. F., Gonçalves, J. P., Mira, N. P., Dos Santos, S. C., Cabrito, T. R., Palma, M., Costa, C., Francisco, A. P., Madeira, S. C., Oliveira, A. L., Freitas, A. T. and Sá-Correia, I. (2014) The YEASTRACT database: an upgraded information system for the analysis of gene and genomic transcription regulation in Saccharomyces cerevisiae. *Nucleic Acids Res* **42**, 1, D161–D166.

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana *Nature* **408**, 6814, 796–815.

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L. A., Rhee, S. Y. and Stitt, M. (2004) mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes *Plant J* **37**, 6, 914–939.

Thomas, M., Yang, L., Carter, B. and Klaper, R. (2011) Gene set enrichment analysis of microarray data from Pimephales promelas (Rafinesque), a non-mammalian model organism *BMC Genomics* **12**, 1, 66.

Thomma, B. P., Eggermont, K., Penninckx, I. A., Mauch-Mani, B., Vogelsang, R., Cammue, B. P. and Broekaert, W. F. (1998) 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.

Tice-Baldwin, K., Fink, G. and KT, A. (1989) BAS1 has a Myb motif and activates HIS4 transcription only in combination with BAS2 *Science* **246**, 931–935.

Ton, J., Flors, V. and Mauch-Mani, B. (2009) The multifaceted role of ABA in disease resistance. *Trends Plant Sci* **14**, 6, 310–317.

Untergasser, A., Nijveen, H., Rao, X., Bisseling, T., Geurts, R. and Leunissen, J. A. M. (2007) Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res* **35**, Web Server issue, W71–W74.

Usadel, B., Poree, F., Nagel, A., Lohse, M., Czedik-Eysenberg, A. and Stitt, M. (2009) A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species, Maize *Plant, cell & environment* **32**, 9, 1211–1229.

van der Graaff, E., Schwacke, R., Schneider, A., Desimone, M., Flügge, U.-I. and Kunze, R. (2006) Transcription analysis of Arabidopsis membrane transporters and hormone pathways during developmental and induced leaf senescence *Plant Physiol* **141**, 2, 776–792.

Van Helden, J. (2003) Regulatory sequence analysis tools *Nucleic Acids Res* **31**, 13, 3593–3596.

Van Huffel, S., Cheng, C.-L., Mastronardi, N., Paige, C. and Kukush, A. (2007) Total least squares and errors-in-variables modeling *Computational Statistics and Data Analysis* **52**, 1076–1079.

van Kan, J. A. L. (2006) Licensed to kill: the lifestyle of a necrotrophic plant pathogen. *Trends Plant Sci* **11**, 5, 247–253.

van Loon, L. C., Geraats, B. P. J. and Linthorst, H. J. M. (2006) Ethylene as a modulator of disease resistance in plants. *Trends Plant Sci* **11**, 4, 184–191.

Van Rossum, G. and Drake Jr, F. L. (1995) *Python reference manual* Centrum voor Wiskunde en Informatica Amsterdam.

van Verk, M. C., Bol, J. F. and Linthorst, H. J. (2011) WRKY transcription factors involved in activation of SA biosynthesis genes *BMC Plant Biol* **11**, 1, 89.

Vert, G., Walcher, C. L., Chory, J. and Nemhauser, J. L. (2008) Integration of auxin and brassinosteroid pathways by Auxin Response Factor 2 *Proc Natl Acad Sci USA* **105**, 28, 9829–9834.

Wang, H., Liu, G., Li, C., Powell, A. L. T., Reid, M. S., Zhang, Z. and Jiang, C.-Z. (2013) Defence responses regulated by jasmonate and delayed senescence caused by ethylene receptor mutation contribute to the tolerance of petunia to Botrytis cinerea *Mol Plant Pathol* **14**, 5, 453–469.

Wang, K., Ng, S. and McLachlan, G. (2012) Clustering of time-course gene expression profiles using normal mixture models with autoregressive random effects *BMC Bioinformatics* **13**, 1, 1–14.

Wang, W., Barnaby, J. Y., Tada, Y., Li, H., Tör, M., Caldelari, D., Lee, D.-u., Fu, X.-D. and Dong, X. (2011) Timing of plant immune responses by a central circadian regulator *Nature* **470**, 7332, 110–114.

Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics *Nat Rev Genet* **10**, 1, 57–63.

Ward, E., Stössel, P. and Lazarovits, G. (1981) Similarities between age-related and race-specific resistance of soybean hypocotyls to Phytophthora megasperma var. sojae *Phytopathology* **71**, 504–508.

Watkins, P. and Venables, W. N. (2006) Non-linear regression for optimising the separation of carboxylic acids *R News* **6**, 2–7.

Weaver, L. M., Gan, S., Quirino, B. and Amasino, R. M. (1998) A comparison of the expression patterns of several senescence-associated genes in response to stress and hormone treatment *Plant Mol Biol* **37**, 3, 455–469.

Weston, D. J., Gunter, L. E., Rogers, A. and Wullschleger, S. D. (2008) Connecting genes, coexpression modules, and molecular signatures to environmental stress phenotypes in plants *BMC Systems Biology* **2**, 1, 16.

Whalen, M. C. (2005) Host defence in a developmental context *Mol Plant Pathol* **6**, 3, 347–360.

Williams, E. J. B. and Bowles, D. J. (2004) Coexpression of neighboring genes in the genome of Arabidopsis thaliana. *Genome Res* **14**, 6, 1060–1067.

Williamson, B., Tudzynski, B., Tudzynski, P. and van Kan, J. A. (2007) Botrytis cinerea: the cause of grey mould disease *Mol Plant Pathol* **8**, 5, 561–580.

Windram, O., Madhou, P., McHattie, S., Hill, C., Hickman, R., Cooke, E., Jenkins, D. J., Penfold, C. A., Baxter, L., Breeze, E., Kiddle, S. J., Rhodes, J., Atwell, S., Kliebenstein, D. J., Kim, Y.-S., Stegle, O., Borgwardt, K., Zhang, C., Tabrett, A., Legaie, R., Moore, J., Finkenstadt, B., Wild, D. L., Mead, A., Rand, D., Beynon, J., Ott, S., Buchanan-Wollaston, V. and Denby, K. J. (2012) Arabidopsis defense against Botrytis cinerea: chronology and regulation deciphered by high-resolution temporal transcriptomic analysis *Plant Cell* **24**, 9, 3530–3557.

Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Prüss, M., Reuter, I. and Schacherer, F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* **28**, 1, 316–319.

Woolcock, A. J., Salome, C. M. and Yan, K. (1984) The shape of the dose-response curve to histamine in asthmatic and normal subjects *Am Rev Respir Dis* **130**, 1, 71–75.

Wright, P. (1999) Premature senescence of cotton (Gossypium hirsutum L.)–Predominantly a potassium disorder caused by an imbalance of source and sink *Plant Soil* **211**, 2, 231–239.

Wu, H., Kerr, M. K., Cui, X. and Churchill, G. A. (2003) *The analysis of gene expression data: methods and software* chapter MAANOVA: a software package for the analysis of spotted cDNA microarray experiments, 313–341 Springer.

Xin, Z., Zhao, Y. and Zheng, Z.-L. (2005) Transcriptome analysis reveals specific modulation of abscisic acid signaling by ROP10 small GTPase in Arabidopsis *Plant Physiol* **139**, 3, 1350–1365.

Xu, Y., Chang, P.-F. L., Liu, D., Narasimhan, M. L., Raghothama, K. G., Hasegawa, P. M. and Bressan, R. A. (1994) Plant defense genes are synergistically induced by ethylene and methyl jasmonate. *Plant Cell* **6**, 8, 1077–1085.

Xu, Z. and Norris, D. (1998) The SFP1 gene product of Saccharomyces cerevisiae regulates G2/M transitions during the mitotic cell cycle and DNA-damage response. *Genetics* **150**, 4, 1419–1428.

Xue, G.-P. (2005) A CELD-fusion method for rapid determination of the DNA-binding sequence specificity of novel plant DNA-binding proteins *Plant J* **41**, 4, 638–649.

Xue, G.-P., Way, H. M., Richardson, T., Drenth, J., Joyce, P. A. and McIntyre, C. L. (2011) Overexpression of TaNAC69 leads to enhanced transcript levels of stress up-regulated genes and dehydration tolerance in bread wheat *Mol Plant* **4**, 4, 697–712.

Yanagisawa, S. (2000) Dof1 and Dof2 transcription factors are associated with expression of multiple genes involved in carbon metabolism in maize *Plant J* **21**, 3, 281–288.

Yanagisawa, S. and Sheen, J. (1998) Involvement of maize Dof zinc finger proteins in tissue-specific and light-regulated gene expression *Plant Cell* **10**, 1, 75–89.

Yang, C.-Y., Hsu, F.-C., Li, J.-P., Wang, N.-N. and Shih, M.-C. (2011) The AP2/ERF transcription factor AtERF73/HRE1 modulates ethylene responses during hypoxia in Arabidopsis *Plant Physiol* **156**, 1, 202–212.

Yang, R. and Su, Z. (2010) Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation *Bioinformatics* **26**, 12, i168–74.

Yeung, K. Y., Haynor, D. R. and Ruzzo, W. L. (2001) Validating clustering for gene expression data *Bioinformatics* **17**, 4, 309–318.

Yoshida, T., Fujita, Y., Sayama, H., Kidokoro, S., Maruyama, K., Mizoi, J., Shinozaki, K. and Yamaguchi-Shinozaki, K. (2010) AREB1, AREB2, and ABF3 are master transcription factors that cooperatively regulate ABRE-dependent ABA signaling involved in drought stress tolerance and require ABA for full activation *Plant J* **61**, 4, 672–685.

Zavaleta-Mancera, H., Thomas, B., Thomas, H. and Scott, I. (1999) Regreening of senescent Nicotiana leaves II. Redifferentiation of plastids *J Exp Bot* **50**, 340, 1683–1689.

Zhang, C., Xie, Q., Anderson, R. G., Ng, G., Seitz, N. C., Peterson, T., McClung, C. R., McDowell, J. M., Kong, D., Kwak, J. M. *et al.* (2013) Crosstalk between the Circadian Clock and Innate Immunity in Arabidopsis *PLoS Pathog* **9**, 6, e1003370.

Zhang, K. and Gan, S.-S. (2012) An abscisic acid-AtNAP transcription factor-SAG113 protein phosphatase 2C regulatory chain for controlling dehydration in senescing Arabidopsis leaves *Plant Physiol* **158**, 2, 961–969.

Zhao, Y., Wei, T., Yin, K.-Q., Chen, Z., Gu, H., Qu, L.-J. and Qin, G. (2012) Arabidopsis RAP2.2 plays an important role in plant resistance to Botrytis cinerea and ethylene responses *New Phytologist* **195**, 2, 450–460.

Zhong, S. and Xie, D. (2007) Gene Ontology analysis in multiple gene clusters under multiple hypothesis testing framework *Artif Intell Med* **41**, 2, 105–115.

Zhong, S., Zhao, M., Shi, T., Shi, H., An, F., Zhao, Q. and Guo, H. (2009) EIN3/EIL1 cooperate with PIF1 to prevent photo-oxidation and to promote greening of Arabidopsis seedlings. *Proc Natl Acad Sci U S A* **106**, 50, 21431–21436.

Zhou, J.-M., Trifa, Y., Silva, H., Pontier, D., Lam, E., Shah, J. and Klessig, D. F. (2000) NPR1 differentially interacts with members of the TGA/OBF family of transcription factors that bind an element of the PR-1 gene required for induction by salicylic acid *Mol Plant Microbe Int* **13**, 2, 191–202.

Zourelidou, M., Torres-Zabala, D., Smith, C., Bevan, M. W. *et al.* (2002) Storekeeper defines a new class of plant-specific DNA-binding proteins and is a putative regulator of patatin expression *Plant J* **30**, 4, 489–497.