

University of Warwick institutional repository: <http://go.warwick.ac.uk/wrap>

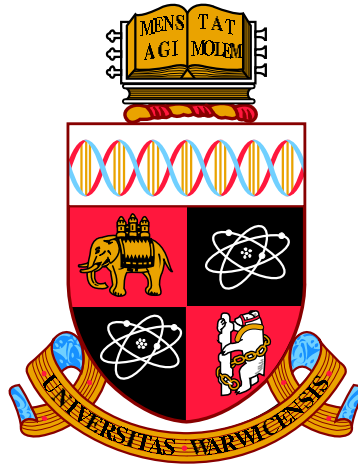
A Thesis Submitted for the Degree of PhD at the University of Warwick

<http://go.warwick.ac.uk/wrap/45731>

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.



**Modelling transcriptional networks
in plant senescence**

by

Stuart David James McHattie

Thesis

Submitted to the University of Warwick
for the degree of

Doctor of Philosophy

Department of Systems Biology

August 2011

THE UNIVERSITY OF
WARWICK

Contents

List of Tables	vi
List of Figures	viii
Acknowledgments	xii
Declarations	xiii
Abstract	xiv
Abbreviations	xv
Chapter 1 Introduction	1
1.1 The model organism: <i>Arabidopsis thaliana</i>	1
1.2 Leaf Senescence in <i>Arabidopsis</i>	4
1.2.1 Causes of Senescence	5
1.2.2 The Senescence Response	8
1.3 A Systems Biology Approach to Finding a Senescence Transcriptional Network	10
1.4 Gene Expression Data during Natural Senescence	11
1.4.1 CATMA Microarrays	13
1.4.2 Experimental Design	13
1.4.3 Microarray Scanning	15
1.4.4 Scan Analysis	15
1.5 Microarray Data Analysis Software (MAANOVA)	17
1.5.1 Purpose and Justification	17
1.5.2 Original MAANOVA Implementation	18
1.6 Analysis of Gene Expression Data	29
1.6.1 Clustering	29
1.6.2 Identifying Over-represented Gene Ontologies	32
1.6.3 Identifying Network Topology Using Variational Bayesian State Space Modelling (VBSSM)	33
1.7 Biological Validation	37
1.7.1 Altered-Expression Mutants	37

1.7.2	Yeast-1-Hybrid	38
1.8	Aims of the Project	39
Chapter 2	Materials and Methods	41
2.1	Chapter Summary	41
2.2	Microarray Analysis	41
2.3	Gene Expression Data Analysis	42
2.3.1	SplineCluster	42
2.3.2	BiNGO	44
2.3.3	Qian Similarity Score	45
2.4	Theoretical Network Modelling	47
2.4.1	Variational Bayesian State-Space Modelling	47
2.4.2	Cytoscape	48
2.4.3	Multi-modelling	49
2.4.4	Metropolis-like VBSSM	50
2.5	Plant Preparation and Phenotype Screening	54
2.5.1	Plant Growth	54
2.5.2	Phenotype Screening	55
2.6	Plant Transcriptomics	60
2.6.1	DNA Extraction	60
2.6.2	T-DNA Insertion Screening	61
2.6.3	RNA Extraction	62
2.6.4	RNA Concentration Measurement with a NanoDrop	65
2.6.5	Bioanalysis of total RNA Samples	65
2.6.6	DNase Treatment	66
2.6.7	RT-PCR Expression Analysis	66
2.7	Cycle Sequencing of DNA using BigDye®	69
2.8	Yeast-1-Hybrid	70
2.8.1	Growth Media Preparation	70
2.8.2	Amplifying Upstream Fragments of the Target Gene	71
2.8.3	Inserting Fragments into Plasmid Vectors	72
2.8.4	Bacterial Transformation	74
2.8.5	Confirming the Presence of the Insert	74
2.8.6	Plasmid Cloning	75
2.8.7	Sequencing the Insert	75
2.8.8	Yeast Transformation	75
2.8.9	Screening Against a Transcription Factor cDNA Library	77
2.9	Microarray Analysis of Reduced-Expression Mutants	80
2.9.1	Preparing and Harvesting Samples	80
2.9.2	aRNA Synthesis and Purification	80
2.9.3	Labelling with Fluorescent Dyes	82
2.9.4	Preparation of a Balanced Experimental Design	82

2.9.5	Microarray Solution Preparations	83
2.9.6	Microarray Hybridisation	86
2.9.7	Microarray Scanning and Data Extraction	87
Chapter 3	Microarray Data Analysis	89
3.1	Chapter Summary	89
3.2	Introduction	90
3.3	Development of MAANOVA	90
3.3.1	Quality Control	90
3.3.2	Data Transformation	97
3.3.3	Model Fitting	108
3.3.4	Identifying Differentially Expressed Genes	109
3.3.5	Extracting Predicted Gene Expression Data	110
3.4	Application of MAANOVA to Microarray Data	115
3.4.1	Quality Control	115
3.4.2	Application of TechRepCheck and Correction of Poor Microarrays	126
3.4.3	LOWESS Transformation	131
3.4.4	Generating a Mixed Model of Gene Expression Data	132
3.4.5	Selection of Differentially Expressed Genes	135
3.4.6	Extraction of Predicted Gene Expression Data	137
3.5	Conclusions	139
3.5.1	Application to Other Datasets	139
3.5.2	Future Development	140
3.5.3	Chapter Synopsis	141
Chapter 4	Analysis of Gene Expression Data	143
4.1	Chapter Summary	143
4.2	Initial Analysis of Gene Expression Data	144
4.2.1	Gene Expression Clustering	144
4.2.2	Gene Ontology Analysis	148
4.2.3	Preliminary Network Modelling	154
4.3	Identification of Module Centres	158
4.4	Biological Validation of Module Centres	165
4.4.1	Phenotype Screening	166
4.4.2	Homozygous T-DNA Insertion Screening	169
4.4.3	Gene Expression Level Testing	171
4.4.4	Known Gene Functions	174
4.5	Predicting Local Networks for Module Centres	178
4.6	Identifying a Predicted Transcriptional Network Associated with Senes- cence	182
4.7	Conclusions	183
4.7.1	Efficiency of Modelling in Identifying Valid Hubs	184

4.7.2	Future Development	184
Chapter 5	Biological Validation of Theoretical Models	187
5.1	Chapter Summary	187
5.2	Yeast-1-Hybrid Analysis	189
5.2.1	Identifying Conserved Regions Upstream of Genes	189
5.2.2	Selecting Fragments to be Screened	189
5.2.3	Amplifying Plasmid Inserts	193
5.2.4	Inserting Fragments into Plasmid Vectors	193
5.2.5	Bacterial Transformation and Plasmid Cloning	193
5.2.6	Screening Against the Transcription Factor Library	194
5.3	Reduced-Expression Microarray Experiment	200
5.3.1	Sample Harvesting	201
5.3.2	Identifying Homozygous T-DNA Insertions	202
5.3.3	Screening Expression Levels	205
5.3.4	Preparing Samples for Hybridisation	208
5.3.5	Microarray Data Analysis	211
5.4	Updating the Putative Transcriptional Network Model	221
5.5	Conclusions	224
5.5.1	Future Development	224
Chapter 6	General Discussion	227
6.1	Microarray Data Analysis	227
6.1.1	Future Developments	228
6.2	Analysis of Gene Expression Data	229
6.2.1	Future Developments	230
6.3	Biological Validation of Theoretical Models	231
6.3.1	Future Developments	231
6.4	Closing Remarks	232
Appendix A	VBSSM Controlling Script	235
Appendix B	WHRIMAANOVA Function Definitions	236
Appendix C	MAANOVA Analysis Script	243
Appendix D	DrawGOGraphs() Source Code	245
Appendix E	Supplementary GO term results	248
Appendix F	Multimodelling Source Code	251
Appendix G	Metropolis-like VBSSM Source Code	254
Appendix H	White Balance Transformation Script	261

Appendix I	Logistic Curve Model Fitting Script	264
Appendix J	Hub Genes Identified by Multi-modelling	270
Appendix K	Logistic Curve Plots for Quantitative Phenotype Screens	278
Appendix L	Yeast-1-Hybrid Sequencing Results	285
Appendix M	Balanced Design Source Code	290

List of Tables

2.1	An example of Qian similarity scores output by TCAP	46
2.2	The four sections of the Metropolis-like VBSSM processing script	52
2.3	An optimised experimental design for microarray hybridisations	84
3.1	Statistics produced by GridCheck summarising each sub-plot	93
3.2	Summary statistics for distributions of \log_2 ratio and sum \log_2 intensity data produced by ArrayView	100
3.3	Technical replicate comparisons identifiable by TechRepCheck	104
3.4	Top 45 worst GridCheck sub-plots ranked by t-tests of the major axis for outlying probes above or below the regression	119
3.5	Classification of low quality microarrays after analysis by MAANOVA	129
3.6	Sample allocations for low quality microarrays	130
4.1	Table of cluster sizes returned by SplineCluster	148
4.2	Highest Significance Over-represented Biological Process GO Terms in Senescence Clusters	149
4.3	A sample of raw results returned by multi-modelling	164
4.4	A sample of hub genes found by multi-modelling	165
4.5	SALK lines which were biologically tested as putative hubs	167
4.6	Genes demonstrating an altered phenotype after quantification of the rate of dark-induced senescence	169
4.7	Nanodrop results for total RNA extracted from mature leaves of reduced- expression lines	172
4.8	Metropolis-like VBSSM modelling runs performed for genes identified as module centres	179
4.9	Numbers of predicted upstream and downstream interactions for each module centre	181
5.1	Conserved transcription factor binding sites in regions to be cloned for yeast-1-hybrid	190
5.2	Length and position of fragments used as inserts for yeast-1-hybrid cloning	190
5.3	Transcription factors proven able to bind specific DNA sequences and cause downstream transcription	195

5.4	SALK lines considered for the model validation microarray experiment .	201
5.5	Senescence values of harvested samples from reduced-expression lines for microarray analysis	203
5.6	Results of a T-DNA insertion screen performed on eleven reduced-expression mutants	204
5.7	Total RNA concentrations obtained by extraction from harvested leaf tissue	206
5.8	Concentrations of aRNA samples to be labelled with fluorescent dyes determined by NanoDrop	210
5.9	A comparison of expression levels for reduced-expression genes in T-DNA insertion mutants	213
5.10	Results of a BiNGO analysis of the 58 genes differentially expressed in both IM137 and IM188	216
5.11	Genes found to be altered in expression during reduced expression of <i>STZ</i> and <i>ANAC092</i>	218
J.1	Hub genes identified by multi-modelling for the gene pool generated using SplineCluster	271

List of Figures

1.1	Phenotypical Stages in the Development of <i>Arabidopsis thaliana</i> in 16 hour long day conditions	2
1.2	The harvesting schedule of the natural senescence microarray experiment	12
1.3	The loop design for microarrays hybridised with samples of the same time of day	15
1.4	The loop design for microarrays hybridised with samples of different times of day	16
1.5	Log ₂ transformation of microarray data allows the data to meet a number of expectations	20
1.6	Examples of graphical output produced by functions of MAANOVA . .	26
1.7	Types of gene profile similarity which are ranked highly by the Qian similarity score	32
1.8	Probabilistic graphical Bayesian network model representation of VBSSM	35
2.1	An example of histograms returned by the script which interprets BiNGO's output	45
2.2	An example of a typical F vs <i>k</i> plot produced by VBSSM for the senescence experiment data	51
2.3	Alignment marks used as a background during phenotype screenings . .	55
2.4	Illustration of colour histogram analysis performed in ImageJ before and after dark-induced senescence	57
2.5	An example of a logistic curve with model parameters shown	59
2.6	An illustration of the location of primers used in testing for homozygous T-DNA insertion	61
2.7	Features of vectors used in yeast-1-hybrid screens	72
3.1	Examples of improved graphical output produced by GridCheck and RIPlot	95
3.2	Examples of improved graphical output produced by ArrayView	96
3.3	Examples of additional graphical output produced by ArrayView	98
3.4	Histogram plots produced by ArrayView	99
3.5	TechRepCheck output demonstrating an obvious outlying replicate . . .	102
3.6	ArrayView ratio heat-maps before and after TechRepCheck probe estimation	107

3.7	Gene expression profile plots demonstrating inclusion of experimental terms of the model	112
3.8	Examples of normalisation applied by inclusion of mean terms of the model	114
3.9	GridCheck sub-plots indicating typical artefacts of CATMA microarrays	116
3.10	ROC curves demonstrating the power of ranking sub-plots using statistics	117
3.11	RIPlot output indicating typical artefacts of CATMA microarrays . . .	121
3.12	Cropped ArrayView heat-maps for sub-grids flagged by GridCheck . . .	122
3.13	ArrayView heat-maps for sub-grids flagged by RIPlot	124
3.14	ArrayView with TechRepCheck estimate indicators	127
3.15	RIPlots before and after the application of LOWESS transformations . .	133
3.16	ArrayView heat-maps of a microarray with spatial bias before, during and after Transform MAData	134
3.17	A comparison of the p -value transformation produced by family-wise error-rate controls of MAANOVA applied to the term Day	136
3.18	Venn diagram summarising numbers of genes showing significant differential expression for each combination of treatment terms in the MAANOVA fixed model	138
4.1	Heatmaps providing evidence of improvement after merger sweeping in SplineCluster	145
4.2	Plots of gene profiles for each cluster output by SplineCluster	147
4.3	Hormone Levels During Leaf Development	151
4.4	VBSSM network models of similar sets of genes demonstrate drastic differences in individual relationships	155
4.5	A comparison of F vs k plots for data with and without biological replicates	157
4.6	An artificial network demonstrating types of node present in a typical regulatory network	158
4.7	Histograms showing the frequency of pairs being presented for multi-modelling	162
4.8	Plots of senescence profiles for knockout mutant lines showing altered phenotype during dark-induced senescence	168
4.9	Analysis of logistic curve models reveal the significance of altered phenotypical responses during dark-induced senescence	170
4.10	Electrophoresis gel demonstrating the presence of homozygous T-DNA inserts for reduced-expression mutants	171
4.11	Agilent 2100 bioanalyser results for total RNA samples from leaves of reduced-expression lines	173
4.12	RT-PCR results demonstrating expression levels of three reduced-expression lines	174
4.13	Expression levels of biologically validated hub genes during senescence .	176
4.14	Convergence of Metropolis-like VBSSM for HAT3	180
4.15	The final VBSSM model showing interactions to be confirmed <i>in vivo</i> .	183

5.1	Elements of the theoretical network model which have been tested <i>in vivo</i>	188
5.2	Yeast-1-hybrid fragments within the 1,000 bp region upstream of transcriptional start sites	191
5.3	Primer sequences used to amplify inserts for yeast-1-hybrid cloning . . .	192
5.4	Comparing digested vector and insert concentrations by electrophoresis gel	194
5.5	Gene expression profiles for transcription factors identified by yeast-1-hybrid and their targets	197
5.6	Primers used to screen for T-DNA insertions in reduced-expression mutants	204
5.7	Bioanalyser results for total RNA samples extracted using two alternative methods	206
5.8	Comparison of cDNA concentrations by PCR amplification of a highly expressed gene	207
5.9	Relative expression levels of genes containing T-DNA in reduced-expression lines	208
5.10	Pseudo-electrophoresis gel produced during bioanalysis of total RNA samples for microarrays	209
5.11	Pseudo-electrophoresis gel produced during bioanalysis of aRNA samples for microarrays	211
5.12	A 4-way Venn diagram showing numbers of genes with altered expression for each line	213
5.13	Binding sites for CATMA probes specific to reduced-expression genes .	214
5.14	A 2-way Venn diagram showing numbers of genes with altered expression for IM137 and IM188 only	215
5.15	An updated putative transcriptional network model based on results of yeast-1-hybrid screens, literature resources and microarray expression analysis	221

List of Procedures

2.1	Function definition for the SplineCluster wrapper	43
2.2	Pseudo-code of the VBSSM control script	47
2.3	Transformation of photographs to obtain white balance	58

Acknowledgments

A number of people have been pivotal to the completion of these studies and, to all of them, I would like to express my deepest gratitude. I would like to thank my supervisors Vicky Buchanan-Wollaston, David Wild and Andrew Mead, who have been supportive, motivating and inspirational during this time. I would also like to thank Jim Beynon, Katherine Denby and Miriam Gifford who have provided their support during this work.

Despite my persistent and mundane questions Emily Breeze, Claire Hill, Jesper Grønlund, Laura Butler, Matthew Watson and Sanjeev Kumar have been tremendously supportive in laboratory techniques. Similarly, Laura Baxter, Christopher Penfold, Richard Savage and Dafyd Jenkins have given their support in matters of theoretical application. My thanks go out to all of them.

Fellow students Linda Hughes, Mary Coates, Jo Rhodes, Laura Lewis, Polly Downton, Ed Morrissey and Richard Hickman have all given them time on more than one occasion to discuss my work and help me to keep it on track. They may not realise it, but the mutual moral support between us has allowed the good times to shine and the bad times to pale into insignificance.

I am particularly grateful to Steven Kiddle who has become one of my best friends and has always given up his time to explain mathematical quandaries in an understandable way. With such a rare gift as this, I know that his future success is inevitable! Another of my best friends is Oliver Windram who has shown me how it is possible to work hard whilst still maintaining a healthy social life. His motivational personality is second to none and I genuinely hope to stay in contact for long to come.

My girlfriend, Annette Taylor, has put up with a lot of neglect during my studies and I am very grateful for her patience with me. Finally, my parents, Carol McHattie and David McHattie, have been the driving force behind my successes in life and I could never thank them enough for the support and determination they have given me to be the best that I can be. Even after death, my father remains an inspiration to me.

Declarations

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

Abstract

Senescence is a highly regulated developmental process in plants in which nutrients are remobilised from organs which are no longer required or are stressed so that they may be used by organs which are only just developing. Whilst much is known about the causes of and the resulting outcomes of this process, very little is known about the genetic machinery which link them. Some genes have been identified as having a regulatory role in the senescence process, but many of these have been determined by a forward genetics approach whereby mutants are randomly screened for phenotypical effects. A much better approach, where possible, is the reverse genetics approach whereby mutants are sought for testing as they are suspected to demonstrate a phenotypical effect.

It was the purpose of this study to find novel ways of identifying those genes which may be highly regulatory of the senescence process and to determine how they are able to lead from several different known causes of senescence through to the senescence response itself. It was hypothesised that, using measurements of the expression levels of a very large number of genes throughout the senescence process, theoretical models of regulation between those genes could be determined and that these theoretical models would allow specific interactions to be identified and explained using biological validation techniques.

A large microarray experiment, performed prior to the start of this project, measured the expression of over 30,000 genes in the *Arabidopsis thaliana* genome over a period of 21 days during natural senescence. By cleaning this data and fitting an ANOVA driven model to the resulting intensity measurements, it has been possible to separate effects leading to observed expression changes. The levels of each of these effects were tested by F-tests and this has allowed the identification of 8,878 genes which are significantly differentially expressed during senescence.

By first using theoretical models to find genes amongst the set of 8,878 which demonstrate highly robust regulatory behaviour on other genes in the set, 118 genes were able to be isolated for further study. A senescence phenotype screen was developed to assess reduced-expression mutants of many of those 118 genes and 8 were shown to have a significantly altered timing of senescence when compared with wild-type plants. The surrounding networks of each of those 8 genes were formed by applying theoretical regulatory network modelling in another novel manner similar to a Metropolis-Hastings approach which identified a set of 75 genes providing a testable regulatory network model.

The resulting network model has been tested biologically to establish the accuracy of the predictions. Whilst many of the predictions were not confirmed, a vast network has been identified surrounding two of the highly regulatory genes indicating a junction of two separate pathways leading to the senescence response and providing a network structure which could be used in another round of theory and validation. Additionally, these results introduce new interesting questions about how the senescence network may have evolved to respond to so many inputs.

Abbreviations

A_{340nm} absorbance at wavelength 340nm

AGI *Arabidopsis* genome identifier

ANOVA analysis of variance

APPLES Analysis of Plant Promoter-Linked Elements

aRNA anti-sense ribonucleic acid

bp base pair

CATMA Complete Arabidopsis Transcriptome Micro Array

cm centimetres

CO₂ carbon dioxide

DAS days after sowing

DMSO dimethyl sulfoxide

DNA deoxyribonucleic acid

DNase deoxyribonuclease

E. coli *Escherichia coli*

EDA edge attribute file

EDTA ethylenediaminetetraacetic acid

EtBr ethidium bromide

EtOH ethanol

FDR false-discovery-rate

FWER family-wise error-rate

g gram

g centrifugal force (relative to gravitational pull)

GO gene ontology

GST gene-specific sequence tag

H₂O water

His histidine

JA jasmonic acid

kb kilobase

L litre

LB lysogeny broth

Leu leucine

LOWESS locally weighted sum of squares

μ g microgram

μ L microlitre

μ M micromolar

M molar (mols per litre)

m metres

mL millilitre

mM millimolar

mg milligram

mm millimetres

MAANOVA MicroArray ANalysis Of VAriance

MCMC Markov chain Monte Carlo

MCS multiple cloning site

MOPS 3-(N-morpholino)propanesulfonic acid

nm nanometres

NOA node attribute file

ORF open reading frame

PCA principal component analysis

PCR polymerase chain reaction

PEG polyethylene glycol

pH $-\log_{10}[\text{H}^+]$

RNA ribonucleic acid

MAR major axis regression

mRNA messenger ribonucleic acid

RPM revolutions per minute

RT-PCR reverse transcriptase polymerase chain reaction

SA salicylic acid

SAG senescence associated gene

SD synthetic defined (base media for yeast)

SDS sodium dodecyl sulphate

SIF simple interaction format

SIGnaL Salk institute genome analysis laboratory

SNP single nucleotide polymorphism

SOC media Super Optimal broth with Catabolite repression

T-DNA transfer DNA

Tris tris(hydroxymethyl)aminomethane

Trp tryptophan

UV ultra violet

V volts

VBSSM Variational Bayesian State Space Modelling

W Watts

Y1H Yeast-1-hybrid

YPDA yeast peptone dextrose adenine

Chapter 1

Introduction

Senescence is a process familiar to most as the onset of old age. The etymology of the word senescence is from the Latin for old man, *senex*, which is also the source of the word senile in modern English. In the animal kingdom, senescence is often thought of as affecting the whole organism and associated with deterioration of strength, wrinkling of the skin and eventually death. In contrast, senescence in the plant kingdom is a highly important developmental process in which nutrients are remobilised from organs which are no longer needed or are stressed and transferred to those organs which are just developing.

The study of plant senescence is not only an interesting area of research from a biological perspective, but also one of important social impact given the need for food security in a future of ever increasing populations. Those crops which suffer wastage either because of unavoidably poor growing conditions, undefended pathogenic attack and/or poor handling of processed food crops severely impact the food industry. If the mechanisms which lead to senescence were better understood, it could be possible to delay its onset and improve yield from harvests.

This project has used a system-wide approach to identifying a key set of genes thought to be associated with the senescence processes of *Arabidopsis thaliana*. By identifying these, the topology of small areas of the associated transcriptional networks have then been established.

1.1 The model organism: *Arabidopsis thaliana*

Arabidopsis thaliana is a spring annual plant found in areas of Asia, Europe and North-West Africa. It is a dicotyledon and a member of the *Brassicaceae* family. *Arabidopsis* is widely used as the model organism in areas of genomic research of plants and, in some cases, other eukaryotic organisms. A number of reasons exist for this and are discussed:

Small genome *Arabidopsis* has one of the smallest genomes in its kingdom with only 115.4 Mbp over 5 chromosomes (Arabidopsis Genome Initiative, 2000). The entire

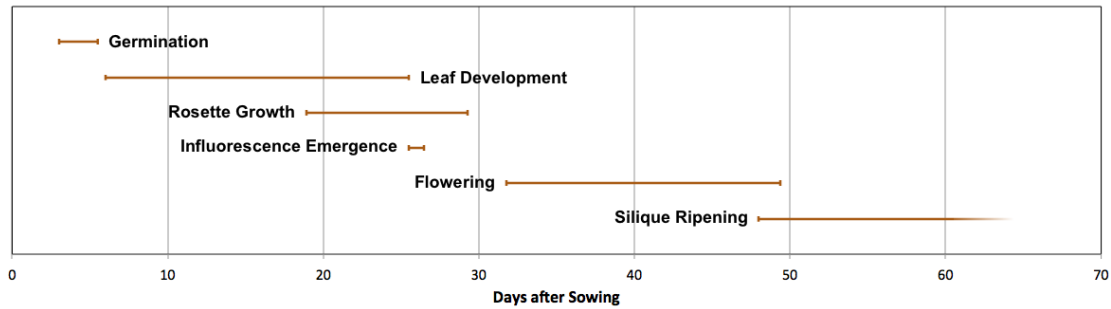


Figure 1.1 – Phenotypical Stages in the Development of *Arabidopsis thaliana* in 16 hour long day conditions

The phenotypical stages in the development of *Arabidopsis* subjected to 16 hours of daylight each day were measured by Boyes *et al.* (2001). It can be seen that during the first 4 weeks after sowing, the plants undergo differentiation, growth and expansion. After this, the plants are fully matured and begin to flower and fill their siliques. The point of completed silique ripening and complete senescence were not determined in the study because plants were removed before seeds were dispersed to prevent cross contamination.

genome was sequenced and analysed simultaneously in all 5 chromosomes by members of the Arabidopsis Genome Initiative (AGI) (Theologis *et al.*, 2000; Lin *et al.*, 1999; Salanoubat *et al.*, 2000; Mayer *et al.*, 1999; Tabata *et al.*, 2000). This extensive knowledge about the sequence of *Arabidopsis* led to the construction of The *Arabidopsis* Information Resource (TAIR) (Rhee *et al.*, 2003), which has allowed researchers to collaborate by providing annotations and other information about identified genes. All these collaborative resources have given geneticists a head start when using *Arabidopsis* as a model organism in their research.

Little ‘Junk’ DNA During analysis of the *Arabidopsis* genome, soon after it was fully sequenced, it was estimated that 25,498 genes exist in the genome (Arabidopsis Genome Initiative, 2000). When compared with other species such as maize, which is also used as a model organism in genetics, it can be seen that the maize genome is made up of 2.3 Gbp over 10 chromosomes and contains only 32,540 protein encoding genes (Schnable *et al.*, 2009). This indicates that gene density is more than 15 times higher in *Arabidopsis*, making random mutations much more likely to be part of a gene sequence.

Since the initial estimates of genes in the *Arabidopsis* genome, further genes have been identified and the current number of genes identified is 27,406, meaning even less of the genome is made up of ‘junk’ DNA (http://www.ebi.ac.uk/integr8/OrganismStatsAction.do?orgProteomeId=3¤tclicked=GENOME_STATISTICS).

Rapid Development *Arabidopsis* is able to complete a whole generation, from seed-to-seed, in around 5 to 7 weeks in optimum conditions. This fast development allows many generations to be observed in a relatively short period of time so that mutations can be easily established, including double and triple mutations. The developmental timeline of Col-0 *Arabidopsis*, treated with 16 hours of daylight each day, has been

quantified by Boyes *et al.* (2001) from initial sowing through to silique ripening and senescence. An adaptation of their timeline can be found in Figure 1.1. This was determined to be accurate because the coefficient of variation between replicates is less than 15%.

Plants are Small *Arabidopsis* is a relatively small plant and can be grown in large quantities with only limited available space. Col-0 rosettes, at their largest, only measure 82.3×73.1 mm (Boyes *et al.*, 2001) and can be grown adjacent to one another. Furthermore, seed stocks need only be stored in small tubes or packets in a cool room, thousands of seeds for many thousands of genotypes taking up only a modestly sized box.

High Seed Yields On average, each mature *Arabidopsis* plant will shed 29.9 seeds from each of 160.4 siliques, resulting in an average of 4,796 seeds per plant (Boyes *et al.*, 2001). Due to variations between individuals, a small minority can produce in excess of 10,000 seeds. This high yield of seeds allows a single genotype to be maintained over many experiments and therefore for studies to remain consistent.

Simple to Mutate There are many methods for generating *Arabidopsis* mutants including transformation by transfer DNA (T-DNA) inserted using *Agrobacterium tumefaciens* to introduce new genes or over-express existing genes (Lloyd *et al.*, 1986) and random disruption of *Arabidopsis* genes by the insertion of large sequences of ‘junk’ T-DNA (Feldmann, 1991). Because it is simple to generate mutants, it is easy to perturb the system in *Arabidopsis* and therefore to identify the role of specific genes by disrupting or enhancing their expression.

Self Pollination Most members of the *Brassicaceae* family actively prevent self-pollination by action of *S*-locus receptor kinases (SRK) which mediate recognition of self-pollen on the female side (Takasaki *et al.*, 2000) and *S*-locus cysteine-rich (SCR) proteins which are located on the coat of the pollen and bind directly to SRK to cause its phosphorylation and prevent pollination (Stephenson *et al.*, 1997; Schopfer *et al.*, 1999). *Arabidopsis*, however, has a mutation of those genes which means that it is able to self-pollinate. This allows recessive mutations to quickly become homozygous and therefore be expressed.

Despite its use as a model organism, it is unlikely that *Arabidopsis* will provide a perfect model for researching processes in other plants with economic and cultural importance as processes do vary between species. However, because of the reasons shown above, it does provide an excellent system to gain a good grasp of the most important aspects of those processes which are largely transferrable to the target plant systems, giving *Arabidopsis* the title of “model organism”.

1.2 Leaf Senescence in *Arabidopsis*

Senescence in plants is an age-dependent, highly co-ordinated and regulated developmental process during which nutrients are recovered from organs which are soon to be discarded so as to recycle them rather than wastefully discard them along with the organ. Senescence can have many causes, both biotic and abiotic, but natural senescence in plants is as a result of maturation and reproduction in monocarpic plants such as *Arabidopsis* or in response to changing environmental conditions in the case of the Autumnal senescence observed in deciduous species. In the latter case, scenes of trees with leaves changing from green to yellow, orange and red are those typically recognised by many as a change in the season as the summer draws to a close and autumn begins. However, behind those phenotypical changes seen during leaf senescence, a complex co-ordination of thousands of genes and many signalling pathways is occurring (Buchanan-Wollaston *et al.*, 2005; van der Graaff *et al.*, 2006). The highly co-ordinated nature of the senescence response is most obvious in a study of *Nicotiana rustica* in which senescent leaves were seen to regain structural features of chloroplasts, cease protease activity, re-green and continue to photosynthesise when the plant was decapitated above the single senescent leaf (Zavaleta-Mancera *et al.*, 1999; Zavatele-Mancera *et al.*, 1999). This indicates that the process is reversible and that the plant is in full control of the senescence process throughout.

Leaf senescence in *Arabidopsis* broadly falls into two main categories: developmental and stress-related. In the case of developmental senescence, silique filling is accompanied by the re-mobilisation of nutrients such as nitrogen, phosphorus, and some metal ions from the leaves so that they may be relocated into the developing seeds (Lim *et al.*, 2007, Reviewed). Once seed has been produced, the leaves serve no further purpose and so the relocation of these nutrients is beneficial to the plant as a whole. During stress-induced senescence, the phenotype is similar, but the gene expression levels may be different from those observed during developmental senescence. One example of this showed that only half the number of genes were altered in expression during dark-induced senescence, during which *Arabidopsis* plants are made to senesce by being subjected to constant and complete darkness, compared with those of developmental senescence (van der Graaff *et al.*, 2006). Further differences between developmental senescence and stress-induced senescence are that, during the latter, the senescence response can be seen to affect only the part of the plant or leaf affected by the stress. This is very noticeable in dark-induced senescence where symptoms such as chlorophyll degradation can be seen to act almost uniformly across the leaf, whereas senescing leaves usually demonstrate a loss of chlorophyll at the distal areas of the leaf first, the cells around the veins of the leaf remaining active the longest to promote transport out of the leaf (Buchanan-Wollaston *et al.*, 2005).

1.2.1 Causes of Senescence

Many of the factors which alter the onset or rate of leaf senescence are known and have been researched extensively to try to identify the mechanism by which they are able to affect the senescence process. However, despite these efforts, those mechanisms remain poorly defined. Each of the known causes of senescence can be categorised as either internal or external factors (Lim *et al.*, 2007). Internal factors are those signals generated by the plant, in response to both developmental age and forms of stress, which can cause a senescence response. External factors are those stresses exerted on the plant which result in a senescence response.

Although the exact mechanisms by which senescence is caused are not known, a number of developmental triggers have been identified. For example, the metabolic rate of cells was suspected to be a trigger when *oresara 4-1* (*ore4-1*) mutants, which are partially defective in chloroplast function, showed a delayed senescence response in developmental senescence, but not in dark-induced or hormone senescence (Woo *et al.*, 2002). Another example is that the lowered activity of processes such as photosynthesis appear to act as a trigger for senescence. This is supported by the observation that increased concentrations of sugars lower photosynthetic activity and that under these circumstances, senescence is induced (Dai *et al.*, 1999; Quirino *et al.*, 2000).

1.2.1.1 Phytohormones

Phytohormones are plant hormones which occur in extremely low concentrations and affect the growth and development of the plant. A number of these have been linked with senescence and their exogenous application and/or the disruption of their associated pathways have been found to alter both the timing and progression of senescence at all stages:

Cytokinins have functions in cell proliferation, shoot formation and shoot branching. The first of these, Kinetin, was discovered by Miller *et al.* (1955) who found that its application caused cells to enlarge and gain weight. This discovery was closely followed up by observations that senescence was delayed by its application to detached leaf samples (Richmond & Lang, 1957).

The most interesting demonstration of the ability of cytokinins to prevent senescence was demonstrated by Gan & Amasino (1995) where the promoter of *SAG12*, a gene known to be highly up-regulated during senescence, was fused with the *IPT* gene which catalyses the rate-limiting step in cytokinin biosynthesis. When senescence began in these plants, cytokinins were synthesised and senescence was halted. Recent studies of genome-wide changes in gene expression have identified that, during senescence, cytokinin synthesis genes (*AtIPT3*) are down-regulated and cytokinin degradation genes such as cytokinin oxidase are up-regulated (Buchanan-Wollaston *et al.*, 2005) further strengthening the association of cytokinin pathways as playing a part in the trigger of senescence.

The precise mechanisms of cytokinins in altering leaf senescence are unknown, but through the observation of a gain-of-function mutant caused by a missense mutation of AHK3 and also a loss-of-function mutant in *AHK3*, it has been found that AHK3 plays a major role in the phosphorylation of ARR2, an *Arabidopsis* response regulator, which in turn leads to delayed leaf senescence (Kim *et al.*, 2006).

Auxin is well established as playing a central role in the formation of plant organs, found in concentration gradients at the tips of emerging primordia (Benková *et al.*, 2003) and also found to be responsible for sculpting intricate shapes in flowers by promoting cell division and expansion in an ordered manner (Aloni *et al.*, 2006). Indole-3-acetic acid (IAA), the compound described as auxin in plants, is found to repress the transcription of SAGs such as TAPG1 (Hong *et al.*, 2000) and SAG12 (Noh & Amasino, 1999) but at the same time, IAA levels are increased during the senescence process (Quirino *et al.*, 1999). This appears to demonstrate that IAA is a controlling factor causing a negative feedback loop during the senescence process.

In the study by van der Graaff *et al.* (2006), six of the thirteen transport proteins suspected to be involved in auxin transport were massively down-regulated during senescence and this is suspected to play a part in the theory of auxin gradients across the leaf causing the senescence response (Addicott *et al.*, 1955). It was found that, where the auxin levels in a leaf fall in comparison to the stalk's auxin levels, abscission of the leaf occurs in beans and cotton. If this also applies to *Arabidopsis*, then a reduction in transport of auxins may help to maintain auxin gradients.

Ethylene was first identified in 1917 as being the biologically active component of illuminating gas which caused trees and plants to shed their foliage upon exposure (Burg, 1968). Ethylene gas is a simple molecule, but exposure to it can be a cause of seed germination, seedling growth, fruit ripening, flower senescence and abscission as well as leaf senescence (Grbić & Bleecker, 1995). However, ethylene has been seen to be incapable of causing senescence in young leaves and can only catalyse the senescence process once it has begun (Jing *et al.*, 2005). However, its role is not insignificant since a mutation of the *ethylene-insensitive 2 (ein2)* gene which renders the plant deficient in ethylene signal transduction resulted in delayed leaf senescence indicating that ethylene does play a part in the developmental senescence response (Alonso *et al.*, 1999).

The mechanism by which ethylene causes enhanced rates of senescence appears to be closely linked to the function of *onset of leaf death (OLD)* genes since a mutation of *old1* allows senescence to begin sooner and be yet further accelerated by ethylene exposure. This suggests that the presence of OLD1 is inhibiting the integration of ethylene. These conclusions, amongst others, have been used to produce a proposed model for the actions of ethylene in the onset of senescence (Jing *et al.*, 2002).

Jasmonic acid (JA) is the precursor of methyl jasmonate (MeJA), both of which were linked with the promotion of leaf senescence (Ueda & Kato, 1980). Other plant

defence responses associated with JA and its related compounds include microbial pathogens (Vijayan *et al.*, 1998; Xie *et al.*, 1998), herbivores and damage by UV-B and UV-C (Schaller, 2001; Berger, 2002). MeJA applied exogenously to *Arabidopsis* leaves has been shown to result in the rapid loss of chlorophyll and a decrease in photosynthesis-related activities (Jung, 2004), giving a senescence-like response. Furthermore, treatment with JA or MeJA leads to the induced expression of a number of SAG genes including *SEN1*, *SEN4*, *SEN5*, *rVPE*, *SAG12*, *SAG14* and *SAG15* (Park *et al.*, 1998; Kinoshita *et al.*, 1999; Schenk *et al.*, 2000).

An interesting study of a nuclear-localised CCCH-type zinc finger, OsDOS, demonstrates an interesting role for the protein as an integrator of the JA and MeJA pathways in the onset of senescence (Kong *et al.*, 2006). When *OsDOS* is over-expressed, senescence is repressed and when it is reduced in expression, senescence is enhanced, but genome-wide expression analysis demonstrated that the JA signalling genes are those which are most affected by the altered expression. This implies that OsDOS is not only a negative regulator of senescence, but also possibly an integrator of the JA pathway as a developmental cue to senescence.

Absciscic acid (ABA) has functions which have been closely linked with drought tolerance (Zhang *et al.*, 1987) as well as being shown to cause abscission of leaves in some species. Additionally, increased levels of ABA have been identified during the onset of senescence (Gepstein & Thimann, 1980). It is thought that ABA is key to the senescence response of a number of plant stresses. This is demonstrated by the exogenous application of ABA which induces several known senescence associated genes (*SAGs*), including *SAG12*, *SAG13*, *SAG17*, *ERD1* and *RD21* (Weaver *et al.*, 1998). Drought, high salt and low temperature stresses are all seen to be concurrent with increased levels of ABA which may indicate that ABA is the cause of the senescence response during those stresses.

ABA biosynthesis gene, 9-*cis*-Epoxy-carotenoid dioxygenase (*NECD*) has been shown to be increased in expression during leaf senescence alongside aldehyde oxidase genes *AAO1*, *AAO3* and *AAO4* which are also involved in the biosynthesis of ABA (Seo *et al.*, 2000; Buchanan-Wollaston *et al.*, 2005; van der Graaff *et al.*, 2006). Additionally, increased expression of *ABI1* and *ABI2* can be seen during senescence which have been identified as forming a negative feedback loop of the signalling pathways of ABA (Merlot *et al.*, 2001). These results demonstrate the highly complex relationship between ABA and senescence.

Interestingly ABA has been linked with both the generation of reactive oxygen species such as hydrogen peroxide (Hung & Kao, 2004) and the protection against them through increased expression of superoxide dismutase (SOD), ascorbate peroxidase (APOD) and catalase (CAT) (Hung & Kao, 2003). The imbalance between these two processes is presumably the eventual cause of the onset of ABA-induced senescence and is likely to be driven by other age-related factors since ABA is unable to induce a

senescence response in younger leaves (Lee *et al.*, 2011).

Salicylic acid (SA) has been identified as a key signalling pathway in the response of plants to pathogen invasion (Gaffney *et al.*, 1993), UV-B oxidation (Surplus *et al.*, 1998) and exposure to ozone (Rao & Davis, 1999). However, Morris *et al.* (2000) have also established a number of important characteristics of SA in developmental senescence. Whilst observing mutants defective in the biosynthesis of SA (*NahG* transgenic plants (Gaffney *et al.*, 1993)) and those with lesions in the SA signalling pathway (*npr1* (Cao *et al.*, 1994) and *pad4* (Glazebrook *et al.*, 1997)), it was noted that the expression of a number of SAGs, including *chitinase* and *SAG12* are considerably reduced and that the expression of these therefore relies on the SA pathway. Additional findings were that the *NahG* transgenic plants showed delayed developmental senescence, but that artificially-induced senescence timings were unaffected by the mutations, implying that SA pathways are developmental senescence specific.

1.2.2 The Senescence Response

Although the response of specific genes has been shown to be different depending on the induction method of senescence (e.g. developmental senescence, biotic stress, abiotic stress) (Buchanan-Wollaston *et al.*, 2005), the processes which constitute the senescence response are similar and well established.

Chlorophyll degradation. Chlorophyll catabolism is the primary cause of yellowing in senescing leaves as pigments are degraded into their amino acids for transport out of the leaf. A review of the pathways of chlorophyll degradation during senescence was produced by Hörtensteiner & Feller (2002). During the catabolism of chlorophyll, pheophorbide *a* is cleaved by pheophorbide *a* oxygenase (PaO) to produce red chlorophyll catabolite (RCC) (Hörtensteiner *et al.*, 1998). This step appears to be rate controlling during senescence as the activity of PaO is dramatically increased at this time whilst the expression of other genes in the chlorophyll catabolism pathway remain constant before and during senescence (Takamiya *et al.*, 2000, Reviewed). RCC is reduced by RCC reductase (RCCR) (Rodoni *et al.*, 1997), leading to the production of non-fluorescent chlorophyll catabolites (NCCs) which are disposed of in the tonoplast without first recycling their nitrogen content (Hinder *et al.*, 1996; Tommasini *et al.*, 1998). The energy expensive process of chlorophyll catabolism is therefore not a method of recovery of nutrients from chlorophyll, but instead a method of disposal for the toxic chlorophyll, which is otherwise highly reactive, to maintain viability of leaf cells during senescence. This is demonstrated by a defective *accelerated cell death 2* (*acd2*) gene, whose product is RCCR and results in the accumulation of RCC causing rapid cell death (Mach *et al.*, 2001).

Protein degradation. Leaf protein, up to 75% of which can be found in the chloroplasts, is degraded and remobilised during senescence. However, the action by which this occurs is still only theoretical and based around a number of observations about protease accumulation and localisation during senescence. Although a number of protease genes are seen to be induced during senescence, these appear to be mostly localised to the tonoplast and therefore cannot act on the proteins of the chloroplasts until membranes are later disrupted (Buchanan-Wollaston *et al.*, 2003, Reviewed). One thylakoid protein, LHCP II, appears to be stabilised in the *acd2* mutant which constitutes a lesion to the chlorophyll degradation pathways, despite other proteins being degraded similarly to the wild type (Gay *et al.*, 2008). This indicates that LHCP II, which forms a complex with chlorophyll, is protected from degradation until it is separated from the protein complex, after which it can be degraded by chloroplast proteases. This may, then, indicate that it is prior processes of senescence that control the timing of protein degradation rather than the enhanced expression of proteases in the chloroplasts.

Yamada *et al.* (2001) describe findings of a cysteine protease (RD21) with a granulin domain which is found to accumulate in the tonoplast of cells prior to and during senescence. At the onset of senescence, the granulin domain is slowly degraded until the protease becomes soluble and is then able to disperse throughout the cell. Ubiquitin-mediated degradation of specific proteins has been linked to senescence by the findings of Woo *et al.* (2001) in which an *ore9* mutant of *Arabidopsis* was found to exhibit delayed senescence. Their analysis of ORE9 (since renamed MAX2) revealed that it contains an F-box domain which are known to interact in SCF complexes (Skowyra *et al.*, 1997) resulting in selective ubiquitination of proteins destined for proteolysis.

Lipid degradation. Membrane lipids provide a plentiful source of energy in the latter stages of senescence. A number of genes encoding enzymes involved in lipid decomposition are enhanced during senescence including phospholipase D (Ryu & Wang, 1995; Thompson *et al.*, 1998), phosphatic acid phosphatase (Ryu & Wang, 1995), lytic acyl hydrolase (He & Gan, 2002) and lipoxygenase (Thompson *et al.*, 1998). Acetyl CoA produced during fatty acid breakdown can either be used for respiration by the plant or otherwise used to generate sugars via the glyoxylate pathway for export from the leaf. *SFP1*, a monosaccharide transporter, is up-regulated during senescence (Quirino *et al.*, 2001), demonstrating that the latter is probably occurring.

Mobilisation of Nutrients. Nutrients are remobilised and removed from senescing leaves for relocation to developing sinks in the plant. A study of nutrient remobilization in senescing *Arabidopsis* leaves (Himmelblau & Amasino, 2001) found that compounds such as zinc, iron, copper, sulphur, chromium and molybdenum were reduced in senescent leaves by over 50% when compared with green leaves. Even more valuable nutrients, nitrogen, potassium and phosphorus were reduced by over 80%. A valuable source of phosphorus comes from the degradation of RNA and DNA by nucleases en-

hanced in expression during senescence such as *RNS2* (Taylor *et al.*, 1993) and *BFN1* (Pérez-Amador *et al.*, 2000). Kamachi *et al.* (1992) reported increased levels of *cytosolic glutamine synthetase (GS1)* during senescence of *Oriza sativa* plants which was also observed by Finnemann & Schjoerring (2000) in *Brassica napus* leaves and phloem. It is thought that this plays a role in the conversion of amino acids to glutamine to enhance nitrogen transport out of the leaves.

1.3 A Systems Biology Approach to Finding a Senescence Transcriptional Network

It is the similarities between developmental and stress-related senescence and the large-scale gene regulation involved in both these cases that make senescence an interesting area of scientific research. However, the economic benefits that could be gained by controlling senescence in commercially valuable crops mean that senescence is also of public interest. However, the likely scale and complexity of transcriptional networks involved in senescence means that they will only be truly understood when the whole biological system is considered at once.

One complication arising from the study of developmental senescence is that it is a natural process rather than a response to an imposed stimulus and so no true mock response can exist. This means that expression of genes which are thought to cause senescence can only be compared with the expression of the same gene at a time before senescence is thought to have occurred. However, with so many genes having pleiotropic effects throughout plant development and senescence having many stages of progress, this may not be a reliable comparison unless it is made at a number of points throughout plant development. Additionally, He *et al.* (2001) have demonstrated that the cross-talk between various forms of stress applied to *Arabidopsis* resulting in leaf senescence is extensive. The confounding effects of so many potential pathways to senescence means that it is nearly impossible to completely isolate individual networks of genes and that only by considering the interconnected networks of all the pathways will the whole system be fully understood.

Past studies have focussed on finding genes with key influence over the senescence response by using either a ‘forward genetic approach’ in which mutants are randomly screened to identify those with altered phenotypes or by using a ‘reverse genetic approach’ in which mutants are produced for genes with known altered expression during senescence to find out whether they demonstrate an altered phenotype. Mutants demonstrating altered leaf senescence are, in the majority of cases, identified by the ‘forward genetic approach’ with some examples being the discoveries that:

- Ethylene insensitive mutant *etr1-1* exhibits delayed leaf senescence (Grbić & Bleecker, 1995).
- Cytokinin insensitive mutant *ahk3* exhibits advanced leaf senescence (Kim *et al.*,

2006).

- Reduced expression mutant of *ore9/max2* exhibits delayed leaf senescence by failing to degrade a repressor of senescence (Woo *et al.*, 2001); and
- *ore4* mutant also delays senescence by reducing photosynthetic activity (Woo *et al.*, 2002).

Far fewer examples can be identified for the ‘reverse genetic approach’ though notable examples include the study of WRKY53, a zinc-finger-type transcription factor, which was thought to play a regulatory role in several senescence associated genes and was then demonstrated to accelerate senescence when over-expressed and delay it when knocked-out (Miao *et al.*, 2004). Study of WRKY53 has since shown it may be a crosstalk element between the response of JA and SA (Miao & Zentgraf, 2007). Further application of the ‘reverse genetic approach’ has lead to the discovery of stress-related transcription factors which can influence the expression of *WRKY53*, including GATA4 which is induced by darkness (Zentgraf *et al.*, 2010). AtNAP is another transcription factor for which mutants were produced because of its enhanced expression during the senescence response (Guo & Gan, 2006) and it was found that knockout mutants exhibit delayed senescence. Breeze *et al.* (2008) used a GeneChip experiment to identify over 800 genes with significantly increased expression during leaf senescence and isolated 2 transcription factors, HSF1 and SAP12, to study knockout mutants and find downstream effects. The results of this analysis were inconclusive since the downstream effects were vast and no direct interactions could be inferred. The authors describe a systems biology approach to identifying gene networks.

By observing the expression levels of all genes in senescing leaves over the entire mature developmental process using microarrays, a high-resolution temporal dataset would be collected which may be used to produce theoretical models of gene regulation during senescence. By considering the expression of many genes changing in expression during senescence, pleiotropic effects may be separated from one another and cross talk between different stress pathways may be defined. The results of such putative models would then be able to guide the biological research to the most likely solutions, instead of blanket screening without cause and trying to identify singular interactions between genes as in the non-systems biology approach. Biologically validated interactions could then be used to refine the prior assumptions of the theoretical model and predict further interactions on a system-wide scale.

1.4 Gene Expression Data during Natural Senescence

Prior to this PhD project, a microarray experiment was performed to provide high-resolution temporal gene expression data from leaf samples undergoing natural senescence with no imposed external stresses. The data obtained by scanning the microarrays

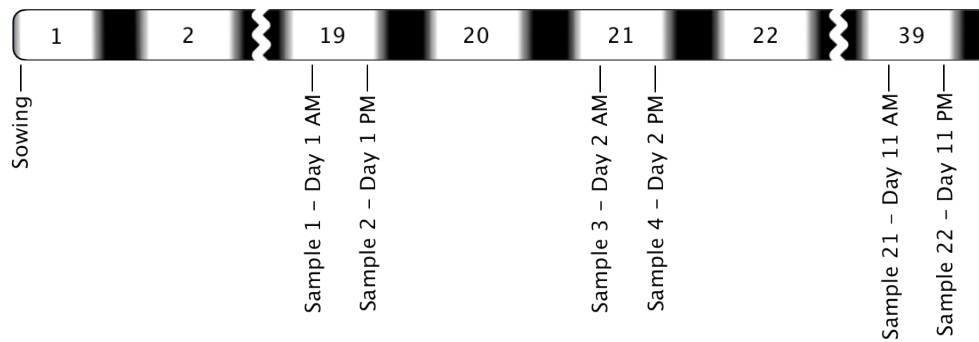


Figure 1.2 – The harvesting schedule of the natural senescence microarray experiment

The schedule for harvesting samples is shown as a timeline. The white areas of the timeline are 16 hour periods of daylight whilst the black areas are 8 hour periods of darkness. Numbers shown within the timeline represent the number of days since sowing. Whilst sowing occurred on day 1, the sample labelled as Day 1 AM was harvested 7 hours into the light cycle 19 days after sowing. The second sample, labelled Day 1 PM was then collected on the same day, 14 hours into the light cycle. The 20th day after sowing was skipped, as was every day that came after a harvesting day. On the 21st day after sowing, samples 3 and 4, labelled Day 2 AM and Day 2 PM respectively, were harvested and this regular schedule was finally completed on the 39th day after sowing when samples 21 and 22 were collected on the morning and afternoon of that day.

of this experiment were normalised and analysed during this PhD and then used to infer possible transcriptional networks between selected genes. A full description of the experiment and associated analysis methods, parts of which were performed by me and therefore also shown in this thesis, can be found in Breeze *et al.* (2011). Specific details about the experiment which are important to the interpretation of results in this thesis are also shown in this section.

Columbia (Col-0) *Arabidopsis* plants were stratified to ensure synchronised and successful germination after sowing. The plants were grown under controlled conditions with a 16 hour photoperiod until 19 days after sowing (DAS) when leaf 7, the 7th full leaf to emerge, had reached approximately half of its maximum size. From day 19 onwards, leaf 7 samples were harvested from randomly selected individuals and immediately frozen in liquid nitrogen so as to avoid wounding response gene expression changes which can be detectably induced in as few as 2 minutes after detachment (Glauser *et al.*, 2008). The choice to harvest leaf 7 was based on the senescence and remobilisation of nutrients from this leaf occurring at the same time as flowering and silique filling. The schedule for harvesting involved collecting samples every other day at 7 and 14 hours (hereafter referred to as morning and afternoon respectively) into the light cycle until a total of 22 sampling occasions had occurred. A visual representation of this harvesting schedule can be found in Figure 1.2.

RNA isolation was performed on four biological replicates collected from independent plants at each time point, providing 88 samples. Those four biological replicates were arbitrarily labelled A, B, C and D, although this does not insinuate association

between similarly labelled replicates at alternative time points. These were amplified using an Ambion MessageAmp II kit. The resulting aRNA was then used as a template to synthesise cDNA labelled with cytosine bound Cy3 and Cy5 fluorescent dyes before being hybridised to 176 2-channel CATMA microarrays in an experimental design intended to minimise experimental biases.

1.4.1 CATMA Microarrays

Microarrays were first introduced by Schena *et al.* (1995), closely followed by two-colour hybridisations (Shalon *et al.*, 1996) so that two samples could be compared directly on the same microarray. By providing a genome-scale quantification of specific mRNA abundance in a sample at an affordable cost, microarrays have revolutionised the field of genomics and are still widely used 16 years later.

Two microarray production methods exist today, robotic spotting and in-situ synthesis (Stekel, 2003). Companies such as Affymetrix have commercialised the production of microarrays by using in-situ synthesis to produce highly reproducible single-channel microarrays based on oligonucleotides synthesised in-situ on a suitable surface using a sequence of light masks that allow the customisation of each feature of the array. Microarray experiments performed prior to and during this PhD, however, use robotically spotted microarrays. In this case, a library of gene-specific sequence tags were synthesised by highly parallel PCR methods. These tags are based on the CATMA (Complete *Arabidopsis* Transcriptome MicroArray) library (Crowe *et al.*, 2003; Hilson *et al.*, 2004; Sclep *et al.*, 2007).

CATMA version 3 microarrays, which were used to analyse the 88 samples in the natural senescence experiment, consist of 31,144 PCR amplicons spotted and immobilised onto the surface of glass slides using a robotic bank of 48 pins arranged as 12 rows and 4 columns. Each pin is responsible for placing a grid of 26×26 probes onto the slide, resulting in 48 pin-tip groups consisting of 676 probes each. Each probe has a sequence length of between 150 and 500 bp and has been designed to provide no more than 70% identity to any other part of the *Arabidopsis* genome (Crowe *et al.*, 2003).

In comparisons using biological samples spiked with calibrated quantities of in vitro synthesised poly(A) RNAs, it was found that CATMA microarrays perform equally as well as Agilent or Affymetrix microarrays in measurements of sensitivity, specificity and the prevention of detecting false positives, but also superseded them in reducing signal saturation for high target concentrations (Allemeersch *et al.*, 2005). This, as well as the much lower expense of CATMA microarrays, makes them an excellent choice for such a high-resolution time course experiment.

1.4.2 Experimental Design

In order to make best use of the microarrays, it was essential to produce an experimental design based around specific principles which avoid bias and ensure connectivity throughout the samples (Mead, in preparation). Those principles were:

- Each sample should be labelled equally often with both dyes.
- Each sample should be present an equal number of times throughout the experiment (technical replicates).
- Each day sample should be co-hybridised in a consistent manner to prevent inconsistent variation between microarrays (e.g. with the day before and the day after).
- Each time of day within each day should be compared equally frequently with all times of day (i.e. $AM \rightarrow AM = AM \rightarrow PM = PM \rightarrow AM = PM \rightarrow PM$)
- Each biological replicate should be compared equally frequently with all biological replicates to improve the consideration of between biological replicate variability. (i.e. $A \rightarrow A = A \rightarrow B = A \rightarrow C = A \rightarrow D = B \rightarrow A = \dots D \rightarrow D$)

The design of the experiment was complicated given the vast number of time points and provides a high degree of connectivity between samples. Whilst one half of the microarrays form eight separate loops within arbitrary biological replicate labels, the other half form a single loop of all samples, connecting across biological replicates. This allows a good consideration of the between biological replicate variability whilst also maintaining smaller loops which add strength to comparisons between pairs of samples.

The design is best described in two diagrams because this allows the connectivity of the design to be more easily interpreted. Figure 1.3 illustrates the co-hybridisation of ‘same time of day’ samples. The direction of the arrows indicates the dye used to label each sample with Cy3 dyes at the tail and Cy5 at the head of each arrow. Morning samples are represented by green lozenges, whilst afternoon samples are represented in yellow. As each row represents the days on which harvesting was performed, it can be seen that comparisons only exist between adjacent days with the loops completed by the co-hybridisation of samples from days 11 and 1. In total, eight loops of eleven comparisons exist within this diagram, four in each of morning and afternoon samples. These smaller loops within the design provide short pathways between samples collected at the same time of day.

The other half of the experimental design is illustrated in Figure 1.4 where the co-hybridisation of ‘between time of day’ samples can be seen. Whilst each row of the figure represents the biological replicates of a particular time point, adjacent rows are of alternate times of day. A consistent comparison pattern exists in the design, for example morning samples of biological replicate A are always hybridised to afternoon samples of biological replicate B. This does not allow every replicate comparison (i.e. $A \rightarrow B, B \rightarrow C$) to be made directly but does ensure that a single loop exists through all 88 samples, providing a connected design. This large loop combined with that of the eight separate loops in Figure 1.3 provides connectivity between all samples to shorten the comparisons between pairs of samples.

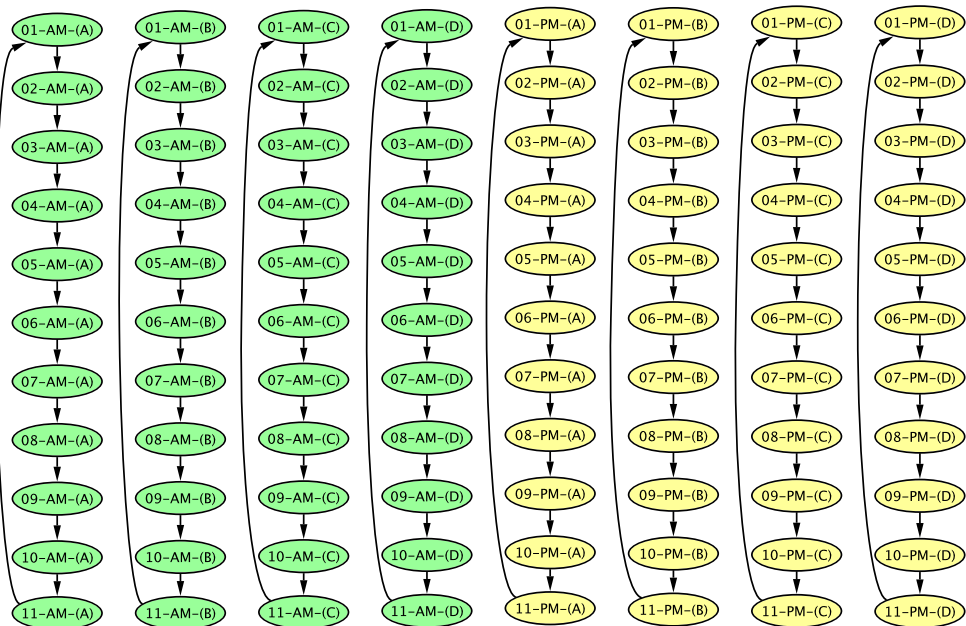


Figure 1.3 – The loop design for microarrays hybridised with samples of the same time of day

The first of two loop designs which form the experimental design compares samples collected on subsequent days but at the same time of day only. Samples bound by Cy3 (Cy5) dyes lie at the tail (head) of each arrow. For clarity, morning samples are shown in green whilst evening samples are shown in yellow. The two digits shown at the beginning of each lozenge indicates the day on which the collection was made (from 1 to 11). A total of eight loops exist containing eleven comparisons each, accounting for half of the total microarrays in the experiment. These smaller loops within the experiment provide short comparisons between morning or afternoon samples.

1.4.3 Microarray Scanning

After hybridisation, the microarrays were scanned using an Affymetrix 428 Array Scanner. Cy3 labelled samples were scanned with a laser of wavelength 532nm whilst Cy5 labelled samples were scanned independently with a laser of wavelength 635nm. Each microarray was scanned at a gain to ensure that the brightness of the spots was maximised whilst minimising saturation as much as possible. Scans were saved as two 16-bit TIFF images per microarray providing $2^{16} = 65,536$ levels of intensity for each channel.

1.4.4 Scan Analysis

In order to analyse the microarray scans and convert them to numeric data describing the intensity of the channels and their background noise, BioDiscovery ImaGene version 7.0 (BioDiscovery) was used. By providing a grid template to identify the probes by Gene ID and location, ImaGene is capable of outputting a table of values in the range of 1 to 65,536 describing, amongst other measurements, the mean signal, mean background, median signal, median background and flag of each channel of each probe of the microarray. Where the probe fails quality measures imposed by ImaGene, usually due to a low signal, the flag is returned as an integer greater than zero to assist in their

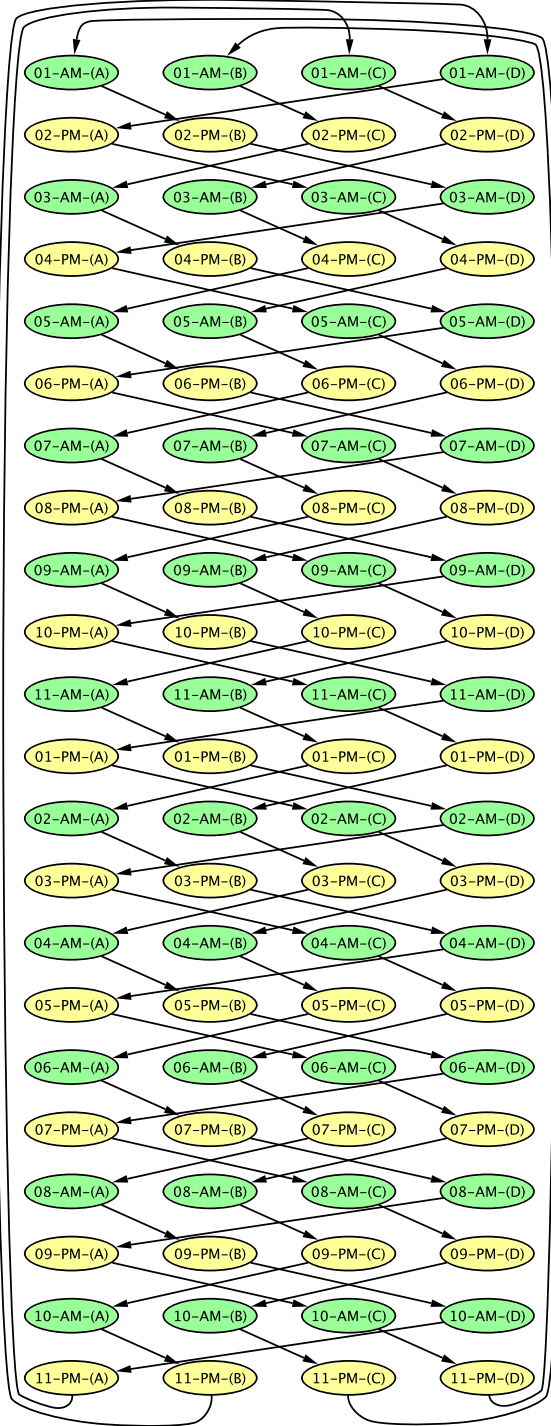


Figure 1.4 – The loop design for microarrays hybridised with samples of different times of day

The second of two loop designs, forming half of the overall experimental design, compares samples collected on subsequent days but at different times of day only. Samples bound by Cy3 (Cy5) dyes lie at the tail (head) of each arrow. For clarity, morning samples are shown in green whilst evening samples are shown in yellow. The two digits shown at the beginning of each lozenge indicates the day on which the collection was made (from 1 to 11). A single loop of 88 comparisons exists, accounting for half of the total microarrays in the experiment.

identification.

Background subtraction was performed on the data to reduce the impact of background signal on the intensity of features on the microarray (Stekel, 2003). This enables the natural fluorescence of the glass and any non-specific hybridisation of labelled DNA to the microarray to be removed. This was performed by subtracting the background median from the signal median for each probe. A typical treatment of microarray data is to log transform the signal, which is mathematically intangible if the signal is less than or equal to zero. For this reason, during background subtraction, those probes where the background median is greater than the signal median were given a value of 1, the lowest possible intensity above 0.

1.5 Microarray Data Analysis Software (MAANOVA)

MicroArray ANalysis Of VAriance (MAANOVA) (Wu *et al.*, 2003) is a free and open source package for the analysis of microarray experiments developed by the Churchill group at the Jackson Laboratory, Maine, USA. The package provides functionality for performing data quality control, data transformation, mixed model fitting, F-test based identification of significantly differentially expressed genes and extraction of estimated gene expression data. The version focussed on in this thesis is that written to operate under the R project (R Development Core Team, 2010). However, MAANOVA was originally written for MATLAB (MATLAB, 2010) but this has been deprecated. A Java graphical user interface is also available to complement the R version of MAANOVA which otherwise requires users to enter text commands to operate the software.

1.5.1 Purpose and Justification

Traditionally, two-channel spotted microarrays are considered to be of a lower quality and reproducibility than some other microarray types such as in-situ synthesised oligonucleotide Affymetrix GeneChips (Park *et al.*, 2004). This is thought to be in part due to the spotting process providing microarrays with differently sized spots (Brody *et al.*, 2002; Jenssen *et al.*, 2002; Tran *et al.*, 2002). However, correct processing of the raw data will result in artefacts being minimised by quality control and repetition of unreliable data. Systematic artefacts of two-channel spotted microarrays can be minimised by transformation of the data to avoid issues such as the incorrect identification of differentially expressed genes at lower intensities because of high ratios caused by background noise (Park *et al.*, 2004).

Although CATMA microarrays specifically have been shown to be of a high quality when compared to technologies like the aforementioned Affymetrix GeneChips (Allemeersch *et al.*, 2005), they are still not without flaw and suffer from problems such as dye-biased hybridisation, spot size variability, non-specific binding and occasionally high background-fluorescence. These problems are typical of two-channel spotted microarrays and inherent to the process of manufacture, labelling efficiency of samples

with different dyes and handling of the microarrays during and after hybridisation. One method of normalising for these sources of variability is to balance their effects across the experiment using a well designed hybridisation strategy. Several different experimental designs have previously been proposed such as the reference design and loop design (Kerr & Churchill, 2001) although each of these are not easily scalable to the 88 samples of the senescence experiment being analysed and so the design in Section 1.4.2 was developed by Andrew Mead of Warwick HRI to satisfy these requirements. The difficulty comes in providing an analysis for these comparisons in which some pairs of samples are not directly compared on the same microarray. Additionally and specifically with the senescence dataset, samples were collected at two times of day, allowing diurnal expression patterns of some genes to be identified when comparing adjacent time points, despite the primary objective being to find genes which are differentially expressed as a cause of or in response to senescence.

MAANOVA provides a solution to these problems. It not only incorporates quality analysis and data transformation steps prior to analysing the data, but it also fits a model to the data to identify which sources of variation exist and provides a decomposition of those such that the significance of their effect upon the expression of each gene can be independently tested. This modelling stage is actually two models with different scopes. The first, a fixed model, identifies the effects of dye and array across the entire experiment. By estimating these and retaining only the residuals, the resulting data is effectively normalised for these effects, discarding overall intensity differences between each dye and across all arrays. The second, a mixed model, further decomposes the residual variability of the first model into gene-specific effects for dye and array, but also for treatment terms which are defined by the experimental design, allowing the separation of the effects of the day and the time of day the sample was collected at as well as identifying variability due to an interaction of those terms and due to the variability between biological replicates.

The second model is labelled as mixed because it contains some terms which are fixed and some which are random. Fixed terms can be defined as those which, upon a repetition of the experiment, would have a predictable effect on the expression of the genes, such as the time point of the sample. Random terms can be defined as those which, upon a repetition of the experiment, would have an unpredictable effect on the expression of the genes, such as the microarrays used for hybridisation (Churchill, 2004). Effectively, the underlying trend of the gene expression is entirely explainable if all the fixed terms were identified, whilst the random terms provide a level of noise to that trend.

1.5.2 Original MAANOVA Implementation

The original implementation of the MAANOVA package provides the functionality described above. These functions include three microarray quality control visualisations, data transformation via a number of user selectable methods, fitting of models to the

microarray data and identification of differentially expressed genes. So that the improvements made to MAANOVA during the course of this PhD can be more easily identified, and to give an overview of the whole package, this section presents the purpose of each original function and, where appropriate, an example of the output it would produce.

1.5.2.1 Loading Data

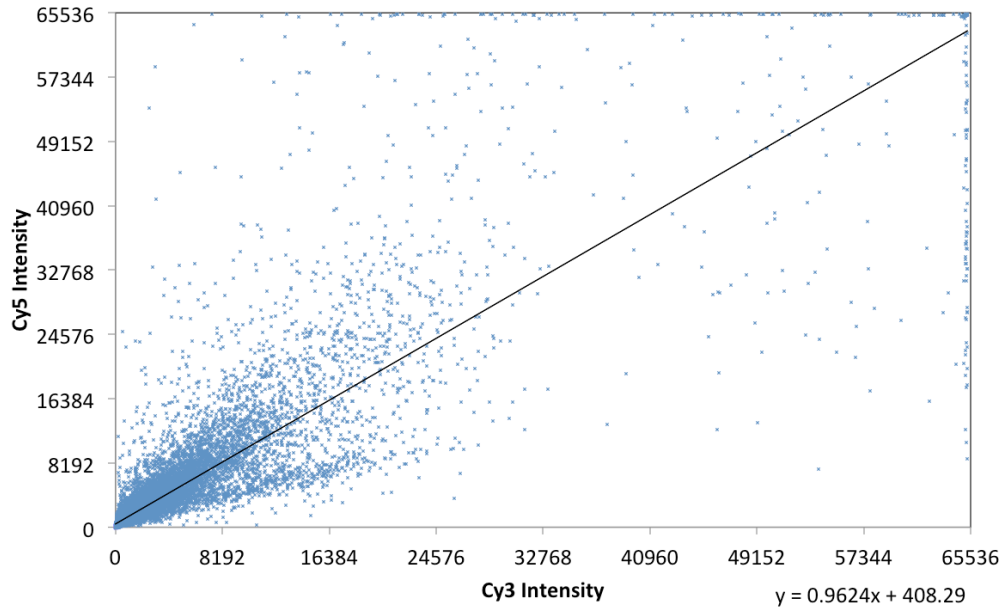
Data from the scan-analysis is loaded into MAANOVA as a single matrix stored in a tab-delimited text file. This file, with each probe occupying one line, contains information regarding the name of the probe, its position on the microarrays, the intensity of the fluorescence for each of the channels on each microarray and whether it was flagged during the scan-analysis. For each of the microarrays in the senescence experiment, three columns of information describe the fluorescence of Cy5, the fluorescence of Cy3 and the flag status. Fluorescence is measured from 1 to 65536 and the flag status is greater than zero where a problem was identified.

MAANOVA, by default, transforms all the expression data to a \log_2 scale on loading it. The purpose of this transformation is to provide a reasonable spread of features across the intensity range, to provide a constant variability at all intensity levels, to transform experimental errors into a normal distribution and to transform the intensity distribution to be approximately bell-shaped (Stekel, 2003). The added benefit of using a \log_2 scale is that a single unit of change on the \log_2 scale represents an example doubling or halving of the absolute gene expression which is easily comparable between genes. An example of the transformation achieved by applying a \log_2 of the expression data can be found in Figure 1.5.

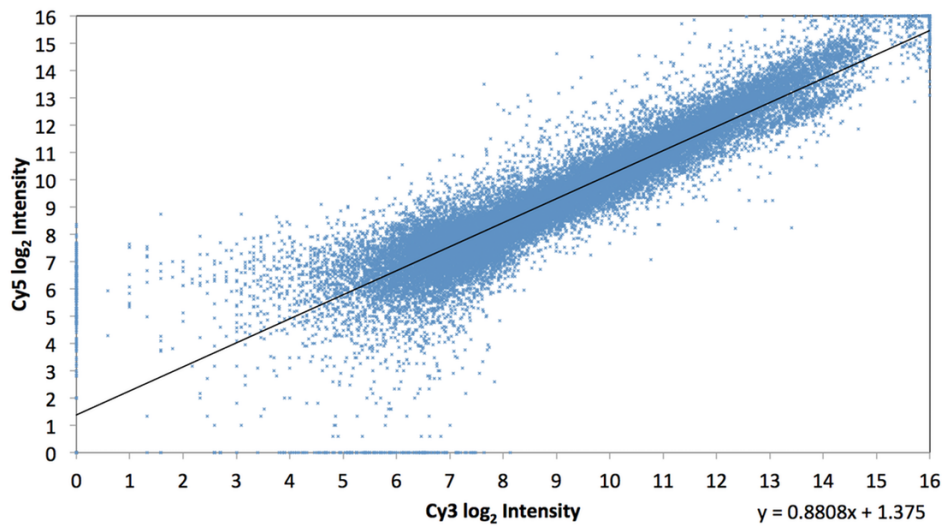
1.5.2.2 GridCheck

GridCheck provides a method of comparing \log_2 intensity levels between the two channels of each pin-tip group of each microarray by plotting them against each other in a scatterplot. For each CATMA microarray, 48 sub-plots are produced in the same 12 row and 4 column arrangement which is present in the sub-grids of the microarrays. These plots use colours to indicate the status of each probe as defined by the flag of each probe when analysed using ImaGene (BioDiscovery). Where the probe was not flagged, the point is plotted in blue, but where the probe is flagged for any reason, the point is plotted in red. This gives a visual indication of the quality of the pin-tip group, as those with high levels of background or large numbers of poorly printed spots will also have high numbers of flagged probe data.

The plots which are output from GridCheck are presented as a separate plotting window per microarray, displayed directly on the screen. In the case of the example microarray data provided with MAANOVA, this results in 16 sub-plots per window in a 4×4 arrangement. However, with 48 sub-plots required to display the CATMA microarray data, these are very cramped and an example of 4 of these sub-plots can be found in Figure 1.6(a). Additionally, no permanent storage of the plots is made



(a) Cy5 vs Cy3 Intensities before \log_2 Transformation



(b) Cy5 vs Cy3 Intensities after \log_2 Transformation

Figure 1.5 – \log_2 transformation of microarray data allows the data to meet a number of expectations

Transforming raw microarray data (a) by taking a log (base 2) of the intensities in both channels (b), allows the data to meet a number of expectations that will be required to allow further transformation to remove systematic bias introduced from experimental sources. The transformed data provides a reasonable spread of features across the intensity range and a reasonably constant variability at all intensity levels. It can be seen, from the trend line, that a small amount of bias still exists towards Cy5 at higher intensities, but this can usually be removed by further transformation steps.

available without laboriously saving the contents of each plotting window, still resulting in low resolution plots.

1.5.2.3 RIPlot

RIPlot is a method for producing scatterplots of the \log_2 ratio of the two channels for each probe of each microarray against the sum \log_2 intensity. This type of visualisation for two-channel microarrays allows for biases imposed by experimental factors to be more easily identified. This plot is essentially identical to those of GridCheck but with a 45° rotation in a clockwise direction providing the added visual benefit that the ratio is now one of the axes (Cui *et al.*, 2003). Where a bias causes a linear regression of the data to have a gradient other than zero, the human eye and brain can better determine this since they are better at processing horizontal lines than they are the diagonal lines of GridCheck (Stekel, 2003). Another characteristic of RIPlot scatterplots is that they include the probes of the entire microarray rather than those of specific pin-tip groups, allowing the identification of intensity specific dye biases. Colours are used within the plots to indicate the flag status of the probes being plotted. Where a probe has not been flagged by ImaGene (BioDiscovery), its point is plotted in blue. If a probe has been annotated with a flag, it is plotted using a red point.

The plots output by RIPlot are presented as a separate plotting window per microarray, displayed directly on the screen. This provides a lot of room for the plots, unlike the plots of GridCheck, but still results in many overlapping windows which can only be permanently stored by saving each plotting window by hand. An example of a plot produced by RIPlot can be found in Figure 1.6(b).

1.5.2.4 ArrayView

When analysing datasets such as those of microarray experiments, it can be frustrating to biologists to have converted optical images of microarray scans into numerical data, removing the visual aspect of the data. Yet, before normalisation and annotation provided by the analysis process, there is little information to be gained from the observation of the microarray scans directly. For this purpose, ArrayView produces grid-like false-colour heat maps of the ratios for each probe, maintaining their context by plotting them in the same positions as the probes of the original microarray. This provides a very intuitive interpretation of the ratios and can help to visualise the causes of specific artefacts seen in both GridCheck and RIPlot, before and after data transformation techniques have been used. This can help to identify the efficiency of transformations being applied and complements the plots produced by other functions of MAANOVA. Probes which are higher intensity in one channel are displayed as either a red or green dot dependent on the more intense channel. The colours used to indicate these ratios between the channels form a gradient from red through black and then green in a linear scale with black exactly half way between maximum red and maximum green.

Heat-maps produced by ArrayView are displayed in a separate plotting window per microarray, directly on the screen. This poses a problem for CATMA microarrays because they contain 312 rows of probes which means each probe has only 4 pixels of height on even a very good resolution screen. Additionally, the default dimensions of an R plotting window are approximately square despite the microarray having a 1:3 aspect ratio, resulting in each probe being represented by thin bars rather than more aptly shaped squares. Once correctly dimensioned, each window for each microarray requires saving by hand to permanently capture the heat-maps.

The brightest colours used to plot the heat-map of ArrayView are defined by the greatest ratio between the channels, which, on most occasions, can be found in the probes used to align the microarray grid, resulting in gene-specific probe ratios being plotted as artificially low as seen in Figure 1.6(d), making it impossible to identify outliers. Additionally, because the ratio being plotted as bright red is not the mathematical inverse of the ratio being plotted as bright green, black does not necessarily represent a ratio of one. As such, the plot can appear to demonstrate a dye-bias on those occasions when a ratio of one is represented by a shade of green or red, as can be seen in Figure 1.6(e).

1.5.2.5 Data Transformation

TransformMADData is a function provided for the purpose of normalising within-arrays. This is necessary to avoid biases from experimental sources introduced at the time of scanning (Stekel, 2003):

- Cy3 and Cy5 labels may be incorporated into the same DNA sequences in different abundances.
- Cy3 and Cy5 dyes may emit different response wavelengths dependent on their abundance.
- Cy3 and Cy5 emissions may be inconsistently measured by the scanner at different abundances.
- Cy3 and Cy5 may be inconsistently focussed if the microarray is not perfectly horizontal during scanning.

To remove these biases, TransformMADData uses regressions of the data to identify non-conformity with expected ideals and then transforms the data to coerce the data to conform with those ideals. In a perfect microarray dataset in which none of the sources of variance above can be found, a linear regression of ratio against intensity would provide an intercept of zero, a gradient of zero and the data would lie along a straight regression. A number of alternate methods are available to choose from and are presented by Cui *et al.* (2003) where the theory of each method and the intended application is explained:

- **Shift** - is a transformation applied to the raw intensity data prior to being \log_2 transformed so as to effectively move the origin of the RIPlot along the vertical (\log_2 ratio) axis and minimise the deviation of the mean \log_2 ratio from zero across all intensities. This is done by the simple addition of a constant to one channel whilst subtracting the same constant from the other channel:

$$\begin{cases} Z_{rk} = \log_2(Y_{rk} + C) \\ Z_{gk} = \log_2(Y_{gk} - C) \end{cases} \quad (1.1)$$

where C is the constant, Y_{rk} and Y_{gk} are the raw intensity values in the red and green channels of probe k , respectively. Z_{rk} and Z_{gk} are therefore the transformed intensity values for each of the channels of probe k .

This is appropriate where one channel has a higher intensity across all probes than the other channel, causing the linear regression of Y_r versus Y_g to have a slope ≈ 1 , but an intercept $\neq 0$.

- **Linear Log** - is a transformation in which the data is separated into a lower proportion of intensities which are to be transformed by an additive linear function and an upper proportion of intensities which are to be transformed by a multiplicative log function. This type of transformation is appropriate for data where the low intensity probes are not affected by an intensity dependent effect, but higher intensity ratios become biased towards one dye due to multiplicative effects that may be introduced by the first three experimental sources of variation shown at the beginning of this section. The transformation is defined by the following functions:

$$Z_{ik} = \begin{cases} \log_2(d_i) - \frac{1}{\ln 2} + \frac{Y_{ik}}{d_i \times \ln 2} & Y_{ik} < d_i \\ \log_2(Y_{ik}) & Y_{ik} \geq d_i \end{cases} \quad (1.2)$$

where the index i refers to either channel of the microarray, d_i is the threshold between the linear and the log transformation functions, Y_{ik} refers to the raw intensity of probe k in channel i and hence Z_{ik} is the transformed intensity value for probe k in channel i . Whilst d_i is calculated from the distribution of the intensity data, it often lies at a value which places 25-30% of the data below the threshold.

- **Linear Log Shift** - as its name suggests, is a combination of both the linear log method and the shift method above. The data is first processed by the shift method to minimise the deviation of the mean \log_2 ratio from zero, and then transformed by the linear log method.
- **Global LOWESS (glowess)** - is a curve fitting transformation which fits a local regression line to the \log_2 ratio of the probes via a locally weighted least squares

estimate which represents genes not differentially expressed. Locality used in the regression is based upon the \log_2 intensity of the probes across the entire microarray and therefore the regression is an estimate of the \log_2 ratio for all probes of similar \log_2 intensity. The LOWESS regression is, effectively, the fit of many linear regressions to the data over small subsets within that data which are then smoothed into a single curve. Locality defines the window of data over which the linear regressions should be performed and is defined by a span parameter, α . Increasing this parameter to 1 effectively provides a single linear regression fit to the data, whilst all values below 1 use subsets of the data to fit the curve for each data point and, as α tends to zero, the curve becomes an exact fit for the data and all ratios would tend towards zero in a LOWESS transformation. The span value provides a tricubic weighting for the adjacent \log_2 intensities and can be defined, for $\alpha < 1$, as:

$$\text{weighting} \propto \left(1 - \left(\frac{\text{dist}}{\alpha \times \text{maxdist}}\right)^3\right)^3 \quad (1.3)$$

where *dist* and *maxdist* refer to the numeric difference in the predictor variable (\log_2 intensity) and the range of that variable respectively. The fitted values are then used as a spot-specific constant to transform the channels of the microarray using the following functions:

$$\begin{cases} Z_{rk} = \log_2(Y_{rk}) + \frac{C_k}{2} \\ Z_{gk} = \log_2(Y_{gk}) - \frac{C_k}{2} \end{cases} \quad (1.4)$$

where C_k is the spot-specific constant obtained from the LOWESS regression, Y_{rk} and Y_{gk} are the raw intensity values of probe k in the red and green channels, respectively, and hence Z_{rk} and Z_{gk} are the transformed intensity of probe k in each channel.

This type of transformation is particularly well suited to microarray data because it is rarely obvious what types of variability exist in the data in order to choose from other transformations, whereas the LOWESS curve fit is driven by the data, causing greater fit where the data is most affected by variability. One criticism is that it is a very strong fitting method and can easily over-fit the curve and substantially reduce the significance of some spot ratios if the span parameter is set too low. As a conservative estimate, 0.1 is the default value for span and should only be decreased if it is the opinion of the user that the data has not been sufficiently transformed to meet the expectations given earlier in this section.

- **Joint or Regional LOWESS (rLOWESS)** - is also a curve fitting transformation based around the same theory as shown for the intensity-based LOWESS, except that locality is defined by the combined predictor variables intensity, spot-row

and spot-column rather than just the intensity. This has the effect of putting constraints on the LOWESS to give priority to spots in the same, or nearby, rows and columns of the microarray when establishing the curve fit and hence providing spatial awareness to the regression. This type of transformation isolates and rectifies the problem caused by the fourth experimental source of variation described at the beginning of this section.

An example of the plots produced by the Joint LOWESS method can be seen in Figure 1.6(c) and these are typical of all the transformation methods. For each microarray, a separate plotting window is presented on screen containing an RIPlot before transformation above, and an RIPlot after transformation below. The red line in the upper plot of Figure 1.6(c) is specific to the LOWESS curve fitting methods and represents the fitted curve. The curve crosses over any probes which are then transformed to a log ratio of zero, effectively defining those genes which are to be considered non-differentially expressed. Once the transformation has occurred, this red line would lie along the length of the horizontal axis at a log ratio of zero.

1.5.2.6 Fitting a Model to the Data

MAANOVA provides a function, `fitmaanova`, which fits a model of gene expression to the microarray data which have just been quality controlled and transformed to minimise systematic artefacts of the microarrays. This function accepts the transformed data object as an input along with a formula describing the model to fit. The formula is composed of experimental terms specified in the design as factors contributing to gene expression. In other words, the most simple formula is “`~ Array + Dye + Sample`” which would indicate that expression of a particular probe on a particular array for a specified sample labelled with a particular dye can be directly identified by the effect contributed by the **Array** used for hybridisation, the effect contributed by the **Dye** used for labelling and the effects contributed by the **Sample** hybridised to that array and labelled with that dye. The result returned from this is an object which shows all the fitted parameters of the model describing the transformed data for each gene probe on the microarrays. Any additional variability not captured by the model terms is assigned to an error term, which might account for unobserved factors such as small differences in laboratory technique.

Whilst **Array** and **Dye** are essential components of the formula, **Sample** can be further partitioned into characteristics of each sample, such as the time at which it was collected and/or the treatment it received. In this case, the model returned provides the effects of each characteristic on gene expression as model parameters, allowing the separation of otherwise complex interactions of the model terms. A typical example formula in this case may be “`~ Array + Dye + Time * Treatment`” whereby it might be anticipated that the **Time** of sample collection will have an effect on gene expression,

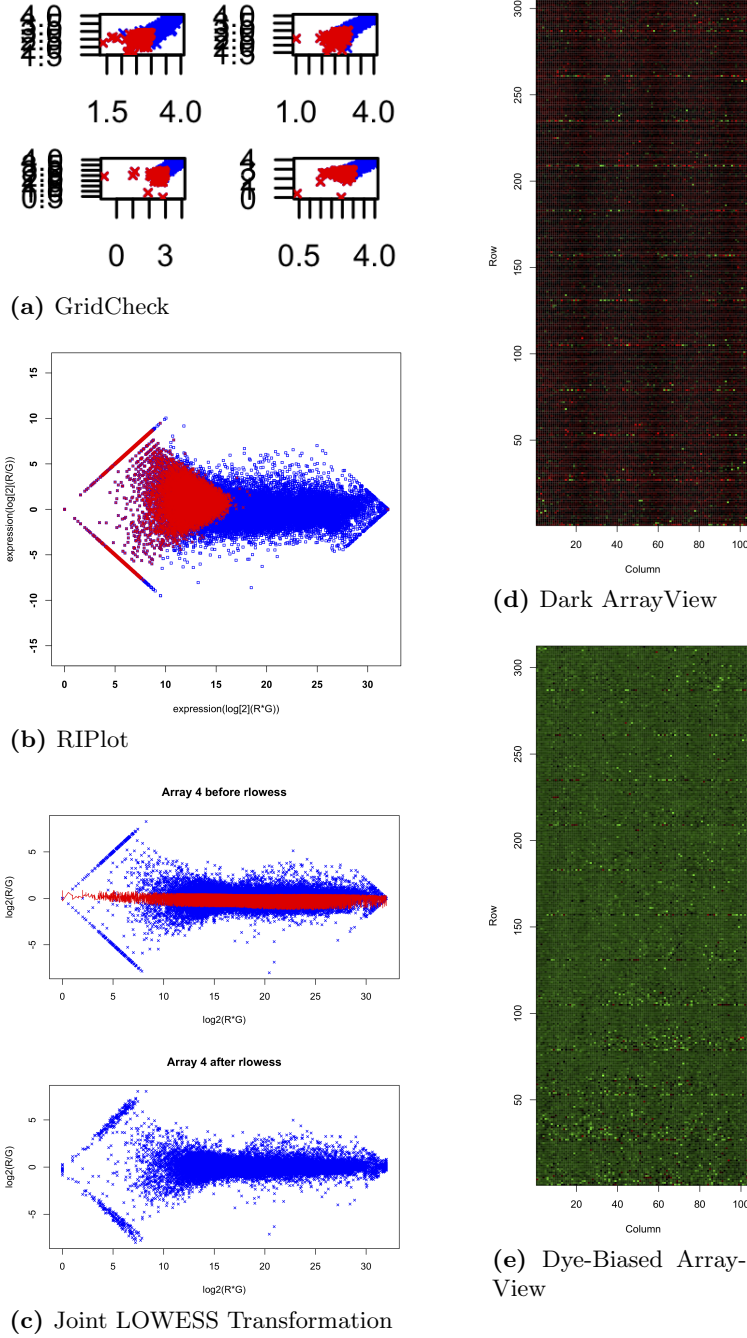


Figure 1.6 – Examples of graphical output produced by functions of MAANOVA

GridCheck, shown in (a), produces scatterplots of \log_2 intensities between the two-channels of the microarrays. Where each microarray has a large number of sub-grids, the plots are almost too small to read, as shown. RIPlot, shown in (b), produces a scatterplot of ratio vs. sum intensity of the channels of the probes of each microarray. All probes are plotted in blue whilst flagged probes are overlaid in red. ArrayView, shown in (d) and (e), plots the ratio of the probes as a heat-map representing the layout of the probes on the microarray. The most extreme ratios define extent of the colour range, resulting in some dark plots as seen in (d) and some plots where neutral ratios are not plotted as black as seen in (e). TransformMADData, shown in (c), plots the RIPlot before and after transformation. The red line in the upper plot defines the transformation to be applied and helps to identify the level of influence applied by the transformation.

as may the **Treatment** of the sample. By also specifying that there is an interaction between them, this defines a third term which can capture different effects of one factor at different levels of another. For example, if there were three treatments and two time points and expression of a gene was consistent in the first time point of all treatments, but up-regulated at the second time point of only one treatment, this change in expression would be captured by the interaction term because it is specific to only one **Time** and **Treatment** combination.

The description of the `fitmaanova` function explains that, whilst interactions are supported, they cannot be more complex than two-way interactions such as that shown above. This proves to be a constraint for typical experimental practices, such as including biological replicates, because these would then be nested within the interaction term and are therefore a three-way interaction term.

Where some terms can be considered as random, as defined in Section 1.5.1, it is possible to provide this information to `fitmaanova` so that a mixed model analysis can be applied. This information is passed to the function as another formula describing the random terms. If it is assumed that **Array** and **Dye** are random, an appropriate random term formula would then be “ $\sim \text{Array} + \text{Dye}$ ”. This complements the formula describing all the terms of the model.

Although **Dye** has been interpreted in this study as a random effect, it is also often treated as a fixed effect. The choice does not, however, make an enormous difference to the outcome of this study. The full dataset was analysed twice: once with **Dye** as a random effect and once with it as a fixed effect. The outcome of this analysis was that, although changes associated with each effect at each time on each gene were altered, they were consistent between the effects of the model and so the F-test scores for differential expression were unaltered and gene expression profiles were not altered when exporting to mean normalised expression data. The choice of how to interpret the **Dye** effects would be entirely down to user preference in this case.

1.5.2.7 Identifying Differentially Expressed Genes

`MATest` is a function of `MAANOVA` for performing F-tests on effects of the model returned by `fitmaanova`. The function requires that the microarray data used for modelling and the fitted model are passed to it. The user must also define the term they wish to test which will be one of the terms used to generate the fitted model. `MATest` uses the information provided to generate F-statistics for each gene indicating the variation of that term, across all levels, on the gene. Where a contrast matrix is also supplied to `MATest`, defining which levels of the term should be considered by independent t-tests, it is able to perform the test for specific comparisons rather than testing across all levels with an F-test. This can be appropriate when the levels are independent rather than part of a series because it allows the identification of genes altered in expression between specific levels.

The F-statistic is a calculation of the variance associated with the term divided

by the variance associated with technical replicates (or the error term) as shown:

$$F = \frac{\text{explained (term) variance}}{\text{unexplained (error) variance}} \quad (1.5)$$

and hence the p -value associated with the F-statistic is obtained from the F-distribution with the degrees of freedom associated with those two variances. Where the p -value is below a selected significance threshold, this indicates that those genes are statistically significantly differentially expressed as a direct consequence of that term of the model.

Two alternative calculations of the F-statistic are returned by MAtest: the F1 statistic which is calculated as shown above and could be considered the ‘usual’ F-statistic; and the Fs statistic which is based on the James-Stein shrinkage estimate of the error variance (James & Stein, 1961).

A second function, adjPval, is presented as part of the MAANOVA package for the consideration of multiple-testing corrections. The p -values from the F-tests are indicative of the statistical significance of differential expression across the levels of the chosen terms for each gene probe considered separately, but suffer from the inflation of multiple testing Type I family-wise error-rate (FWER) which manifests itself when performing many separate tests such as those performed for each of the genes of the microarrays (Dudoit *et al.*, 2002). Type I errors are a rejection of the null hypothesis when it should have been accepted. It is probable that many Type I errors have occurred where many genes are declared as differentially expressed when they are, in fact, not. adjPval applies a false discovery rate (FDR) correction to the p -values, effectively manipulating the scale of the p -values, without changing the rank order of the genes (Cui & Churchill, 2003). This approach is an alternative to adjusting the significance threshold by controlling the FWER and provides more power (Benjamini & Hochberg, 1995) because the change in the scale of the p -values is driven by the data, rather than simply by the number of tests. In effect, control of the FWER is to decide the significance threshold to use in advance of looking at the data, whilst FDR is an adjustment made to the p -values because of observations made in the data. The result is to reduce the number of Type I errors by increasing the stringency of the test sufficiently to push all false positives past the significance threshold such that they are no longer designated as significant.

Four alternative FDR methods are implemented and one must be selected by the user, depending on their requirements. The StepUp procedure was proposed by Benjamini & Hochberg (1995) as an alternative to FWER, increasing the power of Type I error avoidance by basing the adjustment on the distribution of the p -values rather than the number of tests performed which can be too stringent for large numbers of tests. As the name of the procedure suggests, the p -values are stepped through in rank order from the largest p -value to the smallest p -value. Each is compared with the corresponding critical constant ($q \frac{i}{m}$) where q is a constant derived from the distribution of the p -values, i is the index of the p -value and m is the total number of tests, until a

p -value is found where $p_{(i)} \leq q \frac{i}{m}$, after which all tests are denoted significant.

The StepDown procedure was later proposed by Benjamini & Liu (1999) as an alternative to the StepUp procedure and is clearly defined as neither dominating or being dominated by that procedure, instead being more appropriate when the number of tests is small and/or a large number of the hypotheses are false. In this case, the p -values are stepped through in rank order, but from the smallest p -value through to the largest p -value. So long as the p -value satisfies $p_{(i)} \leq \delta_{(i)}$ where

$$\delta_{(i)} \equiv 1 - \left[1 - \min \left(1, \frac{m}{m-i+1} q \right) \right]^{\frac{1}{m-i+1}}, \quad 1 \leq i \leq m \quad (1.6)$$

all hypotheses are rejected, until a point where $p_{(i)} > \delta_{(i)}$, after which all hypotheses are accepted.

The Adaptive procedure (Benjamini & Hochberg, 2000) is a StepUp like procedure, but the critical constant is calculated using a prior step in which the slope of the p -values are assessed to find a change in the rate of their increase when placed in order. This allows the estimation of the number of true null hypotheses in the tests and therefore a more accurate FDR adjustment by optimising the expectation of the data.

The jsFDR procedure was proposed by Storey (2002) and attempts to fix the rejection region of the p -values and estimate the corresponding error rate rather than do as the stepwise methods above do, which is to fix the error rate and estimate the rejection region. By doing this, Storey claims a potential eight-fold increase in power by the jsFDR procedure when compared with the StepUp procedure.

1.6 Analysis of Gene Expression Data

A number of methods were chosen for use in analysing gene expression data obtained from microarray experiments. Reasons for choosing them to analyse the data in this thesis and explanations of the way in which they process the data are shown in this section.

1.6.1 Clustering

A number of alternative clustering methods exist, but they usually fall into two main classes: hierarchical clustering and partitional clustering. For simplicity of explanation, the term ‘cluster’ is used here as a noun describing sets of similar genes, even if that set only contains a single gene. Both types of clustering use a distance metric in their process and it is this measure which allows the similarity of clusters to be determined.

Partitional clustering algorithms determine all the clusters at once and then iteratively refine the membership of the clusters until no further change occurs. The algorithm best known for this approach to clustering is k -means clustering (Steinhaus, 1957; MacQueen, 1967).

Hierarchical clustering algorithms act by joining or dividing clusters that were established in the previous step. In agglomerative, bottom-up algorithms, (Eisen *et al.*, 1998) each element to be clustered is assigned to a cluster of size one, and, at each step, the most similar pair of clusters are merged. The opposite approach to this process is divisive, top-down algorithms, in which all elements initially exist as a single cluster which is then subsequently divided into smaller clusters by splitting least similar members of an existing cluster into two new clusters.

1.6.1.1 The k -means Algorithm

When performing k -means clustering, initially, a set of cluster centres are defined, either by random or by using a prior heuristic assignment of the genes to be clustered. The algorithm then continues by alternating between two steps, referred to as the assignment step and the update step. In the assignment step, each gene is assigned to the cluster whose centre it is closest to. Once this step is completed, the update step then redefines the cluster centre as the mean of its members. This process of alternating between the two steps causes gradual changes in the membership of the clusters until no change is made to the assignments of the genes to the clusters. Usually, this is a quick process, and can be repeated a number of times to avoid local maxima caused by poor selection of the initial condition. However, when clustering very large populations, the k -means algorithm becomes exponentially slower (Vattani, 2009). Another artefact of k -means is that it tends to produce equal sized clusters which may not be appropriate to the data being observed.

1.6.1.2 SplineCluster

SplineCluster (Heard *et al.*, 2006) provides a Bayesian modelling approach to clustering gene expression, allowing a time-dependent curve fitting process for each of the clusters to aid in the identification of the most-beneficial merger at each stage of an agglomerative hierarchical process. It provides a number of desirable features suitable to large scale, high-dimensional, time-dependent data sets similar to the senescence experiment. This is in contrast to recently developed approaches to time-series clustering such as the full Markov chain Monte Carlo (MCMC) Bayesian approach of Wakefield *et al.* (2003) which uses basis function coefficients to approximate the marginal likelihood at each stage of the hierarchy, which is not computationally feasible for large datasets. Another example is MCLUST (Fraley & Raftery, 1998) which is a generic Bayesian clustering tool implemented by Yeung *et al.* (2001) to cluster gene expression profiles. MCLUST is insensitive to the order of the time-series and will give the same clustering results no matter which order the time points are presented. SplineCluster, due to the use of a model to define a parametric form of the covariance function for each gene, will provide alternative results if the time-series were to be reordered, as should be expected given the time-dependency of the gene expression data from the senescence experiment.

Each gene, assigned to separate clusters, is defined by a curve fit produced using a truncated power spline basis in which each unobserved region of the time series, the time which passes between sampling periods, is represented by a series of functions with varying order to best describe the continuous nature of the changes in gene expression. Each merger at each level of the hierarchy is then identified by a maximisation of a similarity measure based on similarity of covariance terms, preserving the importance of time ordering. At each level, the newly formed cluster is assigned a curve fit for future comparisons of similarity between clusters. Once only one cluster remains, the most likely number of clusters is defined by a maximisation of the marginal likelihood calculated at each step of the hierarchy.

SplineCluster is used as the primary clustering tool in a number of recently published studies of gene expression data (Monnier *et al.*, 2010; Mutarelli *et al.*, 2008; Edwards *et al.*, 2006) and is also cited regularly by reviews of Bayesian approaches to time-series clustering. Its application to the senescence experiment data is highly appropriate given its speed and time-dependent interpretation of gene expression.

1.6.1.3 Qian Similarity Scores

Clustering algorithms often use a distance metric comparing the relative differences between observations of the same time point, which results in clusters only having members with near identical expression profiles. Whilst this allows the identification of genes which may be co-regulated or co-expressed, this approach is unlikely to identify genes which may have a regulatory effect on each other. Neither does it identify genes which may be co-regulated but with opposite expression profiles where some genes are positively regulated whilst others are negatively regulated.

Qian *et al.* (2001) provide a method for scoring genes in a pairwise fashion such that genes can be identified as similar even when their profiles are:

- only matched over a window of time,
- inversely matched; and/or
- best matched after being time shifted.

Examples of these types of match can be found in Figure 1.7. The method achieves these matches by producing a scoring matrix which compares the observation of gene x at each time point with all the observations of gene y . The same process is repeated in a second matrix with one gene profile inverted. By considering the scores of the diagonals in the matrices, each representing a time shift with the diagonal that passes through the centre of the matrix represents no time shift, the best fitting comparison is regarded as the most appropriate match between the genes.

The Qian similarity score provides an alternative, more biologically relevant, approach to finding genes which form clusters, or functional modules, and is the metric used in the implementation of some methods for clustering by affinity propagation (Frey & Dueck, 2007; Kiddle *et al.*, 2010).

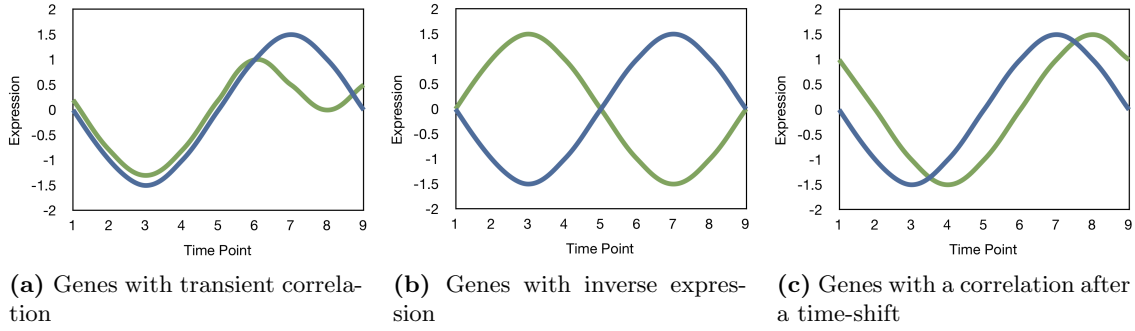


Figure 1.7 – Types of gene profile similarity which are ranked highly by the Qian similarity score

Three examples of similarity between genes which can be highly ranked by the Qian similarity score are shown. In each plot, the horizontal axis represents time whilst the vertical axis represents gene expression. In (a), the green gene is only correlated with the blue gene in time points 1 through to 6. This may indicate that these two genes are co-regulated, but that the green gene has a second component to this regulation which is missing after time point 6. In (b), the green gene is a perfect match to the blue gene, but only if one profile is first inverted. This may indicate that both genes are regulated by a common transcription factor, but that the effect is positive regulation for one gene and negative regulation for the other. In (c), the green gene is again a perfect match to the blue gene, but only if it is shifted back one time point. This may indicate that the blue gene is a regulator of the green gene, and hence why the green gene is later to respond than the blue gene.

1.6.2 Identifying Over-represented Gene Ontologies

Gene Ontology (GO) annotations were a part of the Gene Ontology project (The Gene Ontology Consortium, 2000) as a way of describing gene products as a series of attributes across all species. These attributes, referred to as GO terms, may describe the cellular component, biological process or molecular function of the gene product and may only be submitted with evidence such as experimental results, literature sources or computational analysis. Each gene is tagged with any number of terms and these terms form a hierarchical tree, such that some terms are parents or children of other terms.

During statistical analysis of GO terms, performed by packages such as GOSTats (Falcon & Gentleman, 2007), hypergeometric tests are used to identify whether clusters, or modules, of genes contain a larger proportion of members assigned with specific GO terms than the expected number, given the abundance of that GO term throughout the genome. A p -value is produced which identifies the probability of finding x_c number of genes assigned with a GO term in a cluster of size N when x_g genes are assigned that GO term in the entire genome of *Arabidopsis thaliana*. Where this p -value is below a threshold, typically ($p < 0.05$), the GO term is considered over-represented.

1.6.2.1 BiNGO

BiNGO (Maere *et al.*, 2005) is the method used throughout this thesis for finding over-represented GO terms within clusters of genes. It is Java-based and written as a plugin

to Cytoscape (Shannon *et al.*, 2003) which is typically used for the graphical representation of networks and hierarchical trees such as those provided by GO terms. BiNGO provides functionality for entering the gene lists by simply pasting them into an input dialog box and also provides alternative multiple testing corrections. One of the outputs is a graphical representation of the over-represented GO terms using a hierarchy of circles with varying sizes to represent the significance of the over-representation. The over-represented GO terms are also output as tabular data so that they can be processed and it this data which is used throughout the thesis.

1.6.3 Identifying Network Topology Using Variational Bayesian State Space Modelling (VBSSM)

Reverse engineering the underlying topology of the regulatory network of a set of genes is not a trivial task. Many published methods make the assumption that all the possible interacting elements have been observed and that they have been included in the gene set (Beal *et al.*, 2005). In reality, microarrays do not provide complete information about the regulatory network because they:

- may be missing probes for some genes,
- provide noisy data for some probes,
- do not make observations of metabolites and hormones which may form part of the network; and,
- do not make observations of RNA and protein degradation which can result in altered response times.

This is confounded by the gene set selected for modelling often being identified by clustering methods which are unlikely to associate every element of a given network. A proposed solution to this problem might be to include all observed genes to identify the overall network, but this approach has its own problems since the number of gene pair permutations grow exponentially with the number of genes being modelled. This leads to an unidentifiable model because the information about each gene is finite and restricted to the number of time points and replicates. As the number of genes to be modelled increases, so do the number of parameters to be estimated by the model. Since the number of data points about each gene remains fixed, a number of alternative models become equally likely and it becomes impossible to identify which represents the most likely relationship between genes.

Beal *et al.* (2005) have developed VBSSM which is a State Space Modelling approach to the reverse engineering of the regulatory network for a small set of genes. Linear Gaussian state space models (SSMs) have been known by several previous names in the past, including Linear Dynamical Systems (Roweis & Ghahramani, 1999) and Kalman filter models (Brown & Hwang, 1997). All are a subclass of dynamic Bayesian

networks which are suited to the modelling of time series data. SSMs are particularly suited to the modelling of data collected from gene expression microarrays since they assume the existence of a number of hidden states which evolve with Markovian dynamics and can be used to model the effects of unmeasured variables such as missing gene expression data or protein degradation rates.

By providing a variational Bayesian treatment of SSMs, a novel approach has been provided to ensure that the dimensionality of the state space can be correctly identified without holding out data from that used to train the model, as in Beal (2003). This approach therefore leaves more data available for training the model and can more accurately infer the gene interactions as a consequence.

Beal *et al.* (2005) states that a sequence of p -dimensional real-valued observation vectors $(\mathbf{y}_1, \dots, \mathbf{y}_T)$ are modelled by assuming that at each time step t , \mathbf{y}_t was generated from a k -dimensional real-valued hidden state vector \mathbf{x}_t .

By focussing on models in which the dynamics and the output functions are linear and time-invariant whilst the distributions of the state evolution and noise variables are Gaussian, the following linear-Gaussian SSM equations can be used:

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + \mathbf{w}_t, \quad \mathbf{w}_t \sim N(\mathbf{0}, Q) \quad (1.7)$$

$$\mathbf{y}_t = C\mathbf{x}_t + \mathbf{v}_t, \quad \mathbf{v}_t \sim N(\mathbf{0}, R) \quad (1.8)$$

where A is the $(k \times k)$ state dynamics matrix and C is the $(p \times k)$ observation matrix. These matrices effectively capture the interaction between hidden states at adjacent time steps and influences caused by hidden states upon observations of the same time step, respectively. Q and R are covariance matrices for the state and output noise variables \mathbf{w}_t and \mathbf{v}_t .

By allowing the model to include driving inputs $\mathbf{u}_{1:D}$ which allow the control of the model by an external influence, much in the way that a remote controlled car can be influenced by telling it to move forward, backward or to steer, the model equations become:

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + B\mathbf{u}_t + \mathbf{w}_t \quad (1.9)$$

$$\mathbf{y}_t = C\mathbf{x}_t + D\mathbf{u}_t + \mathbf{v}_t \quad (1.10)$$

where B is a $(k \times d)$ input-to-state matrix and D is a $(p \times d)$ input-to-observation matrix. The driving inputs can be replaced by feedback from gene expression measurements at the previous time step in an attempt to discover the gene–gene interactions across time steps. This allows Equations 1.9 and 1.10 to be rewritten as:

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + B\mathbf{y}_{t-1} + \mathbf{w}_t \quad (1.11)$$

$$\mathbf{y}_t = C\mathbf{x}_t + D\mathbf{y}_{t-1} + \mathbf{v}_t \quad (1.12)$$

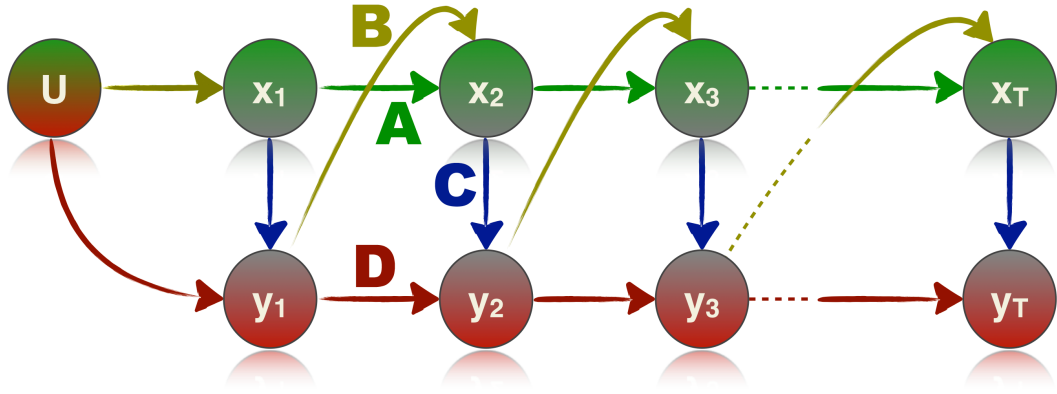


Figure 1.8 – Probabilistic graphical Bayesian network model representation of VBSSM

The VBSSM model can be summarised by plotting a network of matrices (edges) which define the transition between states (nodes). x_t and y_t represent the hidden state and observed genes, respectively, at time t . U is an input to the model and can be used to define known interactions as priors of the model. Green arrows (A) are the state dynamics matrix which captures the transition of the hidden states between time points. Yellow arrows (B) are the matrix which models the influence of observed genes on hidden states of the next time point. Blue arrows (C) are the matrix modelling the influence of the hidden states on the observed genes at each time point. Red arrows (D) are the matrix which captures the observed gene expression level influences on other observed genes at adjacent time points. A combination of the yellow, blue and red matrices can be used to directly describe the expression of each observed gene as a function of only the observed gene expressions at the previous time point ($CB + D$), therefore inferring the influence of each gene on each other gene in the network.

and in turn, since the driving input vector is now p -dimensional, the dimensions of matrix B become $(k \times p)$ and the dimensions of matrix D become $(p \times p)$. The graphical representation of this model can be seen in Figure 1.8 which illustrates these state space vectors as circles connected by arrows which correspond to the four matrices in the model, A, B, C and D .

Under this model, \mathbf{y}_t denotes the gene expression levels at time step t whilst \mathbf{x}_t represents the unobserved hidden factors. D is the matrix which captures gene–gene influences at adjacent time points, C is the matrix which captures the influence of hidden factors on gene expression at the same time point, B is the matrix capturing the influence of gene expression on hidden factors of the following time point and A is the matrix which captures the state dynamics between hidden states. In order to identify the level of influence which exists between genes, the two equations can be rewritten so that \mathbf{y}_t is a function of only gene expression at the previous time step, \mathbf{y}_{t-1} :

$$\mathbf{y}_t = (CB + D)\mathbf{y}_{t-1} + \mathbf{r}_t \quad (1.13)$$

where $\mathbf{r}_t = \mathbf{v}_t + C\mathbf{w}_t + CA\mathbf{x}_{t-1}$ and includes all the contributions from noise and previous hidden states. This leads to the ability to characterise the interaction between gene j and gene i by observing element ij of the matrix $[CB + D]$ which is describing the

influence of gene expression observations at the previous time step upon gene expression observations at the following time step whilst also accounting for all the hidden factors.

Once a model has been inferred, the ability of the model to explain the experimental data is returned as a log marginal likelihood. This score will be maximised when a model is a good fit for the data used to train it. This score can therefore be used to compare two models and identify which is the best explanation of the data provided.

1.6.3.1 Limitations

Whilst VBSSM demonstrates a huge step forward in the elucidation of regulatory networks and has been proven to work successfully on artificially simulated data and data collected in a longitudinal manner, it is unfortunate that it was not possible to gather the senescence experiment samples in the same way. In order to obtain RNA from the leaves of the *Arabidopsis* plants being studied, it was necessary that they were destroyed and biological replicates, despite being labelled similarly at different time points, were in fact collected from separate individuals in a cross-sectional manner.

The models produced by VBSSM use the biological replicates in a longitudinal manner, associating the observations of each replicate label as providing a discrete observation of the entire time series. Efforts were made to determine the importance of this mis-interpretation by randomly reassigning the biological replicate labels within time points to see what effect it may have on the resulting model. It was concluded that although the resulting models demonstrated some differences, the variance between the replicates will still remain constant and as such, the underlying network should still be identifiable.

Another limitation of VBSSM is that its models are linear and assume time-invariant interactions between the genes. It is known that not all gene interactions are a linear relationship and that, on some occasions, unmeasured variables such as phosphorylation can alter the downstream regulatory effects of some genes. However, these are computationally intensive areas for improvement in VBSSM, and could not feasibly be added at this time. No alternative better suited method of modelling could be identified and as such, VBSSM was the sole software used for the reverse engineering of regulatory networks. A recent review provides benchmarking comparisons of VBSSM and other methods (Penfold & Wild, 2011, in print). In this review, four alternative algorithms were compared:

- A time-series networks identification (TSNI) algorithm (Bansal *et al.*, 2006).
- A granger causality analysis (GCA) method (Seth, 2010).
- The G1DBN dynamic Bayesian network (DBN) package (Lèbre, 2009).
- The VBSSM DBN with hidden states, as shown here.

- A non-parametric non-linear dynamical system (NDS) found in the Matlab package GP4GRN (Äijö & Lähdesmäki, 2009).
- An implementation of algorithms proposed by Klemm (2008) to provide a causal structure identification (CSI) method.

Of the comparisons made, it was found that the VBSSM was slightly better than the TSNI and GCA methods at identifying the underlying network topology, it was similarly accurate to the other DBN method tested, but was not as accurate as the NDS and CSI methods shown. However, the latter methods are far more computationally intensive, taking 48 times as long to produce a result when compared with VBSSM and were not available when they were required for this PhD project.

1.7 Biological Validation

1.7.1 Altered-Expression Mutants

Where theoretical approaches had suggested that a particular gene was an important regulator of either senescence or a particular group of genes, altered-expression mutants were used to observe the effects of reducing or enhancing the expression of that gene. Altered-expression mutants are plant lines which have been genetically modified to alter the expression behaviour of a single gene and these may also be pyramided to provide mutations in multiple genes. Mutants can then be observed for phenotypical changes, verifying the importance of a gene in a particular process, or analysis of expression in genes suspected to be regulated by the mutated gene can verify whether an interaction exists. Those mutant types used during the project are shown in this section.

1.7.1.1 Reduced Expression

The Salk Institute Genome Analysis Laboratory (SIGnaL) produced a series of *Arabidopsis thaliana* mutagenesis lines (Alonso *et al.*, 2003), hereafter referred to as SALK lines. SALK lines are Col-0 wild-type *Arabidopsis* lines which have had a large piece of T-DNA inserted into their genome by *Agrobacterium tumefaciens* which, when located within critical transcriptional machinery of a gene such as the upstream region containing the transcription factor binding sites or within the exons of the gene, can disrupt the expression of that gene to a lower level than normal. These mutants are often referred to as knockout mutants, although their expression may only be reduced rather than entirely quenched. SALK lines are listed in a database maintained by SIGnaL (<http://signal.salk.edu/cgi-bin/tdnaexpress>) and can be ordered as seed stock which can then be sown and used in experiments. The actual expression level of the mutated gene must be verified experimentally since the insertions made by SIGnaL were at random and many of the seed stock either have a non-homozygous insertion of T-DNA or are still able to produce a partial protein which maintains some function.

1.7.1.2 Enhanced Expression

Jesper Grønlund and Sanjeev Kumar of The University of Warwick have produced a number of over-expression mutants in which the expression of the mutated gene is either constitutively induced or induced by application of an exogenous influence such as ethanol. Constitutive over-expression mutants have been generated by fusing the open reading frame (ORF) of a target gene to the 35S promoter of the cauliflower mosaic virus (CaMV35S) and then inserting this into the genome of Col-0 plants using T-DNA of *Agrobacterium tumefaciens*. Induced over-expression mutants have also been generated using a promoter for a gene known to be up-regulated in response to a specific trigger such as ethanol.

Although it is convenient to be able to enhance the expression of a selected gene at any selected point in the experiment, any results determined from an inducible over-expressor may be confounded by other responses to the trigger not specific to natural senescence. Equally important, constitutive over-expressors are expressed from very early on in the life of the plant and often cause major physiological changes in the immature stages of plant development. Because of this, many other development related processes have already been altered by the time senescent samples are taken from these lines and they are therefore harder to compare with the wild type.

1.7.2 Yeast-1-Hybrid

Where theoretical methods have proposed that a transcription factor may regulate the expression of a target gene, one method of testing this for validity is by the yeast-1-hybrid system. This system was developed almost 20 years ago to answer a specific question about the role of *ORC6* in the process of yeast cell replication (Li & Herskowitz, 1993). In its simplest form, yeast-1-hybrid consists of a library of *Saccharomyces cerevisiae* transformed with cDNAs fused to an N-terminal activation domain: the prey. The suspected promoter region of the target gene is inserted into the promoter region of a reporter gene vector which is used to transform further yeast cells: the bait. After co-transforming or mating the cDNA library with the bait transformed yeast cells, if the reporter gene is transcribed, this indicates that the transcription factor was able to bind to the target gene promoter and therefore has the potential to play a regulatory role as predicted.

The use of cDNA fragments as a screening library is not an optimisation of the screening process since the only proteins which should have a regulatory role for gene expression are transcription factors. For this reason, the use of a transcription factor cDNA library might be preferential, otherwise transcription factors are not well represented by libraries of cDNA fragments.

One benefit of this system is that it is high-throughput and therefore will identify all transcription factors in the library that are able to bind the designated promoter sequence, rather than only verifying a single prediction. One disadvantage of the system

is that it only verifies that the screened promoter can be bound by elements of the cDNA library and not that it would constitute a regulatory interaction in *Arabidopsis*, nor that the interaction would not be blocked by other factors such as bound micro-RNAs.

1.8 Aims of the Project

The research associated with this project is divided into three chapters of results. The aims of each are as follows.

Chapter 3 - Microarray Data Analysis: To further develop and customise the MAANOVA microarray analysis package so that the design of the senescence microarray experiment described above can be used in an optimal manner as well as providing new quality control and normalisation techniques to maximise the benefit of technical replication. To perform statistical tests of temporal differential expression and identify those genes altered in expression during the senescence period. To obtain temporal gene expression data representing biological changes in gene expression, removing effects of experimental and technical variation.

Chapter 4 - Analysis of Gene Expression Data: To analyse the gene expression data of those genes altered in expression during senescence and identify functional groups of genes. To develop a standard method of quantifying the senescence response for the quantification of altered phenotypes in mutant lines. To use theoretical modelling of transcriptional networks for subsets of genes and identify those which regulate the transcription of many downstream genes before building theoretical networks of likely direct interactions with those strong regulators.

Chapter 5 - Biological Validation of Theoretical Models: To experimentally test predicted interactions of theoretical network models using established laboratory techniques. To use the results of such tests to produce a new theoretical network model which could be used in further research to find the true transcriptional networks of senescence.

Chapter 2

Materials and Methods

2.1 Chapter Summary

This chapter introduces the methods used throughout this thesis. Amongst those methods used is a microarray experiment analysis package MAANOVA. This was used to find inadequacies throughout microarray experiments using dual-channel microarrays and to identify the expression changes in genes throughout those experiments. Where gene expression data was obtained, that data was clustered using SplineCluster and then analysed for over-represented gene ontologies by BiNGO.

Theoretical models of regulation between genes were produced using Variational Bayesian State-Space Modelling (VBSSM) which uses hidden states to account for unmeasurable regulatory aspects and missing gene expression data. Novel approaches to the use of VBSSM includes multi-modelling where many small sets of genes are produced from a larger set in an attempt to identify those genes which are highly regulatory of others. Another approach is Metropolis-like VBSSM in which the surrounding network of a given gene is identified by replacing small numbers of genes around the central gene until a model with maximum likelihood of explaining the gene expression data has been found.

A novel senescence phenotype screen has been developed, using digital photography and quantitative methods to assess the yellowing of leaves over a period of time. This has allowed the identification of altered senescence phenotypes in mutant lines. Yeast-1-hybrid has provided a method for identifying which transcription factors are able to bind to the upstream region of gene sequences and potentially alter their expression. Additionally, further microarrays have provided a method for identifying the downstream effect of reducing the expression of genes key to senescence.

2.2 Microarray Analysis

Analysis of the microarrays which were produced during the senescence experiment was performed using the MAANOVA software package. The original implementation of this software has been described in Chapter 1 but the software was heavily modified to

better suit the requirements of this experiment and hence its further development and application is described in detail within Chapter 3 rather than here.

2.3 Gene Expression Data Analysis

2.3.1 SplineCluster

To cluster selected genes into putative functional groups based on similar expression profiles, SplineCluster (Heard *et al.*, 2005) was used as this provides a hybrid between Bayesian hierarchical clustering and k -means clustering by offering a sweep mergers function as described in Heard (2011). To make the implementation of SplineCluster more simple, an R script was developed which automated the transformation of input data into a format suitable for SplineCluster, called the command line version of SplineCluster and then automatically processed the output data into a spreadsheet once SplineCluster was complete. This script is included as a function in the whrimaanova package developed in Chapter 3. A definition of the inputs necessary for this function can be found in Procedure 2.1. Vectors and matrices required by the function must be R objects which can be loaded from tab-delimited text files using the `read.table()` function.

Not all of the inputs shown are necessary for the processing of the senescence dataset and are included only for completeness. Many of the inputs have defaults, as defined, and, in most cases, can be accepted without need to specify a value for those inputs. For all SplineCluster analyses performed on the senescence experiment data:

- **genefilter** provided a list of genes which were determined to be differentially expressed.
- **fullgenelist** provided a list of all the observed genes of the CATMA microarrays.
- **expression** provided a matrix of expression data for all observed genes across 11 time points.
- **priorprecision** was allowed to remain as the default of 10^{-4} .
- **annotation** provided annotations for all the observed genes.

The **priorprecision** value defines the precision to which the genes must fit the cluster profile to be considered a member of that cluster. As the prior precision is reduced, the fit needs to be less precise and hence fewer clusters are identified. If the prior precision is increased, the fit of the gene expression to the cluster must be more precise to be considered a member of that cluster and hence the number of clusters identified is increased.

Following communications with Nicholas Heard, the developer of SplineCluster, it became apparent that he had recently implemented a method of sweeping the clusters at each merger to identify outliers and reallocate them to more appropriate clusters,

Procedure 2.1 Function definition for the SplineCluster wrapper

<pre>splinecluster(genefilter, fullgenelist, expression, priorprecision = 1e-04, annotation = NULL, datasplit = NULL, gramschmidt = FALSE, sweepmergers=0, xvalues=NULL, errorcovariances=TRUE)</pre>	
genefilter	A vector of gene IDs which are to be clustered.
fullgenelist	A vector of gene IDs for all the rows of expression .
expression	A matrix of expression data in which the rows correspond to the gene IDs in fullgenelist and the columns correspond to the time series.
priorprecision	A floating point value defining the prior precision to use for defining the number of clusters.
annotation	A matrix of annotations in which the first column contains gene IDs and other columns contain annotation information to be included in the SplineCluster output.
datasplit	Where multiple experiments are simultaneously clustered, a vector of number of columns associated to each experiment. If the expression matrix contained 30 columns, the first 19 from one experiment and the last 11 from another experiment, this vector would be <code>c(19,11)</code> .
gramschmidt	A logical value defining whether to use Gramschmidt orthogonalisation of the design matrix to increase the speed of clustering.
sweepmergers	An integer representing the maximum number of times to sweep at each merger and reallocate outlying members. 0 represents no sweeping at each merger.
xvalues	A vector of integers equal in length to the number of columns in expression specifying a true metric of the time at each observation. Only to be specified where observations were not made at equal intervals.
errorcovariances	A logical value specifying whether to plot heat-maps of error covariances when datasplit has been specified.

now published in Heard (2011). Under normal circumstance in bottom-up hierarchical clustering, the merger between two existing clusters can lead to some members of the resulting cluster straying significantly from the new cluster profile. Rather than continuing to completion, as would normally be the case, SplineCluster can be instructed to sweep the current memberships at each merger and reallocate outliers to more appropriate clusters. This process dramatically increases the time taken to reach completion and therefore the user can limit the number of iterations of this reallocation at each merger by specifying an integer for the **sweepmergers** parameter of the SplineCluster wrapper.

Once it was established that **sweepmergers** could be used to reallocated genes which had become an outlier for their cluster due to the nature of the hierarchical process, this was always set a value of 10000 and **gramschmidt** was set to **TRUE** to help speed the process up because merger sweeping increased the processing time to several

days compared with an equivalent number of genes taking approximately 30 minutes without it.

2.3.2 BiNGO

To find over-represented gene ontology (GO) annotations in selected groups of genes, the software package BiNGO (Maere *et al.*, 2005), a java-based plug-in for Cytoscape (Shannon *et al.*, 2003), was used. In order to identify these over-represented genes, a text file was prepared which listed the AGI numbers of the genes in each cluster following the format suggested by the BiNGO User Guide (Flanders Interuniversity Institute for Biotechnology, 2010):

```
Cluster_001
At1g10101
At1g10102
batch
Cluster_002
At2g20202
At2g20203
batch
Cluster_003
:
At5g50505
```

BiNGO was instructed to perform hypergeometric tests to identify the existence of over-represented GO terms from the Biological Process annotations of differentially expressed genes in *Arabidopsis thaliana*. A reference set containing only the genes which were identified as differentially expressed ensures that the significance of over represented GO terms are relative to the abundance of those terms in the genes which were clustered. Only those GO terms which were significantly over-represented at ($p \leq 0.05$) after the Bonferroni family-wise error rate (FWER) correction were returned.

The output of BiNGO was a folder of BGO files which each contained information about the over-represented GO terms of a specific cluster. These were processed using a custom script, written in R, which can be found in Appendix D. The script identifies the GO terms and then plots how significantly over-represented they are as a histogram of the negative log of their p -value whereby more significant results have taller bars. In addition to this, the percentage of genes in a cluster which are annotated with a specific GO term is shown. An example of these plots is shown in Figure 2.1. This allows the rapid interpretation of over-represented terms in each cluster, whilst comma-separated value (CSV) files are also returned to provide a full statistical report of those terms in each cluster.

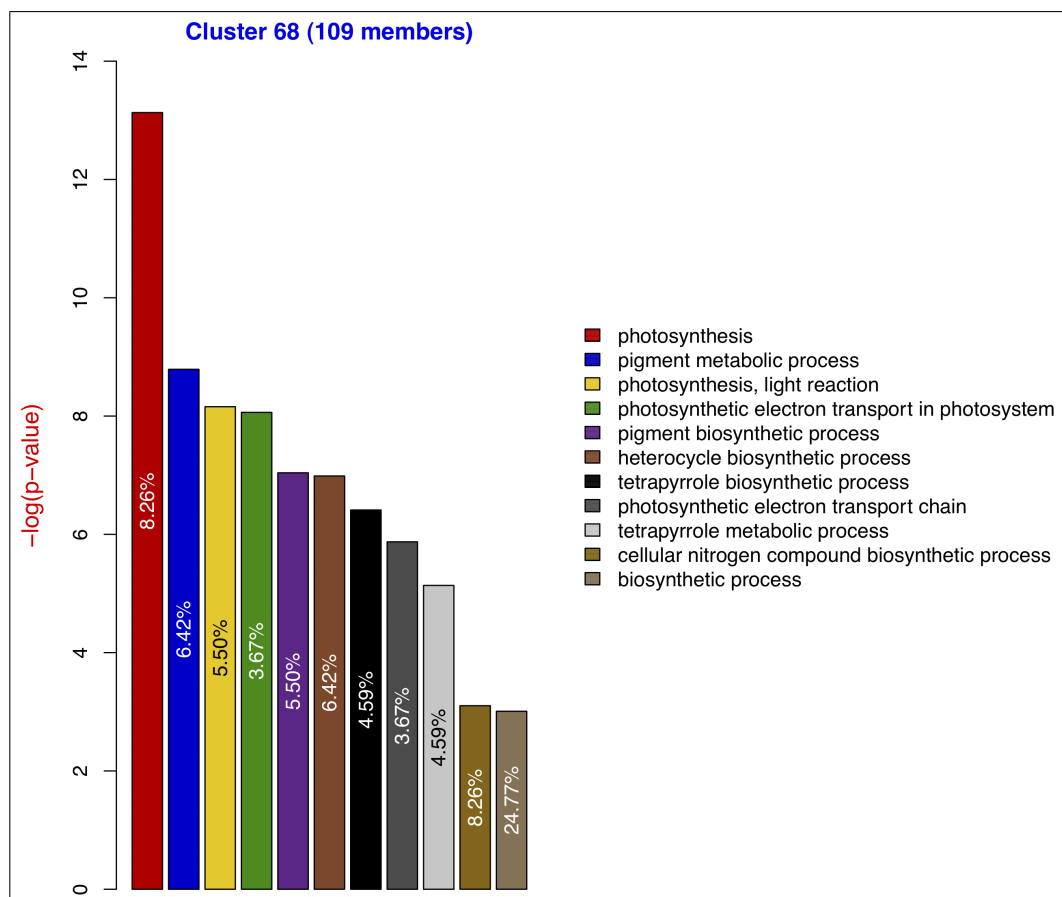


Figure 2.1 – An example of histograms returned by the script which interprets BiNGO’s output

The plot shows significantly over-represented terms of a cluster presented to BiNGO for GO analysis. The height of the bars indicate the negative log of the likelihood of finding those many similarly annotated genes in a cluster of size 109 given the abundance of that annotation throughout the *Arabidopsis thaliana* genome. Where the bar is tall, this indicates a highly unlikely occurrence and hence a significantly over-represented GO term. The percentages indicate the number of genes within the cluster annotated with this term and the names of the terms are identifiable by the key shown to the right of the plot.

2.3.3 Qian Similarity Score

In order to obtain the similarity score provided by the methods of Qian *et al.* (2001), the implementation of the method by clustering software TCAP (Kiddle *et al.*, 2010) was used. This required that the TCAP package was installed into MATLAB[®] (MATLAB, 2010) after which the function `tcap_score()` was used to identify the similarity score between every pair of a given set of genes. Data presented for scoring was a series of expression values, one per time point, for each gene to be scored. Each gene had a mean expression of zero and standard deviation normalised to 1. This expression data was loaded into MATLAB as a matrix with each row representing a gene and each column representing a time point and then passed to the `tcap_score()` function which produced three vectors: `psi_star_vector`, `inverts_vector` and `delays_vector`. These vectors

CATMA ID	Rank	Qian Score	Delay	Inverted
CATMA2a26240	1	-0.11014	0	True
CATMA1b41990	2	-0.12828	0	False
CATMA1a10030	3	-0.15335	0	True
CATMA4a20290	4	-0.16211	0	True
CATMA5a18270	5	-0.17747	0	True
CATMA4a40195	6	-0.19882	0	True
CATMA2a15100	7	-0.20294	0	True
CATMA5a37600	8	-0.20469	0	True
CATMA1a45880	9	-0.20506	0	True
CATMA2a34420	10	-0.21401	0	False
CATMA1a51570	11	-0.22007	0	False
CATMA5a24170	12	-0.22333	0	False
CATMA4a24686	13	-0.22440	0	True
CATMA5a20375	14	-0.23283	0	True
CATMA1a56690	15	-0.23513	0	False
CATMA1a25320	16	-0.24302	0	False
CATMA3a25675	17	-0.24343	0	True
CATMA3a53450	18	-0.24453	0	True
CATMA3a44120	19	-0.24498	0	False
CATMA4a06050	20	-0.24543	0	False

Table 2.1 – An example of Qian similarity scores output by TCAP

Once exported, the Qian similarity scores were presented in a table, as shown, with the highest ranking genes offering the greatest similarity to the target gene. The first row of this table states that the gene associated with CATMA probe ‘CATMA2a26240’ was the closest match to the target gene, ‘CATMA1c71331’. This match was direct without the need for time-shifts as indicated by the 0 in the Delay column, but the genes are only correlated when an inversion is made in one of their profiles.

referred to each pairwise comparison between the provided gene expression data.

- **psi_star_vector** contained the optimum similarity scores.
- **inverts_vector** contained 0 where the similarity was without the need for an inversion and 1 where the greatest similarity was only obtained with an inversion of the profile of one gene.
- **delays_vector** indicated the number of time points gene x must be shifted by to obtain maximum similarity with gene y .

These vectors were then converted into three matrices of $N \times N$ genes, each intersection representing a pairwise comparison between genes.

Another function then used the three matrices above to return only the comparisons of a specified gene with all other genes in the data set. These were sorted so that genes with greatest similarity to the specified gene are listed first. An example of the list output can be found in Table 2.1.

2.4 Theoretical Network Modelling

Where gene expression data was used to produce putative network models of the underlying transcriptional relationships between them, the methods applied were those shown in this section.

2.4.1 Variational Bayesian State-Space Modelling

Putative transcriptional network modelling was performed using VBSSM (Beal *et al.*, 2005) which required MATLAB[®] (MATLAB, 2010) as it is a series of MATLAB scripts which process the gene expression data. The data provided to VBSSM was in the form of three files:

- `data.txt` which contained the expression data for the genes to be modelled. Each row contained the expression data for one gene with all the observations from one replicate adjacent to one another and in the correct order whilst further replicates followed in adjacent blocks.
- `names.txt` which listed the identifiers for each gene to be modelled, one per row.
- `description.txt` which listed the annotations for each gene to be modelled, one annotation per row.

These were then processed with a script which locates the VBSSM toolbox, loads the data into MATLAB, informs VBSSM of the data parameters whilst normalising the expression data, instructs VBSSM to start multiple modelling seeds across nodes of a cluster, plots log marginal likelihood (F) versus the dimensionality of the hidden states (k) of the seeds and then deletes model files of non-optimal seeds. The source code of this script is shown in Appendix A and a description of the process is shown in Procedure 2.2:

Procedure 2.2 Pseudo-code of the VBSSM control script presented in Appendix A.

Line(s)	Description
1–3	Produce three new folders for VBSSM output in the current directory.
4	Loads in the expression data from <code>data.txt</code> .
5–7	Normalises 11 time point expression data to a mean of 0 and a standard deviation of 1 before saving two <code>.mat</code> files containing the data in the correct format for modelling.
8–9	Uses the <code>dfeval</code> command of the MATLAB Distributed Toolbox to request that 10 MATLAB workers process 1 of 10 modelling seeds each.
10–13	Produce a plot of F vs k across the 10 seeds at each k -value and then saves the maximum mean F to a <code>.txt</code> file and the plot to an <code>.eps</code> file.
14	Converts the 10 seeds for the k -value with maximum mean F into three Cytoscape model files per significance level. The levels of significance which are automatically returned are 90.0%, 92.5%, 95.0%, 97.5%, 99.0% and 99.9%.
15–23	Iterates through the <code>.mat</code> files in the <code>vbnet</code> folder and removes all files which do not relate to the seeds with maximum mean F.

The files generated for Cytoscape consist of a `.sif` file which contains information about which genes interact with which others, a `.noa` file which contains annotations for the nodes of the network, i.e. the gene identities and/or functional annotations, and an `.eda` file which contains annotations for the edges of the network, i.e. the frequency of that interaction. Interaction frequency was calculated as a percentage of modelling seeds exceeding the significance level set for the model for a given interaction whilst negative numbers were used where an interaction was a negative regulation or inhibition.

2.4.2 Cytoscape

The three Cytoscape model files returned by VBSSM in Section 2.4.1 were used to produce a graphical interpretation of the interactions that exist between genes. For this, the `.sif` file was used to generate a new model via the “File > Import > Network (multiple file types) ...” menu option. The given model was then populated with annotations by loading the appropriate `.eda` and `.noa` files into the model.

The defaults of Cytoscape only provide a very basic overview of the network and so the VizMapper™ settings were adjusted such that:

- **Edge Color** was set to a continuous mapper based upon the interaction frequency whereby those values less than zero are plotted in red whilst those greater than zero are plotted in green.
- **Make Arrow Color Match Edge Color** was checked to ensure the colour was carried through to the arrow head.
- **Edge Target Arrow Shape** was set to a continuous mapper based upon the interaction frequency whereby those values less than zero are plotted with a T-shaped head whilst those greater than zero are plotted with a delta shaped head.
- **Lock Node Width/Height** was unchecked to allow non-square/non-circular nodes.
- **Node Shape** was set to either ellipse or rounded-rectangle.
- **Node Width** was set to 120.

All other settings were left as defaults. In order to filter the edges of the network to leave only those which were significant in at least 60% of the seeds, the filter tool was used in Cytoscape to highlight those edges with an interaction frequency between -60 and 60. These edges were then hidden from the network view using the “Select > Edges > Hide Edge Selection” menu option. The nodes were then sorted into three areas of the screen. Those which have lots of connections, those with less than 5 connections and those with no remaining connections. The nodes with no connections were then highlighted and hidden from view using the “Select > Nodes > Hide Node Selection” menu option.

The remaining nodes after this process were laid out, by hand, so that the nodes with many connections (typically the hub genes), were assembled around the centre of the network, whilst those with one connection back to the hubs were clustered around the outside of them and those nodes connected to multiple hubs were placed in a position where all their connections could be easily identified. The final network was then exported from Cytoscape as a .png file using the “File > Export > Network View as Graphic...” menu option.

2.4.3 Multi-modelling

To analyse larger sets of genes and try to pick out those which appear to be highly regulatory, a process hereafter referred to as multi-modelling was used. In this process, a large set of genes, typically between 500 and 1000 genes, is selected by hand or by some qualifying criteria because they are thought to show relevance to the biological process being studied and are thought to form a large proportion of the regulatory network being discovered. This process attempts to find those genes in the large set which demonstrate robust regulatory activity on other genes in the set. To do this, subsets are made from the larger set, so that models of a size which are known to converge upon a single solution can be produced. The only information harvested from each model is the number of and the consistency of interactions from the upstream gene of the interaction.

In order to optimise this approach, the selection method of the genes must ensure that each possible pair between the large set is fairly represented by the completed models. To do this, rather than selecting the genes for modelling as individuals, the possible pairs between them were selected instead, focussing more on those pairs which had been represented the least by the models so far. Each new model, then, was produced by analysing the frequency of pairs throughout the models already processed and then by randomly selecting from those with the minimum frequencies. As will be explained in Section 4.3, it is desirable to select 88 genes for each model, and so 44 pairs were selected at random in this way. The genes which exist in those pairs make up the set for the next model. Where some of the selected pairs contained genes already selected in other pairs, the set may not consist of exactly 88 genes, and in this case, further pairs are selected at random and genes added to the set until the threshold of 88 has been met.

Once every pair had been presented at least five times, although this number could be changed if more time were available, the process was stopped and the models were collated to identify the outcome of regulatory interactions between all genes. Once again, the genes were considered in pairs and thresholds were set that signify whether the interactions between those pairs are significant or not. Each pair was given a posterior probability for demonstrating an interaction and the threshold used here was that the posterior probability must be more than 95% in at least one modelling seed on at least 50% of the occasions when that pair was presented in a modelling set. The upstream

member of the pair was then classified as a hub gene if it met this criteria with more than one downstream gene.

It is impossible to make inferences about the regulatory behaviour of specific pairs of genes through this process because each model produced can be anywhere from a very good to a very poor fit depending on the exact genes in the set. Where a gene set contains genes that are a poor fit with each other, regulations will still be inferred, many of which are very tentative. Only regulations in the same model can be directly compared for likelihood and only whole models can be compared with each other for fit to the expression data provided, and so these results can only identify that a gene may be highly regulatory of others, and not that it has better regulatory behaviour on any one given gene.

To assist in the automation of this process so that the time taken to reach the given threshold could be minimised, an R script, shown in Appendix F, was produced which managed the selection of genes, compilation of expression data for those genes, initiation of the VBSSM toolbox and then compilation of the final results.

2.4.4 Metropolis-like VBSSM

Whilst multi-modelling was able to identify hub genes in a large pool of genes, it was not able to identify robust interactions in networks surrounding those genes because the F score for each model was not being considered. By applying a stochastic optimisation approach inspired by the Metropolis algorithm (Metropolis *et al.*, 1953) (Metropolis-like VBSSM) to the selection of genes from a large pool, it is possible to converge on a set of genes whose expression levels are accurately described by a VBSSM model, possibly indicating that the putative network has much similarity to the real network. This approach requires that models are run sequentially and that only small updates are made to the list of genes on each occasion, finding combinations of genes which result in better fitting models.

Unlike multi-modelling it was essential to run models sequentially as the fit of each model must be known before the following set of genes could be selected. For this reason, it was desirable that each model be performed quickly without a substantial loss of accuracy. Two modifications to the VBSSM procedures were made to enable models to be produced in substantially shorter times:

- The number of seeds for each VBSSM model were reduced to 5. As can be seen in the typical F vs k plot shown in Figure 2.2, 10 seeds for models of 88 genes from the senescence experiment were always so similar for any given k value that it was unnecessary to produce so many in order to establish an accurate interpretation of the best fitting model for the provided expression data.
- The number of hidden states were reduced by initially only training the model at a hidden state dimensionality of 8 and 9 and then identifying which has the maximum mean F. The dimensionality was then progressively incremented or

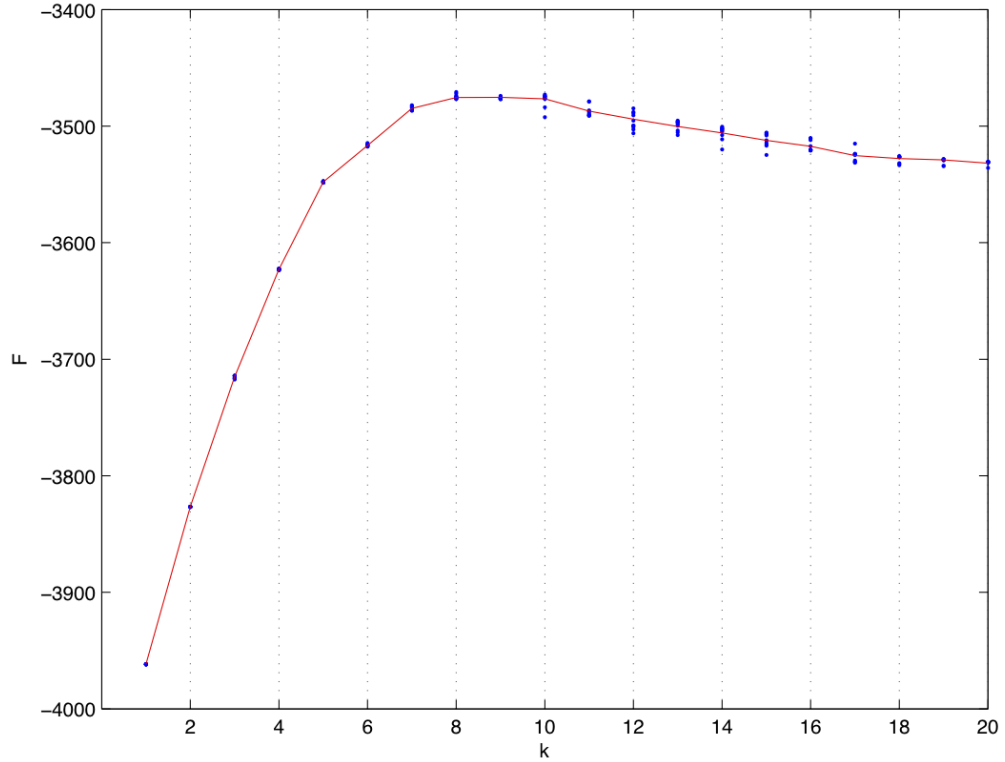


Figure 2.2 – An example of a typical F vs k plot produced by VBSSM for the senescence experiment data

The negative log marginal likelihood (F) is plotted in the vertical axis against the dimensionality of the hidden states (k) plotted in the horizontal axis. The plot shown is based on 88 genes with 11 observations in each of 8 biological replicates for each gene. Each of 20 alternative k -dimensional states were seeded 10 times, requiring 400 total models. The blue points represent the F score of each individual seed whilst the red line plots across the means at each k -dimensional state. The optimum dimensionality of hidden states, in this case 8, can be identified by the peak of the red line. Typically, the red line peaks at a single point.

decremented towards the higher mean F until a lower mean F was encountered indicating that the maximum mean F had been identified. This was possible because of the monotonic convergence on the maximum F .

These two modifications allowed as much as 26-fold reduction in processing time for each model, allowing Metropolis-like VBSSM to progress rapidly through many models in a sequential manner. However, to achieve the selective k sampling, the VBSSM process had to be split into several sections and some modifications made to the plotting script of the F vs k plot. The VBSSM processing script, as shown in Appendix A, was divided into four sections as described in Table 2.2.

To initiate the Metropolis-like VBSSM process, a template folder was produced which contained the VBSSM scripts as described above. A `.txt` file containing the identifier of the gene which was to remain central to the network was supplied along with a second `.txt` file containing a list of identifiers, one per line, for all the genes which are chosen for their similar expression to the central gene. These files are loaded

Script Name	Description
VBSSMPart1.m	Loads the expression data for the genes to be modelled before exiting. The functionality of this script is identical to lines 1–7 in Appendix A.
VBSSMPart2.m	Initiates the production of 5 modelling seeds between two k values as specified by the Metropolis-like VBSSM script. The functionality of this script is similar to lines 8–9 in Appendix A.
VBSSMPart3.m	Plots the F vs k plot for the numbers of hidden states already tested. If the maximum mean F is still ambiguous, further k values will be evaluated in the previous script. The functionality of this script is similar to lines 10–13 in Appendix A.
VBSSMPart4.m	Produces the Cytoscape files for the seeds with the maximum mean F. The functionality of this script is identical to lines 14–23 in Appendix A.

Table 2.2 – The four sections of the Metropolis-like VBSSM processing script

Because of the requirement to sample an indefinite number of k values, the VBSSM processing script could not be defined prior to modelling, and hence it was divided into four sections as detailed above. The source code of these scripts can be found in Appendix G.

by the coordinating script, written in R and found in Appendix G, which ensures that the genes being selected for modelling are chosen by following a set of rules to converge upon that set which maximises the fit of the VBSSM model. This script is very similar to that of multi-modelling with variations in the way that the genes are selected for each model.

The first model of the run consists of the central gene and 87 randomly selected from the pool of similar expression genes. The model is implemented in the four stages of VBSSM as shown above and the script waits for completion of the VBSSM model before continuing with another. The F score associated with this model is accepted as the best yet discovered and hence the score sets the benchmark and this model is labelled “gold”. The remaining models are then always based upon the last known gold model using the following logic:

- A number of genes, referred to as the swap number, are removed from the genes present in the set selected for the last gold model and these are returned to the pool of genes.
- The same number of genes are randomly selected from the pool of genes and placed into the set selected for the last gold model, providing a new set of genes for the next model.
- Once the new model is completed, it is accepted as the new “gold” model if its F score (F') satisfies the condition:

$$\alpha < \exp(-|F' - F|) \quad (2.1)$$

where F is the F score of the most recent gold model and α is a random number drawn from the uniform distribution $U(0, 1)$.

- The swap number controls the granularity of the search through parameter space and is adapted according to the proportion of acceptance for recently conducted models:
 - Where the most recent model was accepted, the swap number is increased by a factor of 1.05^3 to promote wider searches around the new gold model.
 - Where fewer than 20% of the most recent 8 models were accepted, the swap number is reduced by a factor of 1.05 to increase the similarity between subsequent models.
 - Where greater than 30% of the most recent 8 models were accepted, the swap number is increased by a factor of 1.05 to decrease the similarity between subsequent models.
- Once the value of the swap number reaches zero, the search is halted as no further improvements for the model could be found.

It should be noted that this process is not a Markov chain because each new step takes into account more than one previous step's outcomes. Instead, this uses similar logic to, but is not equivalent to, the Metropolis algorithm.

The swap number controlled the likelihood of finding a new model with a better fit than the previous gold model. By increasing this number, the similarity between models was reduced but the likelihood of getting trapped in a local maxima was also reduced. However, after some time had passed and models with better fit were becoming more sparse, searching with large changes between gene sets was inefficient and it is for this reason that the swap number was reduced. By reducing the swap number, the similarity between models was increased and the parameter space was investigated more slowly, allowing fine control over the gene selection to find similar, but improved, models.

The target value of acceptance rate is a compromise between speed and accuracy whereby larger values cause the steps between models to be small and a maximum log marginal likelihood to be found more quickly but at the risk of reducing the breadth of the parameter space search and finding a local maximum as a consequence. Similarly, a smaller target value allows a long time to be spent searching the parameter space by continually making big changes between steps of the process before eventually moving onto smaller steps and finding the true global maximum. However, because of the long time taken for each VBSSM model to complete, it would be impossible to find an optimum value within a reasonable time frame and so the target value of 25% acceptance over the most recent 8 models (represented by a lower and upper bound of 20% and 30% respectively) was chosen as this acceptance rate was optimised by Christopher Penfold

of The University of Warwick in other applications of the Metropolis approach to faster algorithms.

After halting the search, the last gold model was considered to be a good indication of the most likely interactions between the central gene and those genes which were part of the large pool provided. However, most gold models towards the end of the process would have very similar F scores and would, hence, provide interactions that may also be of interest if substantially different from the final gold model.

2.5 Plant Preparation and Phenotype Screening

2.5.1 Plant Growth

Where *Arabidopsis thaliana* plants, either Col-0 wild-type or altered-expression lines, were required for experimental purposes, 100 ~ 150 seeds were placed into labelled 1.5 mL micro-centrifuge tubes and suspended in 1 mL of 0.1% agarose gel solution before being stored at 4 °C for between 48 and 72 hours to stratify the seeds. On the day of sowing, a sufficient number of P24 PlantPak trays were filled with lightly compacted *Arabidopsis* mix compost (Levingtons F2 compost:sand:vermiculite 6:1:1). These trays were soaked by immersion in tap water so as to draw up water through the base of the pots. Once the compost was visibly damp at the surface of the pot, the P24 trays were transferred to solid based trays able to accommodate five P24 trays each. To maintain humidity during germination, a further 2 cm depth of water was added to the large trays. To each pot, five individual seeds were placed with a glass pipette tip; four at the corners of a square with the fifth marking the centre. Labelling sticks pushed into the pots were used to identify the sown seeds and also as a support structure for a single sheet of polythene placed above the pots to increase humidity levels. The trays were then located in a climate controlled environment where the following parameters were maintained:

- 70% relative humidity
- 20 °C ambient temperature
- 350 ppm CO₂ concentration
- 120 $\mu\text{mol.m}^2.\text{s}^{-1}$ daytime light intensity
- 16 hour daylight / 8 hour darkness cycles

One week after sowing, the polythene was removed and discarded and the water topped up to a 2 cm depth once more to prevent dehydration. Two weeks after sowing, where more than one plant existed in a pot, the excess plants were removed using pointed tweezers to ensure that no plants were given a competitive advantage. Three weeks after sowing, the plants had developed up to 12 leaves but not yet begun to bolt and were ready to harvest.

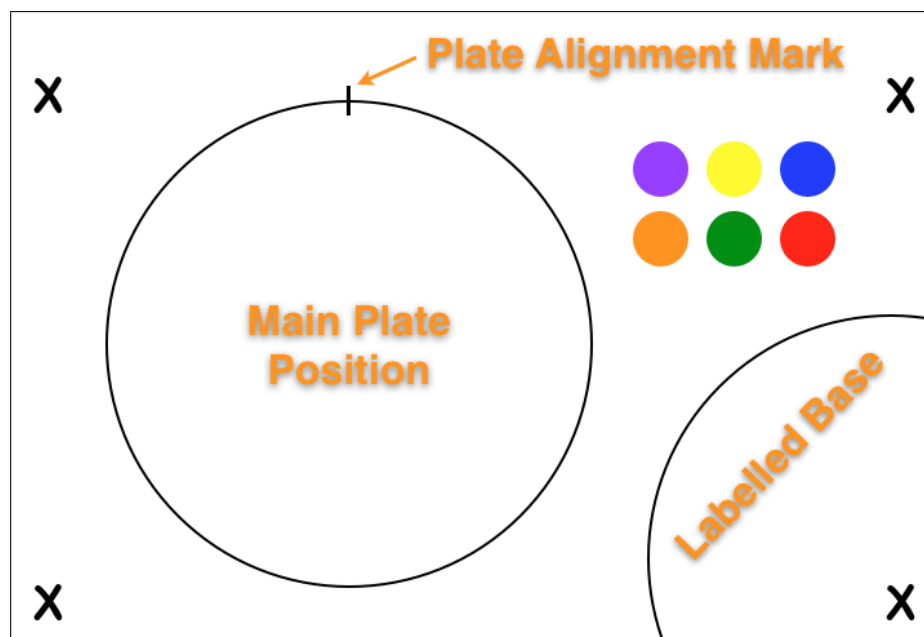


Figure 2.3 – Alignment marks used as a background during phenotype screenings

A piece of paper with markings similar to those shown was used as a background for the plates being photographed. The corners of the viewfinder are aligned with the crosses in each corner of the diagram. The plate with the rosettes was placed in the area marked 'Main Plate Position' using a mark on the edge of the plate to align with the 'Plate Alignment Mark'. The base of the Petri dish is placed in the area designated 'Labelled Base' with the label inside the photographed area. The coloured circles in the top right are present to provide variety of colour for the camera to perform more accurate image normalisation.

2.5.2 Phenotype Screening

Plants were grown using the same methods, as shown in Section 2.5.1. Observation plates were produced by placing 9 cm filter paper discs into single-vented 9 cm Petri dish lids and saturating them with 3 mL of autoclaved water. The Petri dish bases were labelled in one quadrant of the underside with an identifier for the plant line to be observed on the plate. Three healthy plants from each line, prepared using the methods shown in Section 2.5.1, were separated from the compost media by snipping immediately below the attachment point of the petioles so that no root material remains. The three plants were quickly arranged in a triangle formation on the observation plate and transferred to a light-proof box which was subsequently stored in a constant 20 °C environment. Three plates were prepared per line providing nine biological replicates.

At the same time each day, a photography lab was prepared by suspending a Nikon D50 digital SLR camera with an AF-S DX Zoom-Nikkor 18-55mm $f/3.5-5.6$ ED lens above a work surface illuminated only by 4 × 100 W tungsten bulbs arranged in pairs either side. Each of the observation plates were consecutively placed on the work surface and photographed without adjusting or moving the camera. The plates were then stored again in complete darkness at 20 °C until the following day. Once all

lines had demonstrated leaf yellowing, the samples were discarded and the photographs collated for analysis.

In order to ensure that all the plates were treated similarly during photography a place marker was used as a background to ensure consistency both between plates and between days. During the first day of photography, plates were marked on the edge so as to provide a reference point for the orientation of the plate on subsequent days. The background markings are shown in Figure 2.3, where it can be seen that:

- Crosses defined the corners of the photographed area to ensure that the same lens zoom was used each day.
- A consistent white background provided a sample area to normalise images where colour biases existed.
- Plate areas defined the position that the plates and lids should be placed as well as the orientation to maintain consistency between days; and
- A series of coloured dots provided a reference for the camera to help reduce colour biases during the camera’s automatic white-balancing.

In an attempt to further normalise the photography at the experimental stage and reduce the need for post-experimental normalisation, the camera was set to manual settings to capture the plate’s appearance consistently between photographs rather than allowing the camera to optimise the settings for each plate and introduce inconsistencies. The chosen settings were:

Mode: Manual (M)
ISO sensitivity: 200
Lens aperture: $f/7.1$
Shutter timing: $1/200^{\text{th}}$ of a second
Flash mode: Off
White balance: Tungsten lighting
Focussing: Manual - set to the distance of the leaves on the first photograph then left unchanged

Images were captured using an infra-red remote shutter release to prevent camera shake, and all the plates were photographed daily until yellowing had occurred for all lines in the experiment. The photographs were then ready for computational analysis.

The photographs were imported into the photo-cataloging/editing software Aperture (Apple) where they were processed to normalise for white balance and brightness, using the tool ‘Auto Levels (Separate RGB)’. These were then exported into folders, each folder containing all photographs taken of an individual line across all time points, ensuring that the file names allow the time points to be alphanumerically sorted into the correct order. Images were analysed using ImageJ64 (Abramoff *et al.*, 2004) which allowed the selection of a white reference point in the photograph and the area of leaf

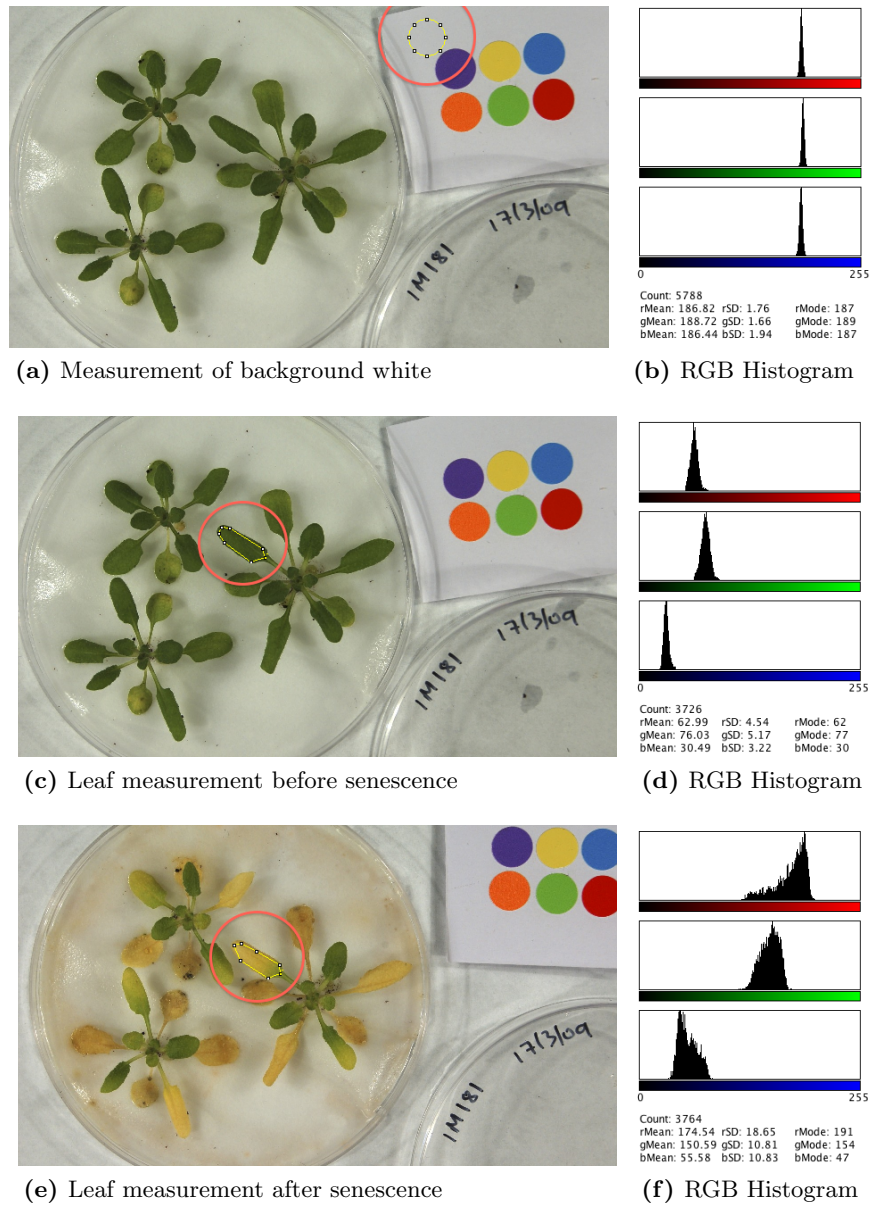


Figure 2.4 – Illustration of colour histogram analysis performed in ImageJ before and after dark-induced senescence

After masking an area of the image in ImageJ, a colour histogram can be obtained describing the red, green and blue components of the pixels in the area. In (a), the measurement of the white background is being made and the lasso area can be seen within the red highlighted area. The consequential histogram for this lasso can be seen in (b) and shows that each of the red, green and blue components are approximately equal as is expected for a white area. (c) shows a leaf prior to senescence being lassoed to obtain the colour histogram shown in (d) whilst (e) shows the same leaf after senescence which gives the colour histogram shown in (f). A brief comparison of the red and green components in both (d) and (f) shows that the ratio between them can be used as a quantification of the progress of senescence

4 as shown in Figure 2.4. Once an area was lassoed, a colour histogram was produced by using the menu option Analyse > Tools > Color Histogram. This allowed the pixel counts in the lassoed area for each intensity (0–255) of red, green and blue to be exported as a text-based contingency table. These tables were collated into text files with one file containing all the time points for a single observation point.

By sourcing the R script shown in Appendix H and using the command:

```
gatherData("path")
```

the data collected in ImageJ64 was normalised using the observed white reference point to calculate the transformation to apply across the whole image. As the paper in the white reference area was not absolute white but averaged 200 for each of red, green and blue over many images, each image was adjusted to bring the mean of each component in the white reference area to 200 by using the procedure shown in Procedure 2.3.

Procedure 2.3 Transformation of photographs to obtain white balance

If the intensity of the components of any given pixel of the image are defined by $\{R, G, B\}$, the mean intensity of the components of all the pixels in a white area of the image are defined by $\{\bar{R}_w, \bar{G}_w, \bar{B}_w\}$ and the expected RGB values for the white area are all 200, the white balance normalised transformation of the pixel can be described as:

$$\begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} = \begin{bmatrix} R \\ G \\ B \end{bmatrix} \times \begin{bmatrix} 200/\bar{R}_w & 0 & 0 \\ 0 & 200/\bar{G}_w & 0 \\ 0 & 0 & 200/\bar{B}_w \end{bmatrix}$$

If the mean intensity of the components of the observed white balance normalised leaf area of the image are then defined as $\{\bar{R}'_{obs}, \bar{G}'_{obs}, \bar{B}'_{obs}\}$, the level of senescence for the observed area can be calculated by:

$$sen = \frac{\bar{R}'_{obs}}{\bar{G}'_{obs}}$$

Once normalisation had been applied to the image, the level of senescence was estimated using the second equation in Procedure 2.3, whereby the ratio between the mean red and the mean green observations of a leaf was used to provide a value which could be used to make a direct comparison of relative senescence between any two rosettes across the experiment. This ratio is typically around a value of 0.5 to 0.7 prior to the onset of yellowing and progresses quickly to a value of around 1.2 after complete yellowing has occurred. At a point where half the leaf has yellowed, the value is almost exactly 1.0.

The senescence ratios were used to plot curves of the onset of senescence against time. Where a visible difference could be identified between the curves of an altered expression line and those of the Col-0 wild type, this indicated that an altered phenotypical response was occurring and justified the continued study of those lines.

A more refined analysis of these differences was obtained by fitting a logistic curve to the level of senescence of each observed rosette over time using the R function,

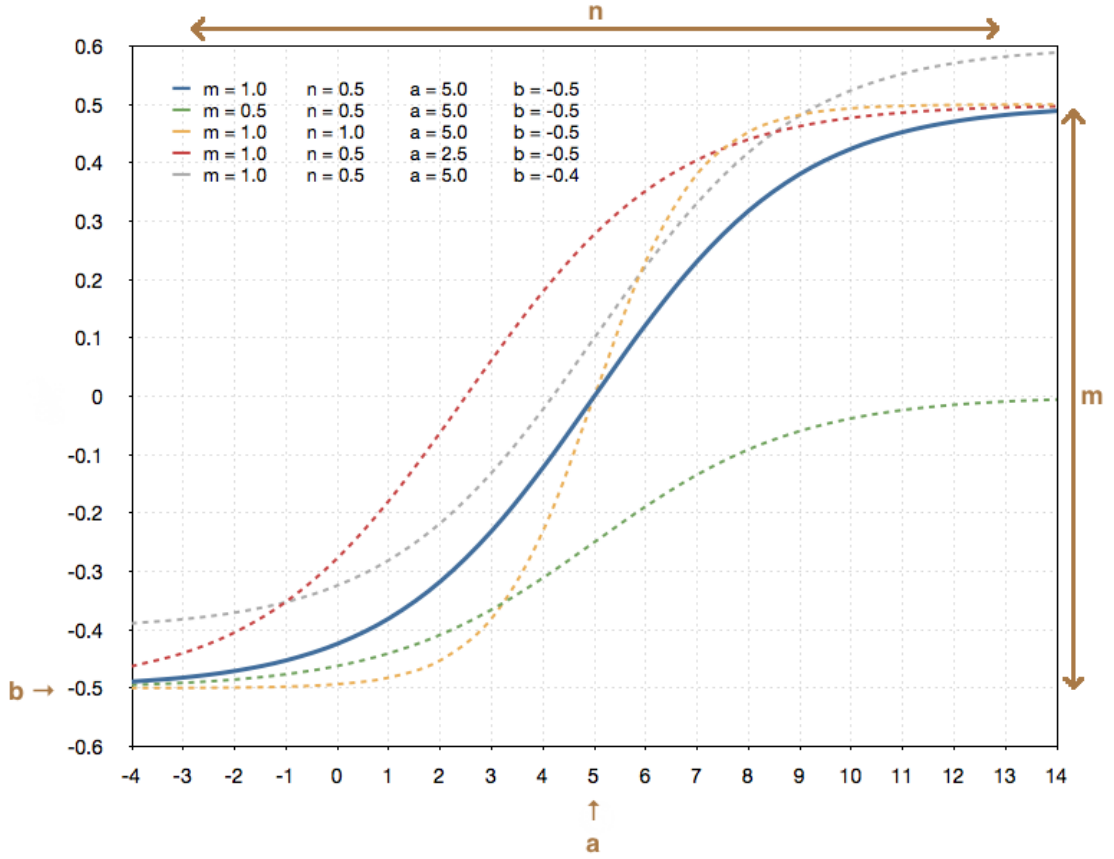


Figure 2.5 – An example of a logistic curve with model parameters shown

This plot shown demonstrates the typical shape of a logistic curve with various parameter adjustments. The four parameters and a representation of their effect on the curve are shown in brown. Using the blue curve as a baseline, the green curve demonstrates a halving of the m parameter which results in half the original amplitude. The orange curve demonstrates a doubling of the n parameter which results in a curve with greater maximum gradient. Reducing the a parameter causes the curve to shift left along the horizontal axis as shown in the red curve, whilst increasing the b parameter causes the curve to shift up the vertical axis as shown in the grey curve.

`nls` to find a non-linear curve for the provided model:

$$y \sim m \times \left(\frac{1}{1 + \exp(n(a - x))} \right) + b \quad (2.2)$$

Where m is the amplitude of the response, n is a parameter of the maximum gradient of the curve, a is the value of x at the point of maximum gradient and b is the value of y at the start point of the curve. The effects of each of these parameters on the shape of the curve are shown in Figure 2.5.

When applied to the senescence screen data, y refers to the level of senescence whilst x refers to time. The models were fitted to the yellowing data of plant lines using an adaptation of the script shown in Appendix I. In this script, the yellowing data of the biological replicates for each plant line, including the wild type, are loaded as matrices. A custom function, `plotLogistic()`, finds the parameters of a logistic curve which fits the profile of the wild type yellowing data across all replicates. These parameters lead

to the calculation of the time and level of yellowing at the point of maximum rate of change in the wild type curve. For each plant line, including the wild type, a logistic curve was fitted to the yellowing data of each biological replicate and the parameters of the curve used to identify key values which compare its shape to that of the mean wild type curve. The values used for this comparisons are:

- The level of yellowing in the altered expression mutant at the time when the mean wild type curve is undergoing maximum rate of change.
- The time at which the altered expression mutant reaches the same level of yellowing as the mean wild type curve during its maximum rate of change.

The two values were then independently processed using an analysis of variance across all lines using the `aov()` function of R. By assuming that the between-replicate variability is common across all lines, the test is therefore more powerful than individual comparisons with the wild type for each line. p -values indicating whether the mean value of each line was different from the wild type were then returned, indicating the significance of the difference observed in a comparison with the wild type. Where ($p \leq 0.05$) in both comparisons this indicates a significant difference between the logistic curves of the two plant lines and therefore a significant difference in their phenotypical response to dark-induced senescence.

Where a logistic curve could not be fitted because insufficient data existed, plant yellowing was graded by eye as an alternative.

2.6 Plant Transcriptomics

This section describes the methods associated with the extraction and consequential analysis of DNA and RNA samples from plant tissue which was harvested, frozen in liquid nitrogen and stored at -80°C .

2.6.1 DNA Extraction

DNA extraction was achieved by using the REDExtract-N-Amp (Sigma) kit. This involved bringing a heating block to 95°C before removing a paper punch-hole sized section of leaf tissue and placing it in a 2 mL micro-centrifuge tube. 100 μL of extraction solution from the REDExtract-N-Amp kit was used to submerge the leaf tissue before the tube was sealed and placed into the pre-heated heating block. Once exactly 10 minutes had passed, the tube was removed from the block and 100 μL of dilution solution was added to the contents before resealing the tube and mixing the contents with a desktop vortexer to halt the reaction. The solution, now containing extracted DNA, was then stored at between $2-8^{\circ}\text{C}$.



Figure 2.6 – An illustration of the location of primers used in testing for homozygous T-DNA insertion

Three primers are used for testing whether the T-DNA insertion is homozygous. Two pairs are made between these primers. If the left and right primers can produce an amplicon, the T-DNA is missing as its presence would provide too much distance between the primers. If the LBb1.3 and right primer can produce an amplicon, the T-DNA is present as the LBb1.3 primer can only bind if the T-DNA sequence exists. After both pairs have been amplified, the line being tested only has a homozygous insertion if the only pair which can produce an amplicon are the LBb1.3 primer and the right primer.

2.6.2 T-DNA Insertion Screening

To test the location and homozygosity of the T-DNA insert in SALK reduced-expression mutants (Alonso *et al.*, 2003), primers are required for simple PCR reactions. For each SALK line, two gene-specific primers are required alongside one primer specific to the inserted T-DNA sequence. The sequences of all three primers are available directly from SALK (<http://signal.salk.edu/tdnaprimers.2.html>) and their relative binding locations are shown in Figure 2.6.

Three PCR reactions per plant line were used to determine the presence and homozygosity of the T-DNA insert. The reactions were prepared in the following combinations of primers and template:

Reaction Index	Template	Forward Primer	Reverse Primer
1	Genomic DNA	Left Gene-specific	Right Gene-specific
2	Plant Line DNA	Left Gene-specific	Right Gene-specific
3	Plant Line DNA	T-DNA Specific	Right Gene-specific

For each reaction, PCR ReadyMix from the REDEExtract-N-Amp kit (Sigma) was used as this accounts for the components of the extraction and dilution solutions used in Section 2.6.1. The protocol for the PCR ReadyMix demands that the reaction components be mixed in a 250 μL PCR tube to a total volume of 20 μL in the following ratios, so that $x + 2y = 6 \mu\text{L}$:

Reagent	Volume
Nuclease-free H ₂ O	$x \mu\text{L}$
REDEExtract-N-Amp PCR Readymix	10 μL
Forward primer	$y \mu\text{L}$
Reverse primer	$y \mu\text{L}$
Leaf disk extract	4 μL
Total volume	20 μL

The prepared tubes were placed in a PCR thermo-cycler using a programme of the following parameters in which the annealing temperature has been optimised for the specific primer pairs:

Step	Temperature	Time	Cycles
Initial denaturation	94 °C	3 minutes	1
Denaturation	94 °C	30 seconds	30–35
Annealing	45–68 °C	30 seconds	
Extension	72 °C	1 minute	
Final extension	72 °C	10 minutes	1
Hold	4 °C	∞	

The products were then separated in a 1.5% electrophoresis gel alongside 1Kb Plus DNA Ladder (Invitrogen) to identify whether the region between the primers had been amplified by the reaction. Only reaction 1 and reaction 3 should be able to produce a product where the T-DNA insert exists and is homozygous. Where reaction 2 produces a product, this indicates that the insert does not exist in at least one strand. The conclusions to be drawn from the results are shown here:

Conclusions	Reaction		
	1	2	3
Check PCR programme and primers	✗	✗	✗
Retry reaction 1 and optimise primers	✗	✗	✓
Retry reaction 1 and optimise primers	✗	✓	✗
Retry reaction 1 and optimise primers	✗	✓	✓
Check extracted DNA	✓	✗	✗
T-DNA insert is present and homozygous	✓	✗	✓
T-DNA insert is not present	✓	✓	✗
T-DNA insert is present but heterozygous	✓	✓	✓

In the case of the final conclusion shown, it may be possible to obtain a homozygous line by screening the next generation in which 1 in 4 individuals should have a homozygous T-DNA insert.

2.6.3 RNA Extraction

Two different methods of RNA extraction were used. Where the tissue to be extracted from was of a mixed sample and was to be used for RT-PCR expression analysis, the Trizol method was used as it is less expensive. Where the tissue was small in volume from a single leaf and the RNA was to be used for microarray analysis, the *mirVana* method was used as this led to better yields from small samples.

2.6.3.1 Trizol Method

So as to create a mixed sample which could be easily aliquoted, a pestle and mortar were used to grind leaves of the same plant line together. Liquid nitrogen was added to the ceramic mortar so as to reduce its temperature to below -80°C whilst the pestle was also submerged in further liquid nitrogen. Once the liquid nitrogen had stopped evaporating violently, a mixed sample of leaves were transferred from their foil pouch into the mortar where they remained in suspension as the liquid nitrogen continued to evaporate. As soon as the liquid nitrogen was completely evaporated, the pre-chilled pestle was used to grind the leaves into a fine powder before they were quickly transferred with a chilled metal spatula into a chilled and labelled 15 mL falcon tube which was then sealed and stored in liquid nitrogen. Approximately 250 μL of the leaf tissue was later transferred from the falcon tube to a chilled 2 mL micro-centrifuge tube for RNA extraction whilst any remaining tissue was stored at -80°C .

Using a fume hood, 1 mL of Trizol reagent was added to each micro-centrifuge tube containing leaf tissue and the leaf tissue was homogenised by inversion of the sealed tubes. The tubes were then kept at room temperature for 5 minutes before 200 μL of chloroform were added and the sealed tubes shaken vigorously for 15 seconds, resulting in an orange solution which was allowed to stand for 3 minutes at room temperature. The tubes were then centrifuged at $8,800 \times g$ for 15 minutes at $2-8^{\circ}\text{C}$. This separated the liquid into three layers: a lower phenol-chloroform layer, a layer of cell tissue and an upper aqueous layer of approximately 600 μL containing RNA and DNA from the cells. The upper layer was carefully removed without disturbing the other layers and transferred to clean 1.5 mL micro-centrifuge tubes before adding 500 μL of isopropanol to each. This solution was incubated at -20°C for between 1 and 2 hours to allow the DNA and RNA to precipitate.

After incubation, the tubes were centrifuged at $8,800 \times g$ for 20 minutes at $2-8^{\circ}\text{C}$ so as to pellet the RNA and DNA at the base of the tubes. The supernatant from each tube was carefully removed and discarded before being replaced with 1 mL of 75% ethanol to clean the pellet. The pellet was resuspended into the ethanol using a vortexer and/or physical agitation of the pellet with a pipette tip. Once suspended, the tubes were centrifuged at $8,800 \times g$ for 10 minutes at $2-8^{\circ}\text{C}$. The supernatant was once again carefully discarded and the pellet was allowed to air dry for 5 minutes before being resuspended in 100 μL of RNase free water.

To purify the total RNA samples, they were then treated with the RNA cleanup procedure of the RNeasy mini kit (Qiagen). These steps were performed quickly, but at room temperature as described in the kit manual. $2 \times 40 \mu\text{L}$ aliquots of RNase free water was then used to elute RNA from the column membrane. Purified RNA samples were stored at -80°C .

2.6.3.2 *mirVana* Method

Where the extraction of higher quality total RNA samples was required, the *mirVana* miRNA isolation kit (Ambion) was used as an alternative to the Trizol method shown above. This kit enabled the extraction of higher yields of total RNA, including a high yield of microRNA fragments, at room temperature, in less time than the Trizol method and with substantially less degradation, making it an excellent choice for microarray samples and samples where a lower yield of total RNA was expected such as those obtained during late stages of senescence.

Where extractions were for a single leaf, samples were ground directly in a 2 mL micro-centrifuge tube using a specially shaped drill bit attached to a high-torque drill. The drill bit was frozen in liquid nitrogen so as to prevent the frozen leaf tissue from defrosting during drilling and the tube was only drilled long enough, for around 45 seconds, to disrupt the cells of the leaf. To lyse the cells, 500 μ L of Lysis/Binding Buffer and 50 μ L of miRNA Homogenate Additive from the *mirVana* kit were added to the tubes and they were mixed with the samples by inversion to homogenise the contents. The tubes were then placed on ice.

After 10 minutes, the tubes were moved to a fume hood and 500 μ L of Acid-Phenol:Chloroform (Ambion) were added to each before vortexing the samples for between 30 and 60 seconds. The samples were then immediately centrifuged at $8,000 \times g$ for 10 minutes at room temperature. The upper aqueous phase was then removed by careful pipetting and transferred to a fresh 1.5 mL micro-centrifuge tube.

600 μ L of 100% ethanol were added to each of the tubes and mixed by pipetting. The sample was pipetted onto a supplied filter cartridge in 700 μ L aliquots and centrifuged at $8,000 \times g$ for 15 seconds and the flow through discarded. Once all the sample had passed through the filter, 700 μ L of miRNA Wash solution 1 were added to each filter column before centrifuging at $8,000 \times g$ for 10 seconds and discarding the flow through. 500 μ L of Wash solution 2/3 were added to the columns before they were centrifuged at $8,000 \times g$ for a further 10 seconds, discarding the flow through. A further 500 μ L of Wash solution 2/3 were then added to the column before centrifuging again at $8,000 \times g$ for 10 seconds and discarding the flow through once more. The filter cartridges were centrifuged a final time at $8,000 \times g$ for 1 minute to remove all traces of ethanol from the filter.

To elute the RNA from the filter cartridges, they were placed in fresh collection tubes, as supplied, and the RNA was eluted from them by placing 50 μ L of nuclease-free water pre-heated to 60 °C on the centre of the filter and then centrifuging at $10,400 \times g$ for 30 seconds. To increase the concentration of the RNA, the flow through was returned to the filter a second time a centrifuged at $10,400 \times g$ for a further 30 seconds. RNA samples were stored at -80 °C.

2.6.4 RNA Concentration Measurement with a NanoDrop

The NanoDrop ND-1000 (Thermo Scientific) offered a cuvette-free method of performing a spectrophotometer analysis of very small RNA samples. For this, 1.5 μL of elution buffer was placed onto the fibre-optic receiver of the NanoDrop and the reading arm lowered to make contact with the drop and cause it to bridge the 1mm gap. The software operating the NanoDrop was then instructed to configure for the RNA-40 settings and to take a blank measurement to calibrate for the background absorbance of the elution buffer. The droplet was removed with absorbent tissue and the pedestals cleaned with 70% ethanol to remove any residues and complete the calibration process.

To make measurements of the RNA samples 1.5 μL of the first sample was placed on the pedestal and the reading arm lowered to make contact with the droplet. The software was instructed to take a measurement and this returned a plot of the absorbances between wavelengths of 220–350 nm. Additionally, the concentration of RNA in the sample was returned as units of $\text{ng}/\mu\text{L}$ as well as a ratio of the absorbance at 260 nm and 280 nm. This ratio signifies the purity of the sample and should be around 2.0 in very pure samples. Once the measurement was taken, the pedestals were cleaned with ethanol once more and the next sample loaded until all samples were analysed.

2.6.5 Bioanalysis of total RNA Samples

To verify the integrity of total RNA samples, they were analysed using the Bioanalyzer 2100 Expert (Agilent) which offers highly reproducible analysis of fragment sizes in very small samples of total RNA over short periods of time using an RNA 6000 Nano Kit (Agilent). In preparation for this analysis, each of the reagents included in the kit were brought to room temperature for 30 minutes. 550 μL of RNA 6000 Nano gel matrix were transferred to a filter column and this was then centrifuged at $3,200 \times g$ for 10 minutes to pass the gel through the filter. Once this was completed, the RNA 6000 Nano dye concentrate was vortexed and then centrifuged for 10 seconds to collect the dye concentrate at the bottom of the tube. 1 μL of the dye concentrate was added to at 65 μL aliquot of the filtered gel matrix and then vortexed to thoroughly mix. This mix was centrifuged at $10,400 \times g$ to complete the preparation of the reagents.

1.5 μL aliquots in 200 μL PCR tubes were taken of each of up to 12 RNA samples to be analysed and of the RNA 6000 Nano ladder. These were then linearised by heating to 70 °C for 2 minutes before promptly cooling on ice. A new RNA Nano chip was placed on the chip priming station and 9 μL of pre-prepared gel-dye mix were pipetted into the well marked ①. The plunger of the chip priming station was pushed down until held by the clip to pressurise the gel-dye mix and released exactly 30 seconds later. A further 9 μL of gel was added to each of the two wells marked ②. 5 μL of RNA 6000 Nano marker was pipetted into each of the 12 sample wells as well as the well marked with a ladder symbol. 1 μL of RNA ladder was added to the well marked with a ladder system, whilst 1 μL of each RNA sample was added to one of the 12 sample wells. A further

1 μL of RNA 6000 Nano marker was added to unused sample wells.

Once the chip was prepared, it was horizontally mounted and vortexed in a vortex mixer at $1,920 \times g$ for 60 seconds. In order to ensure that the pins of the bioanalyzer were not contaminated, they were bathed in RNazap (Ambion) for 60 seconds and then in nuclease free water for a further 60 seconds, using the specially designed cleaning chips provided as part of the RNA 6000 Nano Kit. Once the pins were clean, the vortexed chip was loaded into the receptacle and the bioanalyzer was left for 20 minutes to run all the samples. The output provides a plot per sample of frequency of different fragment sizes and then produces the equivalent data in the form of a pseudo-electrophoresis gel for comparison between samples. Where the highest frequency peaks are the largest fragments, the RNA has not been substantially degraded, whilst high frequencies of small fragments, or a lack of peaks, indicate that the RNA has been degraded and is not suitable for further analysis.

2.6.6 DNase Treatment

DNase treatment of RNA samples requiring no DNA be present was performed using TURBO DNase (Ambion) by the following methods. Aliquots of up to 44 μL of purified RNA samples were transferred to PCR tubes so as to provide approximately 2 μg of RNA per sample. Where this volume was less than 44 μL , the appropriate volume of nuclease free water was also added to bring the total volume to 44 μL . 5 μL of TURBO DNase 10 \times buffer was added to each of the samples as well as 1 μL (2U) of TURBO DNase enzyme. These tubes, once sealed, were allowed to incubate at 37 $^{\circ}\text{C}$ for 30 minutes, after which time a further 1 μL (2U) of TURBO DNase enzyme was added and the tube allowed to incubate for a further 30 minutes at 37 $^{\circ}\text{C}$.

To halt the reaction, DNase inactivation reagent was resuspended by vortexing and flicking its tube and 10 μL of the reagent were then added to each reaction. The reactions were allowed to incubate at room temperature for 2 minutes with occasional mixing. The reaction tubes were then centrifuged at $10,400 \times g$ for 90 seconds to carry the DNase bound to the inactivation agent to the bottom of the tubes. The supernatant, containing only the RNA and no DNA, was carefully removed and transferred to a new tube. The DNA free samples were then measured for RNA concentration on the NanoDrop as described in Section 2.6.4.

2.6.7 RT-PCR Expression Analysis

Two alternative methods were used to perform RT-PCR expression analysis. illustra Ready-To-Go RT-PCR beads (GE Healthcare) provide both reverse transcriptase and *Taq* to allow a single-step reaction from mRNA to PCR product whilst the use of SuperScript[™] II Reverse Transcriptase (Invitrogen) allows the initial transcription of mRNA back into DNA before a standard PCR is used to detect the concentration of specific fragments.

Using either method, primers were designed to amplify a region of the gene being observed. In genes where introns exist, these primers were located either side of an intron as this allowed reverse transcribed mRNA to be differentiated from genomic DNA in the sample. Primers were typically 20 bases long with a CG/AT ratio very close to 1:1. Where a gene had a T-DNA insert approximately in the centre of its coding region, two pairs of primers were designed: one before the T-DNA and one after so that abundance of partially transcribed mRNA fragments could be assessed.

2.6.7.1 Using illustra Ready-To-Go RT-PCR Beads

illustra Ready-To-Go RT-PCR Beads (GE HealthCare) were used as a fast method of providing RT-PCR results for RNA samples. The supplied first strand primer (pd(T)_{12–18}) was pre-prepared to a concentration of 0.5 $\mu\text{g}/\mu\text{L}$ before being stored at -20 °C. For each reaction, 200 ng of total RNA were used as a template. RNA samples for each altered-expression line were compared with those of the Col-0 wild-type and each sample was tested by RT-PCR twice to provide two replicate observations of the expression.

The tubes for the reactions each held a bead containing desiccated Moloney Murine Leukemia Virus (M-MuLV) reverse transcriptase and *Taq* DNA polymerase. In order to calculate the volume of DEPC-treated water to dissolve the beads in, the following table was used to identify the volumes of each component in the individual reactions:

Component	Volume
First-strand primer pd(T) _{12–18}	1 μL
PCR primer 1	2 μL
PCR primer 2	2 μL
200 ng of template RNA	y μL
DEPC-treated water	z μL
Total volume:	50 μL

y was calculated from the RNA concentration obtained with the NanoDrop using the methods of Section 2.6.4. Once this volume was known, z was easily calculated using the equation $z = 45 - y$. The bead was dissolved in the appropriate volume of water by gentle flicking of the tube and incubation on ice for 5 minutes. Template RNA was added to each tube, followed by the addition of each of the primers. The tubes were then transferred to a PCR thermo-cycler and setup with the following programme:

Step	Temperature	Time	Cycles
Reverse Transcriptase Activity	42 °C	30 minutes	1
Reverse Transcriptase Inactivation	95 °C	5 minutes	1
PCR Denaturation	95 °C	30 seconds	50
Primer Annealing	55 °C	30 seconds	
PCR Extension	72 °C	1 minute	

As the programme reached the 20th, 26th, 32nd, 38th and 44th cycle, the reaction was paused so that a 5 μ L sample taken from each of the tubes. Samples were separated on a 2.0% agarose electrophoresis gel to identify the size and abundance of PCR products. A comparison between the brightness of product bands in the Col-0 wild-type samples versus those of the same time points for the altered-expression line allowed the relative transcription levels of the genes to be identified.

2.6.7.2 Using SuperScript™ II Reverse Transcriptase

SuperScript™ II Reverse Transcriptase (Invitrogen) was used to provide cDNA as a separate reaction from the PCR in the RT-PCR process. This allowed less RNA to be consumed by the RT-PCR process when it was to be used for a further purpose such as a microarray experiment. The first-strand cDNA synthesis was performed following the Invitrogen instructions whereby 100 ng of total RNA were used as a template in the following reaction mix:

Component	Volume
Oligo(dT) _{12–18} (500 μ g/ml)	1 μ L
dNTP Mix (10 mM each)	1 μ L
100 ng of total RNA	y μ L
Sterile, distilled water	z μ L
Total volume:	12 μ L

y was calculated using the total RNA concentration obtained from NanoDrop analysis as described in Section 2.6.4. Once this was known, volume z could be calculated by the formula $z = 10 - y$. This mixture was allowed to incubate at 65 °C for 5 minutes and was then quickly chilled on ice, centrifuged briefly to collect the contents of the tubes and then the following components were added to each:

Component	Volume
5 \times First-Strand Buffer	4 μ L
0.1 M DTT	2 μ L
RNaseOUT™ (40 units/ μ L)	1 μ L
Accumulative volume:	19 μ L

The contents of the tubes were mixed by gentle flicking before incubating at 42 °C for 2 minutes. After this time, 1 μ L (200 units) of SuperScript™ II reverse transcriptase was added to each reaction and mixed by pipetting. The tubes were incubated at 42 °C for 50 minutes before being the reverse transcriptase was inactivated by incubation at 70 °C for 15 minutes.

Once cDNA had been synthesised in the above manner, expression levels of specific transcripts could be compared between lines by using PCR reactions with gene-specific primer pairs as described at the beginning of this Section 2.6.7. Each altered-expression line was compared with expression of the altered gene in the Col-0 wild-type

in two replicate PCR reactions. The reaction mix used for each replicate PCR was as shown:

Component	Volume
10× PCR Buffer	5 μL
50 mM MgCl_2	1.5 μL
10 mM dNTP Mix	1 μL
Forward primer (10 μM)	1 μL
Reverse primer (10 μM)	1 μL
<i>Taq</i> DNA polymerase (5 U/ μL)	0.4 μL
cDNA from first-strand reaction	2 μL
Autoclaved, distilled water	38.1 μL
Total volume:	50 μL

The tubes were then placed into a PCR thermo-cycler and a programme used with the following settings:

Step	Temperature	Time	Cycles
Initial denaturation	94 °C	2 minutes	1
Denaturation	94 °C	30 seconds	50
Annealing	55 °C	30 seconds	
Extension	72 °C	1 minute	

5 μL samples were taken from each tube immediately after the 20th, 25th, 30th, 40th and 50th cycles and these samples were separated on a 2.0% agarose electrophoresis gel to compare the brightness of bands, and hence the abundance of transcripts, between the altered-expression line and the Col-0 wild-type. Where a delay existed in the accumulation of product for the altered-expression line, this was considered an indication that expression was reduced for that line. Where the altered-expression line accumulated product more quickly than the Col-0 wild-type, this was considered an indication that the expression was enhanced for that line.

2.7 Cycle Sequencing of DNA using BigDye[®]

In order to provide sequencing of DNA sequences, BigDye[®] (Applied Biosystems) was used by preparing the following mixture of components for each sequencing reaction in a 200 μL PCR tube:

Component	Volume
Ready Reaction Premix	1 μL
BigDye Sequencing Buffer	2.5 μL
Sequencing Primer (10 μM)	0.5 μL
Template	x μL
Nuclease-free Water	y μL
Total Volume	10 μL

Where x can be calculated from the concentration of the DNA template. The total template to be added must be between 150 – 300 ng. The volume of water is therefore calculated as $y = 6 - x$. After brief mixing, the tubes were placed in a PCR thermocycler set up with the following programme:

Step	Temperature	Time	Cycles
Initial denaturation	96 °C	1 minutes	1
Denaturation	96 °C	10 seconds	25
Annealing	50 °C	5 seconds	
Extension	60 °C	4 minutes	
Cooling	4 °C	∞	1

The resulting product was stored at -20 °C until sequenced using a chromatography system.

2.8 Yeast-1-Hybrid

Yeast-1-hybrid was performed using a standard Clontech Matchmaker kit with some modifications made by Claire Hill of The University of Warwick.

2.8.1 Growth Media Preparation

2.8.1.1 Lysogeny broth (LB) Liquid Media with Antibiotic

LB liquid media was prepared by dissolving the appropriate quantity of LB granules in deionized water in a bottle 25% larger than the volume of water. This was performed using an autoclave to sterilise the solution. During cooling, once the solution had reached 60 °C, a $1/1000^{\text{th}}$ volume of an appropriate antibiotic was added to the solution.

2.8.1.2 LB Agar Plates with Antibiotic

Media for LB agar plates were prepared in the same way as shown in Section 2.8.1.1, with the exception that a 2% w/v of agar powder was also added before autoclaving. Once the solution had cooled sufficiently to hold the bottle, plates were poured in a laminar flow cabinet and allowed to set for one hour before being stored upside down in a sterile bag.

2.8.1.3 YPDA Liquid Media

YPDA liquid media was prepared in a similar manner to LB liquid media as described in Section 2.8.1.1 with the exceptions that desiccated YPDA powder was used in substitute of LB granules and no antibiotic was added during cooling.

2.8.1.4 YPDA Agar Plates

For simplicity of preparation, pre-mixed powders (ClonTech) were used in the preparation of YPDA agar plates. For this, a sachet of powder was dissolved in 500 mL of water which was then sterilised using an autoclave. Once cool enough to touch, plates were poured in a laminar flow cabinet and allowed to set for one hour before being stored upside down in a sterile bag.

2.8.1.5 SD Liquid Media

SD liquid media was prepared by dissolving powdered SD base into a sufficient volume of water. This was then supplemented with powdered nutrients lacking one or more essential amino acids. The nutrients and SD base were dissolved and the solution sterilised by autoclaving.

2.8.1.6 SD Selection Agar Plates

In a similar manner as shown in Section 2.8.1.4, pre-mixed powders (ClonTech) were used in the preparation of SD selection plates. Selected drop-out nutrients were either SD-Leu-Trp (catalogue number 630317) or SD-Leu-Trp-His (catalogue number 630319). Where the addition of 3-Amino-1,2,4-triazole (3AT) was desirable this was performed by adding powdered 3AT once the temperature of the solution had reached 60 °C in a WaySafe cabinet to the desired concentration.

2.8.2 Amplifying Upstream Fragments of the Target Gene

To generate the plasmid for Y1H, the insert for the specific bait construct had to be amplified from genomic DNA. For this PCR reaction, primers were designed to amplify the upstream region of the target gene. Amplification of the inserts was prepared by adding the following components to a 200 μ L PCR tube per insert:

Component	Volume
KOD Buffer #1	5 μ L
25 mM MgCl ₂	3 μ L
10 mM dNTP Mix	5 μ L
Forward primer (100 μ M)	1 μ L
Reverse primer (100 μ M)	1 μ L
Col-4 genomic DNA	1 μ L
Dimethyl sulfoxide (DMSO)	0.5 μ L
KOD enzyme	1 μ L
Autoclaved, distilled water	32.5 μ L
Total volume:	50 μ L

Once prepared, the PCR reaction was run using the following programme:

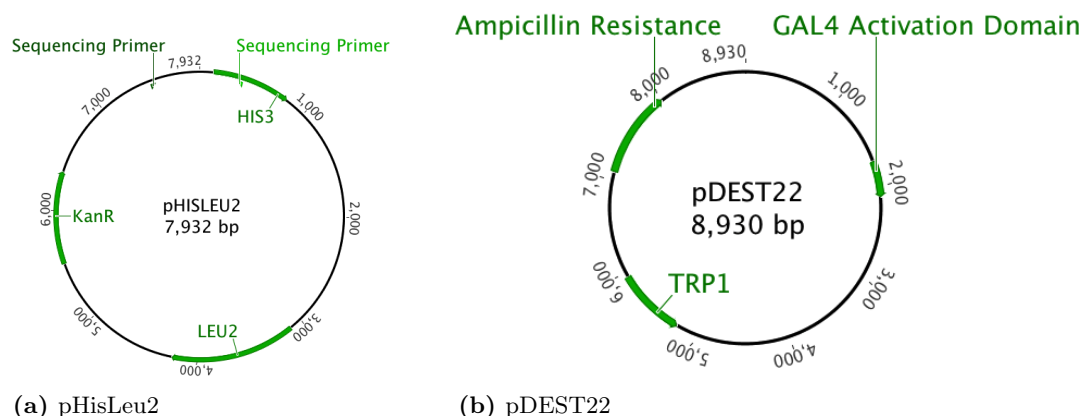


Figure 2.7 – Features of vectors used in yeast-1-hybrid screens

The features of both the bait vector (a) and the prey vector (b), can be seen. The leucine synthesis gene in pHISLeu2 allows transformed cells to survive on media lacking leucine, whilst the tryptophan synthesis gene in pDEST22 allows transformed cells to survive on media lacking tryptophan. Where a transcription factor fused at the N-terminus to the GAL4 activation domain in pDEST22 is able to bind to the sequence in the promoter of the *HIS3* gene in pHISLeu2, the activation domain then causes transcription of *HIS3* and the yeast cell is able to survive on media additionally lacking histidine.

Step	Temperature	Time	Cycles
Initial denaturation	94 °C	2 minutes	1
Denaturation	94 °C	10 seconds	30
Annealing	60 °C	20 seconds	
Extension	72 °C	30 seconds	
Final extension	72 °C	7 minutes	1
Cooling	4 °C	10 minutes	1
Hold	20 °C	∞	

To ensure that proper amplification of the desired products had occurred, 5 μL of each PCR product was run on a 1.5% electrophoresis gel. Where the desired product size existed, the remaining PCR product was separated by being run on a 2.0% electrophoresis gel. The appropriate product was cut from the gel using a clean scalpel blade on a preparatory UV transilluminator and then placed into a clean 2 mL micro-centrifuge tube. DNA was extracted from the gel fragments by following the manufacturers protocols of the QIAquick Gel Extraction kit (Qiagen) to provide 15 μL of clean DNA fragment.

2.8.3 Inserting Fragments into Plasmid Vectors

Inserts were confirmed to be of high enough concentration by running 10 μL of a 1 in 10 dilution on a 1.5% agarose electrophoresis gel to ensure the band was still visible. Adequate inserts were digested in preparation of ligation with an appropriate vector. In the case of the pHISLeu2 vector (see Figure 2.7(a)), restriction sites were EcoRI

and MluI, whilst pBlueScript KS+ vector (Agilent) required either SacI and HindIII or SacI and SpeI depending on which combination of restriction sites were absent in the insert. Fermentas FastDigest enzymes were used, as these provide efficient digestion of the sequence without significant star activity within a short time frame. The digestion reactions were assembled in 200 μ L PCR tubes and contained the following components:

Component	Vector Digestion	Insert Digestion
Nuclease-free Water	14 μ L	11 μ L
Fermentas 10 \times FastDigest Buffer	2 μ L	3 μ L
Vector/Fragment DNA	2 μ L	14 μ L
Restriction Enzyme 1	1 μ L	1 μ L
Restriction Enzyme 2	1 μ L	1 μ L
Total Volume:	20 μ L	30 μ L

These were incubated at 37 $^{\circ}$ C in a PCR thermocycler to allow digestion of the DNA, after which they were heated to 80 $^{\circ}$ C to inactivate the enzymes and prevent star activity. In order to perform the reactions simultaneously the addition of the second enzyme was staggered from the first due to the differing efficiencies. The timings of each pair of enzymes are shown here:

	Vector Digestion			Insert Digestion		
	EcoRI	SacI	SacI	EcoRI	SacI	SacI
	MluI	HindIII	SpeI	MluI	HindIII	SpeI
Minutes @ 37 $^{\circ}$ C after Enzyme 1	0	10	10	15	10	25
Minutes @ 37 $^{\circ}$ C after Enzyme 2	5	5	5	5	20	5
Minutes @ 80 $^{\circ}$ C to Inactivate	5	5	5	5	5	5
Total time (minutes)	10	20	20	25	35	35

In the case of digested vectors, the whole digested product was run on a 0.8% agarose electrophoresis gel to separate the linear and un-cut vector. The linear vector was cut from the gel and extracted using the manufacturer's protocol for the QIAquick Gel Extraction kit (Qiagen). Similarly, the digested inserts were cleaned using the same protocol. A 1 in 10 dilution of both vectors and inserts were run on a 1% agarose electrophoresis gel to compare concentrations. Ligations were prepared at room temperature in 200 μ L PCR tubes by mixing the following:

Component	Volume
5 \times T4 DNA Ligase Buffer	2 μ L
T4 DNA Ligase Enzyme	1 μ L
Digested Vector	x μ L
Digested Insert	y μ L
Nuclease-free Water	z μ L
Total Volume:	10 μ L

Typically, x and y would be between 0.5 and 3 and would ensure that both vector and inserts were present in equal quantities. z could be calculated as $7 - (x + y)$. Once

the ligation was prepared, it was placed in an insulated vessel which was subsequently placed at 2 – 8 °C overnight.

2.8.4 Bacterial Transformation

To clone the plasmids, competent *E. coli* cells were transformed. For this, agar plates with LB and antibiotics were prepared using the preparation described in Section 2.8.1.2. The antibiotic used for selection was kanamycin (50 µg/mL) for cells transformed with pHisLeu2 and carbenicillin (100 µg/mL) for cells transformed with pBlueScript KS+.

To perform the transformation, pre-prepared aliquots of competent *E. coli* cells were defrosted on ice for 10 minutes. The ligation mix from the previous day was added to the cells and mixed gently. The cells were left on ice for a further 30 minutes after which they were heat shocked at 42 °C for 30 seconds before being snapped back on ice for 2 minutes. 500 µL of SOC media (Super Optimal broth with Catabolite repression) was added and the solution incubated at 37 °C for 90 minutes to allow cells to recover.

The cultures were centrifuged at $3,200 \times g$ for 2 minutes to pellet the cells at the bottom of the tube. 300 µL of the media were removed and discarded. The cells were then resuspended in the remaining media by gentle repeat pipetting. All 200 µL were spread over the pre-prepared agar plates. Plates were allowed to dry and were then sealed with Parafilm and stored upside-down overnight at 37 °C so that individual colonies could form.

2.8.5 Confirming the Presence of the Insert

To distinguish between colonies containing uncut vector and those containing a plasmid with an insert, a colony PCR was prepared to examine the size of any insert in the MCS of the vector. Colonies were picked from the LB plate and suspended in 100 µL of sterile water. A master mix was produced which was sufficient for up to 20 PCR reactions and is shown here:

Component	Volume
10× PCR Buffer (Invitrogen)	20 µL
Orange G loading dye	20 µL
10 mM dNTPs	4 µL
50 mM MgCl ₂	6 µL
SABR18 Forward Primer (100mM)	1 µL
SABR19 Reverse Primer (100mM)	1 µL
Taq Polymerase (Invitrogen)	1 µL
Nuclease-free Water	167 µL
Total volume	220 µL

SABR18 and SABR19 primers were designed to bind either side of the MCS and therefore amplify any insert that has successfully ligated into that region. 1 µL of each colony solution was added to 9 µL of the master mix and was set to run on the following PCR programme:

Step	Temperature	Time	Cycles
Initial denaturation	94 °C	2 minutes	1
Denaturation	94 °C	30 seconds	25
Annealing	50 °C	30 seconds	
Extension	72 °C	2 minutes	
Final extension	72 °C	7 minutes	1
Cooling	4 °C	10 minutes	1
Hold	20 °C	∞	

The PCR product was made to run on a 1.5% agarose electrophoresis gel to identify the size of the amplified region. Where an insert was present, the band would be equivalent to 800 bp plus the size of the insert.

2.8.6 Plasmid Cloning

Colonies demonstrating an insert were cultured overnight in 10 mL of LB liquid media, prepared according to Section 2.8.1.1 and containing an appropriate antibiotic as before. Cultures were incubated overnight at 37 °C whilst shaking at 220 RPM.

The following morning, plasmids were extracted from the bulked up colonies by following the manufacturer's protocols for the QIAprep Spin Miniprep (Qiagen). 2 μ L of the resulting plasmid solution was then digested using the restriction enzymes appropriate to the vector and insert and following the same protocol as that for vector digestion shown in Section 2.8.3. 5 μ L of the digested plasmids were made to run on a 1.5% agarose electrophoresis gel to identify that the insert was present in the plasmids and was of the correct size.

To identify the concentration of the extracted plasmids, the NanoDrop was used as shown in Section 2.6.4, but with the exception that the setting for analysing DNA samples was used instead of RNA. This allowed the identification of DNA concentrations in ng/ μ L.

2.8.7 Sequencing the Insert

To check that the insert is of the correct sequence without any polymorphisms, the extracted plasmid was sequenced using insert specific primers to provide sequencing results from both ends of the insert. The protocol shown in Section 2.7 was used and the resulting sequences were combined as a consensus sequence in Geneious (Drummond *et al.*, 2010) to compare with the expected sequence of the *Arabidopsis* genome. Where no polymorphism was found, the plasmid was used to transform yeast cells ready for screening.

2.8.8 Yeast Transformation

During transformation with the bait construct, Y187 strain *Saccharomyces cerevisiae* cells (Clontech) were used. The genotype of this strain is shown:

MAT α , ura3-52, his3-200, ade2-101, trp1-901, leu2-3, 112, gal4 Δ , met, gal80 Δ , URA3 :: GAL1_{UAS}-GAL1_{TATA}-lacZ (Harper *et al.*, 1993)

which indicates that the Y187 strain is an ‘ α ’ mating type and can therefore mate with ‘a’-type cells. It is also deficient in the *leu2*, *trp1* and *his3* genes and therefore cannot survive on media lacking any of these essential amino acids (leucine, tryptophan and histidine) unless complemented by a plasmid containing these genes. The bait plasmids contain the *LEU2* gene and, therefore, successfully transformed cells can then survive on leucine deficient media.

In order to transform yeast cells, a culture of the Y187 cells was grown overnight in YPDA solution (yeast peptone dextrose adenine) prepared according to Section 2.8.1.3. Yeast cells were cultured in a 10 mL aliquot of YPDA liquid media in a 50 mL falcon tube. The culture was grown by incubating at 30 °C and shaking at 200 RPM overnight. The following morning, yeast cultures were divided into 1 mL aliquots in 1.5 mL microcentrifuge tubes before being gently pelleted in a centrifuge at $1,600 \times g$ for 5 minutes.

Once pelleted, the supernatant was removed and the yeast cells resuspended in 1 mL of 0.1 M lithium acetate. The cells were once again pelleted at $1,600 \times g$ for 5 minutes before removing the supernatant and resuspending the yeast cells in a further 1 mL of 0.1 M lithium acetate. The tubes containing the suspended cells were incubated in a 30 °C water bath for between 1 and 2 hours. Whilst incubating, a solution containing the plasmid for transformation was prepared by mixing 290 μ L PEG, 4 μ L of carrier DNA (denatured herring testes DNA) and 3 μ L of plasmid DNA in a 1.5 mL microcentrifuge tube. This was also incubated in the same water bath to allow the solution to reach 30 °C.

Once 1 – 2 hours had passed, the cells were resuspended in the lithium acetate by gentle pipetting and 100 μ L of the cell suspension was mixed into the PEG/DNA solution in a swirling motion. The mixes were incubated at 30 °C for a further 50 minutes and then transferred to a 42 °C water bath to heat shock the cells for 15 – 30 minutes. Immediately after removal from this water bath, the yeast cells were centrifuged at $2,400 \times g$ for 3 minutes. The supernatant was removed and the cells were resuspended in 200 μ L of sterile water. The resuspended cells were spread over an SD-Leu selection plate. Once dry, the plates were sealed and incubated upside down at 30 °C until sufficient growth had occurred.

Once individual yeast colonies could be identified on the selection plates, four colonies were selected for restreaking onto fresh SD-Leu plates divided into four quadrants. These plates were sealed in the same manner as before and incubated upside down at 30 °C until sufficient growth had occurred. After this time, the plates were stored at between 2 – 8 °C to prevent further growth of cells.

2.8.9 Screening Against a Transcription Factor cDNA Library

The library used in these screens was a transcription factor (TF) cDNA library transformed into AH109 strain cells (Clontech) of *Saccharomyces cerevisiae* via the pDEST22 gateway vector (see Figure 2.7(b)) supplied with an Invitrogen ProQuest yeast-2-hybrid system. The genotype of the AH109 cells was:

MATa, trp1-901, leu2-3, 112, ura3-52, his3-200, gal4 Δ , gal80 Δ , LYS2 :: GAL1_{UAS}-GAL1_{TATA}-HIS3, GAL2_{UAS}-GAL2_{TATA}-ADE2, URA3 :: MEL1_{UAS}-MEL1_{TATA}-lacZ (James *et al.*, 1996)

indicating that the AH109 strain is an ‘a’-mating type and will therefore mate with Y187 strain cells as have been transformed above with the bait construct. Disruption to the *trp1*, *leu2* and *his3* genes means that the cells cannot grow on media lacking any of the essential amino acids, tryptophan, leucine or histidine, unless complemented by a plasmid containing those genes. When transformed with the prey construct, the *TRP1* gene found in the pDEST22 vector allows the cells to survive on media lacking tryptophan. If the cells mate with Y187 cells transformed with the bait construct, the pHisLeu2 vector complements the offspring with the *LEU2* gene which then allows survival on media deficient in both leucine and tryptophan. If the transcription factor is able to bind to the promoter region in the bait construct, activation of the *HIS3* gene in the pHisLeu2 vector results in the ability to survive on media lacking each of leucine, tryptophan and histidine.

The TF library consists of approximately 1,400 transformed cultures of *Saccharomyces cerevisiae* and are grouped into wells containing 12 transformants each. Two 96-well plates are used to hold the entire library and a duplicate copy of the library is present in a second pair of plates. The positions of individual transformants vary between copies of the library so that positive results might be interpretable without the need for sequencing the transcription factor found in the pDEST22 vector.

2.8.9.1 Culturing Yeast Colonies

500 μ L of SD-Trp liquid media was added to each of the wells of 4 \times deep 96-well plates. A 96-well colony picker was then used to transfer a droplet of cells from the stock TF library plates into the respective deep well plates. The deep well plates were sealed with a gas permeable membrane and incubated at 30 °C.

Several days later, transformants carrying the promoter ‘bait’ fragments were cultured by placing a scoop of cells into 10 mL of SD-Leu in a 50 mL falcon tube. These were incubated at 30 °C with shaking at 220 RPM.

2.8.9.2 Preparing Yeast Cell Mating

Each of four YPDA plates were prepared for mating by placing 96 \times 3 μ L of the cultured transformant suspension onto their surfaces in the arrangement of a 96-well

plate. Once dried, 3 μ L of each of the wells of the corresponding TF library plate were added directly on top of the dried spots. After drying, the plates were sealed and returned, upside down, to a 30 °C incubator.

2.8.9.3 Transfer of Yeast Cells to Selection Plates

Each YPDA plate, covered in spots of yeast cell growth, was replica plated onto four alternative selection plates using velvet squares. The order of replication was from highest selective pressure to lowest: SD-LTH + 100 mM 3AT, SD-LTH + 50 mM 3AT, SD-LTH, SD-LT. Selection plates were sealed with Parafilm and incubated at 30 °C. Although no longer needed, YPDA plates were stored at between 2 – 8 °C as a precaution in case the screen were to be repeated.

The following day, once the replica selection plates had had enough time to allow yeast cell growth, excess number of cells were removed from the plates by pressing each plate into three fresh sheets of velvet. After this, the plates were resealed and returned to the 30 °C incubator.

2.8.9.4 Restreaking of Growing Colonies

After a number of days of growth, some of the spots had colonies growing upon them. The SD-LT plates were able to support any diploid yeast cells which had been formed by the mating of a-type and α -type cells. In the case of SD-LTH plates, only those cells where the gene product of pDEST22 was able to bind to the DNA inserted into pHisLeu2 were able to grow. If excessive growth was found on the plate with no added 3AT, attention was drawn to increasing levels of 3AT which inhibits the action of the HIS3 gene and therefore allows less cell growth. The plate with highest selective pressure that still maintained cell growth was selected for observation.

The plates were photographed using a gel doc imaging system with upper lighting. After this, 9 cm SD-LT plates were used to restreak colonies which were able to grow into small patches, 50 per plate. Colonies that had grown on the 15 cm selection plates were transferred to a patch on the smaller plate using the tip of a 20 μ L pipette. A record was made of the location from which the colony was picked from the selection plate and of the patch number to which it was restreaked. Once all growing colonies had been transferred, the patch plate was sealed with Parafilm and transferred to a 30 °C incubator.

2.8.9.5 Identification of Interacting Transcription Factors

Since as many as 12 library transformants were present per spot, and to ensure that the one responsible for the transcription of the HIS3 gene in each colony has been reliably identified, the transcription factor was re-amplified and sequenced. Re-amplification was performed by colony PCR, in which 10 μ L of 20 mM NaOH was added to each well of a 96-well PCR plate. For each of the colonies restreaked the previous day, a

colony sized scoop of cells were removed from the patch plate using a pipette tip and placed into one of the wells of the 96-well plate. These colonies were incubated in a PCR thermocycler for 10 minutes at 99 °C to lyse the cells. During this time, a master mix for 96 PCR reactions was prepared consisting of the following components:

Component	Volume
Nuclease-free water	1.45 mL
10× PCR buffer (Invitrogen)	200 μ L
Orange G dye	200 μ L
10 mM dNTPs (Invitrogen)	40 μ L
50 mM MgCl ₂ (Invitrogen)	60 μ L
100 mM Forward primer (SABR447)	5 μ L
100 mM Reverse primer (SABR448)	5 μ L
Taq polymerase (Invitrogen)	10 μ L
Total volume	1,970 μL

A fresh 96-well PCR plate was prepared by transferring 19 μ L of the above mix to each well. Once the yeast cells had been lysed in NaOH, 1.2 μ L from each well were added to the corresponding wells of the PCR mix plate. The plate was run on the following PCR programme:

Step	Temperature	Time	Cycles
Initial denaturation	94 °C	2 minutes	1
Denaturation	94 °C	30 seconds	40
Annealing	55 °C	30 seconds	
Extension	72 °C	3 minutes	
Final extension	72 °C	7 minutes	1
Cooling	4 °C	10 minutes	1
Hold	20 °C	∞	

Once the reaction was completed, 3 μ L of the product were run on a 0.75× TAE pre-cast agarose gel for 10 minutes to identify whether a fragment had been amplified before sequencing.

A 96-well MultiScreen HTS plate (Fisher) was pre-wet by the addition of 50 μ L of 0.1 mM Tris (tris(hydroxymethyl)aminomethane) pH 8 solution. All remaining PCR product for positive colonies was transferred into the corresponding well of the HTS plate and the plate was placed on a vacuum manifold to draw the liquid through the filter. Once fully drawn, a foil seal was used to prevent the escape of liquid from the bottom of the filter and 35 μ L of 0.1 mM Tris pH 8 solution was added to each well and a shaker used to aid the re-suspension of DNA fragments from the filter. The suspensions were then transferred to a final plate for NanoDrop analysis of their concentrations.

A representative sample of the wells were used to identify the mean concentration of DNA in each and an appropriate volume calculated to provide 20 – 50 ng for sequencing. Sequencing was performed using the methods of Section 2.7 where SABR447 was

used as a primer. The results of this were processed in Geneious and MegaBLAST (Altschul *et al.*, 1990) to identify the transcription factors which were able to bind to the DNA of the target gene.

2.9 Microarray Analysis of Reduced-Expression Mutants

Whilst validating the transcriptional network found by theoretical modelling, microarrays were used to identify downstream effects of reduced-expression in key genes of the network.

2.9.1 Preparing and Harvesting Samples

Selected reduced-expression mutants were sown and grown using the methods described in Section 2.5.1, 24 plants per line. Three weeks after sowing, plants were transferred to petri-dish observation plates following the same preparation as that of Section 2.5.2. Observation plates were stored at 20 °C in absolute darkness, but photographed each day between 12:00 and 14:00 following the same procedures as Section 2.5.2.

During the software analysis of the colour of leaf 4 of each rosette, a score was produced indicating the level of senescence. Four randomly selected rosettes of each line were each pre-allocated one of five target senescence scores of either 0.7, 0.8, 0.9, 1.0 or 1.1, leaving four spare plants in each line. Each day, any rosettes which were near to reaching or had only just passed their senescence score were harvested. Harvesting took place at between 17:00 and 19:00 each day. Leaves 3, 4 and 5 were individually detached using clean scissors and placed into screw cap 2 mL tubes before being submerged in liquid nitrogen and later stored at -80 °C.

2.9.2 aRNA Synthesis and Purification

Once selected samples had been isolated and pure total RNA had been extracted from those samples, these were used to produce high concentrations of aRNA using the MessageAmp II kit (Ambion) according to the manufacturers instructions. This consisted of five stages of amplification.

During the first stage, first strand cDNA synthesis was performed by incubating the following mixture at 42 °C for 2 hours:

Component	Volume
RNA sample	11 μL
T7 Oligo (dT) Primer (Ambion)	1 μL
Pre-incubate the above at 70 °C for 10 minutes.	
10 \times First Strand Buffer (Ambion)	2 μL
10 mM dNTP Mix (Ambion)	4 μL
RNase Inhibitor (Ambion)	1 μL
ArrayScript Enzyme (Ambion)	1 μL
Total volume:	20 μL

In the second stage, first strand cDNA was used to synthesise the second strand cDNA by adding the following to the reaction and incubating at 16 °C for 2 hours:

Component	Volume
First strand cDNA from above	20 μL
10 \times Second Strand Buffer (Ambion)	10 μL
Nuclease-free water	63 μL
10 mM dNTP Mix (Ambion)	4 μL
DNA Polymerase (Ambion)	2 μL
RNase H (Ambion)	1 μL
Total volume:	100 μL

The third stage purified the resulting cDNA by binding the DNA to a spin column and washing away contaminants with ethanol solutions before eluting purified cDNA with 20 μL of nuclease-free water pre-heated to between 50–55 °C.

The fourth stage synthesised aRNA from the cDNA by use of transcription enzymes. The following components were mixed and then incubated at 37 °C for 14 hours:

Component	Volume
Purified cDNA from above	20 μL
75mM T7 ATP Solution (Ambion)	4 μL
75mM T7 CTP Solution (Ambion)	4 μL
75mM T7 GTP Solution (Ambion)	4 μL
75mM T7 UTP Solution (Ambion)	4 μL
10 \times T7 Reaction Buffer (Ambion)	4 μL
T7 Enzyme Mix (Ambion)	4 μL
Total volume:	44 μL

After 14 hours, the reaction was quenched by the addition of 60 μL of nuclease-free water. The final stage purified the synthesised aRNA by binding the aRNA to a spin column and washing away contaminants with ethanol based solutions before eluting the purified aRNA with 50 μL of nuclease-free water pre-heated to between 50–55 °C.

2.9.3 Labelling with Fluorescent Dyes

Each sample was required to be labelled by Cy3 and Cy5 dyes. Therefore, for 15 samples, 30 labelling reactions were prepared as shown. The NanoDrop spectrophotometer was used, as described in Section 2.6.4, to determine the concentration of aRNA for each sample so that 5 μg or more could be used in a labelling reaction. This was combined with 0.5 μL of random nonomers (3 $\mu\text{g}/\mu\text{L}$) and 0.5 μL of RNase out (Invitrogen), bringing the total volume to 10.5 μL using nuclease-free water. This mixture was incubated for 10 minutes at 70 $^{\circ}\text{C}$. After this incubation, a labelling reaction was prepared as follows:

Component	Volume
Reaction mix from above	10.5 μL
5 \times SuperScript II First Strand Buffer (Invitrogen)	4 μL
0.1 M DTT (Invitrogen)	2 μL
10 mM dATP, dGTP, dTTP, 2 mM dCTP mix (Invitrogen)	1 μL
SuperScript II reverse transcriptase (Invitrogen)	1 μL
Cy3 or Cy5 labelled dCTP (GE Healthcare)	1.5 μL
Total volume:	20 μL

The reaction was incubated in the dark at 42 $^{\circ}\text{C}$ for 2.5 hours before adding 2 μL of 2.5M NaOH and incubating for a further 15 minutes at 37 $^{\circ}\text{C}$. 10 μL of MOPS buffer (3-(N-morpholino)propanesulfonic acid) was then added and the reactions placed on ice. cDNA from the reaction was purified using QIAQuick PCR Purification spin columns and eluting with two aliquots of 30 μL of buffer EB. Concentrations of the labelled samples were obtained by using the NanoDrop spectrophotometer as described in Section 2.6.4, but by measuring absorbance and fluorescence of the dye rather than cDNA.

2.9.4 Preparation of a Balanced Experimental Design

An experimental design which was as balanced as possible was developed by first finding designs in which the comparisons between different biological replicates, differently dyes and different lines were equally abundant. This was achieved using the two scripts shown in Appendix M. The first of these scripts used a sample skeleton file as a pre-defined pattern which had already been optimised to ensure that each of the 5 lines was co-hybridised with each other line exactly 3 times and to ensure that each line was represented equally by both Cy3 and Cy5 labelling. After fitting all permutations of the 15 samples to this skeleton, each was tested to fit two more criteria: each sample must be equally labelled by Cy3 and Cy5 and each sample must be directly co-hybridised with every other sample. All designs which fit this criteria were returned by the first script as a series of tab-delimited text files.

Each design must also demonstrate balance in the co-hybridisation of the biological replicates to be truly unbiased, and this was optimised by the second script. In

this script, a sample of the balanced sample designs obtained by the first script were selected for testing permutations of biological replicates. Each line was represented by 3 biological replicates, but the allocation of these to the sample numbers, optimised above, was not yet defined. Each possible allocation of the biological replicates to the sample numbers were considered by the script and, after each allocation, two metrics were applied to test the distribution of comparisons between biological replicates: the variance of a contingency table listing frequencies of comparisons between replicates and the variance of the rows of a contingency table listing the frequency of replicate labels co-hybridised with specific samples. Those designs which were able to minimise these two metrics were considered optimally balanced.

A custom computer algorithm was used to identify those experimental designs which would minimise the biases. This was performed in two stages in which the initial stage paired sample numbers in such a way that each sample number was labelled with each dye exactly twice to provide four dye-swapped technical replicates across the experiment. The samples from each line also had to be compared exactly once with every other sample in the experiment (i.e. any of samples 1, 2 and 3, representing IM137, must be compared only once to each of 4 through to 15).

Once the first stage was complete and many possible arrangements had been discovered, many of those arrangements were permutation tested for biological replicates in the second stage. Within each line, the three biological replicate labels, A, B and C would be allocated to the three sample numbers representing that line. Since three elements can be orientated in eight different ways, the size of the problem was $8^5 = 32,768$ possible combinations of biological replicate. After each permutation, the design was tested by comparing the frequency of biological replicate hybridisations with each sample as shown in Table 2.3(b) where it can be seen, for example, that sample 1 was hybridised with 2 replicates of other samples labelled A, 1 replicates labelled B and 1 replicate labelled C. The variances of these frequencies for each sample are summed and a minimisation of this sum indicates a better distribution of between replicate comparisons. The sum of 5.000, as shown, is the minimum achievable for 15 samples, 3 biological replicates and 4 technical replicates. A second metric of the optimisation of the design is the variance of sums across frequencies of co-hybridised biological replicate labels as shown in Table 2.3(c). In this case, if an imbalance of comparisons between similarly labelled replicates existed, the variance would be greater than zero.

The design which was implemented whilst hybridising the microarrays is shown in Table 2.3(a). Wash solutions were pre-prepared as shown in Section 2.9.5 and CATMAv4 microarrays were hybridised according to the methods shown in Section 2.9.6.

2.9.5 Microarray Solution Preparations

In preparation of hybridising labelled samples to microarrays, two hybridisation buffers and three wash solutions were required. These were made to the following specifications. Volumes shown are sufficient for eight hybridisations.

Array Number	Labelled with Cy3			Labelled with Cy5		
	Sample	Line	BioRep	Sample	Line	BioRep
1	1	IM137	C	4	IM188	C
2	2	IM137	B	6	IM188	A
3	3	IM137	A	9	IM244	C
4	1	IM137	C	12	IM330	A
5	2	IM137	B	11	IM330	B
6	3	IM137	A	13	Col-0	A
7	7	IM244	A	1	IM137	C
8	8	IM244	B	2	IM137	B
9	5	IM188	B	3	IM137	A
10	15	Col-0	B	1	IM137	C
11	14	Col-0	C	2	IM137	B
12	10	IM330	C	3	IM137	A
13	6	IM188	A	7	IM244	A
14	4	IM188	C	8	IM244	B
15	4	IM188	C	13	Col-0	A
16	6	IM188	A	14	Col-0	C
17	5	IM188	B	11	IM330	B
18	9	IM244	C	5	IM188	B
19	10	IM330	C	4	IM188	C
20	12	IM330	A	6	IM188	A
21	15	Col-0	B	5	IM188	B
22	9	IM244	C	12	IM330	A
23	8	IM244	B	10	IM330	C
24	7	IM244	A	15	Col-0	B
25	14	Col-0	C	9	IM244	C
26	13	Col-0	A	8	IM244	B
27	11	IM330	B	7	IM244	A
28	11	IM330	B	14	Col-0	C
29	12	IM330	A	15	Col-0	B
30	13	Col-0	A	10	IM330	C

(a) Hybridisation Design

Sample	Line & Rep	Hybridise Alongside			Variance
		A	B	C	
1	IM137 C	2	1	1	0.333
2	IM137 B	1	2	1	0.333
3	IM137 A	1	1	2	0.333
4	IM188 C	1	1	2	0.333
5	IM188 B	1	2	1	0.333
6	IM188 A	2	1	1	0.333
7	IM244 A	1	2	1	0.333
8	IM244 B	1	1	2	0.333
9	IM244 C	2	1	1	0.333
10	IM330 C	2	1	1	0.333
11	IM330 B	1	2	1	0.333
12	IM330 A	1	1	2	0.333
13	Col-0 A	1	1	2	0.333
14	Col-0 C	1	2	1	0.333
15	Col-0 B	2	1	1	0.333
Sum:					5.000

	A	B	C	Sum
A	3	3	4	10
B	3	4	3	10
C	4	3	3	10
Sum	10	10	10	

Variance of "sums": 0.000

(b) Sample Comparisons

(c) BioRep Comparisons

Table 2.3 – An optimised experimental design for microarray hybridisations

The experimental design shown in (a) was produced using a computer algorithm designed to meet the pre-determined criteria of the experiment. The design is shown to be optimised by noticing that each sample is labelled by each dye exactly twice; by considering the frequency of each biological replicate label co-hybridised with each sample, shown in (b), where it can be seen that each sample has a minimised variance across those frequencies and; by minimising variance across the sum of replicate-to-replicate comparisons shown in (c).

2.9.5.1 Pre-hybridisation Buffer

Component	Volume or Mass
Bovine Serum Albumin (Sigma-Aldrich)	1.2 g
20× SSC	30 mL
14% SDS	860 μ L
Sterile deionised water	90 mL
Total volume: ~120 mL	

2.9.5.2 Hybridisation Buffer

Component	Volume or Mass
Formamide	112.5 μ L
20× SSC	112.5 μ L
14% SDS	3.15 μ L
4 μ g/ μ L Yeast tRNA (Invitrogen)	56.25 μ L
Sterile deionised water	165.6 μ L
Total volume: 450 μ L	

2.9.5.3 Wash Solution 1

Component	Volume or Mass
20× SSC	25 mL
14% SDS	1.8 mL
Sterile deionised water	223 mL
Total volume: ~250 mL	

2.9.5.4 Wash Solution 2

Component	Volume or Mass
20× SSC	1.25 mL
14% SDS	1.8 mL
Sterile deionised water	247 mL
Total volume: ~250 mL	

2.9.5.5 Wash Solution 3

Component	Volume or Mass
20× SSC	5 mL
Sterile deionised water	995 mL
Total volume:	1 L

2.9.6 Microarray Hybridisation

Microarray hybridisations were performed according to the procedures used by the PRESTA group at The University of Warwick. For this, the pre-hybridisation buffer, prepared according to Section 2.9.5.1, was pre-heated to 42 °C. Once warmed, a Coplin jar was filled with the buffer and eight CATMA microarrays submerged in it. The jar was maintained at 42 °C by placing in an air incubator for 1 hour.

40 pmol of labelled samples to be co-hybridised to each microarray were combined in PCR tubes and freeze dried in complete darkness for between 60 and 90 minutes. Incubated microarrays were washed by shaking in a hybridisation rack in 5 fresh baths of deionised water. The slides were then briefly washed in isopropanol before being transferred to 50 mL falcon tubes and centrifuged for 1 minute at $1,600 \times g$, taking care that the glue of the barcode would not spread onto the spotted area.

A hybridisation container consisting of a platform raised above water in a sealable box was used to support hybridisation chambers which protected the microarrays. The freeze-dried labelled samples were resuspended in 50 μ L of hybridisation buffer, prepared according to Section 2.9.5.2, and incubated at 95 °C for 5 minutes. The samples were then spotted along the length of a microarray which had been placed in a hybridisation chamber before being sealed by a microarray coverslip and encased by the chamber. Once eight chambers had been filled, damp tissue was placed over the chambers to ensure high humidity in the container and the container was incubated at 42 °C for between 16–20 hours. Wash Solution 1, prepared according to Section 2.9.5.3, was incubated at 42 °C at the same time.

After incubation, the microarrays were submerged in Wash Solution 1 and the coverslip removed. The arrays were then washed in Wash Solution 1 by placing in a hybridisation rack and shaking vigorously for 5 minutes on an orbital shaker. After this time, the rack was transferred to a bath of Wash Solution 2 which was used to wash the microarrays vigorously for 10 minutes on an orbital shaker. Following this, four successive washes were completed whilst submerged in Wash Solution 3 for 1 minute per wash whilst being vigorously shaken on an orbital shaker. The microarrays were then briefly immersed in isopropanol before being placed into 50 mL falcon tubes and centrifuged for 1 minute at $1,600 \times g$. Microarrays were then stored in the dark until scanned.

2.9.7 Microarray Scanning and Data Extraction

After hybridisation, the microarrays were scanned using an Affymetrix 428 Array Scanner. Cy3 labelled samples were scanned with a laser of wavelength 532nm whilst Cy5 labelled samples were scanned independently with a laser of wavelength 635nm. Each microarray was scanned at a gain to ensure that the brightness of the spots was maximised whilst minimising saturation as much as possible. Scans were saved as two 16-bit TIFF images per microarray providing $2^{16} = 65,536$ levels of intensity for each channel.

In order to analyse the microarray scans and convert them to numeric data describing the intensity of the channels and their background noise, BioDiscovery ImaGene version 7.0 (BioDiscovery) was used. By providing a grid template to identify the probes by Gene ID and location, ImaGene was capable of outputting a table of values in the range of 1 to 65,536 describing, amongst other measurements, the mean signal, mean background, median signal, median background and flag of each channel of each probe of the microarray. Where the probe fails quality measures imposed by ImaGene, usually due to a low signal, the flag was returned as an integer greater than zero to assist in their identification.

Chapter 3

Microarray Data Analysis

3.1 Chapter Summary

This chapter discusses the development of the MAANOVA package to enhance its ability to identify inadequacies amongst the microarrays in an microarray experiment. Further development to GridCheck, RIPlot and ArrayView have led to a more statistical analysis of their outputs using a principal component analysis which leads to a major axis regression and the identification of genes which are distant from the mean of the regression. This has allowed specific microarrays to be isolated for investigation and corrected. Additional to the developments above, TechRepCheck has been introduced which makes pairwise comparisons between the probe intensities of same-sample replicates. A major axis regression of these pairwise comparisons has allowed the identification of inconsistent replicates at the probe level and therefore to identify which areas of the microarrays require attention.

The developments above have resulted in the identification of 35 microarrays out of 176 which exhibit inadequacies. Of those, 18 were sufficiently corrected by re-analysing the microarray scans whilst a further 4 were re-hybridised to new microarrays. The remaining 13 were affected in only small areas of the microarrays which did not justify any further action to be taken.

Once the microarrays were up to a sufficiently high standard, the intensities were analysed by an ANOVA driven model fit which successfully separated the four effects known to exist amongst the samples: day of harvest, time of day of harvest, the interaction between those main effects and the effects between biological replicates. The levels of each of these effects were analysed by false discovery rate corrected F-tests and those genes which were significantly differentially expressed for each term were analysed. Those genes which were significantly differentially expressed for the day effect totalled 8,878 and the expression data for those genes were extracted so that it may be used as described in further chapters for the identification of theoretical regulatory network models.

3.2 Introduction

In this chapter, the data obtained from the senescence microarray experiment, introduced in Section 1.4, is analysed using a customised version of MAANOVA. The development of this customised version of MAANOVA is described here, followed by its application to the senescence microarray experiment data. The overall aim of this analysis was to provide high-resolution normalised temporal expression data for every gene on the microarrays so that it may be used to model transcriptional networks as described in Chapter 4. Therefore the development of MAANOVA focussed on providing more rigorous quality control tests and a novel form of analysis which allows information about the technical replicates to be used to provide inconsistent replicates with more consistent intensity measurements.

Although the customised version of MAANOVA has been adapted in response to the need to analyse the senescence microarray dataset, these adaptations could be applied to any microarray dataset and this is discussed further in Section 3.5.1. The installation of the WHRIMAANOVA package is essential for this purpose and this is therefore made available on a data CD accompanying this thesis.

3.3 Development of MAANOVA

The features of MAANOVA form a good building block towards the analysis of the senescence experiment, but a number of improvements have been made in order to increase the usefulness of the package, particularly in the area of providing a stronger statistical basis for the initial data quality control. These new implementations are described below and are complemented by a definition of the function parameters listed in Appendix B.

Some of the adaptations described here were presented for the degree of MSc Systems Biology as part of a mini-project of twelve weeks. Further adaptations, building on those produced during the MSc, are also described here and the distinction between these has been clearly identified in the text.

3.3.1 Quality Control

Three functions were provided for the purpose of quality control in the original MAANOVA package and these theoretically form a robust set of tools for identifying flaws which are typical in two-channel microarrays. Some of the typical flaws which may be identified are:

- Non-specific binding, in which large numbers of probes are showing high intensity hybridisation in one or both channels for some, often adjacent, probes.
- High background levels resulting in over correction by background-subtraction normalisation.

- Spatial biases in the channels whereby mean intensity of one channel may exceed the other channel at some point around the edge of the microarray, whilst the opposite edge is the inverse.
- Physical damage to the microarray which fluoresces and artificially amplifies the observed signal in affected probes.
- Under and over-gained scanning of the microarrays, resulting in saturated or dull intensity and a consequential loss of distinction between probes.

However, there was much room for improvement in the presentation of the results, the long term storage of those results and the type of output presented to the user. In particular, substantial statistical information is potentially available about the distribution of intensities and ratios across the array but is neither being generated nor provided to the user in a way which they can use to make judgements on the quality of the data, currently having to rely on a subjective and qualitative assessment of the plots provided. The improvements presented here ensure that the user can make informed decisions about the quality of the experiment before attempting to extract gene expression data and identify differentially expressed genes. This semi-automated approach is explained further in the results shown in Section 3.4 and discussed briefly in Section 3.5.

3.3.1.1 GridCheck

Initial modifications to the GridCheck function ensured that the plots could be stored for the long term as a PDF file. This not only allowed referral back to the original output when desired, but also stored the plots as a single multi-page file in a vector format, enabling very high zoom levels to show increasingly high levels of detail. So that this file may be printed to A4 paper with an aspect ratio of approximately 2:3, and in consideration of the 1:3 layout of the CATMA microarray sub-grid, the sub-plots are, by default, spread across two pages per microarray with the top half and bottom half of each microarray providing a full page of sub-plots each. The default number of pages used to present each microarray can be defined by the user in the event that microarrays with different sub-grid layouts were analysed.

Minor changes to the output include using orange as an alternative plotting colour for flagged probes and ensuring that the axes of each plot cover the range from zero to the maximum intensity observed throughout the experiment. By doing this, the plots are much more comparable throughout the experiment and do not give a false impression about the anticipated 1-to-1 relationship between the channels.

A desirable improvement to GridCheck was to allow the extraction of statistics which would describe the data in the plots so that high-throughput analysis could be achieved and quality issues identified whilst removing errors inherent with analysis by eye such as fatigue. In order to extract useful statistics about the data plotted in GridCheck sub-plots, an analysis of the relationship between the two channels was

performed, alongside a review of the distribution of probes which deviate from the regression of the plot.

An assumption made about the samples of each microarray could be that the number of genes which increase in expression between samples A and B should be equalled by a similar number of genes which decrease in expression, providing no overall bias towards either channel. Therefore, the probes outlying the regression between samples, representing those genes with altered expression, would not, collectively, be found to lie as a cluster of points in any region of the plot.

Biases in the channels can lead to regressions other than $Y_g = Y_r$ and hence a test of conformity to this relationship would not be unbiased. Therefore, in order to identify those points which do not conform to the expected relationship, a major axis regression (MAR) was used through a principal component analysis (PCA).

A 2-dimensional PCA allows the identification of the major and minor axes through the data based on the eigenvalue decomposition of the variance-covariance matrix. The major axis lies in the direction of maximum variance in the data whilst the minor axis lies 90° to this with the origin between the two being at the mean of the original horizontal and vertical axes. The units of each axis are the number of standard deviations from their origin. Given that the direction of maximum variance defines the major axis regression line, the deviation from that regression could be quantified using the minor axis which is orthogonal to the regression. Those points of each plot which are greater than 1.96 or lesser than -1.96, which define the boundaries for 95% of normally distributed data ($p \leq 0.05$), in the minor axis are considered potential outliers and are therefore highlighted by a small red circle. This provides a highly visual interpretation of the distribution of those points when the plots are viewed, and an example of this can be seen in Figure 3.1(a).

In order that the statistical information collected during the analyses described above could be used to identify GridCheck sub-plots with unusual distributions, without the need to manually check every plot, a series of values summarising each sub-plot are output in the form of a tab-delimited text file which can be opened by spreadsheet software such as Microsoft Excel. An example of the first few lines from the table describing the senescence experiment microarrays can be found in Table 3.1. Clarification on how these statistics may be used to help identify GridCheck sub-plots with underlying issues can be found in Section 3.4.1 and their application to other datasets is discussed in Section 3.5.1.

All of the improvements to GridCheck listed above were devised and implemented during the course of the aforementioned MSc project and were submitted as a partial contribution toward that degree, though bugs and inefficiencies were removed after that submission.

Array	Row	Col	PC1 Var	PC2 Var	PC2 StDev	x Mean	y Mean	# Scat.	% Scat.	Slope	Inter.	Upper PC1 μ	Upper Stat	Upper p-Value	Lower PC1 μ	Lower Stat	Lower p-Value
1	1	1	10.8459	0.5123	0.7157	9.5322	9.0605	35	5.4	0.931	1.094	2.6536	2.0092	0.0722	1.4630	2.1054	0.0463
1	1	2	11.2202	0.4222	0.6497	8.9906	8.8536	25	3.86	0.926	0.793	-2.0257	-1.7364	0.0995	-4.3941	-4.0913	0.0094
1	1	3	9.9009	0.507	0.7120	9.2028	8.7851	32	4.94	0.931	1.021	2.0305	2.9366	0.0078	4.4808	4.5736	0.0013
1	1	4	9.608	0.3842	0.6198	8.9910	8.7960	22	3.4	0.996	0.23	-3.4685	-3.2537	0.0086	-2.2651	-1.6562	0.1286
1	2	1	11.2704	0.3778	0.6146	9.3703	9.2244	29	4.48	0.994	0.203	2.5449	2.1667	0.0467	1.7165	1.0973	0.2940
1	2	2	11.6442	0.5425	0.7365	9.0831	8.9740	29	4.48	0.982	0.271	-2.8002	-2.6801	0.0158	-6.0686	-11.9292	0.0000
1	2	3	8.9864	0.4053	0.6366	8.7236	8.5072	28	4.32	0.97	0.472	4.1744	6.0797	0.0002	1.6234	1.4122	0.1749
1	2	4	11.5988	0.6113	0.7818	9.1550	8.8792	30	4.63	0.925	0.941	-2.2839	-2.1828	0.0569	-2.8496	-2.2938	0.0333
1	3	1	9.1335	0.2712	0.5207	9.4483	9.1823	32	4.94	0.953	0.697	1.0754	1.2069	0.2430	2.0190	1.6048	0.1345
1	3	2	8.4623	0.3299	0.5743	9.0481	9.0016	27	4.17	0.971	0.31	-2.7796	-2.6984	0.0223	-1.4236	-1.6329	0.1232
1	3	3	7.8861	0.2694	0.5190	9.1851	9.0210	30	4.63	0.977	0.373	0.6825	0.8826	0.3874	1.1825	0.6945	0.5097
1	3	4	8.8305	0.2561	0.5060	9.2590	9.1150	28	4.32	0.951	0.589	-1.9366	-2.1781	0.0457	-1.9866	-1.3391	0.2075
1	4	1	9.412	0.3902	0.6246	9.8178	9.5768	21	3.24	1.022	0.029	5.7000	4.4037	0.0010	3.5368	1.3493	0.2141
1	4	2	6.0582	0.2527	0.5026	8.8861	8.6919	30	4.63	0.963	0.516	0.7933	0.9751	0.3411	2.8972	3.5578	0.0074
1	4	3	6.3998	0.4245	0.6515	8.5133	8.3248	24	3.7	1.08	-0.476	-1.6707	-1.8272	0.0890	-3.4219	-1.6227	0.1433
1	4	4	9.4506	0.2881	0.5367	9.3999	9.1886	27	4.17	0.972	0.465	0.3113	0.2540	0.8028	-0.2942	-0.2180	0.8317
1	5	1	9.3078	0.2832	0.5322	9.5817	9.3474	27	4.17	0.975	0.466	0.5457	0.4994	0.6247	0.3023	0.1787	0.8617
1	5	2	6.6683	0.3084	0.5553	8.9066	8.6283	25	3.86	0.976	0.483	-2.2953	-2.0247	0.0639	-2.3355	-2.3480	0.0407
1	5	3	5.012	0.2253	0.4746	8.4383	8.2028	35	5.4	0.957	0.59	0.9893	0.9275	0.3705	1.3387	2.1498	0.0439
1	5	4	8.383	0.2208	0.4698	9.2961	9.0763	30	4.63	0.979	0.407	-0.7300	-0.7079	0.4885	0.3982	0.3573	0.7276
1	6	1	9.4849	0.3438	0.5863	9.6595	9.3949	30	4.63	0.971	0.539	-1.0571	-0.7470	0.4707	-0.8273	-0.7653	0.4545
1	6	2	6.7385	0.2211	0.4702	9.1308	8.9223	33	5.09	1.017	0.059	-0.7197	-0.7774	0.4475	-2.1569	-2.8052	0.0140
1	6	3	5.6799	0.2752	0.5245	8.6059	8.3265	44	6.79	1.013	0.175	-0.3337	-0.5153	0.6112	-2.1172	-4.1305	0.0005
1	6	4	8.4316	0.24	0.4898	9.3381	9.0232	31	4.78	0.989	0.411	-0.6903	-0.5886	0.5661	0.0374	0.0431	0.9661

Table 3.1 – Statistics produced by GridCheck summarising each sub-plot

During the analysis of sub-grids by GridCheck, a table is produced showing statistical analysis of the probes they contain. The first three columns present the number of the microarray and the location of the sub-grid on that array. This is followed by the variance captured by the first principal component, the variance captured by the second principal component and the standard deviation of the second principal component. The origin of the major axis regression is defined by the mean of the original axes, x and y . The number of scattered (outlying) probes is defined and then placed into context by specifying what percentage of the sub-grid this represents. The slope and intercept of the linear regression through the data are provided. The final six columns describe the distribution of outlying probes above and below the regression line by referring back to the first principal component. The first three of these columns show the mean of the first principal component, a t -statistic and p -value indicating the significant deviation from zero of outliers above the regression line in the first principal component. The last three columns show the same, but for outliers below the regression line.

3.3.1.2 RIPlot

To improve the clarity of graphical output and provide permanent storage of this, RI-Plot now generates two vector-based PDF format documents. By being vector-based, increasingly higher levels of detail are available under any enlargement of the plots. The first PDF file provides exactly the same plots, without modification as shown in Figure 3.1(b), as were originally available from the MAANOVA package. The second PDF contains sub-plots, similar to GridCheck and in the same layout, which present the RIPlot for each sub-grid of each microarray.

No statistics are calculated for these plots because they would be essentially equivalent to those produced by the statistical analysis in GridCheck. The sub-plots of the second PDF are essentially identical to GridCheck, but with a 45° clockwise rotation. The plots of the first PDF are unique because they present the data of an entire microarray on one plot, but there are consequently fewer plots to observe. The purpose of using statistics to analyse the plots would be to identify those which show unusual characteristics without manually needing to check each plot. Those plots identified by the statistics would still require human intervention to verify the exact cause of the problem and, in the case of the senescence experiment, only 176 plots are produced and can be easily scanned by eye in a few minutes.

All of the changes made to RIPlot were devised and implemented during the course of the MSc project mentioned on page 90.

3.3.1.3 ArrayView

As with the other quality control functions, one of the first modifications made to ArrayView was to export the heat-maps to a vector-based PDF file as this provides long term storage of the plots as well as the ability to focus on small areas of the array whilst retaining good resolution. To aid in printing the PDF on A4 paper, given the 1:3 aspect ratio of CATMA microarrays, the user can choose to either place 2 microarray side-by-side on each page, as is default, 4 microarrays side-by-side in landscape or 8 microarrays in a layout of 2 rows by 4 columns in portrait.

Where heat-maps were previously often dominated by the intensity of control probes, masking the intensity of gene-specific probes as can be seen in Figure 1.6(d), those control probes are now identified during loading of the microarray data into MAANOVA as those which do not have unique gene identifiers. Where replicate spots for a probe are included on the microarray, they must therefore be named uniquely. This does not prevent replicates from being interpreted correctly by MAANOVA so long as they are specified when the data is loaded. Those probes identified as control probes are now interpreted by ArrayView as having a \log_2 ratio of zero, allowing the identification of ratio data from the gene-specific probes as seen in the new heat-map of the same microarray shown in Figure 3.2(b). Additionally, to aid in the identification of control probe locations, those probes are now marked with a white \boxtimes symbol overlaying their position on the heat-map.

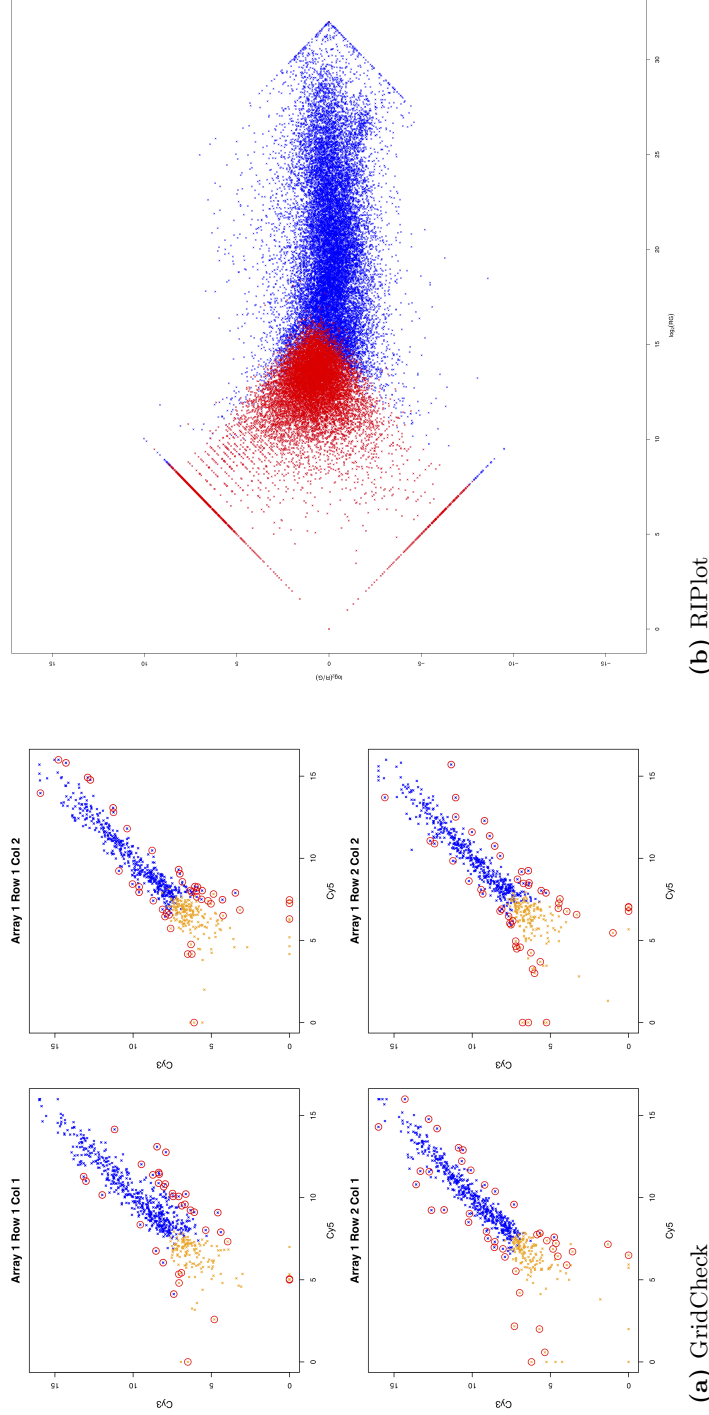


Figure 3.1 – Examples of improved graphical output produced by GridCheck and RIPlot

Once improvements had been made to the functions of GridCheck and RIPlot, the graphical output produced before any changes were made to the functions of to those plots shown in Figures 1.6(a) and 1.6(b) which show the graphical output produced before any changes were made to the functions of MAANOVA. The sub-plots produced by GridCheck, shown in (a), are now well spaced and large enough to view the data comfortably. The scale of both axes of every plot, regardless of the range of the intensity values plotted, are always from zero up to the maximum intensity measured in the experiment. Flagged probes, previously plotted in red, are now plotted in yellow, and those probes which stray from the regression are plotted with a red circle around their point, to aid in identifying the distribution of such probes. The plots of \log_2 ratio vs \log_2 mean intensity of each array produced by RIPlot are now similar to that shown in (b). Given that these plots do not illustrate anything beyond that shown by GridCheck, it was deemed unnecessary to provide similar analysis of outlying probes and hence the only improvement is to the resolution of the plot by storing it in a vector format that can be enlarged to any scale.

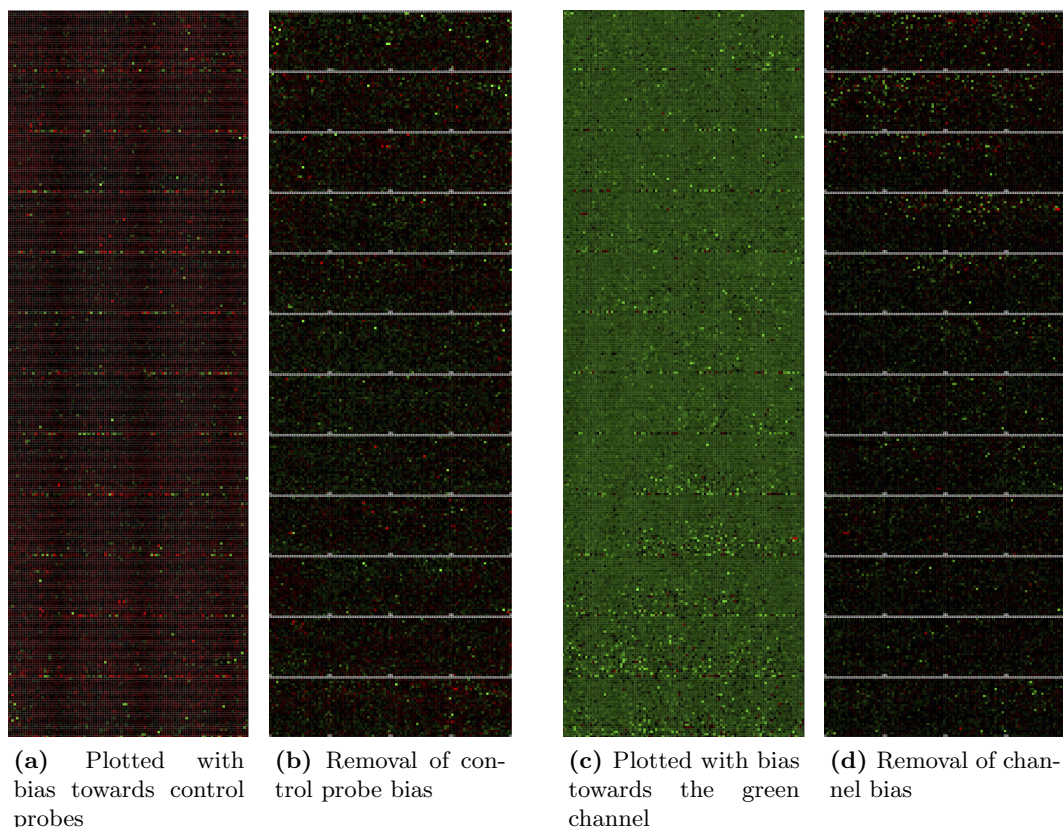


Figure 3.2 – Examples of improved graphical output produced by Array-view

Once improvements had been made to ArrayView, the graphical output is now presented as shown above. The heat-maps shown can be directly compared to the same heat-maps produced with the original MAANOVA implementation, shown in Figures 1.6(d) and 1.6(e). It can be seen that, by removing the previously dominating control spot ratio information from the plot and replacing them with white \boxtimes symbols, ratios in the probes specific to the genome become more pronounced as shown in (b). By controlling the scale of the colour-map so that a zero \log_2 ratio is always displayed in black, as shown in (d), the bias towards one channel is removed. This would make it much easier to see if there is a spatial bias towards either channel.

It was noticed, once the control probe locations were identified, that the rows of probes plotted by ArrayView were inverted in relation to the expected order in which row 1 column 1 exists at the top left of the microarray. The new heat-maps are now plotted to satisfy that expectation.

The issue of channel bias in the colour map used to plot the heat-maps, as demonstrated in Figure 1.6(e), has been corrected so that a \log_2 ratio of zero is always plotted in black. This is achieved by identifying the maximum absolute \log_2 ratio of each microarray and using that value as the maximum for both green and red of the associated heat-map rather than having independent maximum values for each colour. The improvements can be seen in the new heat-map of the same microarray shown in Figure 3.2(d) where a green bias no longer exists.

Additional to the ratio heat-maps already available from ArrayView, an obvious

extension of the graphical output provided by ArrayView is intensity heat-maps, in which the colours of the plot indicate the sum \log_2 intensity of each probe. A second PDF is now produced which is identical in format and layout to the PDF described above, but a new colour map ranging from black for minimum intensity to white for maximum intensity is used to plot the sum \log_2 intensity of both channels instead. An example of this type of plot is presented in Figure 3.3 where the intensity heat-maps of the same microarrays as are displayed in Figure 3.2 can be seen. Since the colour white is used as part of the colour map, the \boxtimes symbols used to identify the location of control probes are plotted in red.

The alterations made to ArrayView, as described to this point, were submitted for the MSc project described on page 90, but the additions described throughout the rest of this section about the methods of ArrayView were devised and developed during the course of this PhD.

One of the primary objectives in the improvement of ArrayView was to provide statistical analysis of the distribution of data for both the \log_2 ratio and the sum \log_2 intensity. This is initially provided in the form of histograms which show the distributions of ratios and intensities, independently. The \log_2 ratio histograms, shown in Figure 3.4(a), should indicate an un-biased distribution of ratios centred around zero, as shown. The colours used to plot the bars are representative of the colours used in the heat-map. The sum \log_2 intensity histograms, shown in Figure 3.4(b), should indicate that the majority of probes are represented by a sum \log_2 intensity of approximately 40% of the maximum or greater, indicating that expression was above the background noise of the microarray, with only a small number below this threshold and very few reaching maximum intensity where saturation has occurred.

These plots are presented as two separate PDF files, one each for the \log_2 ratio and sum \log_2 intensity. Each PDF displays one histogram per page, each representing one microarray. Two further PDF files are also produced, displaying equivalent histograms for each sub-grid of each microarray, mimicking the layout used by GridCheck. Each PDF is accompanied by a table of data, describing the distribution of the data within the histograms, and examples of these can be found in Table 3.2. These statistics could potentially be used to identify arrays with unusual distributions, as will be discussed in Section 3.4.1.3.

3.3.2 Data Transformation

Whilst quality control functions aim to inform the user when the raw data is inadequate for analysis without applying any transformation to the data, the data transformation functions are designed to identify systematic errors of microarray data and remove or reduce them by coercing the data to fit an expectation. This is not to suggest that transforming data is an acceptable strategy to ensure that the data appears adequate because no transformation comes without compromise nor the risk that useful information is being lost. Where costs permit, it is always a better option to re-hybridise samples to new

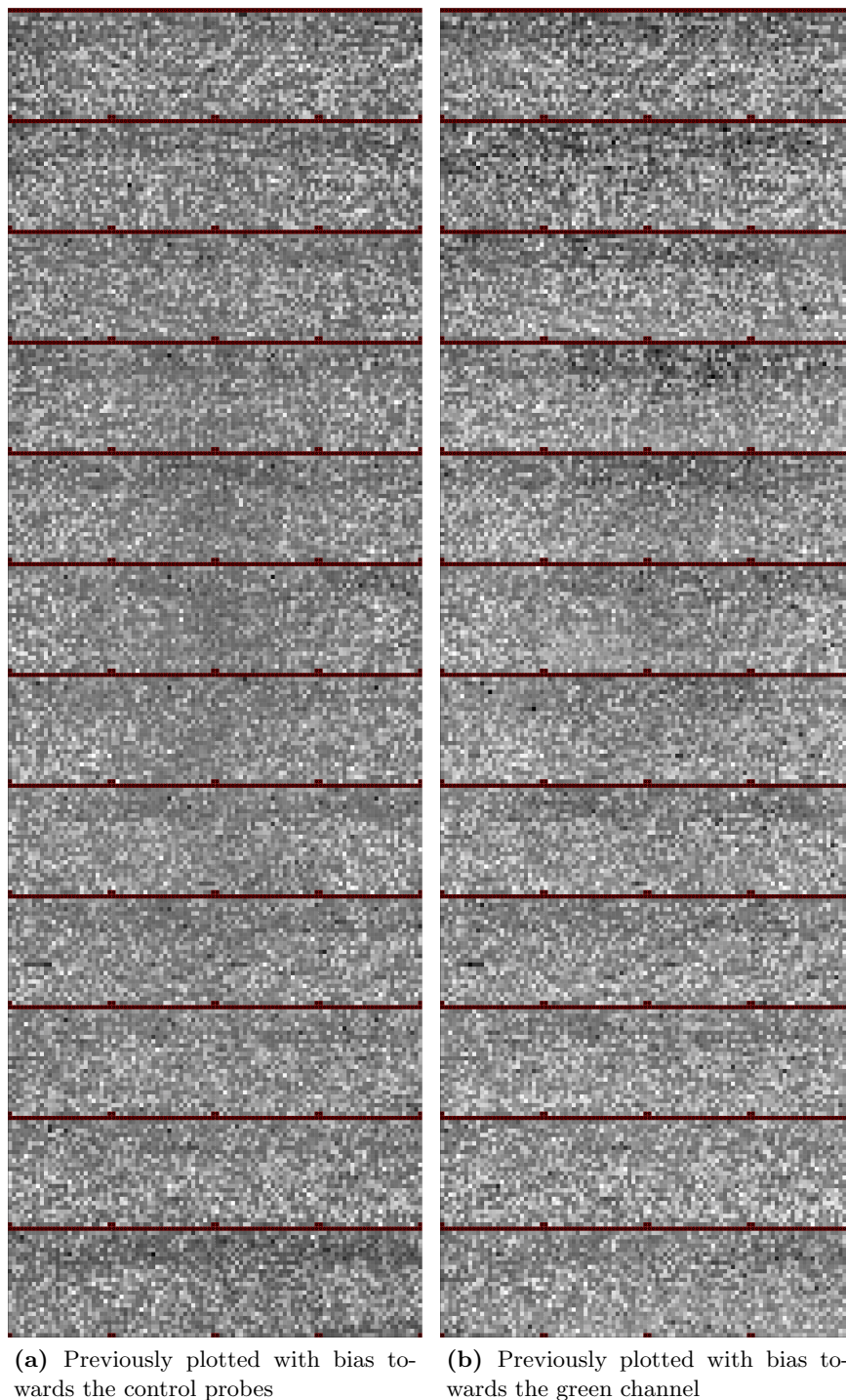
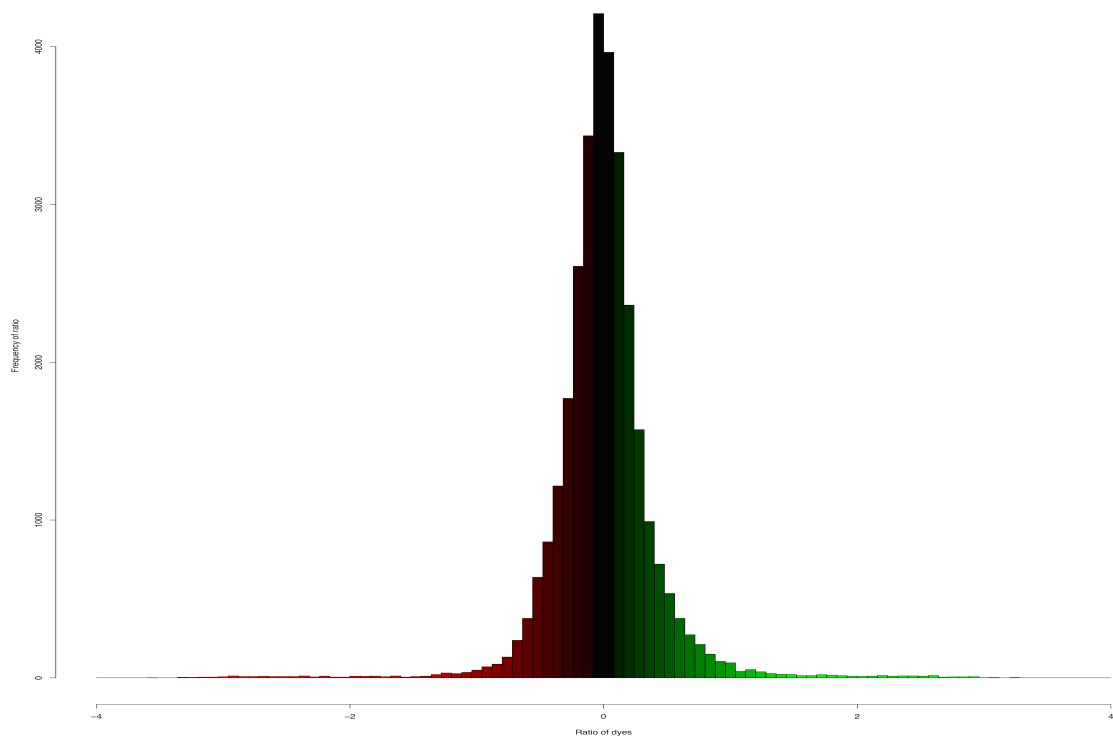
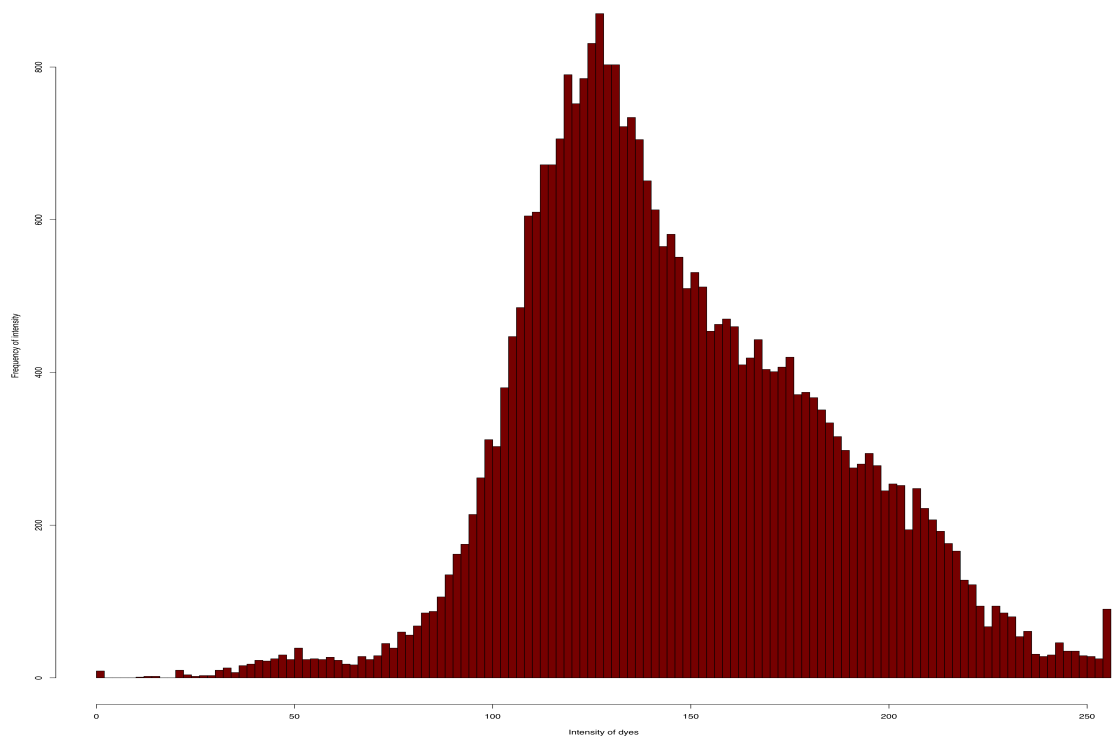


Figure 3.3 – Examples of additional graphical output produced by Array-view

Additional functionality provided to ArrayView allows the presentation of the sum \log_2 intensity of the two channels of the microarray. This is presented in the form of a heat-map with black representing zero intensity and white representing maximum intensity. As the colour white is already used, the \boxtimes symbols representing the control spots are plotted in red as an alternative. The plots shown are of the same microarrays as in Figure 3.2. If a problem existed in which a range of co-located probes exhibited no abnormality in the \log_2 ratio, but their sum \log_2 intensity was abnormal, this type of plot would help to identify the issue.



(a) Histogram of probe ratios



(b) Histogram of probe intensities

Figure 3.4 – Histogram plots produced by ArrayView

Histograms of the ratios, (a), and intensities, (b), in which the vertical axis represents frequency and the horizontal axis represents ratio or intensity, respectively, are plotted as part of the output of ArrayView. Typically, the majority of ratios are close to zero, with only a very small percentage of differentially expressed genes. Also typically, the majority of probes have an intensity of at least 40% with a diminishing number reaching maximum intensity, though the peak of probes at maximum intensity indicate that a number have been saturated. A statistical summary of both types of plots are returned as a table of values.

Array #	Min.	1st Qu.	Median	Mean	3rd Qu.	Max	Standard Deviation
1	-3.560	-0.175	-0.008	0.000	0.161	3.243	0.410
2	-3.329	-0.159	0.000	0.000	0.160	3.675	0.387
3	-6.466	-0.302	-0.012	0.000	0.293	4.197	0.634
4	-3.282	-0.157	0.025	0.000	0.166	3.362	0.375
5	-3.297	-0.244	0.036	0.000	0.264	3.863	0.479
6	-4.326	-0.169	0.023	0.000	0.183	3.266	0.382
7	-3.687	-0.263	-0.017	0.000	0.235	3.123	0.555
8	-3.279	-0.132	-0.011	0.000	0.117	3.182	0.335
9	-3.736	-0.213	-0.014	0.000	0.195	3.481	0.467
10	-4.168	-0.160	-0.004	0.000	0.150	3.122	0.373

(a) Summary statistics for \log_2 ratio distribution

Array #	Min.	1st Qu.	Median	Mean	3rd Qu.	Max	Standard Deviation
1	0	119.5	139.7	146.0	171.6	255	37.7570
2	0	123.3	142.4	148.6	174.0	255	36.9281
3	0	121.6	141.4	145.9	170.1	255	36.0596
4	0	124.4	144.9	150.7	177.4	255	38.2576
5	0	118.8	136.4	141.8	163.4	255	35.0551
6	0	125.9	144.0	149.2	172.2	255	35.1632
7	0	126.2	143.3	148.4	170.8	255	34.2313
8	0	125.6	143.3	150.0	174.8	255	36.4579
9	0	116.1	132.0	138.9	160.6	255	35.2351
10	0	125.4	144.3	150.7	176.2	255	36.9117

(b) Summary statistics for sum \log_2 intensity distribution

Table 3.2 – Summary statistics for distributions of \log_2 ratio and sum \log_2 intensity data produced by ArrayView

For each histogram produced by ArrayView, a series of statistics are produced as shown. For histograms displaying ratios, the table shown in (a) is produced, which presents the minimum, 1st quartile, median, mean, 3rd quartile, maximum and standard deviation of the ratios for each array. For histograms displaying intensities, an identical set of statistics are produced as shown in table (b).

microarrays that are worst impacted by inevitable errors and flaws. However, subtle inadequacies can be corrected at a much lower cost by using the methods described and implemented here.

3.3.2.1 TechRepCheck

TechRepCheck is a new function devised for comparing the technical replicates for each independent sample throughout the experiment to ensure that they are providing consistent levels of hybridisation for each probe, as should be expected. Whilst it would be anticipated that samples hybridised to the same microarray should show some differential expression throughout the probes, because the two samples are not the same, when a sample is hybridised on more than one microarray throughout the experiment, the resulting intensities should be close to identical. Using a process similar to GridCheck, the technical replicates are compared and individual probes can be identified as inconsistent between replicate pairs. Complex rules are then applied to identify which of the technical replicates are outliers on a probe-by-probe basis and a more consistent

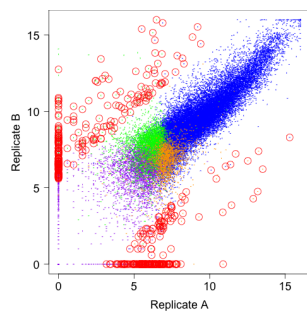
value of the outlying probe(s) is estimated based on the data provided by the remaining technical replicates.

Comparisons between technical replicates are made on a pairwise basis, in which, for any given sample, $\frac{N \times (N-1)}{2}$ pairwise comparisons can be made, where N is the number of technical replicates. The senescence experiment, with four technical replicates, is therefore assessed by making six pairwise comparisons. These comparisons are made by plotting \log_2 intensity values in a similar manner to the way in which GridCheck compares samples bound to the same microarray. An example of these plots and their layout can be found in Figure 3.5. Since dye balancing forms an important principle of the experimental design, it should usually be true that each sample is bound to Cy3 and Cy5 equally often. For four technical replicates, this then provides two same-dye comparisons and four different-dye comparisons. Given that the dye used in labelling can affect the binding efficiency of specific genes (Gresham *et al.*, 2008; Rosenzweig *et al.*, 2004), same-dye comparisons are more likely to provide identical expression levels. By sorting to ensure that replicates A and B are bound to Cy3 and replicates C and D are bound to Cy5, Figures 3.5(a) and 3.5(f) represent same-dye comparisons. Each of the other plots compare a Cy3-bound replicate with a Cy5-bound replicate.

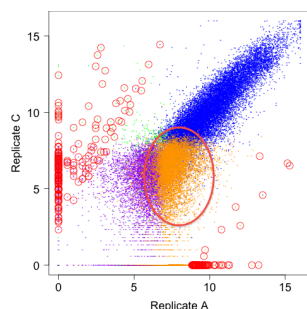
In each plot, four colours are used to plot the points, indicating the flag status of the probe they represent. In GridCheck, the points were either flagged or not flagged by ImaGene (BioDiscovery), since the two samples were hybridised to the same microarray, and so there was only a need for two plotting colours to indicate the flag status. However, TechRepCheck compares observations from two microarrays, and hence the flags can be one of four different permutations. Points plotted using blue represent those which are not flagged in either replicate. Green points are flagged in the replicate plotted on the horizontal axis only whilst orange points are flagged in the replicates plotted on the vertical axis only. If the point was flagged in both replicates, it is plotted using purple.

Similarly to GridCheck, a PCA is implemented to fit a MAR to assist in the identification of probes which are outliers from the expected relationship. Those probes which deviate from the major axis regression line by more than 1.96 standard deviations ($p < 0.05$) in the minor axis are then plotted within a red circle to highlight them. This type of analysis is particularly appropriate since it is based upon the MAR regression of the replicates rather than their conformity to any particular relationship, which allows the replicates to exhibit dissimilar absolute expression values whilst retaining, more relevant, similar expression relative to the rest of the probes. This allows the comparison of replicates without the effects of Array and Dye causing misinterpretation.

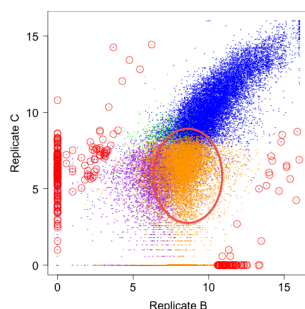
Figure 3.5 presents the output produced by TechRepCheck for a single sample with four technical replicates. Without applying computational analysis to individual probes, it is still possible to identify replicates which are contributing less to the analysis by containing many probes with characteristics not found in the other replicates. For example, Figures 3.5(b) and 3.5(c) exhibit a greater than usual number of orange points, as circled, whilst Figure 3.5(f) exhibits a similar number of green probes, also circled.



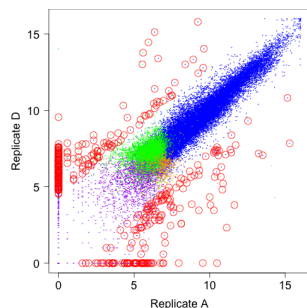
(a) Replicates A and B



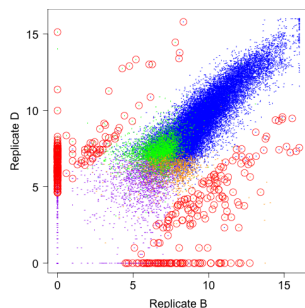
(b) Replicates A and C



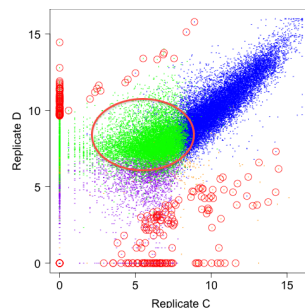
(c) Replicates B and C



(d) Replicates A and D



(e) Replicates B and D



(f) Replicates C and D

Figure 3.5 – TechRepCheck output demonstrating an obvious outlying replicate

The output of TechRepCheck visualises the relationship between all the probes of technical replicates, regardless of the arrays and channels they are associated with in the experiment. Despite being hybridised to different microarrays, it is possible to make comparisons between technical replicates and learn about inconsistencies between them. Assuming two replicates are hybridised to each channel, six plots are produced as shown. (a) and (f) always compare replicates of the same channel whilst (b) through (e) compare only replicates of opposite channels. Blue points represent normal probes, whilst those plotted in orange or green were flagged in one channel and those plotted in purple were flagged in both channels. If probes lie more than 1.96 standard deviations from the mean of the minor axis of a major axis regression, they are identified as outliers with a small red circle around their point. In the case shown, large clusters of probes have been circled in red in (b), (c) and (f). These demonstrate a dissimilarity in the flagged probes of replicates on those plots. Since the only replicate common to all these plots is Replicate C, it is very likely that this replicate has a problem associated with it which, without this analysis, could go unnoticed.

All three of these plots also demonstrate a non-linear relationship between the replicates, indicating that something is influencing the data within them when compared with the other three plots of the same replicate set. The replicate which is common among the three plots is replicate C, and this is consistent with the highly flagged probes too since replicate C is on the vertical axis of Figures 3.5(b) and 3.5(c) and on the horizontal axis of Figure 3.5(f).

This information may inspire the user to investigate the microarray upon which replicate C is hybridised, but a more thorough investigation can be performed by using a similar approach to that shown above but for individual probes of the microarrays. In this case, the processing power of the computer allows the fast and efficient identification of outlying replicates on a probe-by-probe basis. This can be undertaken by analysing the points which were identified as outliers from the MAR analysis. For each probe identified in at least one of the six plots, a summary is produced, identifying which plots present the probe as an outlier.

Table 3.3 offers a visual way of interpreting the otherwise complex output produced by this frequency analysis:

- Where no plot indicates that a probe is outlying, no outlying replicates exist and there is no need to apply any flags.
- Where a single plot identifies a probe as outlying, the two most extreme replicates are inconsistent with one another, but this does not provide evidence to identify which is the actual outlier, so no replicates are flagged.
- Where there are two plots indicating an outlier for a probe, one replicate will still remain similar to all the other replicates, with two replicates on one side of its expression and one replicate the other side. The outlying replicate(s) in this case are identified by considering the distance between this common replicate and both the mean of the paired replicates and the intensity of the isolated replicate. If one distance is three times greater than the other, the replicate(s) with the greater distance are flagged as outlying. Three was chosen as an arbitrary threshold after comparing examples of this situation by hand.
- Where three plots indicate an outlying probe, either one replicate is significantly different from all the others and is flagged as the outlier, or else two replicates will show significant difference from all but one other and these are flagged as outliers.
- Where four plots indicate an outlying probe, if one replicate is identified as being inconsistent to the others, this replicate is flagged as the outlier, otherwise two pairs of replicates exist and the pair with lower expression is flagged as the outlier if its log expression is below 8 which is safely within the background noise of the microarray.
- Where five plots indicate an outlying probe, only two replicates are similar to each other and the remaining two are flagged as outliers.









Plots with Outlying Probe	Technical Replicate Emblems	Description
Zero		All replicates have been identified as consistent with each other and, as such, no outliers exist.
One		A single pair of replicates (A and D) have been identified as inconsistent with each other, but the evidence is not compelling enough to identify an individual replicate as outlying.
Two		The two outlying plots (A vs D and B vs D) indicate that D may be inconsistent. However C is most consistent with other replicates, therefore $\left \frac{C-(A+B)/2}{C-D} \right $ is calculated to identify whether A and B, or D, are least consistent. If the result is less than $\frac{1}{3}$, D is considered the inconsistent replicate; if the result is greater than $\frac{1}{3}$, A and B are considered the inconsistent replicates; otherwise all replicates are considered acceptable.
Three		When all three outlying plots represent a common replicate (D), this replicate is identified as inconsistent.
Four		Where no common replicate exists in all three outlying plots, two replicates (A and D) will only be consistent with one adjacent replicate and these will be considered the inconsistent replicates. If one replicate (D) is represented by three of the outlying plots, it is inconsistent with all of the other three replicates and is flagged for estimation.
Five		If no single replicate is represented by three outlying plots, the replicates exist as two similar pairs and either the lower pair (A and B) have a mean expression below background ($< 2^8$) in which case they are considered the inconsistent replicates, otherwise the evidence isn't conclusive enough to identify which pair to estimate.
Six		Only two replicates are identified as consistent with one another (A and B in the top example or B and C in the bottom example) and therefore the remaining two replicates are flagged as inconsistent.
		No two replicates are similar, therefore not enough consistent data exists to provide an accurate estimate.

Table 3.3 – Technical replicate comparisons identifiable by TechRepCheck

Once TechRepCheck has been performed, each technical replicate has been compared with each other technical replicate, resulting in six comparisons for a four technical replicate experiment. Each probe which strays from the regression by more than 1.96 standard deviations ($p < 0.05$) in the minor axis is identified as outlying in a plot. The number of outlying plots for each probe are used to identify which replicate(s) are responsible as shown above. Red, blue, green and yellow lozenges are used in the emblems shown to represent the values for a given probe in each technical replicates. Overlapping lozenges represent non-significant differences between replicates.

- Where six plots indicate an outlying probe, no replicate is similar to any other and therefore no individual probe can be identified as an outlier. This scenario suggests that a higher level of residual variability exists for this probe.

Limitation: TechRepCheck can, currently, only automatically analyse and identify outlying replicates on a probe-by-probe basis where four technical replicates exist. Three or fewer technical replicates would, in the majority of cases, not provide enough proof to identify the inconsistent replicate, whilst five or more technical replicates would have many more possible outcomes per probe than those shown in Table 3.3, providing a more complicated coding problem. If any number of technical replicates except four are found for a biological sample, the plots of TechRepCheck are still produced, but the identification of outlying replicates is skipped.

This limitation could be lifted by generating a generalised rule set associated with the identification of technical replicates which are outliers. This was beyond the scope of the analysis of the senescence data, and hence its lack of implementation, but theoretically, with the correct rule set, any larger number of technical replicates could be analysed in this manner.

Once a list of flagged replicates for each probe has been produced, TechRepCheck then attempts to estimate the expression levels of those replicates by using data from the other replicates for that probe. This type of inference could be fraught with inaccuracies due to the nature of the technical replicate data, since the mean expression of each microarray may be different due to scanner gain inconsistencies or alternate handling of the microarrays causing binding affinity differences, but by using a mean rank method of estimation, these inconsistencies are accounted for.

Appropriate estimated intensity values for inconsistent replicates are calculated using a mean rank strategy in which the rank intensity of the affected probe is determined across all gene-specific probes for each of the consistent replicates. The mean of these ranks determines an appropriate rank for the inconsistent replicates. The intensity of the probe in each of the inconsistent replicates is adjusted so as to give the affected probe the calculated rank.

The identification of probes as outliers in particular replicates and then consequential estimation of a more appropriately consistent intensity may provide a method for avoiding poor data by using information from other microarrays, but is still a loss of data because the intensity of some probes in some samples are now based on fewer real observations and a number of estimates. These estimates are based on the real observations of other technical replicates used in collaboration with real observations of other probes on the same microarray, but are still only estimates.

In order to identify whether a large number of estimates are being generated in specific regions of the microarrays, which indicates that they might be a physical artefact and therefore avoidable, ArrayView has been adapted to plot the positions of estimated probes after TechRepCheck has been performed so that these areas of the

microarrays are easily identified. Figure 3.6 shows a region of the ArrayView ratio heat-map for the same microarray both before and after TechRepCheck has been used for probe estimation. Gold cross symbols indicate the probes which have been estimated in one or both channels. A + symbol is plotted where the estimate was made in the first channel, whilst a \times symbol is plotted where the estimate was in the second channel. If both channels were estimated, a combination of the two symbols, *, is plotted over the probe. Generally, local clusters composed mainly of either the + or the \times symbol indicate that a problem has occurred with a specific channel of the microarray, but clusters of the * symbol indicate that the probes experienced a problem, either because of damage to the microarray, or because of poor analysis of the microarray scans.

An additional output produced by TechRepCheck is a table showing how many technical replicate sets each probes requires estimation for. This could help to improve probes of the microarrays in future versions as those with the greatest number of estimations are providing inaccurate information about the expression of their specific gene. This information could also be considered during overall model fitting to the data, but the approach by which this could be done is not entirely clear.

The plotting functions of TechRepCheck, including the application of a PCA, were integrated into the MAANOVA package during the time of the MSc project, but all other aspects of TechRepCheck, including the identification of outlying technical replicates for individual probes and the consequential estimation of a more consistent intensity were developed during the time of the PhD.

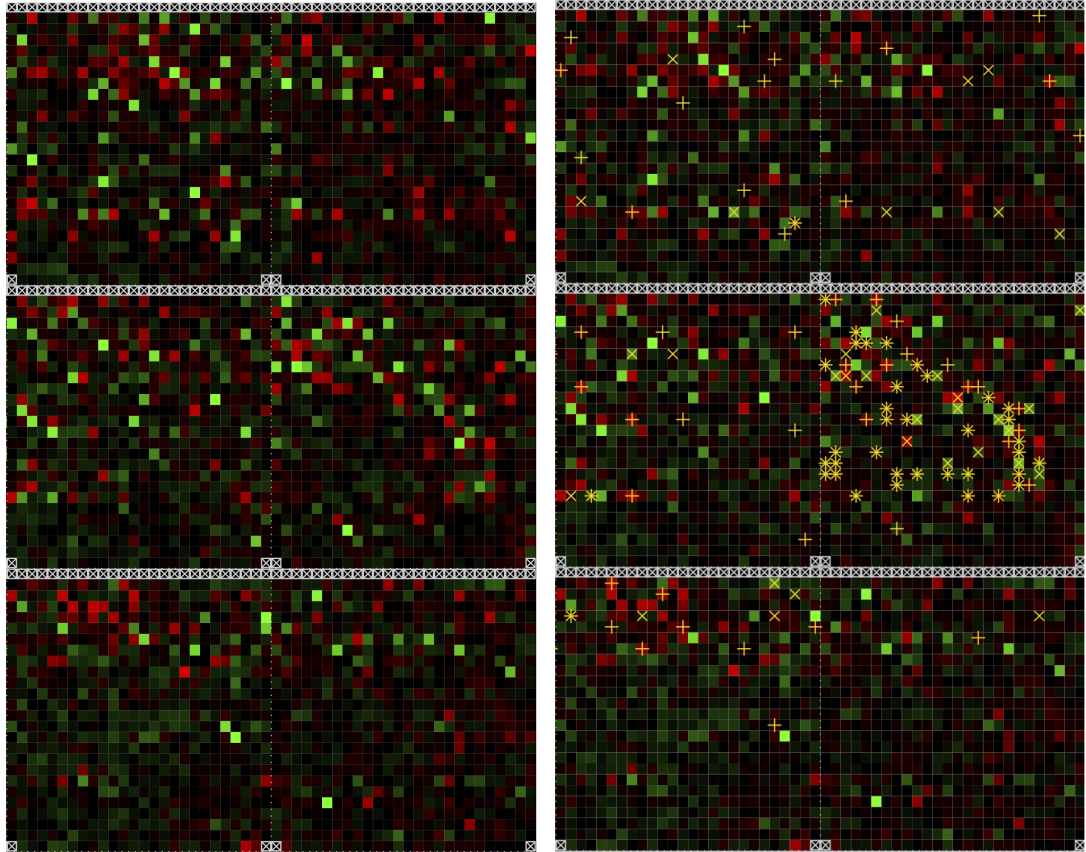
3.3.2.2 Transform MAData

Additional to the transformation methods already implemented in the TransformMAData function as described in Section 1.5.2.5, it was noticed that no method exists to bias the weight of the LOWESS transformation towards probes of the same pin-tip group. This type of weighting makes logical sense because the probes printed by the same pin-tip are assumed to be influenced by a blocking effect provided by that pin. This has been observed in real data as an obvious distinction in the ratio between sub-grids of the microarray. By applying a LOWESS within these sub-grids it is possible to ensure that localised artefacts are reduced in significance before considering normalising across the entire microarray.

By numbering each sub-grid of the microarray and providing both intensity and the sub-grid number of the probes as predictor variables for the LOWESS model, a series of fitted values are obtained which can be used to correct the ratio of those probes using the equation:

$$\begin{cases} Z_{rk} = \log_2(Y_{rk}) + \frac{C_k}{2} \\ Z_{gk} = \log_2(Y_{gk}) - \frac{C_k}{2} \end{cases} \quad (3.1)$$

where C_k is the spot-specific constant obtained from the LOWESS regression, Y_{rk} and Y_{gk} are the raw intensity values of probe k in the red and green channels, respectively,



(a) Before TechRepCheck has been performed (b) After TechRepCheck has been performed

Figure 3.6 – ArrayView ratio heat-maps before and after TechRepCheck probe estimation

Once TechRepCheck has provided estimated intensities for probes identified as technical replicate outliers on a microarray, ArrayView uses gold crosses to identify their location on the microarray. (a) presents the original ArrayView output before intensity estimation whilst (b) shows the same output after TechRepCheck has been applied. Note that the crosses applied to the plot are + shaped where an inconsistency is found in one channel and \times shaped when found in the other channel. Therefore, the combined shape $*$ is shown when both channels are inconsistent. In a large number of cases, the estimates show a \log_2 ratio closer to zero than the original values, which is the expected \log_2 ratio of most probes on the microarray. A small artefact covering a number of probes is clearly identifiable in the middle-right region of the heat-map after the estimation crosses have been plotted, but is less visible, and certainly not prominent, before this.

and hence Z_{rk} and Z_{gk} are the transformed intensity of probe k in each channel.

3.3.3 Model Fitting

In order to fit a mixed model to the senescence microarray data, the terms of that model must be defined. Dye and Array are terms which must always be included as these are always a source of variation in a two-colour microarray experiment. Alongside these are the terms defined in the experimental design: Day, Time of Day (ToD) and Biological Replicates (BioRep). Day and ToD form an interaction as they are both measures of the same variable, time, and it is within the interaction term of those two that BioRep exists. BioReps were arbitrarily labelled within sampling times and are only comparable with BioReps of the same sampling time, but the BioRep labels (A, B, C and D) are consistent across the whole experiment and so must be separated by nesting.

Of the five top level terms listed above, some are fixed and some are random. Those which contribute a random effect to the experiment must be treated as such, and include Array, because no two microarrays would ever hybridise identically and therefore could not be repeated without a random contribution to the expression data. Dye is often defined as fixed because it may be expected that each dye will always inhibit or enhance the hybridisation of specific probes by the same degree, but there is some evidence to suggest that the labelling process itself has a contribution to the observed expression levels even when labelling with the same dye (Klevebring *et al.*, 2009) and on this occasion Dye has been defined as random.

Samples collected on the same Day or Time of Day would be expected to provide a consistent response and so these are both defined as fixed terms. The interaction between Day and Time of Day is also therefore defined as fixed. Biological replicates were treated as fixed in the model because it was desirable to consider their variability whilst determining other experimental effects of the model.

The correct definition of the model formula is therefore:

$$\sim Dye + Array + (Day * ToD)/BioRep$$

where terms written in *italics* are defined as random. ‘*’ indicates that two terms are to be modelled separately but also as the interaction between them. ‘/’ indicates that the proceeding term is nested within the preceding term; in other words biological replicates are only comparable with those at the same time point. The interpretation of this formula is that the expression of each gene can be decomposed into the expression contributed by the dye the sample is labelled with, the array the sample is hybridised to, the Day which the sample was collected on, the Time of Day the sample was collected at, effects specific to the Day and Time of Day combination the sample was collected at (the interaction term) and the biological replicate that the sample represents. When

this is written out in full, it could be written as:

$$Y = \mu + \sim\text{Array} + \sim\text{Dye} + \text{Day} + \text{ToD} \\ + \text{Day:ToD} + \text{Day:ToD:BioRep} + \epsilon \quad (3.2)$$

where \sim represents random terms and $:$ indicates an interaction between terms. ϵ is a measurement of error and is a random term by definition, capturing the effects which cannot be associated with any other term. By including the interaction terms, effects which are dependent upon day and time of day can be captured, as they would otherwise be lost to the measurement error term.

3.3.4 Identifying Differentially Expressed Genes

MATest provides F-tests for individual terms of the model to identify the genes which are differentially expressed over the levels of the term. The F-statistics which are output from the test are determined by the calculation shown in Equation 1.5. This can be used to determine the significance of the variance across the levels of the term by means of a comparison to the variance of the error term of the model, which describes the technical variability of the experiment. This is a typical calculation of F-statistics and ensures that genes are only defined as significantly differentially expressed if their variance due to the selected term is significantly greater than that of the noisy background. However, an alternative analysis can be performed by comparing the variance due to one of the terms of the model to the pooled between biological replicate variability as this allows the more meaningful biological variability to correct the calculated F-statistic rather than the measurement variability which should be consistent throughout the experiment.

In order to make a comparison between the variability of each term of the model and the variability due to the biological replicates, an F-test must first be performed on all the fixed terms of the model, including the biological replicate interaction term. For the senescence dataset, this requires that MATest be used to get F-statistic values for Day, Time of Day, the Day \times Time of Day interaction and the Day \times Time of Day \times Biological Replicate interaction. The F-statistics can then be used to recalculate new F-statistics so as to change the denominator in the calculation to the variance of the biological replicates as shown:

$$F = \frac{F_{term}}{F_{bioreps}} \quad (3.3)$$

$$= \left(\frac{\text{term variance}}{\text{error variance}} \right) \left(\frac{\text{error variance}}{\text{biorep variance}} \right) \quad (3.4)$$

$$= \frac{\text{term variance}}{\text{biorep variance}} \quad (3.5)$$

where F is the new F-statistic from which p -values can be inferred by using the degrees of freedom for both the chosen term and the biological replicate interaction term.

MATest produces two different F-statistics, as described on page 28, and both

can be modified in the same way. A function has been produced, `ConvertMATEst`, to make this conversion with minimal input from the user. Once the conversion has been made to both F-statistics, the corresponding p -values are also identified so that the newly converted F-test object can be used in exactly the same way as unconverted F-test objects, including applying FDR adjustments, where desired.

To accompany the `ConvertMATEst` function, another function, `AnalyseMATEst`, has been produced with the intention of making the production of lists of differentially expressed genes easier. Up to three converted F-test results can be passed to this function and it will output lists of genes corresponding to those which are identified as differentially expressed for each of the terms. The type of F-statistic to use and the choice of whether to use the FDR adjusted p -values or not are all parameters which can be provided. Where three tests are passed to `AnalyseMATEst`, seven lists are produced and exported as text files. These correspond to differential expression in:

- Term 1 only
- Term 2 only
- Term 3 only
- Term 1 and 2 only
- Term 1 and 3 only
- Term 2 and 3 only
- All three terms

`AnalyseMATEst` then uses the produced lists to plot a Venn diagram presenting the size of each of the lists, as well as the number of genes which are not differentially expressed for any of the terms provided. These lists and the Venn diagram are then used to identify the number of genes for further analysis.

3.3.5 Extracting Predicted Gene Expression Data

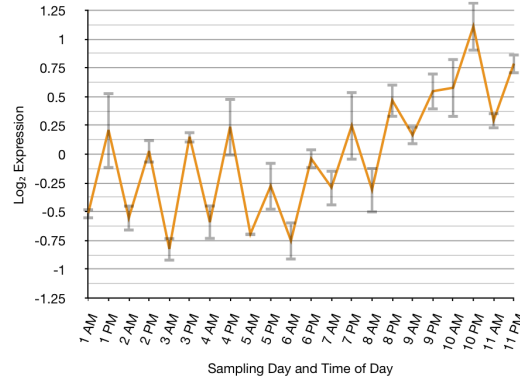
Once a mixed model has been fitted to the data and a list of genes showing differential expression identified, it is possible to obtain predicted expression data for those genes based on the fitted model terms. For each of the terms of the model, a matrix of values is returned with dimensions of the number of spots on the microarrays by the number of levels for the term. These values indicate the deviation from the mean log expression of that gene for each of the levels of each term. These deviations are additive, meaning that if it is desirable to recombine the Day term (11 levels) with the Time of Day term (2 levels), each of the 11 levels for Day can be summed with each of the 2 levels for Time of Day to provide 22 gene expression values, corresponding to the sampling times at which the plants were harvested.

By selectively adding back the effects that are to be further analysed, different forms of the data can be established. Figure 3.7 presents three possible extractions of the expression data for a particular gene with error bars indicating the variation between biological replicates for any given time point. In Figure 3.7(a), the data which have been extracted demonstrate the relative levels of expression for the gene during the harvesting period. The values shown are as they would be observed at those points in time without any experimental factors such as differences caused by arrays or dyes as these are terms which were not included during reconstruction of the expression data. This format of the data is typical of the sort you might receive from conventional methods of microarray analysis which are unable to separate the effects introduced into the experiment.

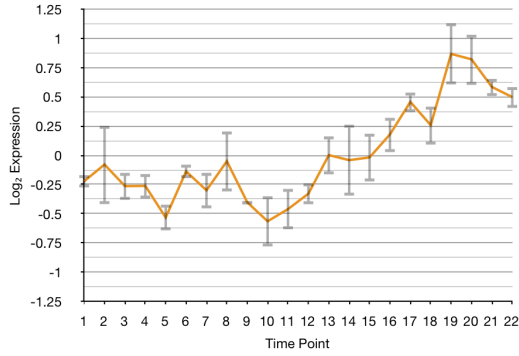
By excluding the Time of Day effects when reconstructing the expression data, this no longer becomes a contributing factor to the profile of the gene, as can be seen in Figure 3.7(b). In this case, the diurnal effects of this gene have been removed, to reveal the underlying change in expression with the onset of senescence. Whilst this may be achieved by simply averaging the morning and evening samples for each day, this approach would not take account of the confounding effect of the interaction between day and time of day. The model, however, does take this into account and provides a more appropriate method of removing time of day effects. The removal of time of day effects leaves a profile which can be better used to compare the differential expression due to senescence of several genes without identifying genes as highly differentially expressed simply because they have a strong diurnal effect.

Whilst removing the diurnal effect alone produces a useful profile, another alternative method of reconstructing the expression data is to completely remove the differences between samples from the same day by also omitting the interaction term between day and time of day. This plot can be seen in Figure 3.7(c) and only has 11 time points, but because the morning and evening samples are identical under this structure, the biological replicates of each can be treated as a single set, providing eight biological replicates per time point instead of the usual four. This not only reduces the size of the error bars since a greater number of observations increases the confidence in the data but also helps to evenly spread the time points. Under the previous model of 22 time points, the samples from the same day are 7 hours apart, whilst the time difference between samples on adjacent sampling days is 41 hours, providing a near 6-fold difference in the time scale between adjacent time points.

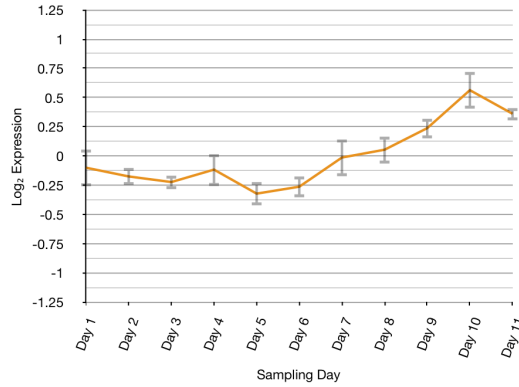
One of the powers of this analysis is that gene expression data can be extracted to suit various forms of analysis. Where senescence is the main focus of the study, the diurnal effects may mask the underlying changes throughout the longer senescence time course and so can be omitted. Where putative transcriptional network modelling will be performed, it is beneficial to only consider observations which were taken at regular time intervals and so morning and evening samples can be treated as being from the same day by omitting the changes which have occurred between those observations. Where



(a) 22 time points including diurnal effects.



(b) 22 time points excluding diurnal effects.



(c) 11 time points.

Figure 3.7 – Gene expression profile plots demonstrating inclusion of experimental terms of the model

The inclusion or exclusion of terms of the model controls the sources of variation contributing to the expression profile of individual genes. All three plots show the expression profile of the same gene and error bars represent one standard error for the biological replicates. In (a) all 22 time points are displayed and diurnal effects of the gene can be seen throughout the time course. This output is typical of microarray analysis as it includes all the measured variables of the experiment. The diurnal effects can be excluded from the profile by simply choosing not to include the time of day term from the model, leading to the plot shown in (b). This plot still contains 22 time points, but only the changes in gene expression due to ageing are present. By pooling the morning and afternoon samples by removing the interaction term from the model, the profile in (c) is obtained and has the equivalent of eight biological replicates but with the advantage that the time points are regularly spaced.

clustering analysis is desirable, it may be appropriate to only provide one observation per time point. In this case, the biological replicate term can be omitted to exclude those effects from the expression data.

Even after the expression data is extracted, further forms of normalisation can be performed on the data in order to suit the analysis method. In the case of transcriptional network modelling, the computational identification of similarity between gene expression profiles can be simplified by subtracting the mean expression and then dividing by the standard deviation across the time course. This ensures that even if the magnitude of the expression changes and the absolute expression of two genes are different, their expression profiles will be identical after normalisation if they are similar in response. This type of normalisation is also well suited to clustering algorithms, which also make comparisons between genes.

Figure 3.8 provides an example of the ways in which the gene profiles can be predicted using the fitted mixed model effects, each providing alternative levels of similarity or differentiation between the expression profiles. In most cases, it is desirable to ensure that all genes lie along a mean expression of zero so as to ensure that if their expression profiles exhibit the same kinetics, their expression values are also equal. In order to do this, the reconstruction of the gene expression must include only the experimental terms. When this is the case, the mean expression of each gene will be zero as shown in Figure 3.8(a).

An alternative to this is to introduce the relative mean expression of each gene causing the profiles of the genes to separate, indicating which genes have the highest overall expression during the senescence period and also providing information about the fold difference between the expression of the profiles. This can be seen in Figure 3.8(b) which also demonstrates that, in this form, the experiment mean is still zero and so the absolute expression of the genes cannot be identified. The actual expression profiles of the genes do not transform, but are merely translated to new gene means.

The final mean which can be introduced into the data is the experiment mean which represents the mean expression of all observed genes. By adding this, the relative distance between the expression profiles of the genes are maintained, but the expression values of each gene are now back to the original scale of the experiment. The resulting plots after the addition of the experiment mean are shown in Figure 3.8(c) which is identical in appearance to Figure 3.8(b) with the exception of the scale of the vertical axis which now indicates an overall mean of 8.7. After this transformation, every gene's profile will be represented by positive values.

Where comparisons must be made between gene profiles without consideration for their relative or absolute expression, the gene expression used should be derived from a similar extraction to that shown in Figure 3.8(a) where genes are centred about a mean of zero. A further transformation can then be applied to ensure that the magnitude of expression changes are uniform throughout all the genes. This transformation was

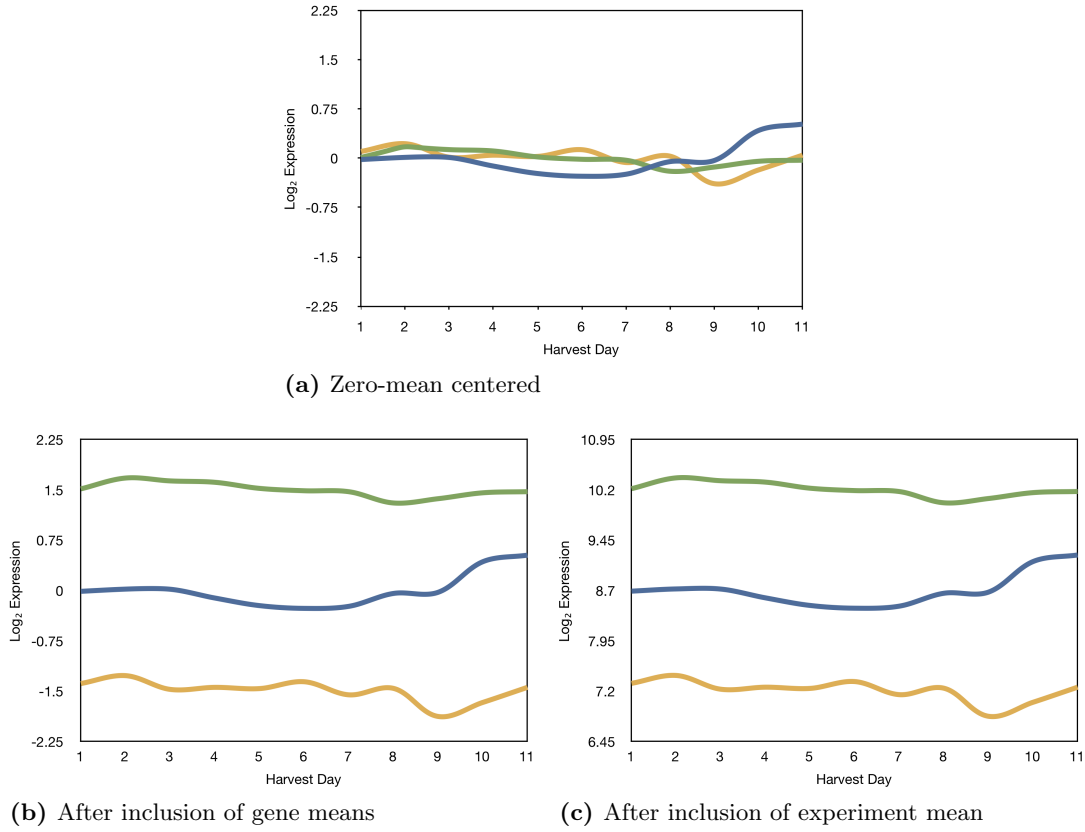


Figure 3.8 – Examples of normalisation applied by inclusion of mean terms of the model

The inclusion of terms related to the relative expression of genes allows data to be extracted in formats suitable for specific purposes. When only the terms of the model related to the input variables of the experiment are summed, the output for the three genes shown above are as the plot (a) in which each profile is centred about a \log_2 expression of zero. By including the gene means for each gene, shown in (b), the shape of the profiles is unaffected, but the relative expression levels between genes are restored, indicating that the gene plotted in green has a higher overall expression than the blue or yellow profiles. In this case, each gene is still centred around a mean of zero over the whole experiment. By opting to also include the experiment mean, shown in (c) the profiles are once again unaffected, as is the relative difference in expression between them. However, all expression values are now greater than zero and have been returned to the same scale as they were measured on the microarrays.

performed by applying the equation:

$$y = \frac{y}{\sqrt{\frac{1}{T-1} \sum_{t=1}^T (y_t - \bar{y})^2}} \quad (3.6)$$

where y is the \log_2 expression of a gene across all time points, T is the number of observations and y_t is the \log_2 expression of the gene at time t . This ensures that even a small change in the expression of a gene can be correlated with a large change in the expression of another gene. This is especially useful when clustering genes based on the shape of their profile rather than their magnitude.

3.4 Application of MAANOVA to Microarray Data

Until this point, a number of functions and processes have been introduced with minimal demonstration of their application with the senescence dataset. This section describes the application of those functions and the consequential results.

3.4.1 Quality Control

3.4.1.1 GridCheck

The initial quality control of the data is provided by GridCheck which compares the intensity of both channels of each microarray. Given the improvements made to the GridCheck function, it is now possible to identify patterns of probes which lie outside the expected regression of the plot through the implementation of a PCA and by circling probes which stray from the mean of the minor axis by more than 1.96 standard deviations ($p < 0.05$) in red.

The presence of red circles in the sub-plots should be expected since the purpose of the microarrays is to identify differential expression, but the presence of excessive numbers of red circles or an unusual distribution of these points, amongst other anomalies in the GridCheck sub-plots might indicate a problem with the collected data. Figure 3.9 demonstrates and explains three common anomalies seen in GridCheck sub-plots.

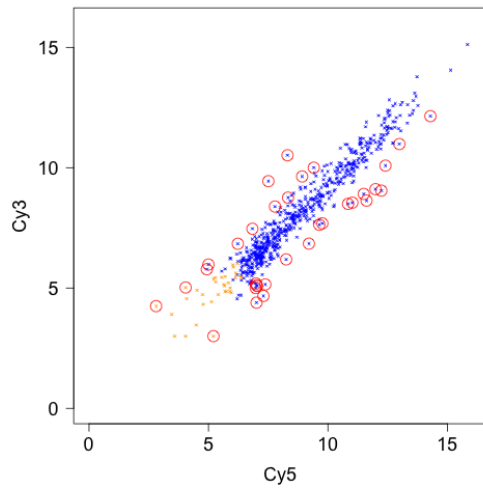
However, the identification of these plots is a labour intensive task which is heavily subjective, and as such the statistics table output by GridCheck to summarise each sub-plot has also been used to help rank those sub-plots so that the worst artefacts can be quickly identified without the need to scan through all 8,448 sub-plots produced during the analysis of the senescence dataset.

Two alternative statistics were used to rank the GridCheck sub-plots. These were:

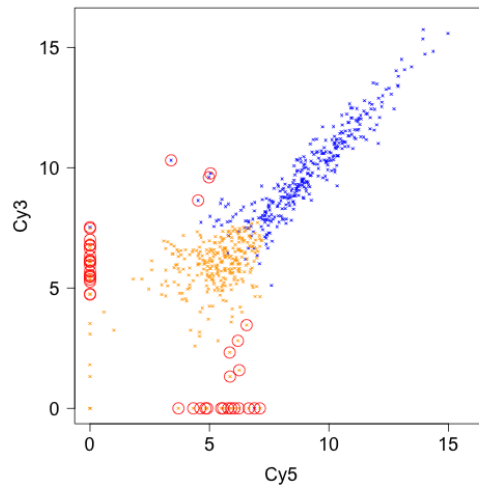
- (a) The p -value generated by a one-sample t-test against a mean of zero of the major axis values for probes which lie more than 1.96 standard deviations ($p \leq 0.05$) above zero in the minor axis (i.e. above the regression).
- (b) The p -value generated by a one-sample t-test against a mean of zero of the major axis values for probes which lie more than 1.96 standard deviations ($p \leq 0.05$) below zero in the minor axis (i.e. below the regression).

The reason for choosing these statistics is that, when the p -value is below a selected confidence threshold, it identifies a non-standard distribution of differentially expressed probes along the major axis and therefore a likely bias amongst those probes. This is true for all of Figures 3.9(b), (c) and (d).

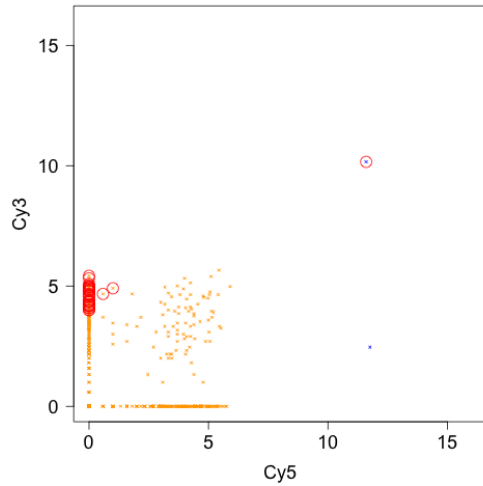
Figure 3.10 illustrates the effectiveness of the two statistics in the identification of problematic pin-tip groups. All 8,448 GridCheck plots were assessed by eye to identify



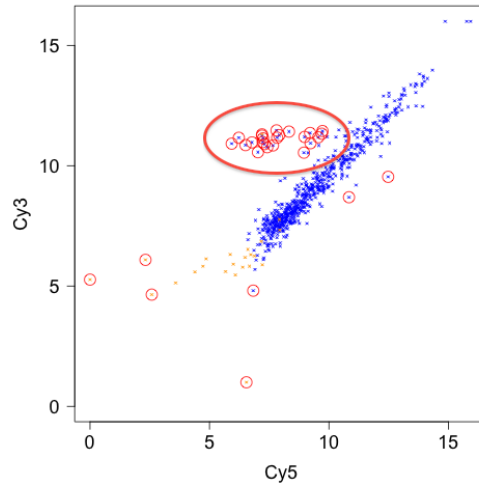
(a) Microarray 11, Sub grid (3,3) demonstrates no noticeable artefacts.



(b) Microarray 171, Sub grid (3,4) demonstrates a number of probes with zero intensity in either channel.



(c) Microarray 44, Sub grid (3,3) demonstrates many probes with zero intensity in either channel.



(d) Microarray 33, Sub grid (3,1) demonstrates an isolated cluster of probes with bias towards one dye.

Figure 3.9 – GridCheck sub-plots indicating typical artefacts of CATMA microarrays

Blue points represent the intensities of probes which are not flagged by ImaGene during microarray scan analysis whilst orange points are those probes which have been flagged. Red circles outline those probes which stray from the mean of the second major axis in a PCA by more than 1.96 standard deviations ($p < 0.05$). The expectation of a GridCheck plot with no noticeable artefacts, (a), is that most probes will be expressed to a similar level in both channels. Those probes which stray from this relationship should not be great in numbers, nor should they be biased to one side of the regression or be grouped into clusters. (b) shows a large number of probes with zero intensity in one or the other channels. This can be a sign of high-background in the array accompanied by background subtraction. (c) shows only two un-flagged probes, indicating a serious problem with the data of that sub-grid, either caused by low hybridisation due to low yield probes during printing, or incorrectly aligned grids during ImaGene analysis. (d) demonstrates a number of probes (as highlighted) which show higher ratios than expected from the remaining data and therefore form a cluster of outlying probes. This is caused by non-specific binding of the Cy3 channel.

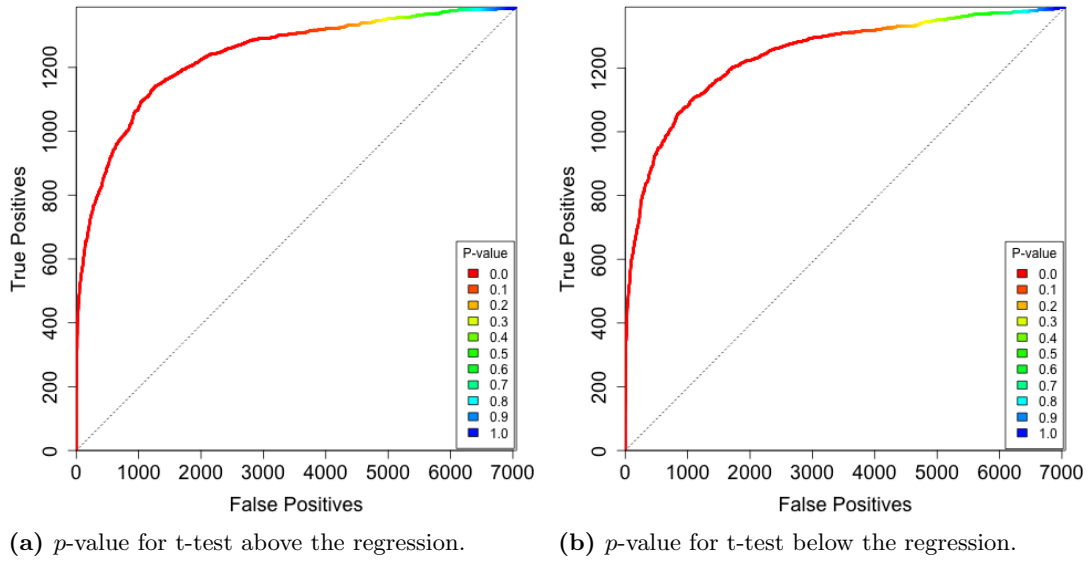


Figure 3.10 – ROC curves demonstrating the power of ranking sub-plots using statistics

A number of sub-grids throughout the senescence experiment were identified by a human expert as being of insufficient quality, based on the visual output of GridCheck. Simultaneously, sub grids were ranked by using statistical analysis of the GridCheck sub-plots. By comparing these ranks against the hand identified sub-grids, ROC curves were produced. The criteria for high ranking sub-grids were (a) the p -value from a t-test of deviation from a mean of zero on the major axis for outlying probes above the regression and (b) the p -value from a t-test of those below the regression.

common artefacts like the ones shown in Figure 3.9. This list was used as the gold-standard to which the automated ranking must aspire. In order to test the automated ranking against the gold-standard, ROC plots were produced in which true positives are plotted against false positives. True and false positives are both sub-plots identified by the automated ranking as having an artefact, but the distinguishing characteristic is that true positive were also identified as having an artefact when assessed by eye, whilst false positives were not.

Starting at the origin of the ROC plot, a list of GridCheck sub-plots ranked by the selected statistic is traversed. If the entry in the list was also identified by eye, the ROC plot curve moves one unit vertically. If it was not identified by eye, the ROC plot curve moves one unit horizontally. The greater the area under the curve, the more effective the strategy is. A dotted line is plotted diagonally across the plot because this defines the path of the curve in the case that random selections are being made. In both cases it can be seen that more than 80% of the true positives can be identified before even 20% of the false positives have been incorrectly ranked highly.

The top 45 worst GridCheck plots for each statistic, are shown in Table 3.4. These have been ranked by the p -value of the t-test to put the lowest p -value at the top of the table. It can be seen that arrays 23, 25, 27, 130 and 172 dominate the top ranks under both statistics and a closer inspection of those sub-plots of GridCheck

shows that they demonstrate a similar arrangement of outlying probes to those shown in Figure 3.9(b). Interpretation of these results is made later in this thesis in Section 3.4.2 where the actions taken to rectify any defects are described.

Although this proves that these ranked statistics can be used to accurately identify the majority of GridCheck sub-plots with a visible artefact, it does not demonstrate how this information may be used when processing alternative datasets, since the GridCheck plots, on this occasion, have already been categorised by eye. However, the p -value threshold which identifies approximately 80% of the true positives in this dataset is ($p \leq 2.0 \times 10^{-4}$) for both statistics. It would not be feasible to correct every issue identified in this way, but it does provide a semi-automated approach to finding the worst areas of the microarray experiment in very little time.

3.4.1.2 RIPlot

RIPlot applied to the senescence dataset identified a few artefacts that are common amongst microarrays of the experiment. One of these artefacts is also identified by GridCheck, whilst the others are very large artefacts affecting only a small number of probes in each of a large number of sub-grids. The analysis of whole microarrays allows RIPlot to identify these deviations from the ideal whilst the small impact they make on individual sub-grids makes them invisible to GridCheck.

Figure 3.11 presents an archetypal plot (Figure 3.11(a)) as would be expected from a microarray with no major flaws. In this case, the ratio is zero mean-centred throughout the entire intensity range with only a small increase in the variance of the ratio at lower intensities. Few probes exist to the far right where probes which are saturated in one or both channels reside and the majority of blue probes exist between combined log intensities of 12 and 30.

A microarray which has been scanned at too high gain would provide data as shown in Figure 3.11(b), where many probes exist in a $>$ shape to the right of the plot at the highest combined intensities. This shape is characteristic of probes which are saturated in one or both channels because saturation in both channels is the only situation that allows a combined \log_2 intensity of 32. At this intensity, the \log_2 ratio must be zero as both channels are the same. As the combined intensity is decreased because the channels are no longer both saturated, the most extreme ratio achievable is when one channel remains saturated. The ratio is therefore always restricted. The same concept is true of the left edge of the plot where a shape resembling $<$ can be found. Having many saturated probes is not good for the data because the distinction between the probes is lost if all have the same intensities.

When a number of probes are substantially distant from a ratio of zero, it is possible for them to form a cluster of points in the RIPlot as shown in Figure 3.11(c). This cluster is also identifiable in the GridCheck sub-plot shown in Figure 3.9(d). The reason for this is often non-specific binding of large numbers of probes as the problem only exists in one channel.

Array	Meta Row	Meta Column	p -value
130	10	4	3.28×10^{-44}
27	2	3	5.42×10^{-43}
130	1	3	5.14×10^{-42}
130	9	1	7.87×10^{-42}
25	3	3	2.51×10^{-41}
25	8	1	1.94×10^{-40}
25	10	3	9.61×10^{-40}
130	8	1	2.45×10^{-39}
25	2	2	2.77×10^{-37}
25	9	3	3.08×10^{-37}
25	9	2	8.19×10^{-37}
130	11	3	1.11×10^{-36}
27	2	2	1.18×10^{-36}
130	4	2	1.76×10^{-36}
25	5	2	1.65×10^{-35}
23	9	1	2.00×10^{-35}
27	7	2	3.37×10^{-35}
130	7	2	4.02×10^{-35}
25	10	1	1.99×10^{-34}
130	10	3	2.98×10^{-34}
23	9	3	5.77×10^{-34}
172	2	2	6.54×10^{-34}
23	1	2	8.85×10^{-34}
130	7	1	9.18×10^{-34}
130	12	3	3.70×10^{-33}
27	6	3	4.98×10^{-33}
27	12	3	1.00×10^{-32}
130	7	4	1.34×10^{-32}
130	4	4	1.38×10^{-32}
172	5	4	2.34×10^{-32}
130	10	2	2.61×10^{-32}
27	4	1	2.99×10^{-32}
25	10	2	5.02×10^{-32}
130	11	2	7.69×10^{-32}
130	8	4	9.54×10^{-32}
27	4	2	1.08×10^{-31}
27	1	3	1.41×10^{-31}
130	1	1	3.66×10^{-31}
25	7	1	3.90×10^{-31}
23	10	4	5.01×10^{-31}
25	9	1	1.21×10^{-30}
23	4	2	1.30×10^{-30}
130	10	1	1.74×10^{-30}
25	7	4	2.52×10^{-30}
25	1	1	3.66×10^{-30}

(a) Sub-grids demonstrating abnormally distributed probes above the regression

Array	Meta Row	Meta Column	p -value
25	4	3	1.71×10^{-43}
25	11	2	2.14×10^{-42}
24	1	3	6.38×10^{-42}
130	9	2	1.36×10^{-41}
130	6	4	1.02×10^{-38}
25	11	4	1.92×10^{-38}
27	5	3	7.76×10^{-38}
130	12	2	7.87×10^{-38}
27	3	2	3.75×10^{-37}
25	11	1	4.65×10^{-37}
23	9	2	6.59×10^{-37}
25	8	4	8.41×10^{-37}
130	2	2	4.07×10^{-36}
25	3	2	4.86×10^{-36}
27	9	2	6.26×10^{-36}
25	2	3	7.38×10^{-36}
130	1	2	4.69×10^{-35}
172	3	2	1.15×10^{-34}
25	2	1	1.47×10^{-34}
25	7	2	2.30×10^{-34}
130	3	2	2.72×10^{-34}
25	9	4	1.69×10^{-33}
130	2	1	8.46×10^{-33}
172	1	3	1.24×10^{-32}
130	12	4	1.51×10^{-32}
26	6	3	3.64×10^{-32}
130	2	4	1.62×10^{-30}
130	1	4	1.95×10^{-30}
27	2	2	2.28×10^{-30}
130	8	2	2.66×10^{-30}
17	6	3	3.83×10^{-30}
130	11	4	5.26×10^{-30}
25	8	2	5.67×10^{-30}
25	1	2	6.06×10^{-30}
25	1	3	1.00×10^{-29}
25	11	3	1.08×10^{-29}
59	7	3	2.49×10^{-29}
23	2	1	3.13×10^{-29}
56	4	2	3.46×10^{-29}
25	2	4	7.02×10^{-29}
25	5	4	3.17×10^{-28}
25	12	2	3.40×10^{-28}
26	3	2	3.55×10^{-28}
23	10	2	5.26×10^{-28}
172	3	3	6.69×10^{-28}

(b) Sub-grids demonstrating abnormally distributed probes below the regression

Table 3.4 – Top 45 worst GridCheck sub-plots ranked by t-tests of the major axis for outlying probes above or below the regression

One-sample t-tests performed on the major axis values of outlying probes in GridCheck sub-plots were performed independently for those probes above, (a), and those below, (b), the regression of the plot. Lower p -values from these tests determine that the sub-plot should be checked for artefacts. It can be seen that microarrays numbered 23, 25, 27, 130 and 172 dominate the lowest observed p -values in both cases, and these were found to have many outlying probes in a similar pattern to the GridCheck sub-plot shown in Figure 3.9(b); the most frequently observed issue with GridCheck sub-plots in the senescence experiment.

Some microarrays show a high variance in the ratio throughout the entire intensity range, as shown in Figure 3.11(d). This could be an indication that one or both of the channels are noisy and as such should be visually verified, but can, in this case, also be caused by two samples being hybridised to the same microarray despite being temporally distant. Some microarrays in the senescence design are hybridised by a day 11 and day 1 sample to complete the loops and, since a lot of gene expression changes have occurred during that time, the resulting RIPlot can show a high level of variance in the ratio.

Given that RIPlot does not generate statistics about the plots and that there are so few plots when compared with the sub-plots of GridCheck, these plots were visually assessed and compared with the output of ArrayView to try to explain observed artefacts.

3.4.1.3 ArrayView

As was described in Section 3.3.1.3, ArrayView allows the visualisation of the spatial distribution of ratios and intensities throughout the probes of the microarrays in the form of a heat-map. This type of visualisation is especially informative where transformations have been applied to the data as a comparison before and after the transformation demonstrates its effectiveness. ArrayView has also been used to provide a visual determination of the cause of artefacts identified by GridCheck and RIPlot, as has been described above. In this way, it is possible to determine the course of action to take to avoid these artefacts during the analysis.

Figure 3.12 presents the ratio and intensity heat-maps for those pin-tip groups which were highlighted by GridCheck in Figure 3.9. Figures 3.12(a) and 3.12(d) are the ratio and intensity heat-maps, respectively, for the pin-tip group exhibiting a large number of probes with zero intensity in one or both channels in Figure 3.9(b). The probes affected cause the intensity heat-map to display an exceptional number of dark spots on the intensity heat-map when compared with those of other microarrays. Those probes are also seen to present large ratios, plotted as bright red and bright green in the ratio heat-map. These large ratios are an artefact of one channel having zero intensity. These artefacts are most likely caused by the background level exceeding that of the signal on the microarray and the consequential background subtraction causing the signal to be reduced to zero.

The pin-tip group with almost entirely zero expression probes identified by GridCheck in Figure 3.9(c) has been plotted as a ratio and an intensity heat-map in Figures 3.12(b) and 3.12(e) respectively. On this occasion, it is obvious that the overall intensity of this pin-tip group is significantly lower than the rest of the microarray, but again this is less obvious in the ratio heat-map. This is a serious issue since the expression data for over 600 probes in these two samples are effectively missing and will most likely cause disruption in the analysis of the whole experiment.

Figure 3.9(d) identified a pin-tip group with an outlying cluster of probes which

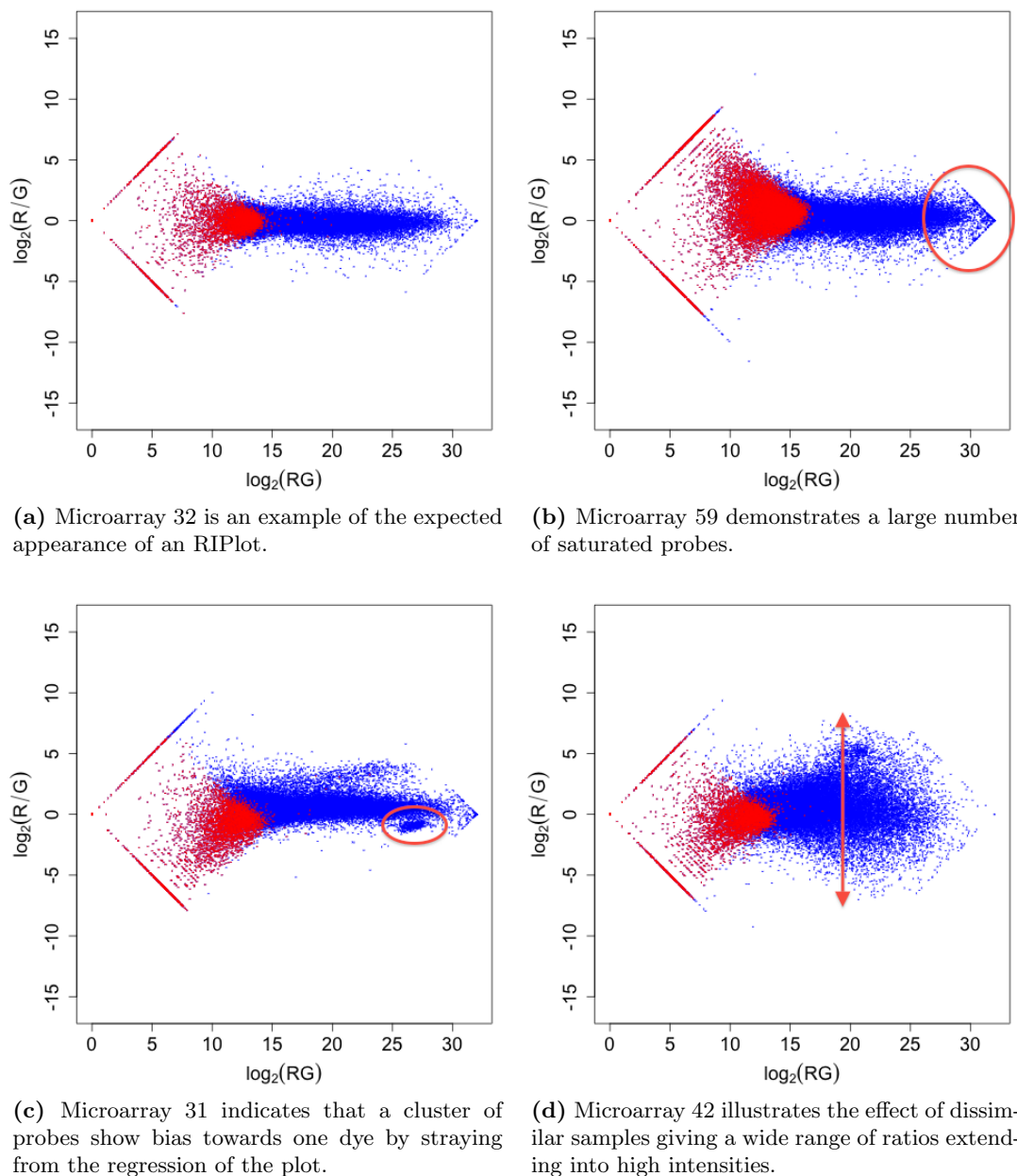


Figure 3.11 – RIPlot output indicating typical artefacts of CATMA microarrays

An archetypal plot, as in (a), exhibits characteristics such as the majority of blue points existing between a combined \log_2 intensity of 12 and 30, no groups of points straying from the main regression of the plot and a tight, uniform range of ratios along the blue points. (b) shows an example plot in which many probes are saturated in one or both channels (highlighted) providing no distinction between them. In (c), a number of probes form a cluster (highlighted) which strays from the main regression, indicating a possible artefact of the microarray. The wide range of ratios in (d) demonstrates that the channels of this microarray have less similarity than on other arrays, requiring that any possible reason for this be investigated as it may be caused by a problem with the microarray itself.

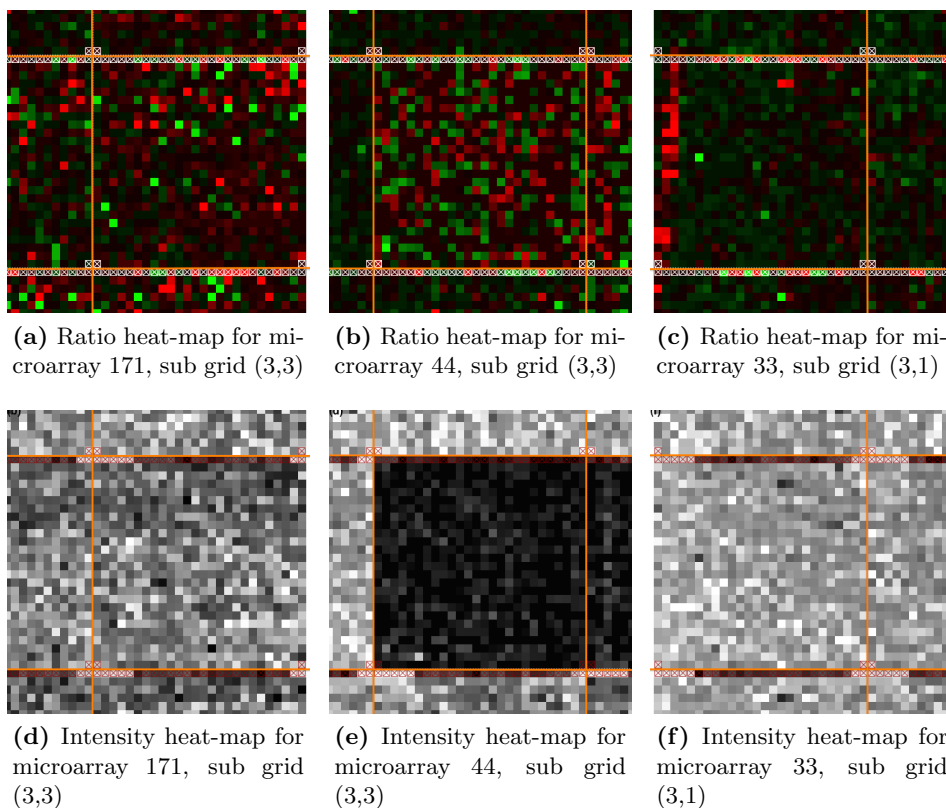


Figure 3.12 – Cropped ArrayView heat-maps for sub-grids flagged by Grid-Check

The sub-figures show cropped areas of ArrayView heat-maps designed to illustrate some of the sub-grids flagged by GridCheck. For clarity, the margins between the pin-tip groups have been defined by orange lines, though these do not appear as boldly on the ArrayView output. In each case, the flagged sub-grid is the one displayed in full, whilst the edges of other sub-grids are shown only as a comparison. (a) and (d) illustrate the ratio and intensity plots respectively for the sub-grid presented by GridCheck in Figure 3.9(b) where many of the outlying probes were zero intensity in at least one channel. It can be seen that this results in more low intensities appearing in the intensity heat-map when compared with that of (f) where the intensities are normal. Where the lower intensities exist, these are often represented by maximum ratio in the ratio heat-map since one channel is demonstrating zero expression. (b) and (e) are the corresponding ratio and intensity heat-maps for the sub-grid flagged by GridCheck in Figure 3.9(c). In this case, the intensity heat-map highlights the severity of the low intensities throughout the sub-grid and this is also identifiable by the increased ratios of the probes in the sub-grid. As has been stated, (f) demonstrates the appearance of an intensity heat-map with no unusual characteristics, despite representing the same sub-grid shown by GridCheck to contain a cluster of outlying probes towards one channel in Figure 3.9(d). (c) is the corresponding ratio heat-map which clearly shows a range of probes which are biased towards the red channel. The thin, long shape of this artefact identifies it as a scratch on the microarray rather than a probe hybridisation problem.

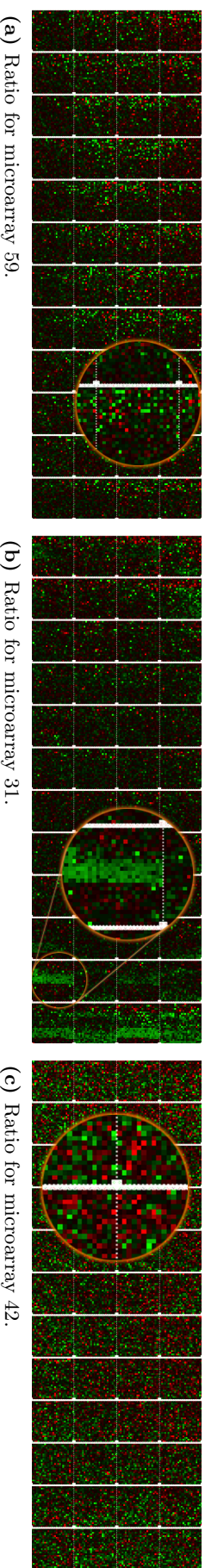
strayed from the main regression, giving rise to the conclusion that some probes have been dye-biased. Figures 3.12(c) and 3.12(f) are the ratio and intensity heat-maps, respectively, for the pin-tip group and it is clear that the probes causing the cluster of points are those plotted in red to the left of the ratio heat-map. Although these are obvious in the ratio heat-map, it appears that they show no identifiable abnormality in the intensity heat-map. This type of artefact can be seen on many different microarrays in the experiment, although it usually only affects a minor number of probes as can be seen here. The fact that the overall intensity is unaffected may indicate that it is coincidence, but it can often be attributed to marks on the scan images which cause high background for those probes.

Three microarrays with concerning characteristics were identified by RIPlot in Figures 3.11(b), 3.11(c) and 3.11(d). The ratio and intensity heat-maps for these microarrays are presented in Figure 3.13 and are compared.

The first concern, shown in Figure 3.11(b), is an example of a microarray in which many of the probes have reached saturation in at least one channel. The ratio and intensity heat-maps for this microarray are presented in Figures 3.13(a) and 3.13(d) respectively. The saturated probes do not appear to have an effect on the overall appearance of the ratio heat-map, but the intensity heat-map does demonstrate a large number of maximum intensity probes, leading to some probes appearing almost white and giving the impression of high contrast across the plot. These heat-maps do not indicate a localisation of the over-saturated probes and hence it should be a case of simply re-scanning the microarray at a lower gain to prevent the observed saturation.

Figure 3.11(c) highlights a number of probes which form a cluster of points which stray from the main regression of the data on that microarray. Figures 3.13(b) and 3.13(e) present the ratio and intensity heat-maps, respectively, for this microarray and the cause is clearly visible in both heat-maps. As has been noticed on CATMA microarrays, sometimes localised groups of probes will exhibit non-specific binding. This results in highly fluorescent groups of probes and consequential loss of gene expression data. The cause is unknown, but the identification is simple with the tools provided by MAANOVA and the most effective solution is to re-hybridise the samples to a new microarray. A discussion of actions taken because of observations made through this analysis is shown in Section 3.4.2 alongside actions taken for other reasons throughout the experiment.

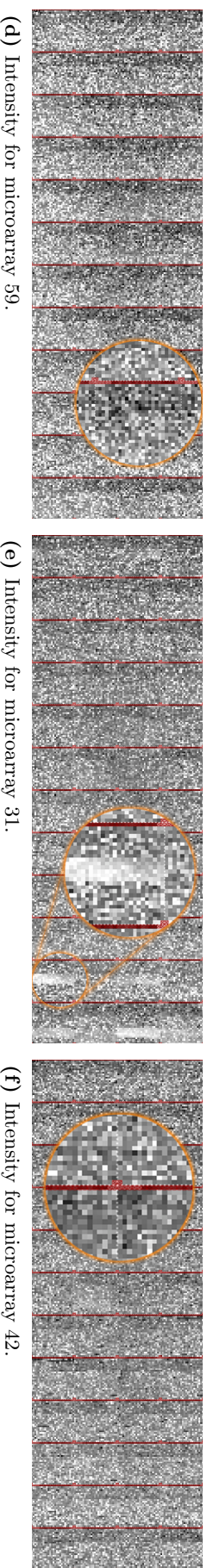
The final concern identified from the plots of RIPlot, was that of a plot with a wide range of ratios extending throughout all intensities, displayed in Figure 3.11(d). The ratio and intensity heat-maps for this microarray are presented in Figures 3.13(c) and 3.13(f) respectively. The only indication of this unusually wide range of ratios is a high number of bright green and bright red probes on the ratio heat-map. As these are not localised to any one area of the microarray, this is most likely caused by the samples themselves being distinctly different. This could be an indication of a poor quality sample, but an investigation of the experimental design shown in Figures 1.3 and 1.4



(a) Ratio for microarray 59.

(b) Ratio for microarray 31.

(c) Ratio for microarray 42.



(d) Intensity for microarray 59.

(e) Intensity for microarray 31.

(f) Intensity for microarray 42.

Figure 3.13 – ArrayView heat-maps for sub-grids flagged by RIPlot

The sub-figures show whole ArrayView heat-maps in which the entire microarray can be seen. Areas of these have been enlarged and circled in orange to assist in their interpretation. (a) and (d) show the ratio and intensity heat-maps, respectively, for the microarray flagged by RIPlot in Figure 3.11(b) where a number of the probes of the microarray are seen to be saturated in at least one channel. This is identifiable in the intensity heat-map because a greater number of probes are exhibiting near-white representation, indicating high intensity. This effect is also manifested by the unusual contrast between the brightest and dimmest probes in the plot. The ratio heat-map for this microarray does not show any related artefacts. A cluster of probes showing unusual ratios was identified during the observation of the RIPlot shown in Figure 3.11(c), and the corresponding ratio and intensity heat-maps are presented in (b) and (e). It can be seen that the artefact is clearly identifiable in both the ratio and the intensity heat-maps, towards one end of the microarray. This pattern of probes with high intensities and extreme ratios is often seen on CATMA microarrays and is assumed to be non-specific binding, but the reason for this is unknown. The wide range of ratios identified by the RIPlot shown in Figure 3.11(d) can be clearly identified in (c) as a larger than usual number of probes plotted in bright green and red, though these are not localised to any particular region of the microarray. The intensities of the probes, shown in (f), identify no artefacts and by referring to the experimental design it becomes clear that the cause of the extreme ratios is that the hybridised samples are from each end of the time series and therefore have substantially different levels of gene expression.

indicates that this microarray has a sample from the first time point and the last time point hybridised to it. As these are the most distant time points of the experiment, a higher degree of differentiation should be expected.

It has been shown that, GridCheck and RIPlot are able to identify certain artefacts each, and that neither is capable of identifying all artefacts. Yet it has also been shown that ArrayView always presents an identifiable pattern when an artefact exists, though some are more obvious when looking at GridCheck or RIPlot, such as saturated probes. In most cases, ArrayView is an adequate visualisation of the data for identifying where improvements could be made to the experiment. However, the analysis is qualitative rather than quantitative, since minor defects such as subtle dye-bias would most likely go unnoticed. Therefore, it would be desirable to provide a statistical analysis of the ArrayView output which can show the user where artefacts are most likely to exist and draw their attention to underlying problems. This was the inspiration for developing the histogram output for the heat-maps so that the distribution of ratios and intensities throughout the microarray can be observed.

Intensity histograms, such as the one shown in Figure 3.4(b), is accompanied by statistics about their distribution such as those shown in Table 3.2(b). The distribution of the intensity data is more complicated and would require further consideration before statistical tests could be applied to decide whether any abnormal intensities are present.

A test implemented with the intention of automating the identification of artefact locations was that of an Analysis of Variance (ANOVA) in which the meta-rows and meta-columns of the sub-grids were used to group the probes. The ANOVA returns two p -values per microarray defining whether the probes in the meta-rows are significantly different and whether the probes in the meta-columns are significantly different. Where an artefact exists and an unusual localisation is present, these tests would therefore identify this by resulting in a significantly low p -value.

The results of the ANOVA tests did appear to provide useful information about the meta-columns, since the comparison is between only four columns with a quarter of the total probes in each, whilst for the meta-rows, the tests were almost always highly significant, most likely due to a small deviation in only one or two rows of the twelve present on each microarray. The test is too sensitive to be of any real use but time constraints limited the time which could be spent on providing an alternative. Instead, the graphical output of ArrayView was enhanced by the new analysis provided by TechRepCheck to ensure it could be interpreted in a semi-quantitative manner.

The graphical output produced by ArrayView, to this point, has been shown as an extension only of the ArrayView function itself. However, the usefulness of ArrayView can be greatly extended by combining it with information obtained from TechRepCheck. Gold crosses, introduced in Section 3.3.2.1, are plotted over probes identified as replicate outliers and consequently estimated in intensity for one or both channels. These are thoroughly discussed and applied throughout Section 3.4.2.

3.4.2 Application of TechRepCheck and Correction of Poor Microarrays

Using TechRepCheck improves the resulting data in two ways. Firstly, the output of ArrayView, already deemed a comprehensive way of visualising the data, is improved by the addition of gold crosses over the probes that are identified as significant outliers in their respective technical replicate sets. These gold crosses draw the attention of the individual looking for artefacts on the microarrays and ensures that those artefacts do not go unnoticed. Where the individual chooses to take no action to correct the artefacts because of time or money constraints, they can then choose to take advantage of the intensity estimation features for the identified probes in which the remaining non-outlying replicates can be used to provide a more consistent intensity value for the outlier, if they wish.

During the analysis of the senescence dataset, TechRepCheck was used, initially to identify artefacts in a semi-quantitative, semi-automated manner. The position of the gold crosses assisted in the identification of outlying intensities when compared with their respective technical replicates. This led to the discovery of artefacts which had, until now, not been identified, and their removal after corrective steps, shown below, had been taken was then verified by TechRepCheck.

With the powerful visualisation provided by TechRepCheck, it was possible to identify which microarrays in the senescence dataset required attention. It was found that, in most cases, either the analysis of the scans had been performed inappropriately for some microarrays or an artefact of the microarray itself had caused a bias in the data which could not be corrected through a process of transformation. The plots presented by ArrayView after TechRepCheck were carefully reviewed to identify where significant improvements could be made for the senescence dataset without introducing excessive financial expenses. Figure 3.14 shows the appearance of regions of ArrayView plots before TechRepCheck, after TechRepCheck and then after a correction has been applied.

Figure 3.14(a) shows the ArrayView output of microarray 44 before the application of TechRepCheck. This microarray was identified by GridCheck as having almost every probe of zero intensity in at least one channel. Once TechRepCheck had been applied to the microarray, the gold crosses were placed as can be seen in Figure 3.14(b). This is very typical of a problem existing in those probes since the majority of the gold crosses indicate a problem in both channels. Upon close inspection, the microarray scan itself was not well represented by the numerical data, leading to the conclusion that a mistake was probably introduced at the stage in which the microarray scans were analysed by ImaGene. After re-analysing the original scan, new data was obtained and the resulting TechRepCheck crosses shown in Figure 3.14(c) indicate that the problem was eradicated.

Figure 3.14(d) is another example of ArrayView output from microarray 164. This microarray has not been identified by any of the previous analyses. Once Tech-

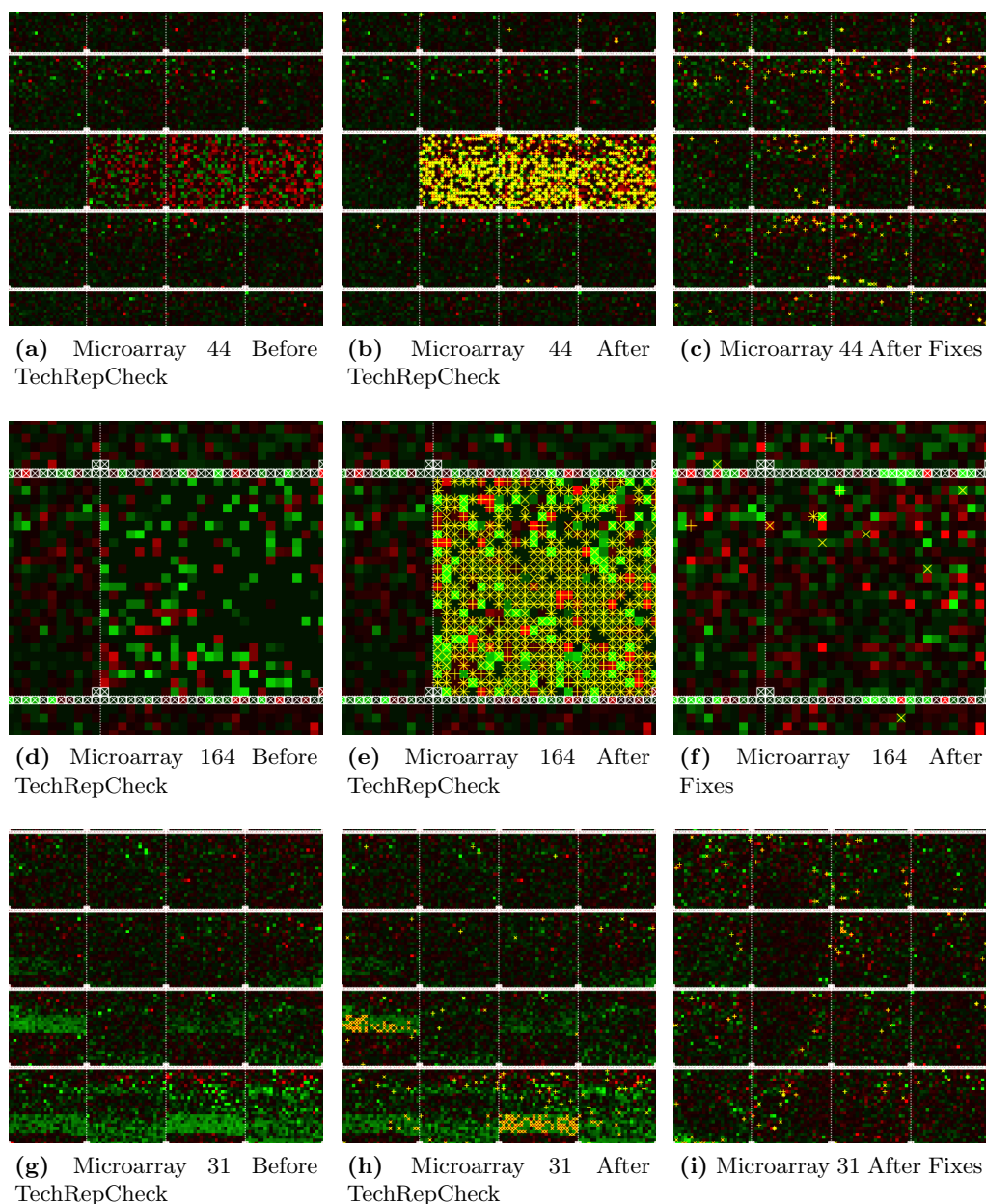


Figure 3.14 – ArrayView with TechRepCheck estimate indicators

In the first example, an artefact can be seen in the ratio plot of ArrayView as an area of brighter colours (a) and this is confirmed by the gold crosses produced by TechRepCheck (b). After reanalysing the array scan using ImaGene, the new data does not contain the same defect (c). In the second example, the artefact is less obvious to the eye (d) but TechRepCheck is able to identify outlying probes throughout a sub grid of the array (e). Reanalysing the array scan using ImaGene once again removes the issue resulting in far fewer probes identified by TechRepCheck (f). In a third example, the artefacts affect only specific probes of the array (g) and TechRepCheck successfully identified a number of these outlying probes (h). Further investigation found that this artefact was also present on the array scan and therefore the samples were hybridised to a new array and the artefact was removed (i).

RepCheck was applied, the crosses were placed as shown in Figure 3.14(e), indicating an almost identical issue to that shown in Figure 3.14(b). The same approach was taken to correct the issue and the output from ArrayView after re-analysing the microarray scan (Figure 3.14(f)) indicates that this was equally as successful.

Figure 3.14(g) is the final example of ArrayView output before TechRepCheck was applied to microarray 31. This microarray was identified by RIPlot as having a cluster of probes biased towards one dye. Figure 3.14(h) shows that the gold cross locations were subtle on this occasion, but a number of crosses are localised to the artefact. The crosses are all the same orientation, indicating that only one channel is affected. By returning to the scan for this microarray, it was clear that the channel in question does show high intensity for these probes which cannot be seen on the other channel and a decision was made to re-hybridise the samples to a new microarray. The data resulting from this new hybridisation can be seen in Figure 3.14(i) where there are no localised clusters.

By mainly observing the ratio heat-map of ArrayView over which TechRepCheck estimates had been plotted, 35 of the 176 microarrays were identified as having artefacts in either the ratios themselves, or, in most cases, having localised groups of probes which have been marked as an outlying replicate in TechRepCheck. Each of these 34 microarrays were then compared with the scans analysed by ImaGene (BioDiscovery) to ensure that the problem lies with the microarray itself before opting to re-hybridise the samples to a new microarray. Each microarray was then placed into one of four classifications:

- The problem does not exist in the scans and the scans should be re-analysed.
- The problem exists in the scans and affects a large proportion of the microarray requiring re-hybridisation of the samples; or
- The problem exists in the scans but only affects a small proportion of the microarray and should be disregarded.
- The problem was identified using GridCheck or RIPlot, but does not have a large impact on the microarray overall.

Although it would be favourable to re-hybridise all the microarrays which fall into the latter three categories, the highly replicated experiment can cope with a small number of poor data points without the final model being influenced significantly, rather than expending resources to perfect the data with very little gain.

The classification of the low quality microarrays and hence the course of action for each is displayed in Table 3.5. As can be seen, eighteen of the microarrays exhibited artefacts in the MAANOVA output which could not be identified on the microarray scans, demanding that those scans be re-analysed. By including the re-analysed data back into the experiment, these eighteen microarrays were corrected without the need to repeat any microarrays. Of the remaining sixteen microarrays with artefacts which

Microarrays	Classification	Action Taken
27 34 38 39 40 44 47 49 55 82 90 91 101 102 112 164 165 174	Microarray scan presents no obvious defect	As the artefact only exists in the MAANOVA output and must therefore lie in the transition from microarray scan to input data, these microarray scans were re-analysed with ImaGene (BioDiscovery).
20 31 62 63	Microarray scan presents a major defect which affects a large number of spots	The loss of data from these microarrays will be strongly influential in the analysis and so the samples were re-hybridised to a new microarray.
19 28 29 97 126 158 161 166	Microarray scan presents a minor defect which affects a small number of spots	As these microarrays mostly contain good data, repeating will be both too costly and without great gain so these artefacts will be overlooked.
23 25 59 130 172	Microarray flagged by GridCheck or RIPlot but not by TechRepCheck	The artefacts are relatively small in magnitude compared to others that have been identified and will be overlooked.

Table 3.5 – Classification of low quality microarrays after analysis by MAANOVA

After the analysis of MAANOVA, 35 of the 176 microarrays in the senescence experiment were identified as containing artefacts which would be influential on the analysis of the experiment. Each was placed into one of four classifications as shown, based on the appearance of the microarray scan. Depending on which classification each microarray received, the action taken to resolve the problem is shown.

could be seen in both the MAANOVA output and on the scans, only four were so bad that they would require re-hybridisation to a new microarray. The remaining thirteen were considered to be poor, but acceptable since the affected probes were only a very small minority of the entire microarray.

Table 3.6 shows the samples hybridised to each of those microarrays shown in Table 3.5, so that the affect of the defects can be associated with samples of the experiment. It should be noted from this that only the thirteen arrays shown at the bottom of the table were left as they were. The overruling reason for this is that those microarrays only demonstrated minor artefacts. However it is also important to note that, amongst those thirteen, five relate to a comparison between days 1 and 2. This was deemed acceptable since it was unlikely that any important changes in expression would happen this early on in the time course.

In almost all cases, the artefacts revealed themselves by offsetting many probes in a localised area which were then highlighted by TechRepCheck. The inclusion of TechRepCheck in the MAANOVA analysis has allowed the identification of scan analysis problems which have otherwise remained invisible. Some pin-tip groups were overwhelmed by estimated spots but without any indication of a problem from any of the current quality control methods because the \log_2 ratios and the log intensities were not

Array Num.	Labelled with Cy3			Labelled with Cy5		
	Day	Time of Day	BioRep	Day	Time of Day	BioRep
27	6	PM	A	7	PM	A
34	8	AM	A	9	AM	B
38	4	PM	C	5	AM	A
39	8	PM	C	9	AM	B
40	11	PM	D	1	PM	D
44	4	AM	D	5	PM	A
47	2	PM	A	3	AM	A
49	11	AM	B	1	PM	C
55	2	PM	A	3	PM	A
82	9	PM	A	10	PM	A
90	9	PM	C	10	PM	B
91	4	AM	C	5	AM	C
101	9	AM	A	10	PM	B
102	3	PM	A	4	AM	C
112	4	PM	B	5	PM	B
164	5	PM	D	6	PM	D
165	7	PM	B	8	AM	A
174	8	AM	C	9	AM	C
20	8	AM	C	9	PM	D
31	3	AM	D	4	AM	D
62	6	PM	B	7	AM	D
63	7	AM	C	8	PM	A
19	8	PM	C	9	PM	C
28	5	AM	B	6	AM	B
29	1	AM	B	2	AM	C
97	11	AM	C	1	PM	D
126	1	PM	B	2	AM	D
158	7	AM	B	8	PM	D
161	2	PM	B	3	AM	D
166	7	AM	D	8	AM	A
23	1	AM	D	2	PM	A
25	1	AM	A	2	PM	B
59	8	PM	B	9	PM	A
130	1	AM	C	2	AM	B
172	4	AM	C	5	PM	D

Table 3.6 – Sample allocations for low quality microarrays

An extraction of the experimental design is shown to identify the samples which are most affected by the microarrays deemed lower quality by MAANOVA analysis. Those in the top section were re-analysed, leading to correction of the identified flaws. Those in the second section were re-hybridised to new microarrays to correct the identified flaws. The third section consists of microarrays which only demonstrate very minor defects whilst the final section shows those microarrays which were identified at the sub grid level by GridCheck, but that only showed minor artefacts in TechRepCheck.

unusual. In the majority of cases, these problems were caused by mis-alignment of the grid in ImaGene, which is why the \log_2 ratios and log intensities did not appear abnormal, their values being obtained from neighbouring spots. TechRepCheck allowed not only the identification but the localisation of this problem so that it may be resolved.

Before the actions shown in Table 3.5 had been applied, 64,734 of the 5,440,512 probes (1.19%) were identified as outliers amongst their respective technical replicates. However, after taking the actions shown, 10,771 probes were restored to an expression level which was no longer outlying amongst the technical replicates. However, a further 6,360 new probes were now identified as an outlier, resulting in a total of 60,296 of the 5,440,512 probes (1.11%) being classed as an outlier. This net improvement of 4,438 probes may only contribute a small percentage of the probes across the entire experiment, but it is very obvious from the resulting ArrayView heat-maps that the most major problems have been removed from the experiment. Although new probes have now been identified as outliers, this is because those probes were previously masked by the extremity of the problems which have now been corrected. These remaining outliers can either be left unmodified and remain an outlier during the model fitting process, or otherwise have a more consistent intensity estimated.

Another alternative strategy to dealing with poor quality replicates could be to exclude those microarrays from the model fitting process. However, since no major flaws could be identified in the microarrays once all repetitions were completed, estimates were produced for the remaining outliers during the analysis of the senescence dataset. The design of the experiment is such that removal of microarrays should minimally impact the results of the overall analysis, but this has not been attempted here.

3.4.3 LOWESS Transformation

Once the data was quality controlled, repetitions completed at the experimental stages and technical replicate outliers estimated, the data was ready for transformation. This step is used as a method of neutralising dye biases across the microarrays and reducing the effects of damage to the microarrays. Whilst attempting to fit each of the transformation functions provided by MAANOVA and comparing the results by observing the changes to the ArrayView plot, it was noted that the joint-LOWESS (rlowess) was the only existing function to successfully remove dye biases which vary in severity across the microarrays. However, this often left identifiable differences between the ratios of adjacent sub-grids, particularly where microarrays appeared to exhibit biased hybridisation in selected probes prior to the transformation. Not wishing to over-fit the whole experiment by using a smaller span value for the LOWESS fitted curve, the method presented in Section 3.3.2.2 was developed as an alternative. This allowed sub-grids to be used as a variable for the LOWESS model and therefore bias normalisation to occur within sub-grids. When used alone, this new method did not provide an effective fit for the data, either, but when used before and in conjunction with the joint-LOWESS, they provided normalisation which appears to eradicate most spatial artefacts typically

seen throughout the experiment.

Figure 3.15 shows the transition of the probes of microarray number 5 in the experiment as the two LOWESS methods described above are consecutively used to transform it. Each of the sub-plots shows an RIPlot of the data as it is transformed. Figure 3.15(a) shows the RIPlot of the untransformed data where it can be seen that the intensities are well distributed with few below 40% intensity or above 90%. The ratio, however, is biased across all intensities, towards the lower channel. Figure 3.15(b) presents the same data, but with a red line indicating the fitted curve of the sub-grid weighted LOWESS. This red line represents those \log_2 ratios throughout the intensity scale which should be zero and therefore no differential expression. Once this fitted curve has been used as an adjustment to the ratios, the transformed data is presented as shown in Figure 3.15(c). Here, the data is more centralised around a mean \log_2 ratio of zero, but has a very slight gradient indicating an unwanted relationship between intensity and ratio. The same data is shown in Figure 3.15(d) with the LOWESS fitted curve based on the joint-LOWESS transformation plotted in red. After transforming the data this second time, the result can be found in Figure 3.15(e). This final form of the data still shows the majority of the variation in the ratio as was seen in the original data, but is also zero mean-centred throughout the intensity range.

The result of such a double transformation can be illustrated more clearly in an ArrayView plot before transformation, after sub-grid LOWESS transformation and after joint-LOWESS transformation. These can be found in Figure 3.16 where the same microarray better exhibits its dye-biases and the removal process is easier to understand. Figure 3.16(a) illustrates the appearance of ArrayView for the biased array before any transformation is applied. The application of the sub-grid weighted LOWESS tends to remove the bias within individual sub-grids, which goes a long way towards removing the obvious spatial bias, as shown in Figure 3.16(b). However, it is obvious that there is a level of bias still present along the length of the microarray, manifesting itself as a patch of red in the top-right area of the microarray. The subsequent application of the joint-LOWESS transformation then results in the data shown in Figure 3.16(c), where no visual bias can be seen, but the ratios have not been overall noticeably reduced. This data would then be ready for mixed model fitting.

Given the reassuring results of applying two LOWESS transformations to the data, it was decided that this would prove the best strategy for transforming the senescence dataset and so the transformation applied was as shown in Figures 3.15 and 3.16 where the sub-grid weight LOWESS was implemented first, with a span of 0.1, and the joint-LOWESS was implemented subsequently, also with a span of 0.1.

3.4.4 Generating a Mixed Model of Gene Expression Data

The mixed model for the senescence dataset was produced using the process shown in Section 3.3.3 whereby the resulting fixed terms were:

- **Day** – The day upon which the sample was collected (11 levels)

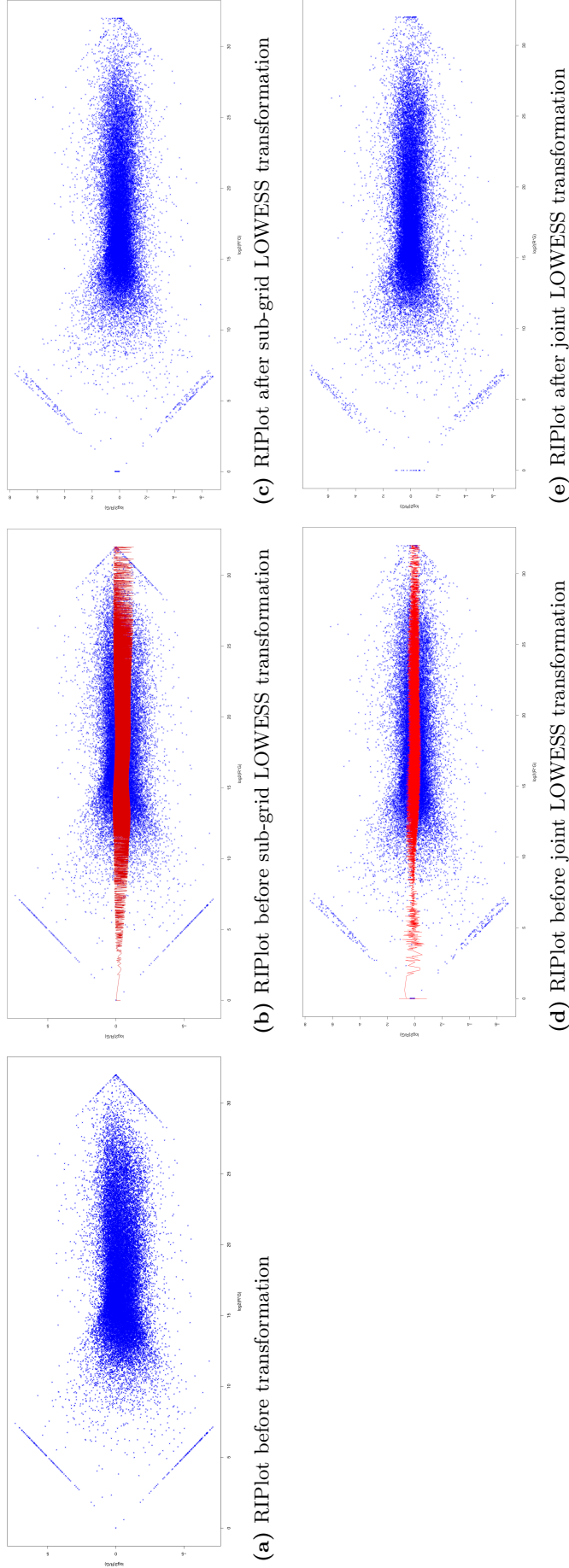


Figure 3.15 – RIPlots before and after the application of LOWESS transformations

Relationships between the \log_2 ratio (vertical axis) and \log_2 intensity (horizontal axis) can be easily identified in RIPlots and therefore these are presented to demonstrate the removal of effects by LOWESS transformations. The microarray shown demonstrates a slight bias towards one channel across all intensities when observed in (a), indicated by a regression below a \log_2 ratio of zero. This bias is more significant in mid-ranged intensity spots than it is at high intensities. This evaluation is verified in a curve fit by the sub-grid LOWESS transformation as shown in red on (b). The transformed data, once this curve has been aligned with a \log_2 ratio of zero, is shown in (c) and demonstrates much less bias than before without losing the majority of variance in the ratio. By then applying the Joint (regional) LOWESS transformation to the data shown in (c), a new curve is fitted to the data and presented as a red line in (d). Once this new curve has been aligned with a \log_2 ratio of zero, the resulting data, shown in (e), provides no evidence of ratio bias with intensity. This data, which has undergone two transformation steps, is now substantially better distributed than the raw data shown in (a) and is ready for modelling.

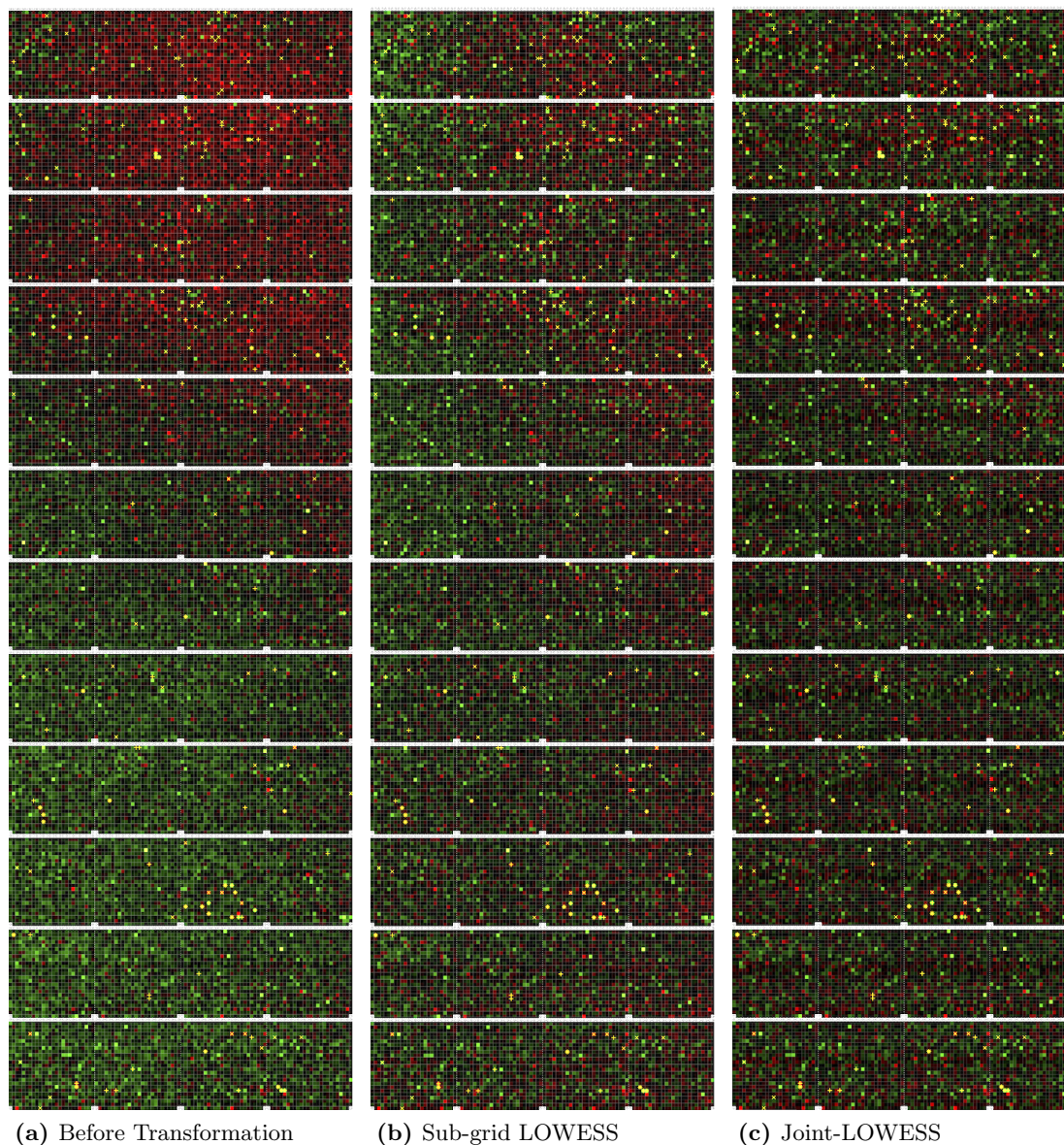


Figure 3.16 – ArrayView heat-maps of a microarray with spatial bias before, during and after Transform MAData

Arrays with spatial bias are most easily identified by the ArrayView ratio heat-map as shown in (a) where the green channel dominates the lower portion whilst the red channel dominates the upper portion. By applying sub-grid LOWESS in (b), the pin-tip groups are normalised which does reduce the bias, but some remains, particularly in the upper-right region. Joint-LOWESS is then successively applied in (c), which indicates that the bias is completely removed.

- **ToD** – The time of day, AM or PM, at which the sample was collected (2 levels)
- **Day:ToD** – The additional changes in gene expression specific to individual harvesting periods (22 levels)
- **Day:ToD:BioRep** – The variation in expression between individual biological replicates within individual harvesting periods (88 levels)

For each of these a matrix of 32,448 rows (the number of spots on the microarray) by x columns was returned, where x is the number of levels for that term, containing the effects at each level. The values of these matrices indicate changes in expression, both positive and negative, of each probe in response to the effects at each level of the term which they describe. This model was then used to identify differentially expressed genes and provide predicted gene expression data for further analysis methods.

3.4.5 Selection of Differentially Expressed Genes

F-tests provided by the MAtest function were used to identify differentially expressed genes. These test results were modified to account for biological replicate variability as described in Section 3.3.4, by using the ConvertMAtest function. This resulted in 32,448 p -values for each of the three terms: Day, ToD and Day:ToD.

To account for the discovery of false positives, each of three false discovery rate (FDR) p -value adjustments were considered. The effect of these adjustments on the distribution of p -values for the term Day can be seen in Figure 3.17 where the genes have been placed into rank order by their p -value, and this p -value is plotted along the vertical axis. It is possible to identify that the un-adjusted data, plotted in black, indicates approximately 21,000 genes with differential expression at a significance level of 5% for the Day term. Applying the StepUp adjustment, plotted in blue, shows that the number of genes identified as differentially expressed at a 5% significance is now approximately 20,000; a small reduction. The Adaptive method, plotted in green, actually increases the number of differentially expressed genes to approximately 22,500 whilst the StepDown method, plotted in red, reduces the number of significantly differentially expressed genes to 8,878.

The rank order of the genes does not change when alternative FDR methods are used, and as such the most significant genes will always be significant, no matter which method is chosen. Given that it seemed unlikely that as many as 20,000 genes were changing in expression significantly throughout the time series, backed up by observations of the gene expression data for non-significant genes, a decision was made to use the StepDown false discovery rate adjustment for the p -values.

Application of the AnalyseMAtest function allowed the extraction of lists of differentially expressed genes at a significance threshold of 5% for each of the three terms and an analysis of the overlaps between those lists. This also resulted in the presentation of a Venn diagram as shown in Figure 3.18, where plots have also been

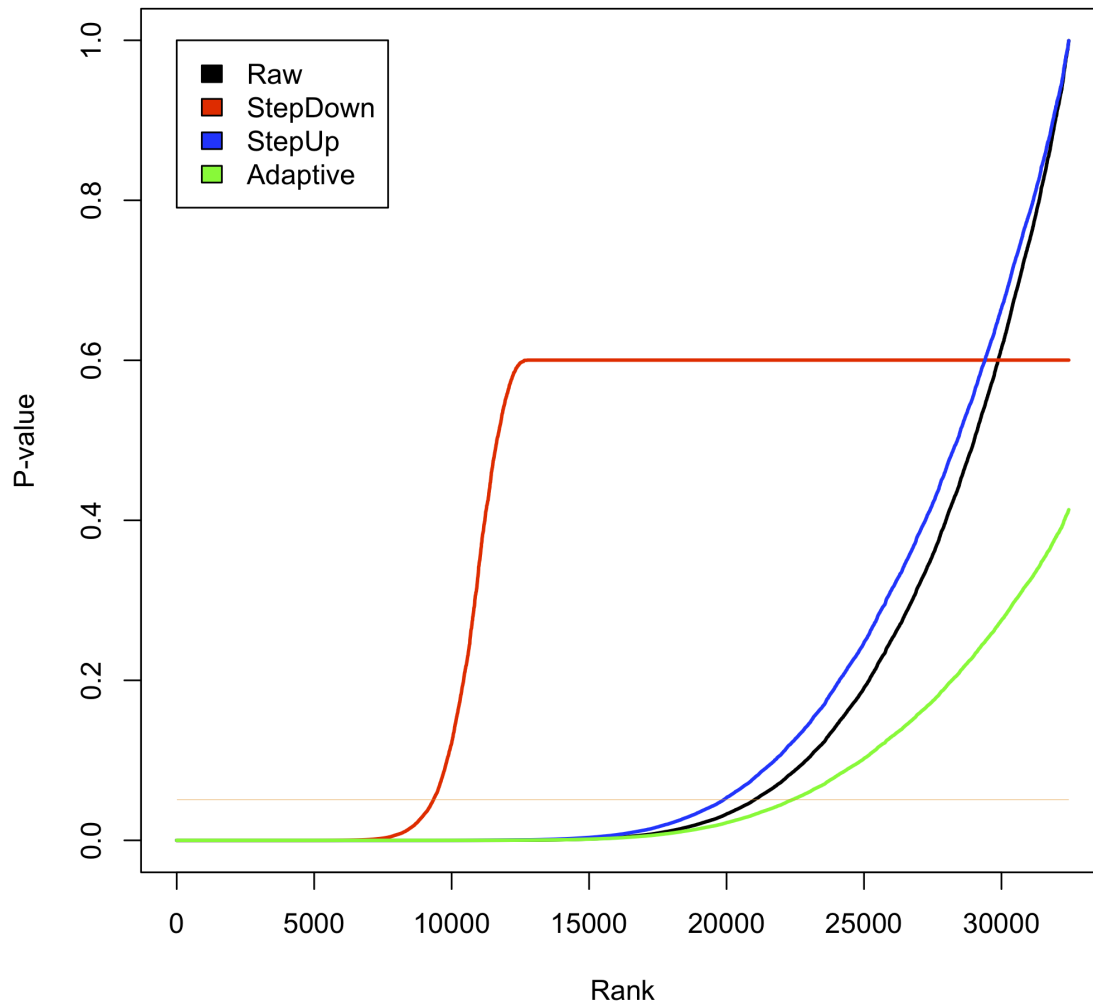


Figure 3.17 – A comparison of the p -value transformation produced by family-wise error-rate controls of MAANOVA applied to the term Day

A plot of p -value against genes ranked by their p -value gives an indication of the distribution of p -values within the microarray. It can be seen that the number of genes declared as differentially expressed at a significance level of 5% is approximately 21,000 in the raw p -value assignments, plotted in black. However, by using one of three FDR controls, this number can be altered by transforming the p -values to include fewer Type I errors in which a gene is declared differentially expressed when it is, in fact, not. The stepup function, plotted in blue, reduces the number of declared differentially expressed genes by approximately 1,000, whilst the adaptive method, plotted in green, increases the number of differentially expressed genes, rather than decreasing them. The stepdown method, plotted in red, is the most stringent, reducing the number of differentially expressed genes by more than half.

overlaid to help demonstrate the type of gene expression profile that may be found under each of the terms and their overlaps.

Genes which are significantly differentially expressed for the term Day show a change in expression across the entire time series, as seen in the top four plots. Those genes which are significantly differentially expressed for the term ToD show a consistent change in expression between morning and afternoon samples, as seen in the lower-left four plots. Where gene are significantly differentially expressed for the term Day:ToD they show one pattern of morning and afternoon samples in one region of the time series and an alternate pattern in another region, as shown in the bottom-right four plots. This altered pattern can only be captured by harvesting-period specific variations because they don't remain consistent over the whole time series. Where a gene is significantly differentially expressed for more than one term, the expression pattern is a combination of the above effects.

3.4.6 Extraction of Predicted Gene Expression Data

A prediction of the expression profiles of every gene on the microarrays can be extracted from the mixed model output using the methods outlined in Section 3.3.5. The choice of which format to use for this data depends almost entirely on the purpose of the data later on. For this reason, two formats were chosen as they best suit the intended further analysis.

For that data which would be used for clustering, the following criteria was required:

- That an average of the biological replicates for each gene be supplied.
- That the magnitude of the expression changes must be similar between genes.
- That, where similar shaped profiles exist, the absolute expression values be similar.
- That changes due to Time of Day be ignored since it is the baseline change in expression over a number of days which defines the gene under senescence.

and so for those reasons, the data for the purpose of clustering was produced by including only variation due to the Day term and normalising the standard deviation of each gene's expression to a value of 1 by using Equation 3.6. Each gene will therefore be zero mean-centred and, where the shape of the profile is similar, will have similar magnitude, making the absolute expression values almost identical for matching pairs of genes. Each gene is described by 11 expression values.

For data which would be used for theoretical network modelling, the criteria was:

- That each biological replicate be separately exported.
- That the magnitude of the expression changes be similar.
- That the time steps be equally spaced so that association between adjacent time steps are not stronger in some regions of the time series than others.

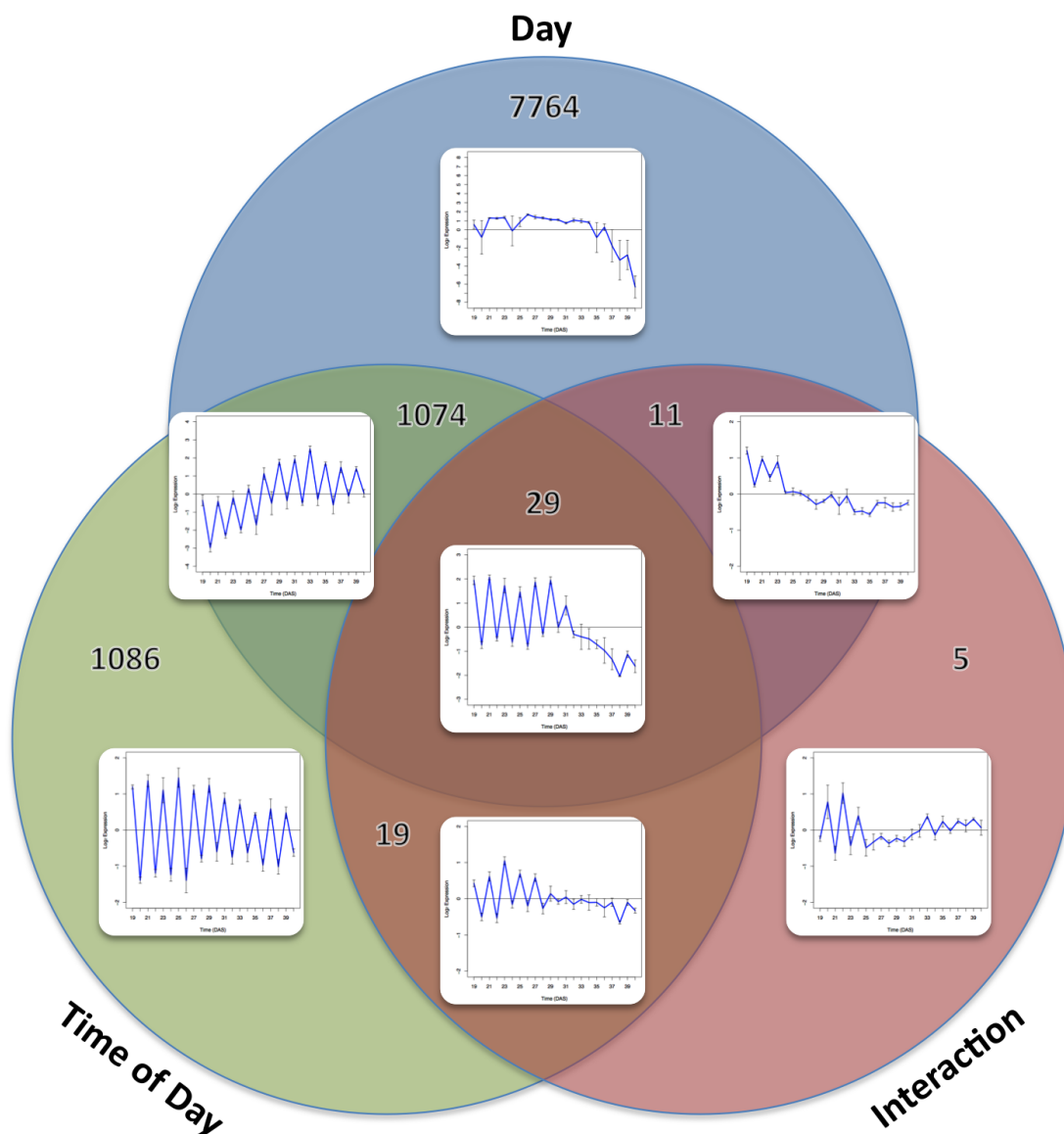


Figure 3.18 – Venn diagram summarising numbers of genes showing significant differential expression for each combination of treatment terms in the MAANOVA fixed model

For each fixed term of the MAANOVA model *Day*, *Time of Day*, and the *Interaction* between these factors, each gene was assessed for differential expression relative to the biological (between-plant) variability using an F-test. After applying a false discovery rate multiple testing correction, the numbers of genes with significant test statistics for each combination of the terms (individual terms only, pairs of terms, all three terms) are shown. The expression profile for an exemplar gene has been plotted for each combination of significant terms. Those with just a significant effect of Day show a smooth pattern within each Day, but either a general trend across the complete time course or a strong change in expression levels at some point during the time course. Those with just a significant effect of Time of Day show a strong diurnal pattern, but no change in expression between days. Those with just a significant Interaction effect have a diurnal pattern for some period during the time course which is not present at other times. Genes with multiple significant terms show the appropriate combinations of these patterns. It should be noted that it is usual in statistics to consider significance in the interaction term to also indicate significance in all main effects and so, whilst they have been plotted separately here, those sections of the diagram with numbers of genes 11, 19 and 5 all fall within the section with 29 genes.

- That changes due to Time of Day would be beneficial where they are appropriate, but due to the harvesting schedule, their inclusion would ensure that time steps are not equally spaced.

and so the data for the purpose of theoretical modelling was produced by including the variation due to the Day term and the Biological Replicates term to provide 88 expression values consisting of 11 time points with 8 biological replicates each (the morning and afternoon samples brought together as a single set of observations for the day). Each gene was normalised to a standard deviation of 1 by using Equation 3.6, ensuring that each gene's profile would be zero mean-centred and of a similar magnitude regardless of actual expression levels.

3.5 Conclusions

3.5.1 Application to Other Datasets

Whilst the methods shown have been optimised to provide useful information relating to the quality, transformation and modelling of the senescence dataset, these methods are very capable of handling a wide range of microarrays and offer both high-throughput analysis of the quality of microarray systems and avoidance of artefacts typically associated with two-channel systems. The quality checking functions are orientated towards two-channel microarrays, but the mixed model approach of decomposing sources of variation in the experiment design is cross compatible with one-channel systems as well.

When analysing other datasets, it is important to note the appropriate path to take during the analysis. Line numbers referred to here relate to those found in the analysis script shown in Appendix C:

- On lines 8–10, the data is loaded into MAANOVA. If the microarray has replicated probes, not only must these be named uniquely, such as Gene0001a and Gene0001b, but MAANOVA must be informed by specifying `n.rep` as a parameter of the function, giving the number of replicates on the microarray.
- If the microarrays are one channel, this must also be defined when loading the data.
- If flag data does not exist, this must be defined when loading the data.
- Once GridCheck, RIPlot, TechRepCheck and ArrayView have been performed, before reaching line 38 where the mixed model is applied, it is essential to ensure that your data is of a quality you are happy with by looking through the output which has been produced. Only once all repetition and re-analysis of scans has occurred, should the mixed model be produced as the model can take several days to produce.

- When the model is fitted, it is important to ensure the model describes your sources of variation correctly. Specification of the terms of the model can be confusing, and it may be beneficial to use the `makeModel` function provided by `MAANOVA` to check the terms which will be fitted before taking the time to complete line 38.
- F-tests should be performed on each of the terms of the model which are of relevance for differentially expressed genes. Though it is optional, it is beneficial to produce the F-test associated with the biological replicates first, because this allows the conversion of further tests, such as shown on lines 50, 55 and 60, and then the removal of the unconverted F-test such as shown on lines 51, 56 and 61, to save memory.
- When applying an FDR, it is worth checking to find out how many p -values are still significant after the adjustment, because some adjustments are more stringent than others.
- When plotting the Venn diagram and exporting the differentially expressed lists, as on line 69, it is important to specify whether to use the F1 or the Fs statistic in the analysis. These are described further in the `MAANOVA` help files. If a significance threshold of 5% is inappropriate, this can be defined whilst calling the `analysematest` function.

3.5.2 Future Development

Although the functionality of `MAANOVA` has been sufficient for the comprehensive analysis of the senescence dataset, there are also a number of areas for improvement, that with further time, would benefit the analysis.

`GridCheck` currently provides statistics to the user but leaves it to them to use those statistics in a meaningful way. It would be sensible for `GridCheck` to use the statistics in an automated way to provide the plots to the user in the order which their attention should be focussed. This would require the user to define which statistics they wish to use for this purpose, but would aid in the observation of the worst microarrays with little effort.

`ArrayView` was extensively tested with different types of statistic to try to identify flaws, but none were appropriate for finding the flaws accurately. With more time, these could be provided to offer a similar functionality to that used in `GridCheck`. Given that `TechRepCheck` has, generally, superseded the need for further statistics within `ArrayView` itself, it might be beneficial for the position of the gold crosses to be analysed automatically as part of the `ArrayView` function to help identify potential problems, rather than leave this task to the user.

`TechRepCheck` is still only able to provide information about outlying replicates on a probe-by-probe basis where four technical replicates exist. This is mostly a result

of the rules behind the identification of replicate outliers being defined so specifically for four technical replicates. If the rules could be generalised, any number of technical replicates could be analysed. This would greatly benefit the analysis, especially where a mix of numbers of technical replicates exist.

The CPU time and memory required to fit a mixed model to the data is substantial with the senescence dataset of 176 microarrays taking approximately 12 hours to complete whilst another similar experiment of 288 microarrays took over 48 hours to completely fit. This indicates that the complexity of the experiment exponentially increases the time to fit the mixed model. Unfortunately, in the event of a failure, the process would have to be restarted as there is no method of recovery currently implemented. Initially, it would be beneficial to provide a restore function to the model fitting process, so that, in the event of a crash, the process could be continued from the restore point. It was noted, whilst reverse engineering the source code of MAANOVA, that a lot of data is made redundant during the process, owing to an excessive amount of memory usage and, occasionally, crashes as a consequence. The structures used during this process could be minimised to prevent this, as well as better handling of memory for variables no longer needed, rather than relying on the garbage collector of R.

3.5.3 Chapter Synopsis

This chapter has focussed specifically on the process of obtaining gene expression data from the CATMA microarrays of the senescence experiment. The methods developed have allowed the identification of artefacts previously unknown to exist amongst the microarrays of the experiment and the subsequent correction of those errors. A new estimation method has been used to ensure that outlying technical replicates are reassigned a more consistent intensity value whilst a new transformation technique has allowed the normalisation of ratios within pin-tip groups of the microarray before then removing spatial dye-biases using the joint-LOWESS function already offered by MAANOVA. The fit of a mixed model and subsequent analysis of the terms of that model have allowed the identification of 8,878 genes as significantly differentially expressed during the observed period of senescence and then provide predicted gene expression data for those genes.

The following chapter takes the expression data extracted from the CATMA microarrays of the senescence experiment and applies it to methods capable of analysing gene expression data collected from any source so that a theoretical gene network can be identified.

Chapter 4

Analysis of Gene Expression Data

4.1 Chapter Summary

In this chapter, the aims were to identify robust regulatory networks amongst genes which appear to have a large influence on the senescence process. In order to do this, the 8,878 genes which were found to be significantly differentially expressed with respect to the day effect of the natural senescence microarray experiment were clustered using SplineCluster. This led to the discovery of 77 unique cluster profiles amongst the genes, many of which demonstrated an over-representation of specific gene ontology annotations that could be related to senescence. Though interesting, there were too many relevant annotations to help reduce the number of genes for modelling.

Initial models were produced using 88 genes as this represents the number of expression data known about each gene and it was found that when larger numbers of genes were simultaneously modelled, the models did not always converge. In order to analyse large sets of genes and identify which of those genes were important to the senescence process, multi-modelling was developed. Under this approach, sets of 88 genes were chosen at random from a hand selected set of 580 genes and those smaller sets were modelled. Once 454 sets had been modelled, 118 genes had demonstrated that they often regulate other genes they are modelled alongside and were selected for further analysis.

Of those 118 genes, 53 were obtainable as T-DNA insertion reduced expression mutants, which were all screened alongside wild-type plants to identify any altered phenotype with respect to senescence. Plants were stored in permanent darkness and photographed daily to obtain a record of their rate of senescence. Custom developed software was then used to analyse the rate of yellowing as a function of the red and green components captured by the digital photographs. The rate of change in the leaf colour was used to fit a sigmoid curve and the parameters of those curves compared between mutants and wild-type to find statistical differences. 8 genes (*ap2*, *stz*, *tcp15*, *hap3a*, *hat3*, *anac092*, *bft* and *MYB59*) were found to demonstrate an altered senescence phenotype when altered in expression.

For each of those 8 genes, a set of around 150 similar genes were identified, based

on the Qian similarity score. Those genes were used in a Metropolis-like VBSSM process in which genes were substituted in and out of successive models, keeping those which demonstrated a better fit to the data provided. This led to the hypothesis of many upstream and downstream regulations for 3 genes analysed in this way. By combining those genes into a set of 75 and modelling that set, a final model is hypothesised which provides a series of interactions which are then biologically validated as described in the following chapter.

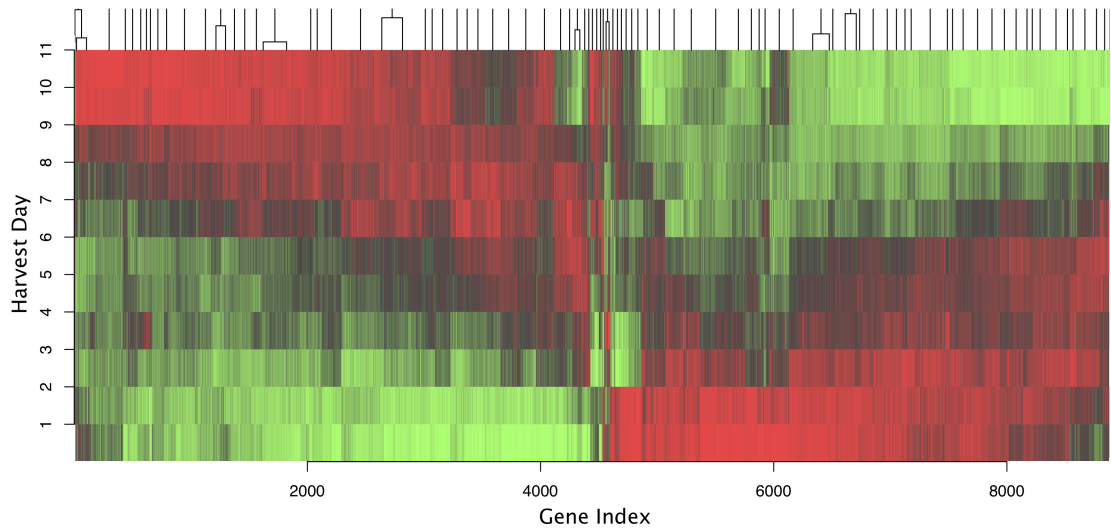
4.2 Initial Analysis of Gene Expression Data

Before developing a strategy for identifying refined network models, a number of standard procedures were used to process the gene expression data and gain an understanding of the range of expression profiles of genes during the senescence process. Once this analysis had been performed, the approach for producing theoretical network models was trialled with a number of gene combinations in order to establish an understanding for the approach and to determine methods of selecting gene modules which form highly likely models of regulation.

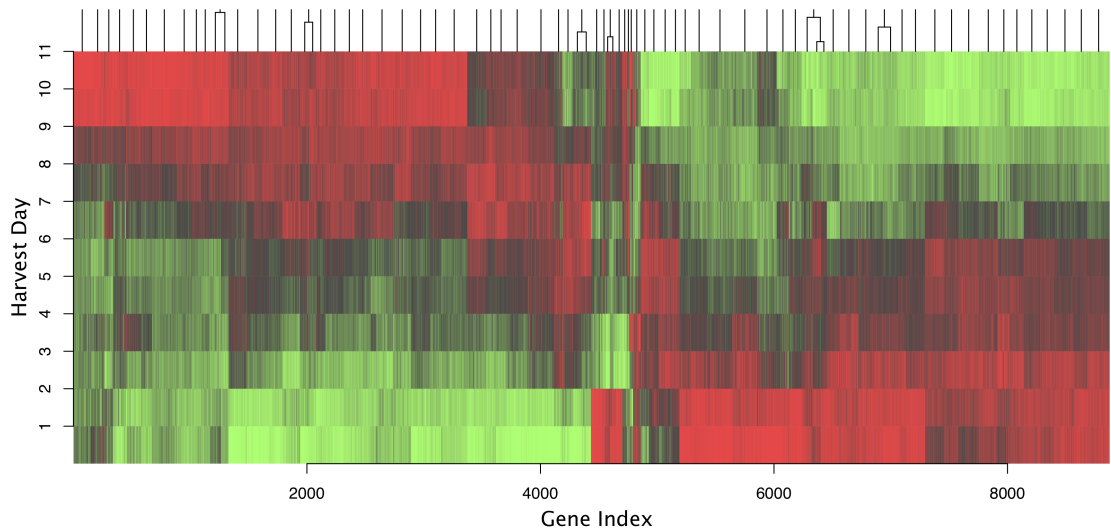
4.2.1 Gene Expression Clustering

A typical analysis of gene expression data is that of clustering genes, where those genes with similar expression profiles are assigned to the same cluster. The purpose of this, in the case of the analysis of the senescence data set, is to provide an overview of typical changes in gene expression observed throughout the senescence process which can lead to the identification of regulatory modules of genes. SplineCluster (Heard *et al.*, 2006) was the method used to perform the clustering. In an initial attempt to cluster the genes, merger sweeping, described in Section 2.3.1, was not yet implemented and the result was 93 clusters which can be presented as a heat-map as shown in Figure 4.1(a). Here it can be seen that, with genes of the same cluster lying horizontally adjacent and the time-series lying in the vertical axis, blocks of genes with similar expression are formed in the horizontal axis, indicating that the genes have mostly been assigned appropriately.

However, by implementing merger sweeping, a reduced set of 77 clusters were formed using the same input data and the heat-map representation of those clusters can be found in Figure 4.1(b). Although there is little visual difference between the two heat-maps, with the exception of the final order of the clusters, it can be seen, by observing part of the dendrogram, that many of the smaller clusters from the centre of the plot before merger sweeping were integrated with other clusters when merger sweeping was used. This indicates that fewer genes have been misallocated by the hierarchical process because of the reallocation of genes when they become outliers for their cluster. Therefore the clusters generated whilst using merger sweeping were used during further analyses.



(a) Before merger sweeping (93 clusters)



(b) After merger sweeping (77 clusters)

Figure 4.1 – Heatmaps providing evidence of improvement after merger sweeping in SplineCluster

Heatmaps produced by SplineCluster illustrate the membership of genes to clusters such that those of the same cluster are plotted adjacently along the horizontal axis. For each gene, the expression at each time point, plotted along the vertical axis, is presented as bright green for low expression, black for mid expression and bright red for high expression. A part of the dendrogram has been left on top of each of the heat maps to indicate the centre of each cluster. It is visible that, in (a), a number of very small clusters exist in the centre of the heatmap, whereas clusters of this size are fewer in (b), hence fewer clusters consist of merely outliers and the merger sweeping has optimised their integration into other clusters.

Plots of the profiles for the 77 clusters can be found in Figure 4.2 which demonstrates that very few genes show more than one sudden change in expression, with the exception of clusters 37 to 44 which show these characteristics. This might form the theory that most genes change expression gradually with the onset of senescence, though these seemingly gradual changes are also likely to be an artefact of the whole-leaf mixed-cell samples which, due to the changes across the leaf being asynchronous, reduces the significance of large changes in expression at specific points across the leaf. The fact that some clusters show sudden changes in expression could indicate that those genes are involved in processes which synchronously affect the whole leaf, such as flowering.

Despite the observations made about the gradual changes in expression, it can be seen that most clusters demonstrate a switching point at which the gradient of the profile is exaggerated in comparison to the rest of the profile. The identification of significant changes in expression was performed and is reported in Breeze *et al.* (2011) which can be found attached to the end of this thesis. This found that the start of significant change for important senescence associated genes such as *SAG12* is after the 6th observation. It is clear from the results of the GO analysis (figure 6 of the publication) that the largest functional changes of the leaf occur after this time.

As can be expected of a transcriptional network, in which some genes are regulators of others, some of the clusters appear to be time-shifted variants of each other. One example of this can be found in cluster 23 and 24 where the major rise in expression is one time point earlier in 23 than it is in 24. However, the identification of these by eye is challenging unless they are adjacent as in this example. A similar trait is that of inverse correlation where the expression profile of one cluster may be the inversion of another which would be typical of a negative regulation in the transcriptional network. An example of this is found by comparison of clusters 32 and 54.

The numbers of genes in each cluster can be found in Table 4.1, where it can be seen that the genes are reasonably evenly spread throughout the clusters with the exception of clusters 38 through to 44 where changes in expression are more distinct than the smooth transition of the majority of clusters. These may contain few members because the rapid changes in expression between adjacent observations are more distinctive, resulting in a lower likelihood that they may be merged with similar but functionally different genes.

Cluster 43 has a similar profile to that seen in cluster 36, with the exception of the first two time points which distinguish them from one another. One possible explanation for this might have been that the first two observations of members of cluster 43 were merely background noise if the expression levels at these times were particularly low. However, by referring to the unnormalised raw data (unpublished), this does not appear to be the case. It was noticed, however, that whilst performing this check, 11 of the 23 members of cluster 43 are without name nor annotation which is a high proportion compared with other clusters. Although this does not affect their expression levels, it may suggest that the members of cluster 43 belong to a small and

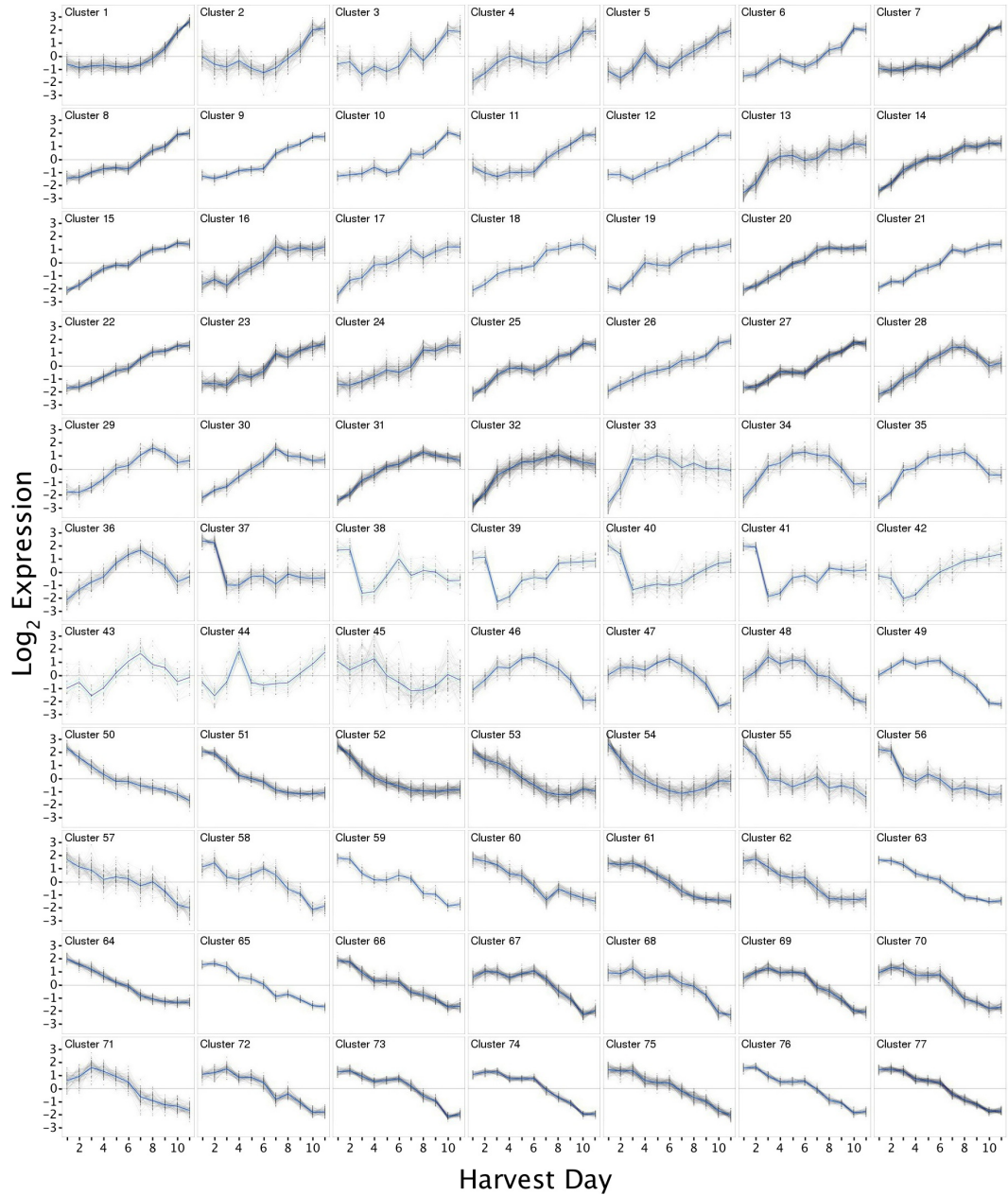


Figure 4.2 – Plots of gene profiles for each cluster output by SplineCluster
Each cluster identified by SplineCluster for the 11 time point dataset for senescence is defined by a plot demonstrating the profile of the genes within the cluster. In each plot, the horizontal axis represents harvest days whilst the vertical axis represents relative log₂ expression levels. Individual genes of the cluster are plotted as dotted lines, whilst the consensus profile is plotted atop of these in blue. All the plots have the same scales with 11 harvest days along the horizontal axis and log₂ expression levels between -3 and 3 on the vertical axis. The horizontal line represents the centre of the change in expression at 0 on the vertical axis.

Cluster	Size	Cluster	Size	Cluster	Size	Cluster	Size	Cluster	Size
1	150	17	79	33	73	49	69	65	67
2	114	18	57	34	89	50	98	66	163
3	77	19	82	35	71	51	141	67	169
4	108	20	158	36	81	52	216	68	117
5	128	21	91	37	98	53	210	69	176
6	96	22	138	38	27	54	169	70	160
7	210	23	193	39	36	55	104	71	106
8	130	24	153	40	53	56	109	72	118
9	79	25	162	41	55	57	93	73	138
10	73	26	95	42	40	58	72	74	125
11	100	27	223	43	23	59	50	75	179
12	63	28	163	44	26	60	109	76	101
13	157	29	79	45	69	61	160	77	196
14	189	30	96	46	69	62	130		
15	116	31	180	47	86	63	82		
16	151	32	228	48	107	64	130		

Table 4.1 – Table of cluster sizes returned by SplineCluster

Each cluster is representative of a number of genes from the data provided to SplineCluster. The number of genes being represented by each of the 77 clusters are shown in the table. It can be seen that, in the majority of cases, the size of the cluster is in the range of 70 – 200 genes and that there are few cases where the cluster is unusually large or small.

poorly studied process of plant development.

The number of genes transitioning from low to high expression and vice versa are approximately equally split with a slight bias towards those genes which increase in expression. Those genes which rise in expression appear to begin doing so earlier than those which fall in expression, indicating that senescence is likely to be actively switched on by positive regulation, rather than being allowed to progress due to the removal of repressive regulation.

In a review by Buchanan-Wollaston (1997) describing the molecular biology of leaf senescence, a number of senescence associated genes, identified from samples of *Brassica*, are categorised in 10 classes of expression. Although it cannot be certain what proportion of genes are senescence specific in the plots of Figure 4.2, it is revealed by the increased number of time points studied that even more profiles than those described by the 10 classes exist within the microarray data. A more recent review (Buchanan-Wollaston *et al.*, 2003) describes some of the processes known to occur during leaf senescence such as chlorophyll degradation, protein degradation, lipid degradation and nuclease action for which some of the clusters identified here may be enriched with genes associated with these processes. This led to the next stage of the analysis: gene ontology analysis.

4.2.2 Gene Ontology Analysis

One method of finding enriched modules of genes specific to biological processes is the analysis of GO annotations. The tool used for this process was BiNGO (Maere

Cluster Number	GO Term Description	Abundance of Term	Abundance in Cluster	Corrected p -value
1	cell wall organization	0.83%	7.21%	1.17×10^{-3}
5	response to jasmonic acid stimulus	0.74%	5.88%	3.11×10^{-2}
6	cellular respiration	0.35%	5.71%	3.41×10^{-2}
10	defense response	3.08%	15.09%	4.17×10^{-2}
12	response to starvation	0.52%	9.52%	1.82×10^{-2}
19	response to water	1.43%	14.49%	1.06×10^{-5}
25	transport	8.72%	20.66%	1.57×10^{-2}
29	pyrimidine nucleotide biosynthetic process	0.06%	5.00%	7.65×10^{-4}
37	cellular process	37.68%	59.74%	3.92×10^{-2}
39	post-translational protein modification	5.38%	28.57%	2.37×10^{-2}
44	response to wounding	0.86%	22.73%	1.67×10^{-4}
45	response to ozone	0.14%	5.88%	1.33×10^{-2}
46	regulation of vegetative phase change	0.03%	3.77%	1.85×10^{-2}
50	translation	2.74%	11.90%	2.65×10^{-2}
51	DNA metabolic process	1.48%	11.58%	5.83×10^{-5}
55	translation	2.74%	39.74%	1.16×10^{-26}
58	photosynthesis	1.51%	15.00%	5.71×10^{-5}
60	anatomical structure development	6.92%	20.48%	1.82×10^{-2}
66	cellular nitrogen compound metabolic process	8.32%	22.58%	4.29×10^{-4}
67	photosynthesis	1.51%	9.38%	1.54×10^{-4}
68	photosynthesis	1.51%	8.42%	3.65×10^{-2}
73	photosynthesis	1.51%	20.75%	3.06×10^{-17}
74	small molecule catabolic process	1.34%	8.25%	1.98×10^{-2}
75	embryonic development	2.54%	9.40%	1.23×10^{-2}
76	photosynthesis	1.51%	15.29%	1.04×10^{-7}
77	pigment biosynthetic process	0.71%	5.41%	4.50×10^{-3}

Table 4.2 – Highest Significance Over-represented Biological Process GO Terms in Senescence Clusters

The highest significance Biological Process GO Term for each cluster is shown. Clusters 1 to 29, in the top section, are generally up-regulated during senescence; clusters 50 to 77, in the bottom section, are down regulated; whilst those in the middle section have expression changes which are more complex. It can be seen that those which are down-regulated show higher levels of significance. This is likely to be because more genes, particularly those relating to photosynthesis, are correctly annotated in the down-regulated category, leading to the discovery of more of them in these clusters.

et al., 2005) which was able to determine those GO terms which are statistically over-represented in a given sets of genes. The sets of genes to be tested were the members of the 77 clusters identified by SplineCluster and these were processed following the methods of Section 2.3.2. Those biological process GO terms which were identified as over-represented within each cluster were processed using a custom script written to output a table of results. For each cluster that returned any over-represented GO terms, the term with highest significance can be found in Table 4.2 whilst the complete list of all returned terms can be found in Appendix E.

As can be seen, by a comparison to the cluster profiles of Figure 4.2, those genes which monotonically increase in expression during senescence are featured in the top region of Table 4.2, whilst those which monotonically decrease in expression during senescence are featured in the bottom region of Table 4.2. A number of clusters are not

so easily defined and tend to both increase and decrease in expression at various stages of senescence. These are featured in the centre of Table 4.2.

The review of leaf senescence by Buchanan-Wollaston (1997) identifies a number of processes known to be up-regulated by the onset of senescence in leaves, such as increased protein degradation, increased nucleic acid breakdown by RNases (Green, 1994), increased lipid remobilisation (Wanner *et al.*, 1991), increased chlorophyll breakdown and increased nitrogen remobilisation, so it is of little surprise that some of the biological processes which are over-represented in the up-regulated clusters relate to the breakdown and transport of cellular components. The highest levels of significance, however, are found in those clusters which are down-regulated during senescence with some offering p -values 10^{21} -fold smaller than those of up-regulated clusters. This would appear to be because processes such as photosynthesis and translation are so heavily researched that a much higher proportion of genes associated with those processes are known and annotated, resulting in greater numbers of associated genes in those clusters and therefore higher levels of significance.

This concept would be best explained by example, and as such, a theoretical cluster of 100 genes may contain 10 genes associated with a given GO term x , 50 genes associated with other GO terms and 40 unknown genes. This then indicates that GO term x is associated with 10% of genes throughout the cluster, but it might also be known that GO term x is only associated with 5% of all genes throughout the genome. This would probably result in GO term x being identified as over-represented in the theoretical cluster. If, at a later date, 10 more of the remaining unknown genes in the cluster were also identified as being associated with GO term x , the percentage of the cluster representing GO term x would then be 20% whilst the abundance of GO term x throughout the genome would only rise very slightly, increasing the significance of the over-representation of GO term x in the cluster. Therefore, a higher proportion of correct annotations for a GO term can positively affect the significance of clusters which are already over-representative of that term.

It is widely accepted that senescence is caused by a combination of environmental stimulus and the synthesis/response to plant hormones such as jasmonic acid (JA), ethylene, cytokinins and abscisic acid (ABA). These hormones are typically also the cause of early senescence during stresses imposed upon the plant and this makes their identification as a non-stress response difficult. Many of the GO terms are annotated as being related specifically to biotic and abiotic stresses, but this is probably due to the greater level of research which has been performed in these areas. Whilst it is simpler to deduce the response of a stress which can be compared with plants in the unstressed state, the study of natural senescence has no mock response to compare with and so annotations specifically associated with natural senescence are limited. However, many of the genes associated with stress related hormones can be thought to also be related to natural senescence and these are the GO terms which have been focussed on here.

Cluster 44 contains a high proportion of genes relating to wound and stress re-

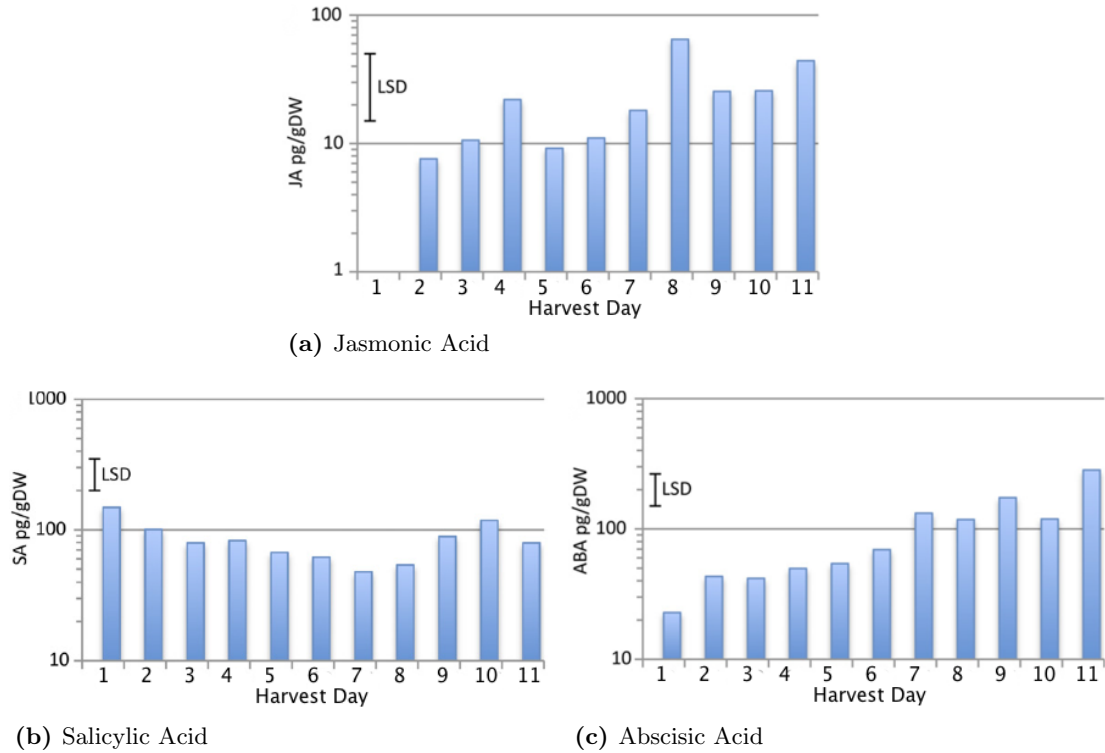


Figure 4.3 – Hormone Levels During Leaf Development

Levels of Jasmonic Acid ((a)), Salicylic Acid ((b)) and Absciscic Acid ((c)) were measured in leaf 7 harvested throughout the course of the senescence experiment. LSD (5%, 41 *df*) for comparisons between pairs of means are shown for each hormone calculated from the ANOVA of log10-transformed data. Values represent the means of five independent biological replicates per time point. DW, dry weight.

sponse, particularly the synthesis of jasmonic acid (JA), a known inducer of senescence. JA is released as a hormonal response to wounding and other stresses (Turner *et al.*, 2002; Wasternack *et al.*, 2006; Wasternack, 2007) which can lead to senescence as a defence mechanism of the plant. The profile of cluster 44 indicates a peak at the fourth observation which coincides with a peak in the levels of JA in the leaves, as shown in Figure 4.3(a) which is taken from figure 2 of Breeze *et al.* (2011), levels of JA resemble the profile of cluster 44 relatively closely. Another cluster with a similar profile is cluster 5 which has a less prominent, but still distinctive, peak at the fourth observation. The over-represented GO Terms for this cluster are also JA biosynthesis related. All JA biosynthesis genes are positively regulated by JA (Pauwels *et al.*, 2008; Sasaki *et al.*, 2001) but it is unclear whether some of the genes in these clusters are purely responsive to JA and not responsible for its synthesis. However, although these interactions are interesting and seemingly connect the profiles of the clusters, interactions involving hormonal regulations would not be easily biologically tested and so this information is not useful to the aims of this PhD.

Cluster 19 is shown to exhibit over-represented GO terms for response to water and also to absciscic acid (ABA) stimulus, a hormone known to be involved in the induction of senescence. The exact role of ABA in senescence is unclear since the

responses to ABA are diverse and the hormone is also known to cause the abscission of leaves and to cause the rapid closure of stomata to limit respiration and consequential transpiration (Munné-Bosch & Alegre, 2004; Leung & Giraudat, 1998). It has been shown that genes induced by ABA, such as RPK1, can also cause early onset of leaf senescence when their expression is artificially increased (Lee *et al.*, 2011), supporting the theory that ABA does play a role in causing senescence. Cluster 19 appears to rise in expression quite early in the time course, between observations 2 and 4, and perhaps even before senescence is thought to truly begin, possibly indicating that some of the genes in cluster 19 are involved in actually causing the senescence response to begin.

Amongst the genes representing ABA response GO terms in cluster 19 are three drought response genes: *ERD10* (At1g20450), *ERD14* (At1g76180) and *RD26* (At4g27410) and two cold response genes: *COR47* (At1g20440) and *LTI30* (At3g50970). These are two stress responses which are typically associated with ABA and can cause the early onset of leaf senescence. However, genes are often later associated with more than one type of stress, such as *ERD10* which is also seen to be up-regulated in response to cold, and their annotations, and hence name, can be skewed by the first study they were identified in. It may be that these genes, associated here by their response to ABA, have more generic functions in senescence.

Cluster 55 has a very large number of genes associated with the process of translation. In particular, most of these genes are ribosomal proteins which form part of the machinery required for translation to occur. A reduction in their abundance early in the time course may be entirely due to the machinery associated with early plant development being halted but a small peak is also present at the sixth and seventh time point, perhaps associated with the rise in expression of late clusters such as cluster 24 which is thought to relate to the senescence response. The continued degradation of the expression of this cluster at the end of the time course is a clear indication of the final subsidence in further gene expression at the end of leaf senescence, supported by the expression level of almost every cluster remaining unchanged or reducing between the penultimate and final observation.

Eight of the remaining down regulated clusters (58, 66, 67, 68, 73, 74, 76 and 77) have over-represented GO terms for genes involved in photosynthesis and the associated organelles. Cluster 66 contains eight genes involved in the synthesis of chloroplasts. Two of these genes, CDP1 (At3g19180) and ARC5 (At3g19720), are involved in the restriction and division of chloroplasts. Mutation in the ARC5 gene causes the number of chloroplasts per cell to be reduced from a mean of 121 in wild-type leaves to only 13 enlarged and dumb-bell shaped chloroplasts in the mutant (Pyke & Leech, 1994; Robertson *et al.*, 1996). This signifies importance in the photosynthetic process which is being degraded over the period of senescence in the leaf. Observation of the profile of this cluster appears to indicate that there is a declining expression leading up to the fourth time point, after which a plateau in expression exists followed by further decline in the expression of the cluster. This may indicate that those genes are subsiding to

a normal level of expression after the initial growth of the leaf but are then degraded further after senescence begins from the sixth time point onwards.

Cluster 77, with an over-represented GO term for pigment biosynthesis, is the only other cluster of these eight which has an over-represented GO term more specific than simply “photosynthesis”. The majority of genes representing this term are directly responsible for the biosynthesis of chlorophyll and hence involved in the same processes as cluster 66, explaining the highly similar profiles of these two clusters. The progressive loss of chlorophyll biosynthesis after the sixth observation during leaf senescence coincides with an observed reduction in chlorophyll content of the leaves as reported in Breeze *et al.* (2011). These key indicators strengthen the belief that senescence becomes the dominant developmental process somewhere between observations 6 and 7.

The remaining six clusters are all related to the function of photosynthesis itself and, together, these over-represented GO terms account for 72 genes. The most interesting aspect of these genes is that they are all from the same GO term and yet they have six separately identifiable expression profiles. This is one of the only GO terms showing such diversity throughout the experiment, most likely to be because of the enormous amount of research which goes into photosynthesis and hence the large number of correctly annotated genes associated with this process. In all cases, the expression of the clusters reduces with time and the most significant drop is at the sixth observation after which the expression is rapidly reduced until the end of senescence. This is obviously well correlated with the yellowing of the leaf, as might be expected. One peculiarity, however, is shown in cluster 58 where it can be seen that the expression drops quickly after the second observation before briefly recovering and then following the same pattern as the other clusters after the sixth observation. There is no obvious cause for this temporary reduction in expression, but an investigation into why these genes show alternative expression to those of the same processes would possibly identify transcriptional networks surrounding photosynthesis.

Cluster 6 is found to contain an over-representation of genes involved in cellular respiration, a process relying on the products of photosynthesis. The profile for this cluster indicates that these genes are only activated very late in the time course, possibly later than most other processes in senescence. Amongst those genes representing the GO term for cellular respiration is *AOX1a* (At3g22370) which has been shown to be up-regulated in response to low nitrogen levels and to be responsible for the negative regulation of cellular respiration (Watanabe *et al.*, 2010). As the energy rich chlorophyll is degraded and amino acids broken down and transported from the leaves of *Arabidopsis* (Munné-Bosch & Alegre, 2004), nitrogen levels will fall as a consequence and it is likely that this causes the late increases seen in this and similar genes. It would seem that these genes may be the response to reduced photosynthesis in the leaves and therefore the need to reduce respiration.

Although a number of senescence related biological processes have been identified by BiNGO analysis, it is apparent, having observed the result of extensive study

in processes such as photosynthesis and translation and comparing these results with those of lesser studied processes, that unless the genes of a given GO term are well annotated, it is difficult to identify that the GO term is over-represented in a cluster. This is detrimental to this study of senescence related transcriptional networks since the intention is to find novel interactions which are likely to be between genes which have not yet been well studied. When combined with the logic that natural senescence processes are harder to study because of the lack of a mock response, a systems biology approach is required in which the expression profiles of the genes are used rather than prior knowledge as is required by BiNGO. The results of tools like BiNGO should only be used as a guide to understanding the range of biological functions which are being affected by the senescence process and their relative timings.

4.2.3 Preliminary Network Modelling

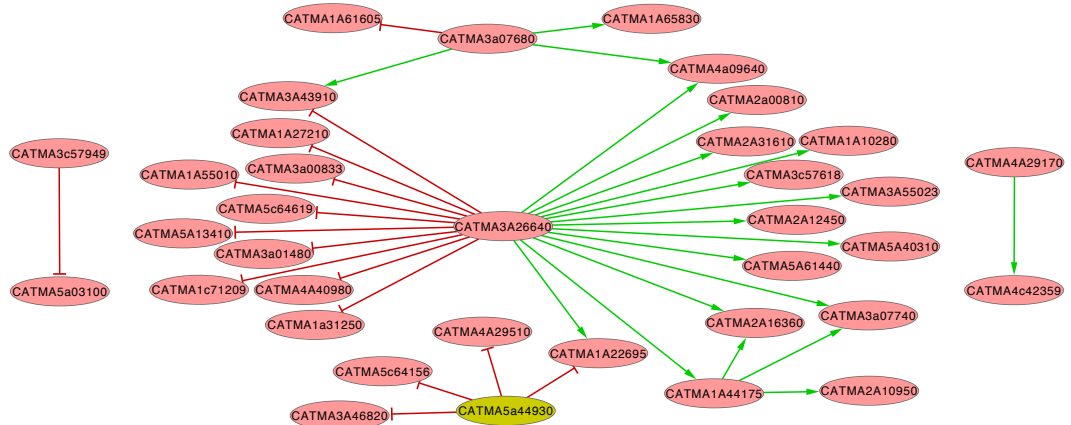
Before continuing the analysis of the gene expression data, it was important to gain an understanding for the way in which the theoretical regulatory network modelling tool worked, what the output meant and how the features of a biological network were structured. In order to do this, subsets of genes had to be selected so as to use their expression data to try to form a network. Initially, selections of genes for modelling were made by either trying to find groups of genes with similarly shaped expression profile or genes which characterised all the observed profiles shown by SplineCluster. Genes assigned to the same cluster could be thought to represent large proportions of a functional network as they are all being expressed at a similar time and therefore may be regulated by a similar upstream entity. For this reason, selections of 50–100 genes within the same cluster were presented together for regulatory network modelling. The data for these models was processed as described in Section 2.4.1.

Although a convincing model could be produced in this way, it was soon identified that small alterations to sets of genes resulted in vastly different network structures. Even the removal or addition of genes which appeared to be peripheral to the network resulted in a new network with relatively little similarity to the original. To demonstrate this, a list of 88 randomly selected genes was produced to give a network model as shown in Figure 4.4(a). The structure of this network appears to indicate four genes which have a large number of downstream interactions. 22 of the original set were then chosen at random, removed and replaced with 22 randomly selected genes from the genome to generate a second set. The new set still, by chance, included the four highly influential genes seen in the model. When a network model was produced from this second set, shown in Figure 4.4(b), only three genes had at least four downstream interactions and, of those, only one (highlighted in yellow) showed highly influential behaviour in both models, despite the other two hub genes existing in the original model's gene set as well.

The outcome shown here could have been a coincidence, however, to ensure this was not the case, four other sets of 88 genes were produced from the original set by randomly substituting 22 genes and this lead to further models with very low structural



(a) VBSSM Putative Regulatory Network Model for Gene Set 1



(b) VBSSM Putative Regulatory Network Model for Gene Set 2

Figure 4.4 – VBSSM network models of similar sets of genes demonstrate drastic differences in individual relationships

Both of the models above were generated from lists of 88 genes with an overlap of 66 genes. Network edges are only shown where at least half the model seeds indicated a confidence level of 95% for its existence. Four genes appear to be key to the structure of the network in (a) by linking with many other genes, whilst only three genes are key to the structure of the network in (b). Only one of those genes is common between the two networks (highlighted in yellow) despite all the key genes being present in both gene sets. This indicates that the selection of periphery genes in the network are very influential on the identified interactions and that no individual model can be trusted to have identified truly likely interactions.

similarity despite each being produced from 75% identical data to the original.

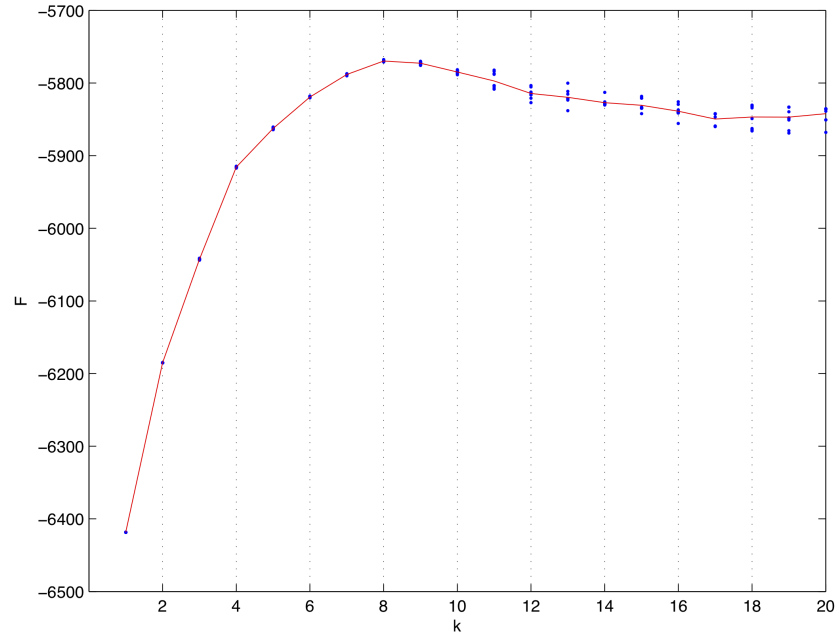
It was thought, knowing that the network models could be so heavily influenced by seemingly unimportant network elements such as peripheral genes, that selecting lots of genes with similar expression profiles from the same cluster would only increase the risk of finding local maxima rather than the true network. For this reason, an initial attempt was made to identify the underlying network structure of the clusters. The mean profile of each cluster was submitted to VBSSM as if it were the expression data of a single gene. The disadvantage with this was that there were obviously no biological replicates and, although an approximation of a cluster's members, the mean profile of the cluster is only a loose definition of all its members and so any interaction discovered between clusters may not be directly transposable for any pair of genes they contain.

The network with the highest mean log marginal likelihood across seeds was that with a hidden state dimensionality of 19 as indicated by the F vs k plot shown in Figure 4.5(b) where k is the dimensionality of the hidden states. However, it can also be seen that the seeds, plotted as blue points, exist at three alternative levels, indicating that at least three possible models are consistent with the data. This is indicative of the quantity of data being an inadequate definition for the number of inferable connections being fitted by VBSSM since there are no biological replicates and only 11 observations made about each of the 77 profiles. The red line, indicating the mean log marginal likelihood across the seeds for each k value, does not demonstrate an obvious maximum with monotonic degradation of the F score either side, making it impossible to be certain of the chosen k value since the log marginal likelihood for all k values above 2 are very similar.

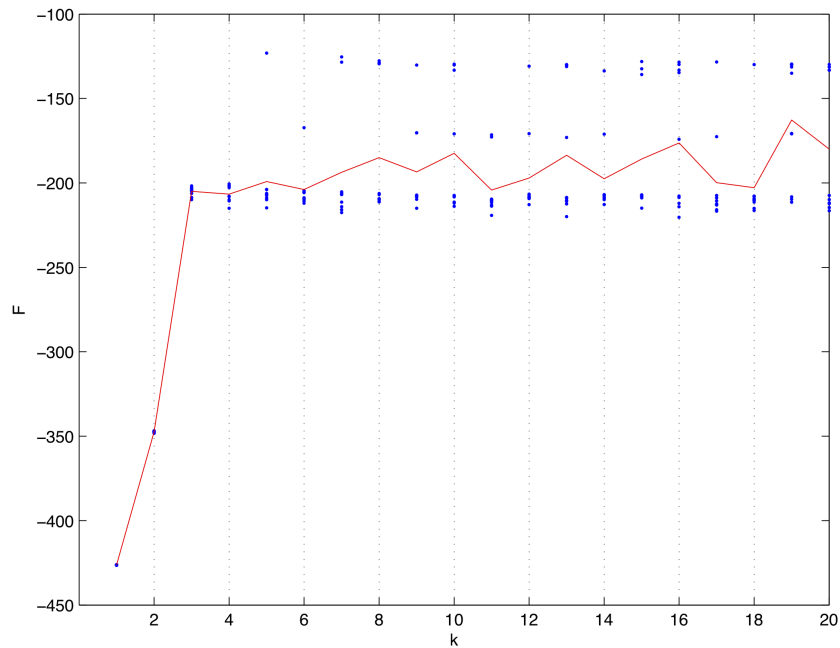
The peculiarities of the F vs k plot in Figure 4.5(b) are mirrored in the resulting network model, which, at the same confidence threshold as the models shown in Figure 4.4, indicate no interactions at all with a k value of 19. This is due to a high level of diversity between the seeds, so that no more than 50% of the seeds are in agreement over the existence of each interaction. A more typical example of an F vs k plot can be found in Figure 4.5(a) which is the one produced for the model shown in Figure 4.4. Here it can be seen that the maximum k value at $k = 8$ is well defined and that all the seeds at that maximum are very similar.

An alternative approach to obtaining the profile of each cluster is to use a singular value decomposition method to identify the profile of an eigengene (Wall *et al.*, 2003). This method was employed by Emma Cooke of The University of Warwick when modelling data of a similar microarray dataset. During Emma's use of this method, she found that, for her data, the eigengene profiles were so similar to that of the mean of the cluster that the resulting model was identical and so this method was not attempted on the senescence data set.

Despite failure to find a method to accurately determine the underlying network structure of such a large number of genes, observation of the types of interactions that exist within networks, such as those found in Figure 4.4, allowed the categorisation of



(a) 88 gene profiles of 11 observations in 8 biological replicates.



(b) 77 gene profiles of 11 observations each with no biological replicates.

Figure 4.5 – A comparison of F vs k plots for data with and without biological replicates

The F vs k plots shown indicate to the user which k value, representing the dimensionality of the hidden states, provides the maximum mean log marginal likelihood (F score) across the seeds. A normal result is shown in (a) where a maximum F score at $k = 8$ is identified using sufficient data to describe the interactions between the chosen number of genes. However, it is also possible to obtain results such as those shown in (b) where the quantity of data about each node of the network is insufficient to identify the interactions between them without finding local maxima. In this example, at least two local maxima have been identified, as indicated by the three alternative F score levels shown in the blue points for each seed.

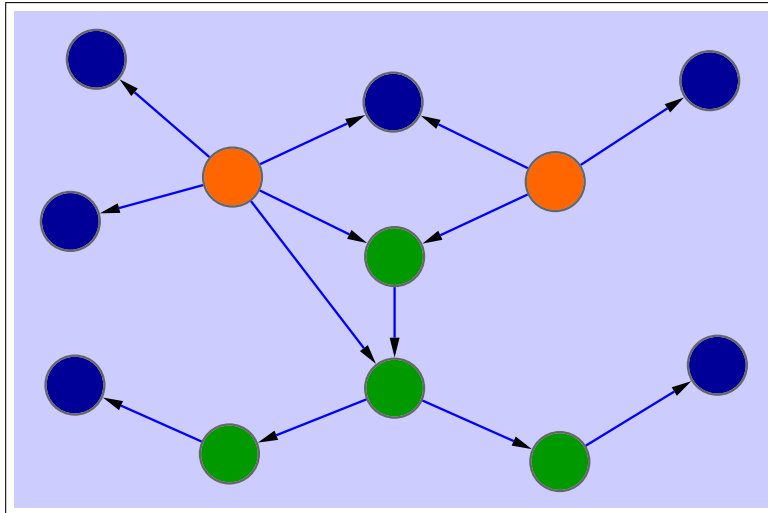


Figure 4.6 – An artificial network demonstrating types of node present in a typical regulatory network

Three major types of node exist in a typical regulatory network. Hubs, represented in orange, have many downstream interactions and few or no upstream interactions. Intermediate nodes, represented in green, have roughly equal numbers of upstream and downstream interactions. Response nodes, represented in blue, have one or more upstream interactions and often no downstream interactions.

genes into three main classifications, as demonstrated in Figure 4.6. There are those genes which form the control of the network and, whilst having many downstream interactions with other genes, are influenced by very few genes, as plotted in orange and hereafter referred to as “hubs”. Those genes which are influenced by upstream genes but also have an influence on downstream genes are plotted in green and are “intermediate” genes. Those genes which have no downstream interactions, but are influenced by upstream genes are plotted in blue and are the “response” genes.

By implementing a systems biology approach to network modelling, it was possible to find a way to identify the hub genes through a series of modelling iterations which allowed the identification of robust network interactions, as described in the following section.

4.3 Identification of Module Centres

By focussing on identifying the module centres, or hub genes, in the networks associated with senescence, it was possible to substantially reduce the number of genes thought to provide strong evidence for being a regulator for senescence related processes. The method used to find those genes was multi-modelling which is described in Section 2.4.3 and involves the consideration of all pairwise associations between a large set of genes. Many smaller models of equally sized gene sets sampled from the larger set were used to provide information about the likely number of interactions each gene has with others in the set. In an ideal situation, it would be informative to produce a network model for every possible combination from the large set, but this is an impractical solution due to

the number of subsets possible.

When comparing the outcome of many smaller VBSSM models in an attempt to understand the regulation between genes of a larger set, it is important to understand the combinatorial problem that rapidly diminishes the effectiveness of the approach. In order to correctly analyse a set of 500 genes by modelling subsets of 88 genes (see below for justification of this number), every possible selection of 88 genes should be considered. The total number of models that would be required for this would therefore be:

$${}_{500}C_{88} = \frac{500!}{88!(500-88)!} = 5.04 \times 10^{99} \text{ models} \quad (4.1)$$

This number of models is clearly impossible to produce even if many millions of models could be produced every second. Instead, a number of models sufficient to determine which genes in the large set are highly regulatory of others in the set were produced: 454 models for a set of 580 genes and 668 models for a set of 722 genes. This allows each gene to be classified by monitoring the number of times pairs containing the gene were present in a set against the number of times those pairs showed a significant interaction. The classifications defined in Figure 4.6 could then be allocated on a per-gene basis.

It was not possible at this stage to infer the interaction of specific pairs of genes, because, although that pair may have been directly presented for modelling on a number of occasions, the likelihood of individual interactions is influenced by other genes in the model resulting in varying log marginal likelihood for each model. Where log marginal likelihoods are not equal, the significance of specific interactions may be biased and can therefore not be compared between models.

The size of small gene sets during multi-modelling was optimised by making observations about the networks which had been selected by hand during preliminary network modelling as described in Section 4.2.3. Those models produced from the expression data of more than 100 genes were, occasionally, unable to converge upon a single solution, producing a plot similar to that of one with too few biological replicates as shown in 4.5(b). In these cases, the quantity of data about the genes was insufficient to identify the network structure. For this reason, the number of genes per model was limited to 88 as this is the number of observations made for each gene (11 time points \times 8 replicates = 88 observations). Where p could represent the number of possible network connections for each gene, including hidden-state relationships, this prevents the number of variables per-gene, p , from exceeding the number of observations per-gene, n : a situation that can result in several equally likely explanations for the given data and the inability to identify which model is correctly describing the interactions between genes.

Although it would be possible to produce models of fewer than 88 genes, it was known that a model of this size could converge on the same solution in several seeds and so the same level of confidence in a combined model could be obtained from this

number of genes. Where genes in the set were not sufficiently likely to form a part of the network structure, they were excluded from the outcome of the modelling and so their presence did not cause harm to the analysis. However, their inclusion made the likelihood of finding non-binary interactions much greater, providing extra relevance to the underlying biological system. Larger network models also make more efficient use of the available computing power since smaller models were observed to take almost as much time as larger models.

In order to select a large gene set from which small sets should be selected, a target size was required. This could be calculated given the amount of time available for the whole multi-modelling process and also the intended goal to reach within that time. Multi-modelling required the use of a shared cluster resource which was capable of 4 concurrent VBSSM models in approximately 1 hour. Initially, an estimate of one week was allowed as it was not certain that useful results would be identified and other methods would have needed to be sought if multi-modelling had failed. A maximum number of models that could be produced in 1 week was therefore $4 \times 168 = 672$. A sensible goal for the process was to present each pair of the large set for modelling a specified number of times. This threshold was set at 5 because fewer than this would leave uncertainty in cases where half the observations were positive whilst the other half were negative, but greater than 5 would reduce the number of genes that could be managed in a week.

It can be calculated that, for a given set of N genes, the number of pairs between them is $\frac{N^2-N}{2}$. Therefore, for each set of 88 genes, $\frac{88^2-88}{2} = 3,828$ pairs would be considered by each model. With these parameters, the size of the large gene set, x , can be calculated as shown:

$$\text{Total models required} < \text{Possible models per week} \quad (4.2)$$

$$\frac{5 \left(\frac{x^2-x}{2} \right)}{3,828} < 672 \quad (4.3)$$

$$x^2 - x < 2 \left(\frac{672 \times 3,828}{5} \right) = 1,028,966.4 \quad (4.4)$$

$$x < 1,014 \text{ genes} \quad (4.5)$$

This leads to the conclusion that all large sets of genes must be fewer than 1,014 genes. An overview of the most important genes amongst all 8,878 differentially expressed was intended and so the first set to be selected was intended to capture all 77 cluster profiles in the large set. A limit of 10 genes per cluster was set and a collaborative effort alongside Vicky Buchanan-Wollaston and Emily Breeze was made to choose those genes from each cluster that showed annotations related to senescence or other potentially interesting pathways. This resulted in a final set of 580 genes which formed the first of two pools of genes for multi-modelling.

A second pool of genes was produced by including only those differentially expressed genes which were annotated by the TAIRv7 (Swarbreck *et al.*, 2008) database as a transcription factor in the *Arabidopsis thaliana* genome. This list was introduced because transcription factors are known to bind directly to the upstream regions of DNA and cause changes to levels of transcription. This type of regulation is more easily confirmed during laboratory tests than other types of interaction between genes, making any discovered networks easier to test. The list produced by this method contained 722 genes which was still manageable. 123 genes existed in both lists, indicating that 457 of the genes in the first list were not transcription factors.

Making selections of genes based on their annotations would most likely provide a bias to the interactions that were observed but, whilst it would have been exciting to identify strong hub-like behaviour from unknown genes, it would also produce difficult dilemmas when attempting to biologically verify any interactions that were identified. It is easier to perform tests for true interactions for genes with known binding sites and/or domains so that a transcriptional network can be established in the biological samples as well as in theory.

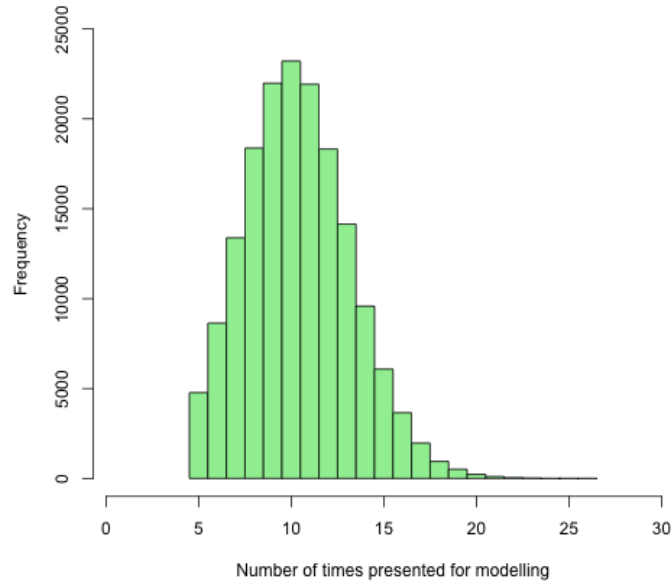
During multi-modelling of the first list of genes, the selection of 88 genes for modelling was entirely at random from the larger pool which, initially, provided a wide selection of pairs of genes. However, despite some pairs being selected many times, it was also noticed that other pairs had not yet been selected at all and that a histogram of the pair selection frequency formed a bell curve. As the aim was to select each pair at least 5 times before the multi-modelling was completed, a new strategy for gene selection was implemented part way through to reduce the number of total models required. The new strategy, as explained in Section 2.4.3 selects genes in pairs rather than individually, ensuring that those pairs which have been selected least frequently are more likely to be selected next than those which have been selected a greater number of times already.

The outcome of this altered approach is easily apparent when comparing the distributions of the pair selection frequencies between the two modelling runs. The latter run, with the pool of transcription factors, used the pairwise selection of genes for models throughout, whilst the former run with the pool of genes selected by cluster assignment was a hybrid of the two selection processes as has been described above. The distributions can be found in the histograms of Figure 4.7.

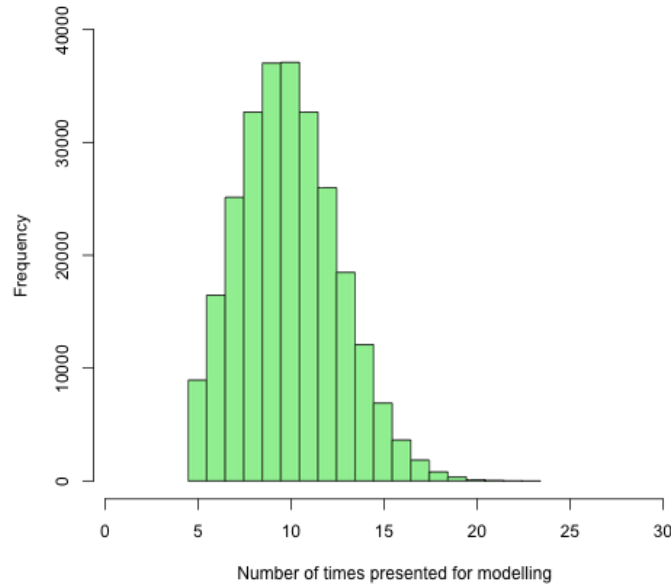
In the first multi-modelling run, shown in Figure 4.7(a), 454 models were required to exceed the threshold of 5 selections for all 167,910 pairs. Each pair was selected an average of 10.38 times with a maximum of 26 times for 3 pairs. The standard deviation of the number of models the pairs were allocated to was 2.84.

In the second multi-modelling run, shown in Figure 4.7(b), 668 models were required to exceed the threshold of 5 selection for all 521,284 pairs. Each pair was selected an average of 9.93 times with a maximum of 23 times for 1 pair. The standard deviation of the number of models the pairs were allocated to was 2.68.

As well as the statistics above regarding the frequency of pair selection, the



(a) SplineCluster selected gene pool



(b) Transcription Factor selected gene pool

Figure 4.7 – Histograms showing the frequency of pairs being presented for multi-modelling

Although the threshold that must be exceeded in order to end multi-modelling was that each pair must be presented at least 5 times, a large number of pairs were presented many more times than this. The histograms demonstrate the distribution of the numbers of times each pair was presented for modelling. It can be seen that the spread of these distributions is smaller in (b) than they are in (a) because the selection process for pairs had been improved to reduce the total number of models required to exceed a threshold of 5 times each.

number of times each gene was selected also shows similar results. In the first multi-modelling run, each gene had been selected in an average of 15.20% of the total models (69.00 models each), with a minimum of 63 selections, a maximum of 80 selections and a standard deviation of 2.620. In the second multi-modelling run, each gene had been selected in an average of 12.25% of the total models (81.83 models each), with a minimum of 76, a maximum of 89 and a standard deviation of 2.190.

These statistics demonstrate that the new selection method is reducing the number of required models by reducing the number of times each pair must be presented to exceed the desired threshold. They also demonstrate that the range of frequencies for the genes and pairs are reduced using the new selection method, allowing the pairwise interactions to be more fairly represented during the analysis of these results.

The results of the multi-modelling are presented in the form of a table showing each pair of genes from the pool alongside the following information:

- The number of times that pair had been selected for a model (No. Times Modelled).
- The number of models with at least one model seed demonstrating an interaction for the pair at the pre-defined significance threshold (No. Times Interacting).
- A percentage indicating the likelihood of an interaction for that gene pair, based on the previous two columns (Interaction Percentage).
- The regulation type, either positive or negative indicating the effect of the upstream gene on the downstream gene (Regulation Type).
- A percentage indicating the number of seeds demonstrating this interaction amongst all those from models where the genes were identified as interacting (Interaction Frequency).

The final percentage is calculated as a mean of the interaction frequencies of all the models demonstrating this interaction. If 5 models demonstrated an interaction and the interaction frequencies of each were: 70%, 80%, 80%, 90% and 70%, the interaction frequency output by multi-modelling would be $\frac{70+80+80+90+70}{5} = 78\%$. An example of the top entries in this table can be found in Table 4.3. All of these entries demonstrate robust interactions at a high significance threshold and the upstream element of these interactions is therefore classified as a putative hub.

Only the putative hubs resulting from the first gene pool are shown in this thesis because these were sufficient for finding small regulatory modules. The definition of a hub, in this case, was a gene which is identified, with greater than 95% confidence, as regulating at least two other genes on more than 50% of the occasions in which the pairs were presented for modelling. This provides a list of 118 genes thought to be regulatory of the others in the gene pool. A list of the 20 genes most likely to regulate others in the pool is shown in Table 4.4 whilst a further 98 also exist and can be found in Appendix J.

Upstream	Downstream	# Times Modelled	# Times Interacting	Interaction Percentage	Regulation Type	Interaction Frequency
At1g68990	At4g04890	12	12	100.00%	Negative	81.67%
At1g68990	At5g04340	10	10	100.00%	Positive	81.00%
At1g27730	At1g50420	10	10	100.00%	Positive	88.00%
At1g68990	At3g50260	9	9	100.00%	Positive	74.44%
At1g68990	At4g18810	9	9	100.00%	Negative	63.33%
At5g26630	At1g77000	8	8	100.00%	Negative	48.75%
At1g68990	At5g60890	8	8	100.00%	Negative	85.00%
At5g39610	At1g49850	6	6	100.00%	Positive	68.33%
At1g68990	At5g16540	5	5	100.00%	Negative	82.00%
At3g05690	At1g23750	5	5	100.00%	Negative	64.00%
At5g26630	At5g56950	13	12	92.31%	Positive	85.00%
At4g00270	At1g75080	12	11	91.67%	Positive	63.64%
At1g68990	At1g28360	12	11	91.67%	Positive	86.36%
At5g03220	At3g62090	11	10	90.91%	Positive	69.00%
At4g00270	At1g26800	10	9	90.00%	Positive	76.67%
At5g26630	At1g18470	10	9	90.00%	Negative	70.00%
At5g03220	At5g26210	10	9	90.00%	Positive	64.44%
At1g27730	At4g26500	9	8	88.89%	Negative	81.25%
At1g68990	At5g14000	9	8	88.89%	Positive	81.25%
At1g28310	At1g74660	9	8	88.89%	Positive	78.75%
At1g27730	At5g08190	9	8	88.89%	Positive	65.00%
At1g68990	At3g57040	9	8	88.89%	Negative	75.00%
At5g26630	At3g02160	8	7	87.50%	Negative	37.14%
At4g00270	At2g18670	8	7	87.50%	Positive	80.00%
At4g00270	At1g44900	8	7	87.50%	Negative	60.00%

Table 4.3 – A sample of raw results returned by multi-modelling

In the first 25 rows of the table output by multi-modelling, the most robust interactions within the large pool of genes are presented. This data is based on a 99.9% confidence threshold, meaning that the last interaction in this table was identified in 87.5% of the models where the gene co-exist, and for those models 60.00% of the seeds were more than 99.9% confident that this negative regulation was real. The choice of a 99.9% threshold is more stringent than the threshold of 95% which was eventually used to define a hub for the purposes of selecting genes to continue studying. It is immediately obvious that the same five genes (highlighted for clarity) are presented upstream in the majority of these interactions, and it is these genes with many downstream interactions which form the hubs being sought by this analysis.

It is important to note that, although these results appear to infer interactions between pairs of genes, that would be an invalid interpretation of these results. It is true that if an interaction is found to be consistent across many models, then it is more likely to appear again in future models, but it is impossible to accurately combine the predictions of a series of models and isolate individual interactions. Interactions in the model are as a result of the whole selection of genes being used to fit the model and those which appear less regularly may simply be more complicated requiring a series of specific genes to be included before they can be given a great enough probability to exceed the applied thresholds.

Instead, the correct interpretation, and the one that is taken here, is to decide that a number of genes in the pool have a higher probability of having many downstream

Hub AGI Number	Number of Genes Downstream	Description
At5g05410	366	DRE-binding protein 2A (DREB2A)
At3g05690	292	nuclear factor Y, subunit A2 (NF-YA2)
At1g27730	266	STZ - Related to Cys2/His2-type zinc-finger proteins
At1g68990	262	male gametophyte defective 3 (MGP3)
At2g47040	249	VANGUARD1 (VGD1)
At2g38880	238	nuclear factor Y, subunit B1 (NF-YB1)
At4g25490	192	C-repeat/DRE binding factor 1 (CBF1)
At2g25900	151	ATCTH
At5g06510	140	nuclear factor Y, subunit A10 (NF-YA10)
At3g15210	136	ethylene responsive element binding factor 4 (ERF4)
At5g62040	135	PEBP (phosphatidylethanolamine-binding protein) family protein
At3g14020	132	nuclear factor Y, subunit A6 (NF-YA6)
At2g34710	117	PHABULOSA (PHB)
At5g18140	101	Chaperone DnaJ-domain superfamily protein
At1g63850	96	BTB/POZ domain-containing protein
At3g25540	95	LONGEVITY ASSURANCE GENE 1 (LAG1)
At5g39610	95	ANAC092 - A NAC-domain transcription factor
At5g13750	87	zinc induced facilitator-like 1 (ZIFL1)
At5g57565	82	Protein kinase superfamily protein
At2g40340	79	DREB2C

Table 4.4 – A sample of hub genes found by multi-modelling

Shown are the 20 most likely hub genes of the 118 which met the threshold of demonstrating an interaction with at least two other genes at more than 95.0% confidence on more than half the occasions where those genes were present in the same model. The other 98 genes also considered a hub but not shown here can be found in Appendix J. The number of genes downstream represents how many other genes of the set of 580 were found to be regularly regulated by the hub gene shown. As was previously stated, this does not infer direct interactions, but instead simply ranks the genes by their regulatory potential so that those which do not regularly regulate others in a putative network model are not studied further.

effects and can therefore be classed as hubs. The list of 118 genes shown in Table 4.4 and continued in Appendix J are those which appear to have a regulatory effect on the largest numbers of genes, qualifying them for further study. Therefore it is those 118 genes which were next tested in a biological context to identify their effect on the senescence phenotype.

4.4 Biological Validation of Module Centres

In order to biologically validate the effect of putative hubs on the senescence process, reduced expression lines were obtained from SALK, as described in Section 1.7.1.1. It was assumed that if a gene were truly central to a senescence related process they would, when artificially reduced in expression, have a large effect on the phenotype of a leaf undergoing senescence. There are situations, such as secondary pathways, in which the reduced expression could be compensated for by other genes, providing little phenotypical change. There should, however, be no circumstances that would lead to a false positive under this assumption.

71 homozygous SALK lines were listed for the 118 putative hub genes, though

some of these were for the same gene, allowing 53 of the putative hub genes to be tested with some replication. Of those ordered, only 64 were delivered due to stock shortages and these are listed in Table 4.5 along with the identifiers used to describe them throughout the remainder of this thesis.

4.4.1 Phenotype Screening

Although senescence is a slow process, often taking more than two months to complete, it can also be induced by a number of different stresses including pathogen attack, hormone response, high and low temperature climates, drought and/or darkness. These can greatly increase the rate of senescence at the risk of also inducing responses in genes specifically associated with those conditions rather than senescence itself. There have been some studies which have found differences between the levels of expression for some genes during stress-induced senescence versus natural senescence (Buchanan-Wollaston, 1997; Becker & Apel, 1993; Gong *et al.*, 2001), but, although this should not be disregarded, time constraints demanded that stress-induced senescence, specifically reduced-light stress, should be used for screening. This was a decision that could also be justified by the fact that the lines being screened were already theoretically central to natural senescence processes, based on observations made during the microarray experiment. The choice of reduced-light stress was in part because it is more easily standardised than any of the other stresses.

The phenotype screen, described in Section 2.5.2, was used to assess whether the plant lines showed an altered phenotypical response to dark-senescence when compared with Col-0 wild-type plants under the same conditions. As well as the 64 ordered from SALK, 11 other reduced expression lines were included which were suspected to be involved in the senescence process from previous studies within the Buchanan-Wollaston group and 19 enhanced expression mutants which were supplied by Jesper Grønlund and Sanjeev Kumar as described in Section 1.7.1.2.

The profiles returned from this screening process describe the level of senescence throughout the observation period. The senescence profiles of each plant line were compared with the profile of the Col-0 wild-type by eye. 8 of the lines which were screened, including 7 of those identified by multi-modelling and 1 enhanced expression mutant (MYB59 - At5g59780), gave the most altered phenotypical response with some lines demonstrating as much as a 2.5 day delay in the onset of senescence. A plot of the senescence profiles of these mutants alongside the profile of the Col-0 wild-type can be found in Figure 4.8. Descriptions and gene identifiers for all 8 lines which have been associated with senescence during the screening process are listed in Table 4.6.

The method of analysis was later developed further to remove the element of subjectivity typical of analysis by eye. This method, involving the fit of a logistic curve to the yellowing data followed by two ANOVAs comparing characteristics of those curves to determine significantly altered phenotypical response, is described at the end of Section 2.5.2 and allowed plots such as those shown in Figure 4.9 to be produced. All

Identifier	SALK ID	AGI Code	Short Description
IM131	SALK_065697c	At1g04250	AUX/IAA transcriptional regulator family protein
IM132	SALK_011820c	At1g04250	AUX/IAA transcriptional regulator family protein
IM133	SALK_062147c	At1g06390	GSK3/SHAGGY-like protein kinase 1
IM134	SALK_043420c	At1g06390	GSK3/SHAGGY-like protein kinase 1
IM135	SALK_081927c	At1g09530	Phytochrome interacting factor 3
IM136	SALK_110897c	At1g22190	AP2 domain containing transcription factor
IM137	SALK_054092c	At1g27730	STZ - Related to Cys2/His2-type zinc-finger proteins
IM138	SALK_050247c	At1g29280	WRKY DNA-binding protein 65
IM139	SALK_099094c	At1g48480	Receptor-like kinase 1
IM140	SALK_142042c	At1g51660	Mitogen-activated protein kinase kinase 4
IM142	SALK_006559c	At1g54160	Nuclear factor Y, subunit A5
IM143	SALK_111422c	At1g56170	Nuclear factor Y, subunit C2
IM144	SALK_052790c	At1g62310	Jumonji (JMJC) domain-containing protein
IM145	SALK_093731c	At1g66390	MYB domain protein 90
IM146	SALK_019747c	At1g69490	NAC-like, activated by AP3/PI
IM147	SALK_011491c	At1g69690	TCP15 - TCP transcription factor
IM148	SALK_124755c	At1g76720	Eukaryotic translation initiation factor 2 family protein
IM149	SALK_143304c	At1g76720	Eukaryotic translation initiation factor 2 family protein
IM150	SALK_141481c	At1g77570	Winged helix-turn-helix transcription repressor
IM151	SALK_059705c	At2g03340	WRKY DNA-binding protein 3
IM152	SALK_000828c	At2g20050	cAMP-dependent protein kinase regulator
IM153	SALK_026551c	At2g22540	K-box region and MADS-box TF family protein
IM154	SALK_072930c	At2g22540	K-box region and MADS-box TF family protein
IM155	SALK_143721c	At2g25900	Zinc finger C-x8-C-x5-C-x3-H type family protein
IM156	SALK_112158c	At2g25900	Zinc finger C-x8-C-x5-C-x3-H type family protein
IM157	SALK_118231c	At2g30360	SOS3-interacting protein 4
IM158	SALK_008924c	At2g34710	Homeobox-leucine zipper family protein
IM159	SALK_003337c	At2g34720	Nuclear factor Y, subunit A4
IM160	SALK_038840c	At2g38880	HAP3A - AtNF-YB1 - Confers drought tolerance
IM161	SALK_036317c	At2g42880	MAP kinase 20
IM162	SALK_123216c	At3g01470	Homeobox 1
IM163	SALK_110045c	At3g04260	Plastid transcriptionally active 3
IM164	SALK_108852c	At3g04260	Plastid transcriptionally active 3
IM165	SALK_129820c	At3g11650	NDR1/HIN1-like 2
IM166	SALK_134653c	At3g12977	NAC (No Apical Meristem) domain transcription factor
IM167	SALK_028169c	At3g14020	Nuclear factor Y, subunit A6
IM168	SALK_143369c	At3g14020	Nuclear factor Y, subunit A6
IM169	SALK_002235c	At3g20910	Nuclear factor Y, subunit A9
IM170	SALK_039707c	At3g25540	TRAM, LAG1 and CLN8 (TLC) lipid-sensing protein
IM171	SALK_057546c	At3g25540	TRAM, LAG1 and CLN8 (TLC) lipid-sensing protein
IM172	SALK_147540c	At3g58040	Seven in absentia of Arabidopsis 2
IM173	SALK_016395c	At3g59400	Enzyme binding; tetrapyrrole binding
IM174	SALK_011461c	At3g59400	Enzyme binding; tetrapyrrole binding
IM175	SALK_056541c	At3g60390	HAT3 - Encodes homeobox protein HAT3
IM176	SALK_014055c	At3g60390	HAT3 - Encodes homeobox protein HAT3
IM177	SALK_016764c	At4g13100	RING/U-box superfamily protein
IM178	SALK_056165c	At4g13100	RING/U-box superfamily protein
IM179	SALK_096411c	At4g13670	Plastid transcriptionally active 5
IM180	SALK_034157c	At4g23810	WRKY family transcription factor
IM181	SALK_095007c	At4g24540	AGAMOUS-like 24
IM182	SALK_091933c	At4g32280	Indole-3-acetic acid inducible 29
IM184	SALK_059819c	At5g11510	Myb domain protein 3r-4
IM185	SALK_064538c	At5g24470	Pseudo-response regulator 5
IM186	SALK_053198c	At5g24470	Pseudo-response regulator 5
IM187	SALK_111899c	At5g24800	Basic leucine zipper 9
IM188	SALK_090154c	At5g39610	ANAC092 - A NAC-domain transcription factor
IM189	SALK_017961c	At5g42190	E3 ubiquitin ligase SCF complex subunit protein
IM190	SALK_030781c	At5g43310	COP1-interacting protein-related
IM191	SALK_059254c	At5g57390	AINTEGUMENTA-like 5
IM192	SALK_151934c	At5g57565	Protein kinase superfamily protein
IM193	SALK_013022c	At5g62040	Brother of FT and TFL1 - PEBP family
IM194	SALK_107954c	At5g64340	Sequence-specific DNA binding transcription factors
IM195	SALK_021930c	At5g64340	Sequence-specific DNA binding transcription factors
IM196	SALK_102383c	At2g38880	Nuclear factor Y, subunit B1
IM197	SALK_116974c	At5g11510	MYB domain protein 3r-4
IM198	SALK_042760c	At1g54160	Nuclear factor Y, subunit A5

Table 4.5 – SALK lines which were biologically tested as putative hubs

Although 118 genes were identified as putative hubs, only 53 of these were available for biological testing. Some genes had more than one reduced-expression line available, and all were ordered and shown above.

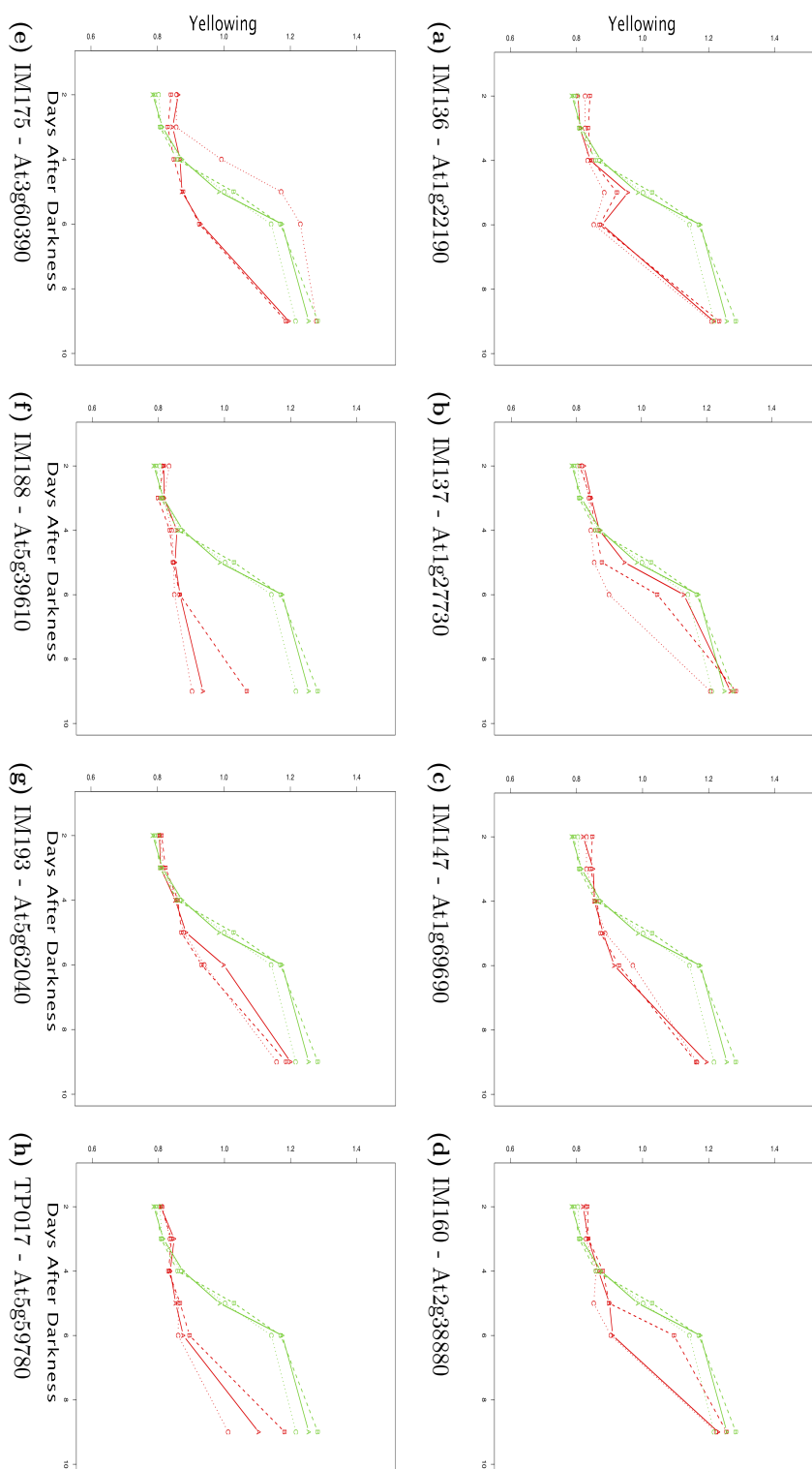


Figure 4.8 – Plots of senescence profiles for knockout mutant lines showing altered phenotype during dark-induced senescence
Each of the senescence profiles for each of the knockout mutant lines were plotted as a comparison to the Col-0 wild-type. Those eight which were identified as having an altered phenotypical response during dark-induced senescence are shown here. In each plot, the three green lines represent the biological replicates of Col-0 whilst the red lines represent the mutant line. The horizontal axis of each plot represents the number of days after darkness whilst the vertical axis represents the level of yellowing in the leaves.

Identifier	AGI	Name/Description
IM136	At1g22190	AP2 domain containing transcription factor
IM137	At1g27730	STZ - Related to Cys2/His2-type zinc-finger proteins
IM147	At1g69690	TCP15 - TCP transcription factor
IM160	At2g38880	HAP3A - AtNF-YB1 - Confers drought tolerance
IM175	At3g60390	HAT3 - Encodes homeobox protein HAT3
IM188	At5g39610	ANAC092 - A NAC-domain transcription factor
IM193	At5g62040	Brother of FT and TFL1 - PEBP family
TP017	At5g59780	MYB59 - Encodes putative transcription factor MYB59

Table 4.6 – Genes demonstrating an altered phenotype after quantification of the rate of dark-induced senescence

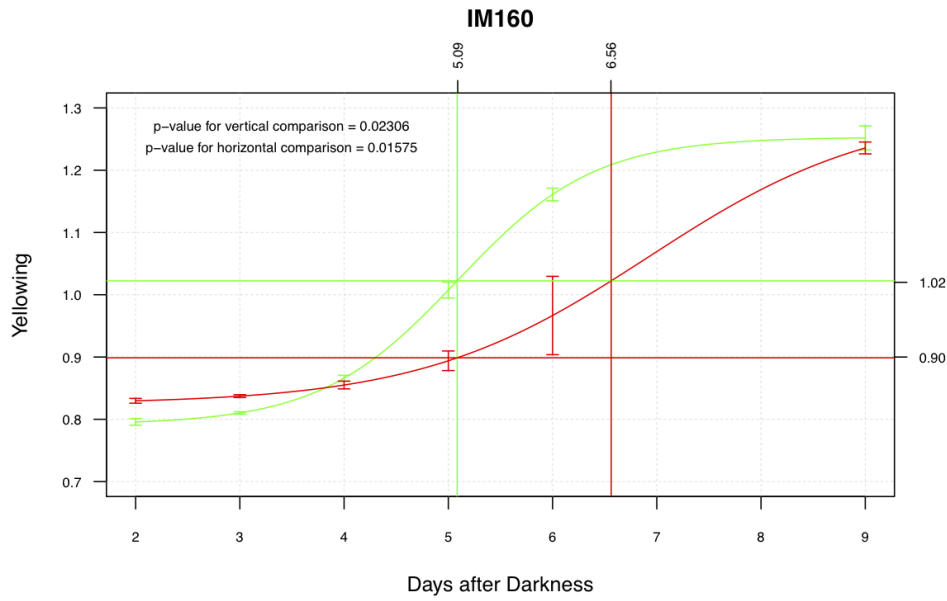
After the introduction of a quantification of the rate of dark-induced senescence, eight genes were shown to have an identifiable difference from the Col-0 wild-type plants.

50 plant lines were re-assessed using this more refined method to verify that the decisions made in selecting the 8 lines for further analysis were correct. Unfortunately, the method relies on being able to fit a logistic curve to the data, and this was not possible for all the plant lines as insufficient observations had been made for the parameters to converge and so only 41 of the lines could be tested in this way. Six of the lines tested gave p -values where $p \leq 0.05$ in both comparisons, indicating significantly altered phenotypical response, and these are shown in Appendix K where they can be reproduced large enough to view properly. Of the 8 lines selected by eye, 3 (IM136, IM188 and TP017) did not allow the fit of a logistic curve, presumably because their delay in senescence was so great that the point of inflection in the curve was not reached within the first 9 days, whilst the other 5 were all identified as significant by the ANOVA tests. The remaining significant plant line, TP010 (At4g05100 - *MYB74*), was the only other line showing altered phenotypical response to dark-induced senescence. TP010 has not been studied further as this analysis was performed some time after the selection of the 8 lines by eye.

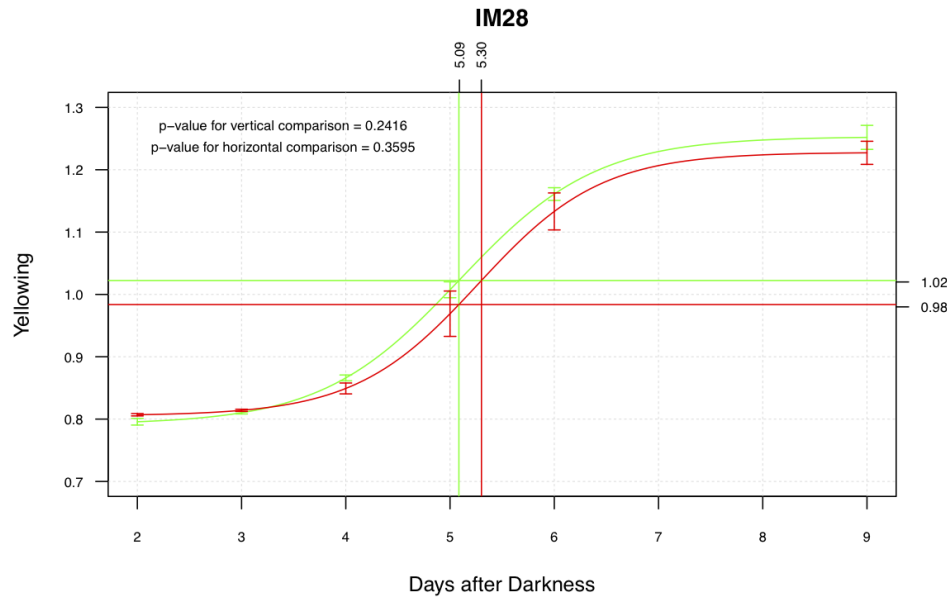
4.4.2 Homozygous T-DNA Insertion Screening

In order to verify the insertion of T-DNA into the intended gene of the 7 SALK plant lines shown to have an altered phenotypical response to dark-induced senescence, DNA was extracted from leaf tissue collected from each of the plant lines using the methods shown in Section 2.6.1. This was then used as a template in PCR reactions designed to verify the presence of T-DNA in the correct region of the genome, as described in Section 2.6.2.

The enhanced-expression line, TP017, was already known to be homozygous since this was screened by Jesper Grønlund and Sanjeev Kumar when it was produced. The enhanced expression of this line was also confirmed at the time the line was produced which was performed by semi-qPCR which determined that, under the assumption that a cycle in PCR results in a doubling of the concentration of the amplified DNA, the expression of MYB59 in TP017 is 19-fold that of the expression of MYB59 in Col-4, the



(a) IM160 - At2g38880 - HAP3A/AtNF-YB1 - Significant difference in phenotypical response



(b) IM28 - At1g27730 - MYB90 - Non-significant difference in phenotypical response

Figure 4.9 – Analysis of logistic curve models reveal the significance of altered phenotypical responses during dark-induced senescence

Fitting logistic curve models to the yellowing data collected from both wild-type (green) and altered expression mutants (red) allows for comparisons to be made between their profiles. The green crosshair is plotted to intersect at the point of maximum change in the wild-type. The horizontal arm of the red crosshair is plotted to identify the level of yellowing in the altered expression mutant at the time of maximum change in the wild type. The vertical arm of the red crosshair plots the time at which the altered expression mutant reaches the same level of yellowing as the wild type at the point of maximum change. The differences between the crosshairs in both axes are individually considered using two ANOVAs and the p -values shown are derived from these. Where $p \leq 0.05$ in both tests, the altered phenotypical response is considered significant as shown in (a), whilst neither p -value meets this criteria in (b) where it is obvious that the response is almost identical. Error bars represent one standard error across biological replicates.

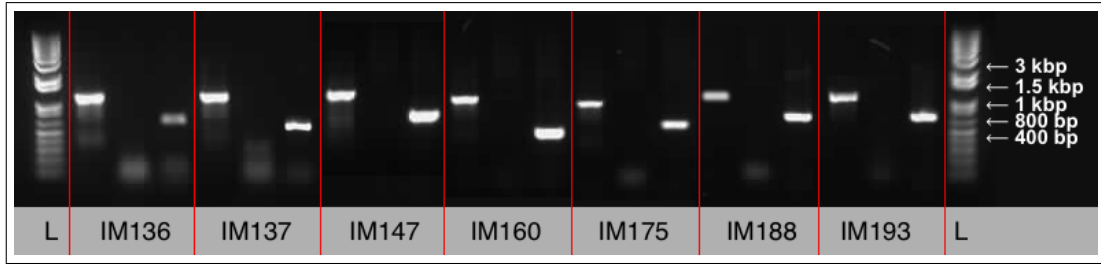


Figure 4.10 – Electrophoresis gel demonstrating the presence of homozygous T-DNA inserts for reduced-expression mutants

For each reduced-expression line, three PCRs were performed. The first lane in each triplet represents gene-specific primers with a genomic DNA template. The second lane represents the same primers with DNA extracted from leaves of the mutant as a template. The last lane represents the right gene-specific primer with a T-DNA specific primer using DNA extracted from leaves of the mutant as a template. Where a product exists in the first and last lane, but not the middle lane, the T-DNA insert is homozygous. The ladders, marked 'L', are 1kb Plus (Invitrogen).

genotype used to produce TP017.

All 7 of the reduced-expression lines were confirmed to have a homozygous T-DNA insert at the expected location in the genome as is demonstrated by the electrophoresis gel shown in Figure 4.10. A product in the last lane of each triplet indicates a T-DNA insert is present in at least one strand whilst a product in the middle lane would indicate at least one strand without the insert. Therefore, the absence of a product in the middle lane indicates that all the mutants are homozygous.

It should be noted, however, that these tests do not confirm that no other T-DNA insertion exists in another part of the genome. This should be considered when interpreting any results.

4.4.3 Gene Expression Level Testing

Reverse Transcriptase PCR (RT-PCR) was used to verify that the reduced-expression lines were reduced in expression when compared with Col-0 wild-type at the same point in the senescence process. This was done by preparing a number of plants and allowing them to grow naturally before harvesting leaves from both the reduced-expression line and Col-0 wild-type plants at a time when the leaves are fully expanded and green. Trizol (Invitrogen) was used in the extraction of total RNA from the leaves, following the methods shown in Section 2.6.3.1, before performing RT-PCR on the samples using the methods shown in Section 2.6.7. The enhanced-expression line, TP017, had been confirmed to show higher levels of expression than that of wild-type plants when it was produced.

After extraction of total RNA for the 7 reduced-expression lines, the nanodrop was able to determine the concentration and purity of the extracted RNA, as shown in Table 4.7. Here, it can be seen that although the total RNA concentrations vary between samples, all have returned a concentration comparable with that of Col-0 and all have a 260/280nm absorbance ratio of more than 2.00, indicating minimal contamination of

Identifier	total RNA Conc. (ng/ μ L)	Absorbance Ratio 260/280nm
Col-0	57.4	2.01
IM136	22.8	2.03
IM137	53.2	2.12
IM147	134.7	2.10
IM160	60.8	2.03
IM175	99.3	2.06
IM188	123.6	2.11
IM193	188.6	2.10

Table 4.7 – Nanodrop results for total RNA extracted from mature leaves of reduced-expression lines

After extraction of total RNA from mature leaf samples, the RNA concentrations were measured on a nanodrop. Each of the extractions returned an acceptable concentration comparable with that of Col-0 and all have a 260/280nm absorbance ratio of more than 2.00, indicating that little contamination exists in the solution.

the samples.

An Agilent 2100 bioanalyser was used to identify the quality and degradation levels of the total RNA obtained from the leaf samples above. The results of this analysis can be found in Figure 4.11. The plots show that each of the samples are of sufficient quality with little degradation with the possible exception of IM136, which does show signs of partial degradation as identified by the taller peaks to the left of the graph. This is also identifiable in the virtual electrophoresis gel which shows a strong band in the smaller sized fragments of lane 2. Despite this degradation, the sample for IM136 was still considered good enough quality for RT-PCR.

Although the total RNA samples were shown to be sufficiently high quality for RT-PCR, the samples also contained vast quantities of genomic DNA. It was essential that the RT-PCR primers were able to amplify a region of cDNA for the specific gene such that it can be distinguished from an amplification of the equivalent genomic DNA. This is possible where introns exist in the sequence as these are spliced from RNA immediately following transcription and therefore shorten the amplified region if the primers lie either side of an intron. Three of the reduced-expression lines, IM136, IM137 and IM147, had a T-DNA insert in genes which feature no introns and therefore the PCR products of both genomic DNA and cDNA would be identical in size. To ensure that any product generated by the RT-PCR would be representative of the abundance of cDNA and therefore the levels of expression for the gene, those three samples were DNase treated to ensure the removal of all traces of genomic DNA.

Three reduced-expression lines were analysed by RT-PCR and the results of this can be found in Figure 4.12. Here it can be seen that whilst IM160 and IM188 both show a reduced level of expression when compared to the Col-0 wild-type, the conclusions to be made from IM193 are less obvious, showing no change in expression due to the T-DNA insert. IM193's T-DNA insert is in the upstream region of the gene rather than the gene sequence itself and, if the transcription factor binding sites have

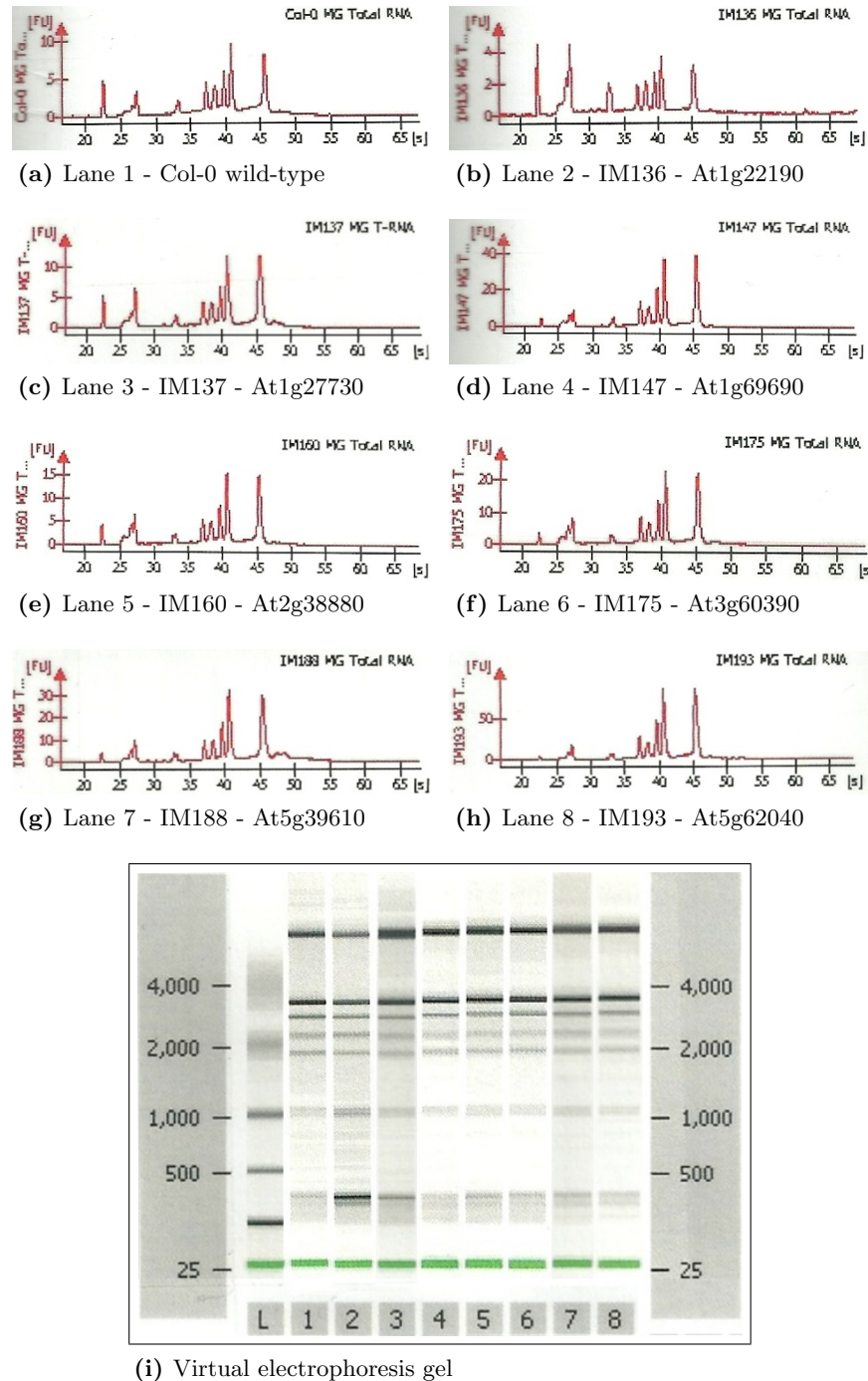


Figure 4.11 – Agilent 2100 bioanalyser results for total RNA samples from leaves of reduced-expression lines

Figures (a) to (h) illustrate the bioanalyser plots of total RNA fragments extracted from mature leaves of Col-0 wild-type and the seven reduced-expression lines. High quality samples with low-levels of degradation demonstrate a high final (right-most) peak on the plot whilst left-most peaks are minimised. All the samples shown are of sufficiently low degradation with the possible exception of IM136 which shows some degradation towards the left of the plot. The virtual electrophoresis gel, shown in (i), is a simulation of how an electrophoresis gel may appear if the samples were separated in this way. In this plot, a grey smear with few distinct bands in the lower area would be an indication of RNA degradation.



Figure 4.12 – RT-PCR results demonstrating expression levels of three reduced-expression lines

In each image, the RT-PCR product of the Col-0 wild-type (WT) and the product of the reduced-expression line (RE) at the same reaction cycle can be seen. IM160, (a), demonstrates that the expression of this gene is substantially lowered by the insertion of T-DNA. This is also demonstrated by the RT-PCR of IM188 as shown in (b). The results are less conclusive for IM193, shown in (c), where expression is approximately equal during the entire reaction. Of these three reduced-expression lines shown, IM193 is the only one to have the T-DNA insert in the upstream region rather than in the gene sequence and as such may be the reason for no obvious change in expression. Another explanation could be that the altered senescence response phenotype shown in IM193 may be caused by another unknown T-DNA insertion in the genome. Given that it is shown here that the expression of IM193 upstream of the insert is seemingly identical to wild-type, this alternative hypothesis could be the cause of the altered phenotype.

not been distanced from the gene sequence by the insert, may explain the unaltered response. It is also quite possible that IM193 also contains a second T-DNA insert at another position in the genome which has caused the senescence response to be altered.

Whilst each of IM136, IM137 and IM147 were successfully treated with DNase to remove traces of genomic DNA and new primers were ordered for IM175, the remaining illustra ready-to-go RT-PCR beads were consumed by individuals of a concurrent project. The suppliers of the beads, GE Healthcare, were also without stock for 3 months making it impossible to complete the analysis of gene expression in the remaining lines. Given that the phenotypes show altered response to dark-induced senescence and that the lines had been confirmed to be homozygous, a decision was made to move forward by predicting the local networks surrounding the affected genes of the 7 reduced-expression lines and the single enhanced-expression line which had been biologically validated.

4.4.4 Known Gene Functions

Research in relevant literature allowed a better understanding of the known functions, if any, of the eight genes identified by phenotype screening and shown in Table 4.6. *AP2* (At1g22190 - IM136) is a dehydration responsive element binding (DREB) transcription factor, the family of which has been shown to be involved in several types of stress, particularly that of dehydration but also as part of the pathway responding to altered levels of plant hormones such as abscisic acid (ABA) (Zhou *et al.*, 2010). The levels of ABA are seen to change during senescence as is shown in Figure 4.3 and a comparison of the gene expression for *AP2* (see Figure 4.13(a)) with that of the levels of ABA

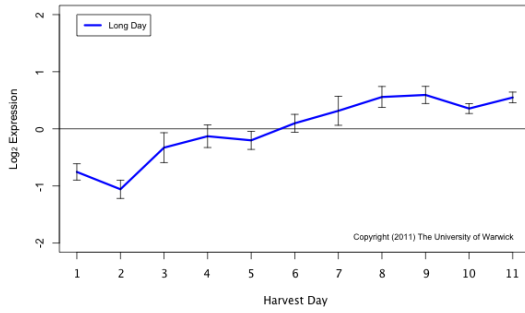
identifies a strong correlation perhaps justifying its role in dehydration response. It has been shown that DREB genes are responsible for the control of stress response genes involved in ABA-dependent pathways (Egawa *et al.*, 2006; Wang *et al.*, 2010; Zhou *et al.*, 2010).

STZ or *ZAT10* (At1g27730 - IM137) is similar to *AP2* in that it is strongly induced by dehydration, high-salt levels, cold stresses and the presence of ABA (Sakamoto *et al.*, 2004). The profile of *STZ*, shown in Figure 4.13(b), can be seen to be very similar to that of *AP2* if not slightly delayed. The biggest rise in expression of *STZ* appears to be between observations 5 and 7 (27 and 31 days after sowing) which is the time at which the senescence process is thought to commence. The level of *STZ* is known to increase rapidly within only 10 minutes of ABA treatment (Sakamoto *et al.*, 2004), though it was then seen to decrease after the first response, perhaps indicating that it is early in the order of events leading to an ABA response and that a later component of the network leads to its inhibition.

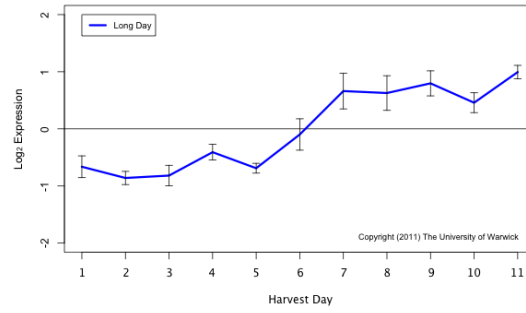
TCP15 (At1g69690 - IM147) encodes a class I TCP protein which has been shown to have similar binding affinities to *TCP11* (At5g08330) which is a regulator of growth in leaves (Viola *et al.*, 2011) and so its reduction in expression throughout the experiment, shown in Figure 4.13(c), may be associated with the halting of leaf growth during the onset of senescence. This would explain the slimmer leaves found on IM147 mutants during the phenotype screening. Further known functions of *TCP15* can be found in Giraud *et al.* (2010) in which it was shown to induce the expression of *PRR5*, a circadian clock gene, which was confirmed by both yeast-2-hybrid analysis and observed reduction in expression of *PRR5* when studying the same SALK institute reduced-expression mutant as was used in this PhD. Some genes of the circadian clock are seen to become disrupted towards the end of senescence whilst others persist their usual circadian rhythms. It could then be conceived that *TCP15* may be a cause for disruption in the rhythm of some circadian genes.

HAP3A (At2g38880 - IM160) encodes a protein which is able to form a three protein complex of HAP2, HAP3 and HAP5 that binds to CCAAT-box motifs (Edwards *et al.*, 1998; Wenkel *et al.*, 2006). Together, this complex, when over-expressed, is able to reduce the expression of *Flowering locus T (FT)* (Wenkel *et al.*, 2006), but no CCAAT-box motif can be found in the promoter region of *BFT* (At5g39610 - IM193) so this would not appear to be the cause of any regulatory relationship identified here. Increased expression of *HAP3A* also shows increased drought tolerance (Nelson *et al.*, 2007) perhaps indicating that the increase in expression is because of other drought responsive genes which are being activated in response to ABA.

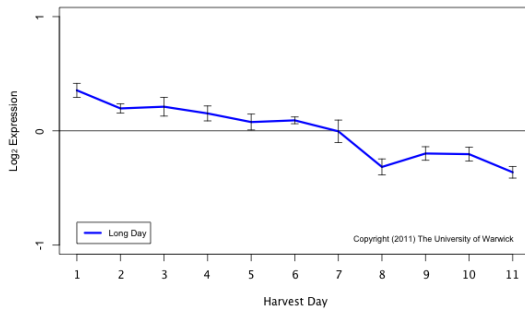
HAT3 (At3g60390 - IM175) is a homeobox-leucine zipper II gene and is part of a family with 9 other HD-Zip genes, though little is known about *HAT3* specifically. This family is unique to plants and has members known to respond to light (Carabelli *et al.*, 1993, 1996; Steindler *et al.*, 1997) and auxin (Sawa *et al.*, 2002) levels, thought therefore to be involved in shade avoidance (Ciarbelli *et al.*, 2008). Most members



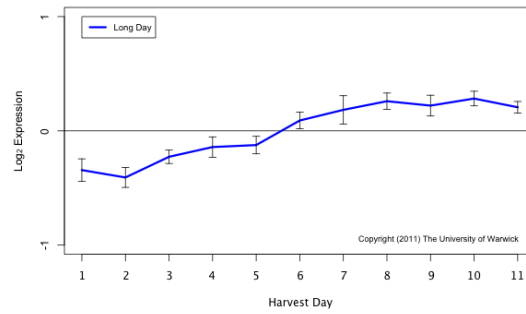
(a) AP2 - At1g22190 (IM136)



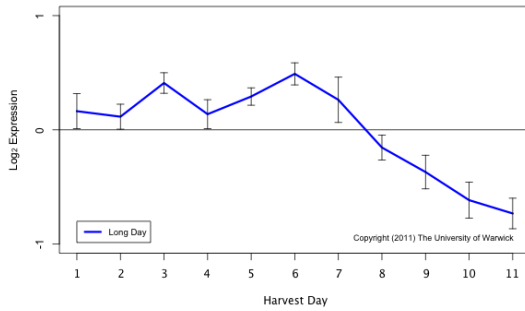
(b) STZ - At1g27730 (IM137)



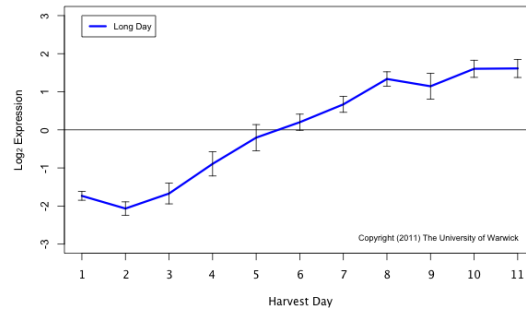
(c) TCP15 - At1g69690 (IM147)



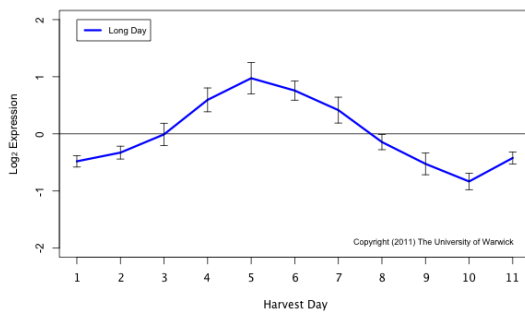
(d) HAP3A - At2g38880 (IM160)



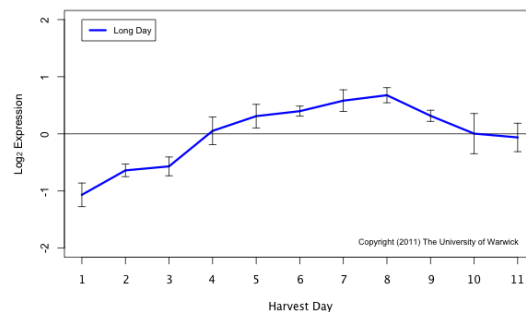
(e) HAT3 - At3g60390 (IM175)



(f) ANAC092 - At5g39610 (IM188)



(g) Brother of FT - At5g62040 (IM193)



(h) MYB59 - At5g59780 (TP017)

Figure 4.13 – Expression levels of biologically validated hub genes during senescence

For each plot, the horizontal axis displays the harvest day whilst the vertical axis displays the expression level which has been mean and sd normalised to 0 and 1 respectively. Each unit of change in expression relates to a doubling or halving of the expression level as these are shown on a \log_2 scale.

of the family increase in expression during leaf senescence, whilst *HAT3* is known to decrease, particularly during the stages of flower development (Ciarbelli *et al.*, 2008) (Figure 4.13(e)). This possibly indicates that its role is only useful prior to the onset of senescence processes or that it suppresses a senescence response from other genes in the family.

ANAC092 (At5g39610 - IM188) is known by many names including *ATNAC2* and *ORE1*, some of which are also used as the name of other genes, which makes literature research challenging. *ANAC092* has previously been shown to have a role in senescence (Ooka *et al.*, 2003; He *et al.*, 2005), its expression suppressed in young leaves by action of the micro-RNA miR164 which is in turn suppressed as the leaf develops by EIN2 (At5g03280) (Kim *et al.*, 2009). This delay in expression of *ANAC092* was also seen during the senescence experiment and can be seen in Figure 4.13(f). Despite large amounts of research being performed on *ANAC092*, its direct downstream targets and physiologically relevant trigger mechanisms are still largely unknown (Balazadeh *et al.*, 2010a, 2011). Response to salt stress is often associated with an increase in expression of *ANAC092* (He *et al.*, 2005; Balazadeh *et al.*, 2010b) and this may indicate that *ANAC092* operates in a associated pathway to STZ which has also been found to show a phenotypical response to dark-induced senescence. Other stresses associated with senescence and *ANAC092* include oxidative stress where it was found that *anac092* mutants were more tolerant to exposure to H₂O₂ or MV than the wild type, perhaps indicating that oxidative stress causes up-regulation of *ANAC092* and consequently the senescence response (Woo *et al.*, 2004).

BFT - Brother of FT and TFL1 (At5g62040 - IM193) encodes for a phosphatidylethanolamine-binding protein similar to the homologues Flowering Locus T (FT) and Terminal Flower 1 (TFL1) which, via a single amino acid substitution, are able to promote or repress flowering respectively (Hanzawa *et al.*, 2005). Despite those similarities, until very recently nothing was known about the function of *BFT* specifically (Carmona *et al.*, 2007). In the last year, it has been observed that the induction pattern of *BFT* is different from those of other members in the FT/TFL1 family although its function is more similar to TFL1 than it is to FT as its over-expression results in the inhibition of flowering (Chung *et al.*, 2010). These findings are despite *BFT* having a higher sequence similarity to *FT*. An extensive study of the activity of *BFT* (Yoo *et al.*, 2010) leads to the same conclusions and also finds that, although thought to be redundant in function, TFL1 has a bigger influence on the timing of flowering than *BFT*. The role of *BFT* during leaf senescence is not immediately obvious, but given its reduction in expression at around the time senescence is commencing (see Figure 4.13(g)), it may be that it possesses the same inhibitory role for the onset of senescence as it does for flowering.

MYB59 (At5g59780 - TP017) was found to be one of many genes down-regulated by *ANAC092* (He *et al.*, 2005) and encodes a MYB domain transcription factor: a superfamily of transcription factors that play regulatory roles in developmental processes

and defence response in plants. With four alternative splice variants, MYB59 has been shown to localise to and become involved in various biological processes (Li *et al.*, 2006) which does inspire the question as to which of these variants are represented by the expression profile shown in Figure 4.13(h). One confirmed role of MYB59 is in the regulation of root growth and the cell cycle (Cominelli & Tonelli, 2009; Mu *et al.*, 2009) but neither of these indicate a role in leaf senescence. Another study found *MYB59* to be noticeably up-regulated by gibberellic acid and jasmonic acid, whilst even larger responses were observed during exposure to salicylic acid and to 100 μ M CdCl₂ (Yanhui *et al.*, 2006). These hormones are typical of those changing in levels during senescence, though this still fails to identify the role MYB59 may play during senescence.

4.5 Predicting Local Networks for Module Centres

Eight genes were now shown to be involved in processes associated with senescence and so VBSSM was now used as a method for identifying genes and their interactions which form a local network around each of the eight module centres. To do this, VBSSM was used in a similar manner to that of multi-modelling, but with gene selection based on a Metropolis-like approach as described in Section 2.4.4. By selecting a pool of genes which are believed to be elements of a network surrounding the eight module centre genes, VBSSM could be provided with randomly selected sets of genes to model and attempt to find the interactions between them. On this occasion, however, rather than changing all the elements of the model, as was the case for multi-modelling, only small numbers of genes are exchanged between models in an attempt to find the set which maximises the log marginal likelihood and is therefore the best gene set for describing the biological data with a predictive model.

Pools of genes were selected separately for each of the eight module centres and were chosen because of their profiles providing a match to the target gene after an optimum time-shift and inversion. The matches were identified using the Qian similarity score (Qian *et al.*, 2001), methods for which are described in Section 2.3.3. After scoring all 8,788 differentially expressed genes for their similarity to each of the genes predicted to be module centres, they were classified into those which required no time-shift for the optimum match and those which did. This is because the similarity score is degraded by decreasing the number of gene expression observations which are compared between two profiles, meaning that those with a time-shift, and hence fewer overlapping observations, are unable to achieve comparably high scores and cannot be ranked with those needing no time-shift.

Each list of genes were ranked according to their similarity with the target gene and a hand-picked cut-off defined for each list based on a comparison of the profiles of genes sampled from the list against the profile of the target gene using the web-based PRESTA Senescence Data Viewer (Legaie & McHattie, 2010). Before the 1,500th ranked gene in each list, substantially reduced similarity could be identified by eye. A separate

Identifier	AGI	Gene Pool Size	Number of Iterations
IM136	At1g22190	156	164
IM137	At1g27730	159	241
IM147	At1g69690	159	309
IM160	At2g38880	152	254
IM175	At3g60390	112	240
IM188	At5g39610	159	170
IM193	At5g62040	134	212
TP017	At5g59780	139	185

Table 4.8 – Metropolis-like VBSSM modelling runs performed for genes identified as module centres

For each of the eight genes identified as module centres, a pool of transcription factors with a similar profile to the module centre, as scored by the Qian similarity score, was produced. The sizes of those pools are as shown alongside the number of iterations required by the Metropolis-like VBSSM approach to find the most likely model between them.

ranked list of genes with a time-shifted match was found to show no obvious profile similarity after the 100th in the list. By using these as a threshold for similarity, all the matching profiles were within a final list of 1,600 genes whilst many false positives had also been identified. As the false positives should not be found to interact with the target gene, their presence was deemed superficial.

Each of the eight lists of 1,600 genes were reduced in size by removing non-transcription factors since the purpose was to find robust networks which were suitable for *in vivo* testing. This provided eight pools of genes which could be combined with the eight predicted module centres for modelling. The sizes of these lists can be found in Table 4.8.

Each module centre, within its own modelling run, was retained across all models whilst other genes of those models were substituted following the procedures of Section 2.4.4. After a number of iterative steps, the number of which for each run is shown in Table 4.8, no improved log marginal likelihood could be found despite substituting as few as a single gene at a time from the last known optimum model. As can be seen, in the majority of cases, the number of models was fewer than 300 which is fewer than should be accepted in finding the optimum network under the Metropolis-like approach. The expectation should be that more improvement can be made given a longer period of time. However, even after applying optimisations to the VBSSM process as described within Section 2.4.4, the time taken to complete each iteration was around 20 minutes, depending on server load. Since the computing cluster was a shared resource and time was limited it was not feasible to continue for longer periods of time. Unlike multi-modelling, in which iterations are not dependent on the previous one, Metropolis-like VBSSM must complete each iteration in series. As well as this, each Metropolis-like VBSSM run required exclusive use of five cores on the shared server, resulting in 40 of 64 cores being used for most of a week.

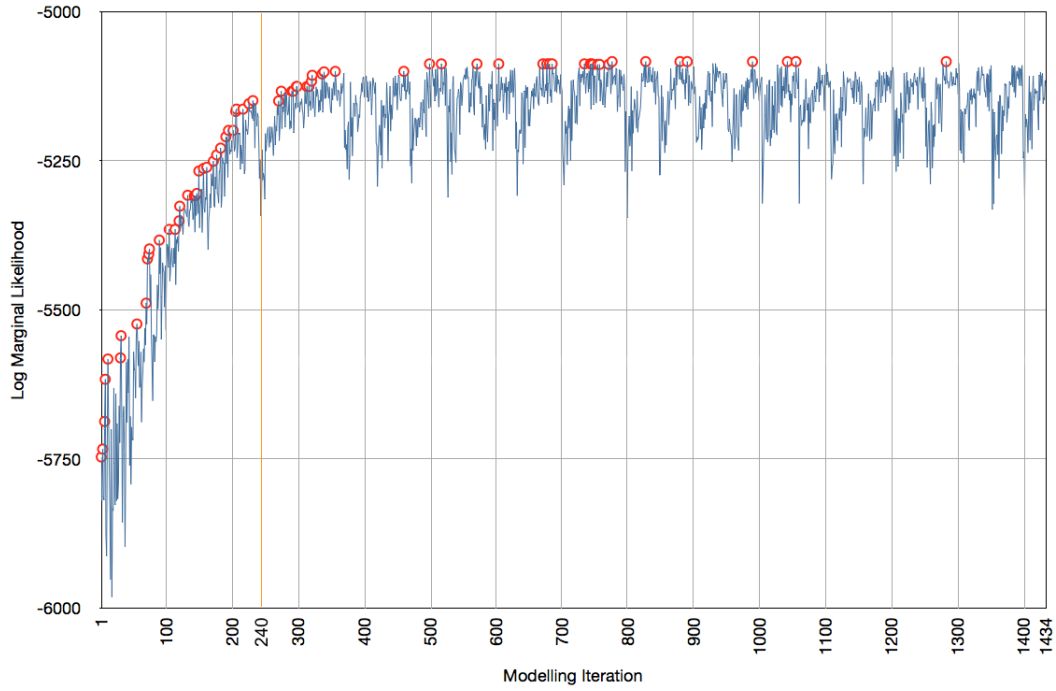


Figure 4.14 – Convergence of Metropolis-like VBSSM for HAT3

The plot shows the convergence of the log marginal likelihood for HAT3 during Metropolis-like VBSSM. The horizontal axis indicates iterations of the process whilst the vertical axis indicates the log marginal likelihood of the model. The blue line shows the log marginal likelihood after each iteration of the process and models which were accepted as better than the last best model are indicated by red circles. The orange vertical line shows the point at which the process was initially halted before, values to the right of that line showing progress in the 30 days after the process was resumed indefinitely. It can be seen that the log marginal likelihood was close to being maximised by the time the process was initially halted at model 240, but that continuing until 360 models were completed would have provided a yet further refined result.

As an assurance that the resulting models were sufficiently converged to be useful, two of the module centres were explored further by allowing them to continue exploring the gene pool to find more optimum gene sets. The Metropolis-like VBSSM was forced into an infinite loop by resetting the number of genes being swapped back to 5 each time that number was reduced below 1. After one month of processing in this way it was found that the models had already converged to at least 95% of the maximum as can be seen in Figure 4.14 where the original stopping point is marked by the vertical orange line.

Whilst researching the possible functions of each of the module centres, it was identified that HAP3A was part of a three protein complex of HAP2, HAP3 and HAP5 that binds to a CCAAT-box motif in yeast (Edwards *et al.*, 1998; Wenkel *et al.*, 2006). By looking back over the list of differentially expressed genes, it was found that HAP2B had also been differentially expressed during the experiment but that HAP5 had not. This inspired the addition of HAP2B as a putative module centre and a ninth Metropolis-like VBSSM run was performed to find if there were any strong interactions with other transcription factors. HAP2B was given a gene pool of 142 genes, selected in

Mutant Identifier	Gene Name	Upstream Interactions	Downstream Interactions
IM136	AP2-domain	1	0
IM137	STZ	2	0
IM147	TCP15	0	0
IM160	HAP3A	0	46
IM176	HAT3	2	4
IM188	ANAC092	3	64
IM193	Brother of FT	1	1
TP017	MYB59	1	0
—	HAP2B	2	56

Table 4.9 – Numbers of predicted upstream and downstream interactions for each module centre

Each of the nine module centres, including the later added HAP2B, provided a final model showing the interactions between the selected gene set. Although the module centre was always present in each model produced, they were not always identified as the centre of the model because other genes which became part of the final model gave stronger, more likely, interactions, resulting in little information about the module centre. Those interactions which were greater than 99.0% likelihood in at least one seed of the model are included above and it can be seen that three of the module centres show strong hub-like behaviour with the majority of the gene set downstream of them. Both HAP2B and HAP3A, known to be involved in a protein complex together (Edwards *et al.*, 1998), show similar numbers of downstream interactions.

the same manner as each of the other module centres, and a run of 142 models provided an optimum log marginal likelihood of -4216.97.

The best model for each of the nine runs was summarised by observing the interactions with the module centre only. Only those interactions which exceeded a 99.0% confidence level in at least one model were included in the interactions that were returned. This threshold was chosen as it allowed only the most robust interactions to be identified, and pruned the networks sufficiently to allow a reasonable number of genes to be identified for a few of the module centres. If this process were to be revisited, consideration of interactions at a lower confidence level could be included so that those module centres with very few interactions could be studied further. Interactions which were predicted to be from upstream genes were separated from those which were predicted to be towards downstream genes. Numbers of each type of interaction are shown in Table 4.9.

The results of this analysis allowed a final list of genes to be produced which was used to identify a final model for biological validation. Given that the only module centres to provide a reasonable number of downstream interactions were the three transcription factors ANAC092, HAP2B and HAP3A, the final list was based on predicted interactions with these alone. The downstream interactions between all three module centres provide a list of over 150 target genes which needed to be reduced to fewer than 88 for a model. The final list was composed of:

- The three module centres, ANAC092, HAP2B and HAP3A.

- Three genes, NYE1 (At4g22920), BFN1 (At1g11190) and SINA (At3g13672), known to be downstream of ANAC092 based on microarray data comparing an ANAC092 enhanced-expression line against the wild-type (Balazadeh *et al.*, 2008, 2010a).
- All upstream regulators of the three module centres.
- All downstream regulators of the three module centres if the interaction were present at greater than 99.0% confidence level in all five models (rather than just one model as included in Table 4.9).

This approach resulted in the production of a list of 75 genes which could be modelled together in order to find a putative biological network.

4.6 Identifying a Predicted Transcriptional Network Associated with Senescence

The 75 genes were used to create one last VBSSM model in an attempt to find the interactions not only for individual module centres, but also between module centres. 20 models were produced with different seeds to ensure that a comprehensive study of the interactions between genes would be performed, and only those interactions which demonstrated a confidence of at least 99.0% in at least 60% of those models were considered for reconstructing the network structure. This threshold was chosen as it allowed the removal of peripheral genes which were less confidently predicted, whilst retaining the structure at the core of the network. The identified interactions are shown in Figure 4.15 where it can be seen that two major hubs, SCL3 (At1g50420) and ANAC092, control the network. SCL3 acts through BLH6 (At2g35940) which then acts upon STZ which is also acted upon by ANAC092 directly. STZ then acts upon SINA and BFN1, which are known to be regulated downstream of ANAC092, providing some evidence for part of the structure of this network.

The network model predicts that ANAC092 enhances the expression of STZ whilst SCL3 indirectly represses the expression of STZ, presumably resulting in an opposing senescence phenotype for altered expression in either of these genes. This is backed up by research which found that SCL3 was markedly up-regulated by gibberellin inactive mutants which also showed delayed senescence (Curtis *et al.*, 2005) whilst it is known from the phenotype screens performed here that an enhanced-expression of ANAC092 causes premature senescence, suggesting an opposite phenotype like the model. Both ANAC092 and STZ have annotations in the TAIR database indicating their association with response to salt stress, which provides a functional link between these two genes.

Given the evidence showing the likely accuracy of the proposed final network model, the network shown is the one that was then validated by *in vivo* study in the following chapter.

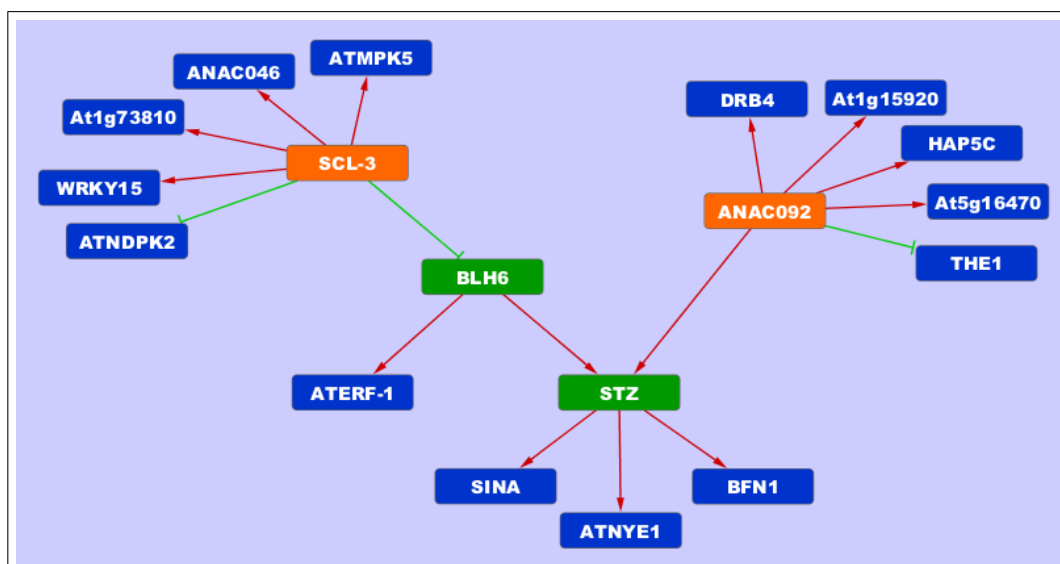


Figure 4.15 – The final VBSSM model showing interactions to be confirmed *in vivo*

A model produced from the list of 75 genes identified by Metropolis-like VBSSM shows those interactions which were of a likelihood greater than 99.0% in at least 60% of the 20 seeds. Note of course that only 18 genes are showing and that the other 57 genes did not have a predicted interaction that met the given threshold for the model. Only the most likely interactions are shown. The colours used for the nodes are reflective of those used in Figure 4.6 and indicate the classification given to each node. Orange indicates a hub gene, green indicates an intermediate gene and blue indicates a response gene. The log marginal likelihood of this network is -3475 which is very high for a list of 75 genes with the senescence data. It can be seen that both ANAC092 and SCL3 are the originating module centres which both pass through STZ to control the downstream effects of SINA and BFN1. Although it cannot be certain that the interactions shown are direct, the hidden state dimensionality was 8 for this model and so many of the interactions are predicted to be direct.

4.7 Conclusions

This chapter has introduced a number of methods which have been used in the analysis of gene expression data. SplineCluster has been used as a method of grouping genes into clusters of similar expression throughout the time course whilst BiNGO has provided a method for identifying which gene ontology terms are over-represented in each cluster. These allowed initial selections of genes to be modelled using VBSSM, but it was discovered that many of the interactions being identified by VBSSM must be false positives since the models could be drastically altered by making small changes to the selection of genes used for the modelling. However, these initial network models did allow the identification of gene roles in the network, including hub genes which are highly regulatory of other genes whilst not regulated by known genes themselves.

As it was hypothesised that hub genes are likely to be the cause of a cascade of interactions providing a senescence response, methods were developed to identify those genes. A form of iterative VBSSM modelling, multi-modelling, was used to select those genes, within a large pool of approximately 700, which robustly regulate multiple

genes in the pool. 118 were identified as being potential hub genes and 53 of these had available seed stock in reduced-expression lines for biological testing. By observing altered phenotypes during dark-induced senescence, seven of the lines were confirmed to have an effect on the senescence response, as well as one enhanced-expression line also tested.

Each of the eight genes were individually processed using another novel iterative VBSSM modelling method, Metropolis-like VBSSM, which allowed the identification of the most robust network of genes within a pool. A ninth gene, HAP2B, was also processed in this way as it was believed it might form a complex with HAP3A which was amongst those eight with a verified altered phenotypical response. By ensuring that the gene of interest from those nine that were processed was never removed from the gene set of its respective sequence of Metropolis-like VBSSM models, the most robust networks were guaranteed to be produced from a gene set including at least one gene of interest.

Three of the nine genes provided information about downstream interactions and these formed the basis for a set of 75 genes to provide a model for biological testing. A final VBSSM model produced from the set of 75 genes indicates a number of interesting and logical interactions which were then biologically tested as described in Chapter 5.

4.7.1 Efficiency of Modelling in Identifying Valid Hubs

The confirmation of 7 lines showing altered phenotype amongst a pool of 53 genes which were designated as putative hubs represents more than 13% accuracy in theoretical identification of hubs using the multi-modelling approach. Unpublished data from the Denby Group indicates that, prior to this approach, an average of 5% of lines tested for a phenotypical effect were biologically validated when alternative approaches were chosen for the selection of genes. This improvement in the selection of genes provides strong evidence that the multi-modelling approach saves time and money when compared with the equivalent time spent testing further lines for their phenotypical response during senescence.

Further testing within the Buchanan-Wollaston group is being performed to provide a baseline level of altered phenotypical responses amongst random selections of reduced-expression mutants. This will provide a better quantification of the improvements being made by the newly developed approaches presented in this chapter.

4.7.2 Future Development

During the prediction of local networks for module centres in Section 4.5 it was noticed that the most suitable network discovered for the majority of gene pools does not demonstrate the intended module centre as the primary hub in the network (see Table 4.9). This may be an indication that a “stronger” hub which can be more easily explained by the gene set dominates the network. Whilst this is not a poor result, it does not provide useful results when the intention is to find the interactions specific to

the module centre being studied, given that these are genes which have been verified to have an altered phenotype during dark-induced senescence.

It was seen and demonstrated in Section 4.2.3, that specific interactions in any given network can be severely disrupted by the replacement of other genes in the gene set even if they form no part of the interactions being disrupted. However, whilst modelling smaller gene sets, it should be true, assuming that true interactions exist between them, that the level of disruption is reduced because the data about each gene remains constant whilst the number of possible interactions between genes is substantially reduced. Christopher Penfold of The University of Warwick developed a Metropolis-Hastings version of VBSSM which allowed not only the control of the scale of variation between gene sets as demonstrated in Metropolis-like VBSSM, but also the control of gene set size during the process. In this way the number of genes removed from the best model in a given iteration does not have to equal the number of genes randomly added to the modelling set. If by changing the number of genes in the network, the log marginal likelihood can be improved, the better model can be accepted with fewer nodes.

Unfortunately, it was found that in every case where this was applied, the optimum model always resulted in only one gene being present: the gene of interest. This is the nature of the log marginal likelihood calculation which favours simpler/sparser models over more complicated/larger models. In order to improve on Metropolis-like VBSSM and/or Metropolis-Hastings VBSSM, it would seem that the acceptance criteria for each model would need to be changed. Suggestions might include scoring each model based on the strength of interactions involving the module centre of interest and/or the number of hidden states required to explain the observed data, these are both factors which are already included in calculating the log marginal likelihood. In this way, the focus remains with the module centre being tested and also in finding interactions which are direct rather than requiring an unobserved intermediate entity.

Chapter 5

Biological Validation of Theoretical Models

5.1 Chapter Summary

In this chapter, the predictions made in the previous chapter were validated *in vivo* so that any further findings could provide prior information to theoretical models and hence improve the outcome of any future models. The basis of these validations is the theoretical network model discovered and introduced in the previous chapter. The core of that model, and the interactions being tested here are shown in Figure 5.1.

Yeast-1-hybrid analysis was used to find which prey from a library of transcription factors would bind to upstream fragments of a number of genes from the model. Those genes tested were *STZ*, *AtNYE1*, *BFN1* and *BLH6*. For each of these genes, an analysis was performed by Laura Baxter using Analysis of Plant Promoter-Linked Elements (APPLES) (Baxter & Ott, unpublished) on their respective upstream region in the genome to find highly conserved regions across species and known transcription factor binding sites within those regions. Where these existed, an effort was made to ensure that those regions were cloned and tested.

The outcome of the yeast-1-hybrid analysis was a list of 14 transcription factors which were found to be able to bind to some of those genes being screened. Amongst those 14 were genes which were already thought to play a part in the senescence process, although the exact mechanism of doing so was not obvious from past research. None of the predicted interactions were confirmed by this analysis.

A second analysis was performed by harvesting senescing leaves from four reduced-expression lines (*anac092*, *stz* and two of *scl-3*) during dark-induced senescence and attempting to synchronise the levels of senescence across lines by using the same photography techniques as were used in phenotype screening in the previous chapter. Expression levels of the majority of the genome were compared between those mutants and the wild-type in a small microarray experiment consisting of 30 microarrays in a balanced design. Only two of the four lines tested demonstrated a reduction in expression

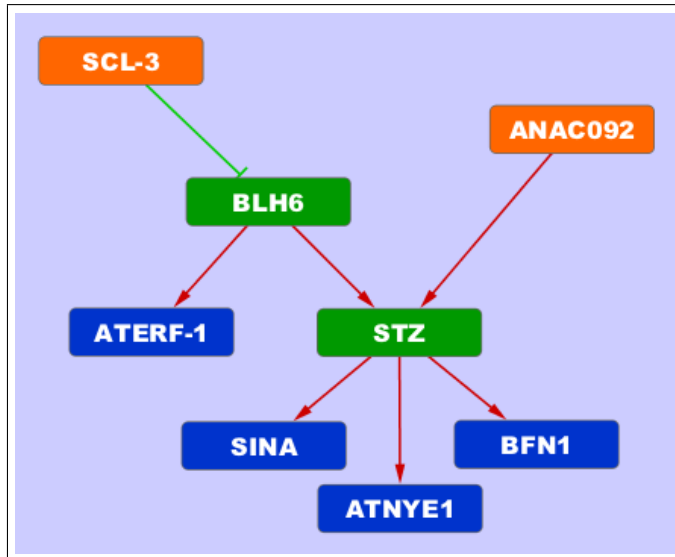


Figure 5.1 – Elements of the theoretical network model which have been tested *in vivo*

The core of the biological network discovered in Chapter 4 is shown and is composed of two hub gene, shown in orange, two intermediary genes, shown in green and four response genes, shown in blue. By application of a microarray experiment comparing mutant lines and application of yeast-1-hybrid to identify which genes are able to bind upstream of the target genes, it has been possible to strengthen and weaken individual interactions of the network to complete the first cycle of the systems biology approach to identifying true transcriptional networks during plant senescence.

(*anac092* and *stz*), but the findings showed that 58 genes were significantly differentially expressed in both lines. Since both lines were thought to be important to the senescence response, a full analysis of those genes was performed. Additionally, it was determined that *STZ* is not regulated by *ANAC092* since its expression level remained unchanged in the *anac092* mutant.

Amongst the 58 genes, three of them (*AtGDPD2*, *AOX1A* and an unknown gene *At2g18690*) demonstrated alternate regulation by *ANAC092* and *STZ* whilst all the others were similarly regulated by both transcription factors. Additionally, the three genes shown to be downstream of *ANAC092* in the theoretical network (*SINA*, *AtNYE1* and *BFN1*) were all shown to be altered in expression by the *anac092* mutant. Several stress response genes were identified as being downstream of *ANAC092* and *STZ* by this analysis, including *SRG* senescence response gene, *COR27* cold response, *ERD10* response to drought and *DIN11* response to darkness. These demonstrate that the network of responses leading to senescence, even at the top of the hierarchy, as is being observed here, is highly complex and diverse.

The findings of both analyses shown here were used to produce a new putative network which is shown at the end of this chapter. The regulations shown may not be direct due to the difficulty in isolating individual interactions, but should provide prior information in any future modelling to help train the model more accurately.

5.2 Yeast-1-Hybrid Analysis

Yeast-1-hybrid was used to identify transcription factors that are able to bind to the upstream region of genes of interest and to test the predictions of the model. This analysis was therefore capable of providing evidence for the identified interactions and also identify further possible direct interactions which were not predicted during network modelling. The genes selected for testing were *BLH6*, *STZ*, *SINA*, *AtNYE1* and *BFN1* as these all have a predicted upstream regulator and form a direct pathway between the predicted hubs and the response genes. As *STZ* is the most central gene amongst these, a more thorough screen was performed for this gene.

5.2.1 Identifying Conserved Regions Upstream of Genes

Before selecting which upstream regions to screen with yeast-1-hybrid, an analysis was performed by Laura Baxter of The University of Warwick using the unpublished AP-PLES software system developed by the Ott group at The University of Warwick. AP-PLES identifies conserved regions within the promoter of genes across species. This is achieved by comparing the promoter region of a chosen gene with its homologues in other plant species to identify sequences which are highly conserved between them. Species included in this analysis were *Arabidopsis thaliana*, *Vitis vinifera* (grape), *Populus trichocarpa* (poplar), *Oryza sativa japonica* (short-grain rice), *Oryza sativa indica* (long-grain rice), *Oryza glaberrima* (African rice), *Sorghum bicolor* (sorghum) and *Zea mays* (maize). During this analysis, the region which lay 1,000 bp before the transcriptional start site was processed for each of the five genes and only two conserved regions were identified. These were:

- ***STZ*** - a 138 bp conserved region starting 226 bp upstream of the transcriptional start site.
- ***BLH6*** - a 180 bp conserved region starting 867 bp upstream of the transcriptional start site.

Searching the TRANSFAC database (Wingender *et al.*, 1996) for known binding sites of transcription factors in these regions identifies two with a very high homology. The identified sites are shown in Table 5.1.

Neither of the transcription factors listed could be found in the genome of *Arabidopsis*, but they may have homologues with alternative names. Both genes obviously contain a conserved region in their promoter and this alone qualifies them as regions which should be studied by yeast-1-hybrid because it is likely that transcription factors other than those shown could bind to these conserved sequences.

5.2.2 Selecting Fragments to be Screened

For each of the genes, pairs of primers were designed to amplify regions of the upstream sequence. These fragments were to be used as the insert for plasmids used in the Y1H



Gene	Site Name	Identity	Sequence	Factors
<i>STZ</i>	XF12\$CONS		ACTTGTCACGCAACT	XF1/2
<i>BLH6</i>	TST1\$CONS_01		GGATTAAGATTA	POU3F1

Table 5.1 – Conserved transcription factor binding sites in regions to be cloned for yeast-1-hybrid

The sites shown are those which demonstrated the highest homology with the upstream sequences of genes to be cloned for yeast-1-hybrid. The identity shows the sequence which must be found to be a 100% match to the known binding sites of the transcription factors shown. The grey background to nucleotides shown in this sequence indicate the actual match with the upstream sequence of the fragment to be cloned.

Fragment Number	Gene	Length (bp)	Primer Pair	Location [†]
1	STZ	1000	11 + 12	1000 → 1
2	STZ	425	11 + 13	1000 → 576
3	STZ	383	14 + 15	777 → 395
4	STZ	435	16 + 17	603 → 169
5	STZ	424	18 + 12	424 → 1
6	BLH6	421	3 + 4	421 → 1
7	BFN1	392	1 + 2	392 → 1
8	AtNYE1	402	7 + 8	402 → 1
9	SINA	397	9 + 10	397 → 1
10	BLH6	472	5 + 6	966 → 495

[†] base pairs upstream of the transcriptional start site.

Table 5.2 – Length and position of fragments used as inserts for yeast-1-hybrid cloning

Information regarding the length and position of fragments to be screened by yeast-1-hybrid are shown. The primer pairs used to produce these inserts are identified by the same numbering used in Figure 5.3.

screen. The fragments and corresponding primers can be seen in Figure 5.2. For all but *STZ*, a fragment of approximately 400 bp directly upstream of the transcriptional start site was selected. As *STZ* was pivotal to the theoretical network model, a number of alternative overlapping fragments were produced within the 1,000 bp region upstream of the transcriptional start site. Within these fragment, one of them, fragment 5, contains the entire previously identified conserved region. As the conserved region of *BLH6* was not within the first 400 bp upstream of the transcriptional start site, a second fragment, fragment 10, was produced further upstream to accommodate the conserved region.

The longer fragment of *STZ* was not of a conventional size for Y1H, but was included as a comparison with the shorter overlapping fragments. The sequences of the primers used to generate these fragments are shown in Figure 5.3. Information about the fragment sizes and their locations relative to the transcriptional start site are shown in Table 5.2.

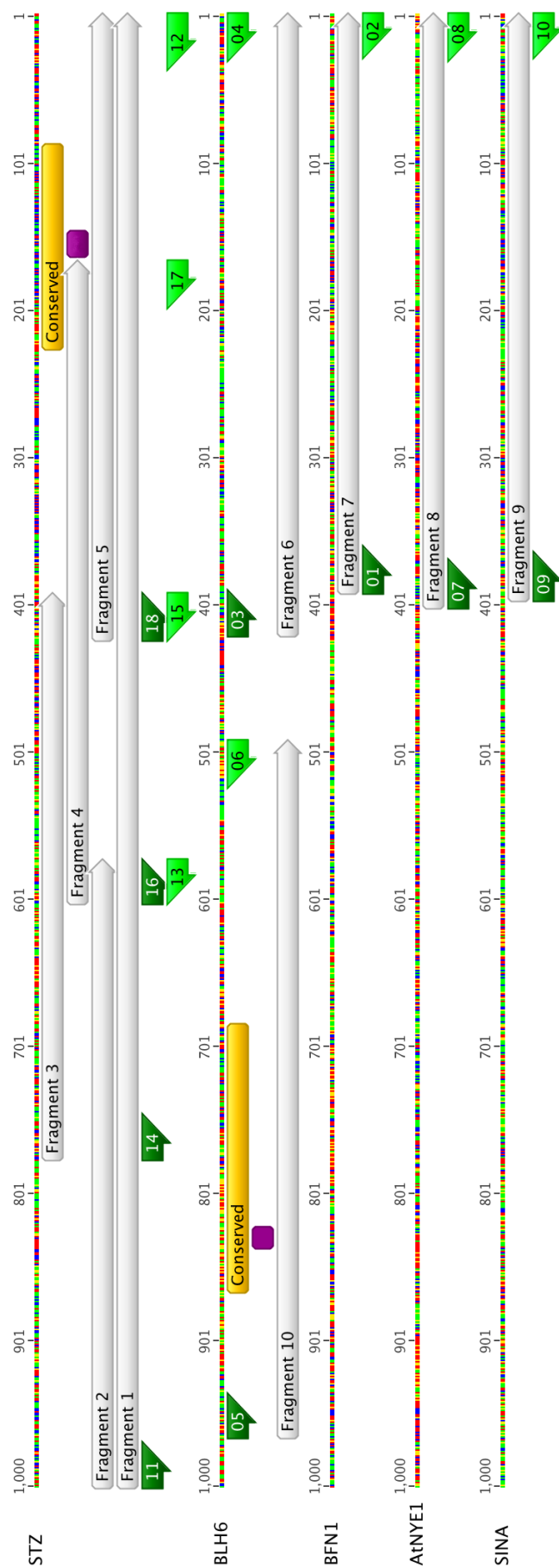


Figure 5.2 – Yeast-1-hybrid fragments within the 1,000 bp region upstream of transcriptional start sites

The sequences shown are of the 1,000 bp region upstream of the gene transcriptional start sites. 10 fragments were selected for screening in yeast-1-hybrid experiments. The position of the fragments is shown and the primers used to generate them are shown with green arrows at each end. Forward primers are shown in dark green whilst reverse primers are shown in light green. Where a conserved region was identified by the APPLS system this is marked in yellow whilst known transcription factor binding sites in these regions are marked by purple squares. Fragments 1, 5 and 10 all contain one of these known binding sites.

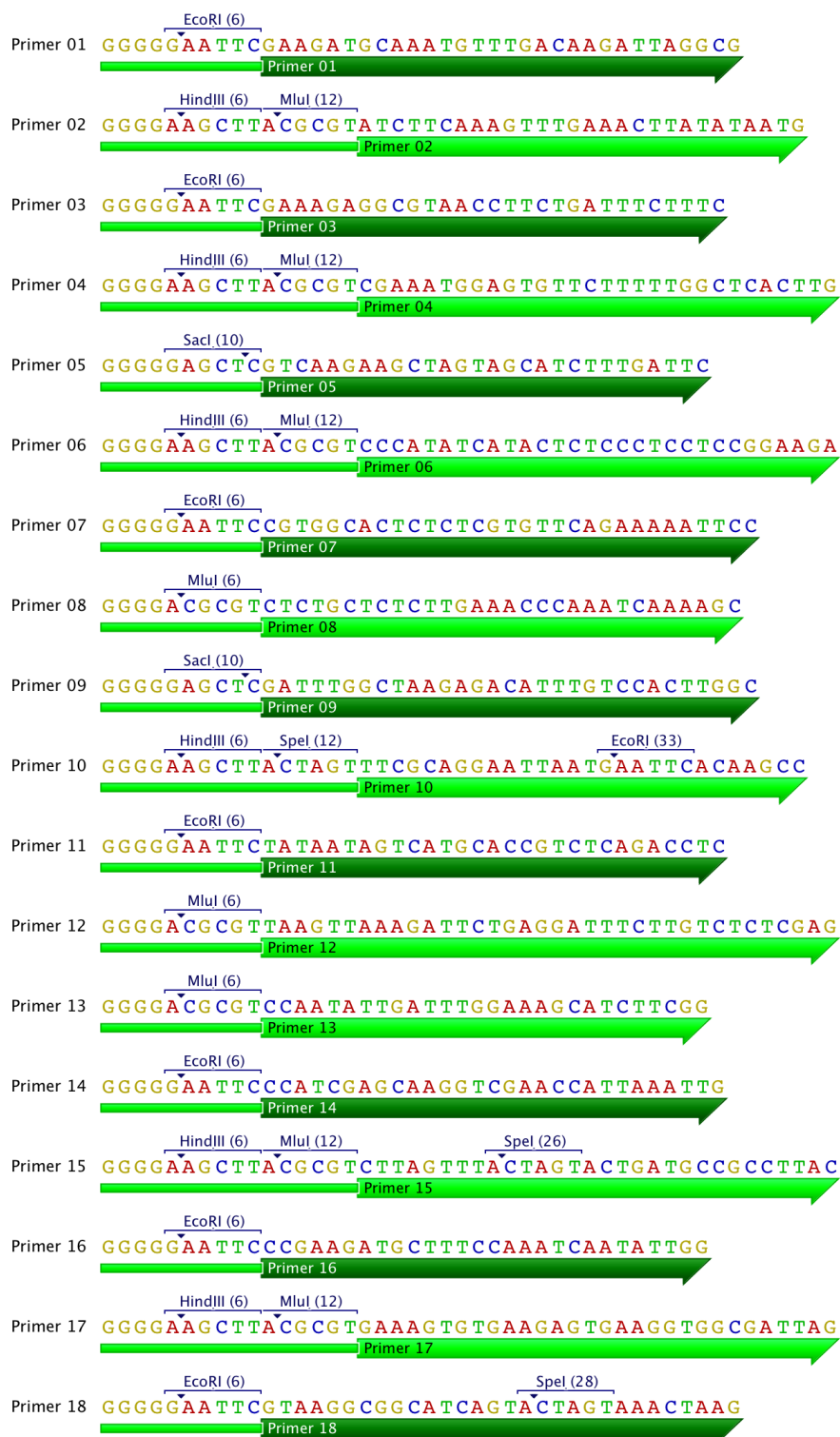


Figure 5.3 – Primer sequences used to amplify inserts for yeast-1-hybrid cloning

The primers shown are those used during amplification of upstream regions of target genes. In each, a poly-G tail begins the 5' end which is followed by one or more restriction sites. The ~30 bp region which follows is complementary to the upstream sequence of the gene and is an approximate 50:50 ratio of AT:CG where possible. Dark green indicates a forward primer whilst light green indicates a reverse primer. Restriction sites shown include EcoRI, MluI, HindIII, SacI and SpeI only. The numbering relates to those shown in Figure 5.2 and Table 5.2.

5.2.3 Amplifying Plasmid Inserts

Each fragment shown in Table 5.2 was amplified by PCR from a Col-4 template by following the methods described in Section 2.8.2. The annealing temperature of the primers was optimised after an initial attempt to amplify all the fragments at 60 °C. It was found that a clean product could be obtained by lowering the temperature to 55 °C for fragments 1 through 5 and by raising the temperature to 65 °C for fragments 6 through 10.

5.2.4 Inserting Fragments into Plasmid Vectors

After amplification and purification, the fragments were digested with restriction enzymes alongside the vectors which were digested with the same enzymes, following the methods described in Section 2.8.3. Fragments 1 through 8 were digested with EcoRI and MluI as these were to be inserted into the pHisLeu2 vector which was also digested with these enzymes. Fragments 9 and 10 were digested with SacI and HindIII as these were to be inserted into the pBlueScript KS+ vector. pBlueScript KS+ was to act as an intermediary step into the pHisLeu2 vector because the restriction sites required to insert into pHisLeu2 directly were present in the fragments themselves and so could not be used.

Once digested, the fragments and vector were cleaned to remove cleaved fragments as described in Section 2.8.3. An electrophoresis gel was used to confirm the concentrations of both vectors and inserts and this can be seen in Figure 5.4. Vector and insert pairs were ligated with one another using T4 DNA ligase in an overnight reaction as described in the legend of that figure.

5.2.5 Bacterial Transformation and Plasmid Cloning

Competent *E. coli* cells were transformed with plasmids containing inserts following the methods described in Section 2.8.4. Colonies were formed on all selection plates with the exception of inserts 9 and 10. Those colonies were then processed by colony PCR, following the methods described in Section 2.8.5, so that the size of the insert could be established. Larger amplified regions indicated that an insert was present. For all fragments except 1 and 5 which had only one positive colony each, at least two of the colonies demonstrated the presence of an insert. By selecting one or two colonies of each fragment for cloning, 14 colonies were cultured overnight in liquid media and the plasmids from those cultures extracted as described in Section 2.8.6.

Plasmids were BigDye (Applied Biosystems) sequenced from both ends of the insert region, following the methods of Section 2.7, which allowed any SNPs introduced by the cloning process to be identified. Sequencing results were processed by Geneious (Drummond *et al.*, 2010) to identify the similarity to the expected genomic sequence. In all cases, the sequence was a perfect identity match with the expected genomic sequence and so all were then used to transform yeast cultures.

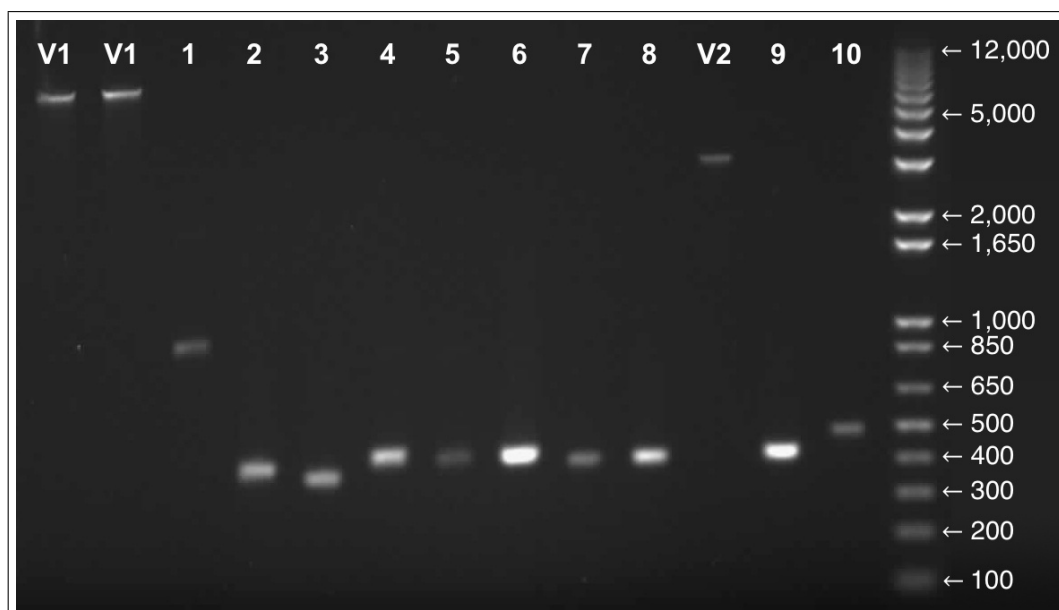


Figure 5.4 – Comparing digested vector and insert concentrations by electrophoresis gel

Vectors and fragments were digested and then compared for concentration using a $1/10$ dilution on an electrophoresis gel as shown. V1 represents two samples of digested pHisLeu2 which was compared with fragments 1 through 8. 2 μ L of vector was used in each ligation reaction with either 3 μ L of fragment 1, 5, or 7; 2 μ L of fragment 2, 3, 4 or 8; or 1 μ L of fragment 6. V2 represents the digested vector pBlueScript KS+ of which 3 μ L was ligated with 1 μ L of fragment 9 and 3 μ L of fragment 10.

Due to a shortage of time, those plasmids containing inserts 9 and 10 were abandoned at this stage. Although they might have provided valuable information towards the validation of the theoretical network model, 8 separately screened inserts consumed the remaining time available for the yeast-1-hybrid screen.

5.2.6 Screening Against the Transcription Factor Library

All 14 samples of amplified plasmids containing inserts of the correct sequence were used to transform yeast by the methods described in Section 2.8.8. Transformed cultures were plated on SD-Leu selective media and individual colonies restreaked onto further SD-Leu plates for storage. One restreaked colony per insert was screened by mating with the transcription factor library following the methods described in Section 2.8.9.

In all screens, after mating, all the mated cells were able to grow on media deficient in leucine and tryptophan as would be expected of cells containing both the pHisLeu2 and the pDEST22 plasmid. This indicated that mating between the bait and the transcription factor strains was successful. However, media deficient in leucine, tryptophan and histidine prevented the growth of most cells into new colonies. Those exceptional colonies which were able to grow must be able to synthesise histidine because of transcription of the *HIS3* gene, caused by an interaction between the transcription factor protein and the DNA sequence inserted into the pHisLeu2 plasmid. Colonies able

		Prey													
		ANAC038	ANAC102	CUC2	ILR3	PIF7	TCP1	TCP3	TCP4	TCP8	TCP14	TCP15	TCP16	TCP20	TCP23
Bait	Fragment 1 - STZ							•	•	•	••	••		••	
	Fragment 2 - STZ							••		•	••	•		•	
	Fragment 3 - STZ							••	•		•	••		••	
	Fragment 4 - STZ		•		•	••	•	••		•	••	••	•	••	
	Fragment 5 - STZ	•		•				••	•	•	••	•	•	••	•
	Fragment 6 - BLH6							•			••	•		••	•
	Fragment 7 - BFN1							••	•		••	••		•	
	Fragment 8 - AtNYE1			•				•	•		••				

Table 5.3 – Transcription factors proven able to bind specific DNA sequences and cause downstream transcription

Sequencing prey transcription factors able to bind to the DNA fragments that were used as bait identified 15 which might play a regulatory role to the bait genes. Two copies of the library were screened simultaneously and so one dot indicates a positive match in one copy of the library whilst two dots indicates a positive match in both copies of the library. Gene abbreviation ANAC stands for *Arabidopsis* Nitrogen Assimilation Control, whilst CUC stands for Cup-Shaped Cotyledon, ILR stands for IAA-Leucine Resistant, PIF stands for phytochrome-interacting factor and TCP is an abbreviation of the three genes which first formed the family, TB1, CYC and PCF1

to survive on the latter plate were therefore sequenced using the methods described in Section 2.8.9.5 to identify the transcription factor causing the *HIS3* gene transcription.

A full table of all the sequenced colonies can be found in Appendix L whilst a summary of the transcription factors thought to be able to bind to each of the DNA sequences and cause transcription of the downstream gene are shown in Table 5.3.

Interestingly, none of these results provided evidence for interactions which were identified by the previously applied theoretical network modelling methods of Chapter 4. This is despite the anticipated upstream transcription factors being present in the library. This may be because the theoretical model does not only predict direct interactions and that other elements, possibly identified here, are needed to make a complete network. Other factors are that the yeast cells only provide a ‘similar’ environment to that of *Arabidopsis* cells and some interactions may involve multiple elements, whereas yeast-1-hybrid only tests one at a time. However, there is still a lot to be learnt from these results which could allow a better selection of genes based around these bait. The rest of this section describes known information about the transcription factors identified here which complement the findings of this screen and offers proposed uses for these results.

More than half of the transcription factors identified by this analysis are from the TCP (TB1 CYC PCF1; named after the first members sharing the same conserved domain) family of transcription factors. This family of plant-specific proteins are known to contain a domain which is predicted to form a basic-Helix-Loop-Helix (bHLH) structure (Cubas *et al.*, 1999). It is thought that the structure of these helices allow an interaction with the main groove of the DNA such that the protein can interact with specific se-

quences. TCP proteins fall into two subclasses and, of the nine TCPs identified above, only TCP3 (At1g53230) and TCP4 (At3g15030) fall into Class II whilst all seven of the others fall into Class I. This is despite the subfamilies being approximately equal (13 members in Class I and 11 in Class II). Five of the class II transcription factors, including TCP3 and TCP4 are known to be regulated by the microRNA 319 (miR319a) and that these TCPs have also been linked with the biosynthesis of jasmonic acid and senescence in *Arabidopsis* (Schommer *et al.*, 2008). TCP3 and TCP4 are shown, here, to be capable of binding the upstream DNA of, and possibly cause regulation of, *STZ* (amongst others) which has been identified as an important gene in the senescence process (see Section 4.4). Therefore, it may be that if these were both included in a network model similar to that of Figure 5.1 an interaction could be found between them.

It is known that both TCP15 (At1g69690) and TCP20 (At3g27010) have similar, although not identical, DNA-binding preferences (Viola *et al.*, 2011). This is complemented by the results of the Y1H screen in which both of these TCPs were able to bind to sequences of *STZ*, *BLH6* and *BFN1*, but not *AtNYE1*. It is interesting that TCP15 shows such a strong ability to bind to the bait sequences screened here as it was one of the eight hubs analysed using Metropolis-like VBSSM (see Section 4.5), but that process was unable to identify any neighbouring transcriptional network nodes. This may be caused by the limitations of that method caused by other strong interactions in a given network as was briefly discussed in Section 4.7.2.

TCP20, also known as PCF1, has been found to cause transcription of *CYCB1* by binding to GCCCR elements in its promotor region (Li *et al.*, 2005), but these elements do not exist in the bait fragments screened here and so this suggests that other elements can be bound by this transcription factor. Other studies have shown that alteration of the expression of TCP20 can cause severe growth alterations (Hervé *et al.*, 2009), but these do not appear to be relevant to senescence and so it may be that TCP20, although capable of binding to the promotor region of *STZ*, is not causing senescence specific responses. Given that *STZ* is known to be part of the pathway of many stress responses (Sakamoto *et al.*, 2004), this would be plausible.

TCP1 (At1g67260), TCP8 (At1g58100) and TCP16 (At3g45150) were the only TCP transcription factors which showed specificity with only *STZ*. A comparison of the profiles of these genes during the natural senescence process with that of *STZ* can be made using Figures 5.5(f), (i), (l) and (o). It can be seen that *TCP8* was not observably altered in expression throughout but that *TCP16* has a slight peak at the fourth time point before falling again. *TCP1* demonstrates a dramatic drop in expression between the sixth and eighth time points before rising again. It is not easy to see, however, how any of these could be regulating the expression of *STZ* when the most significant change in expression for *STZ* is a rise between time points five and seven after which the expression remains high. This may be an example of a combined effect of these TCP genes with others not considered here, but it would seem much more likely that these genes are able to bind upstream of *STZ*, but do not actually regulate its expression

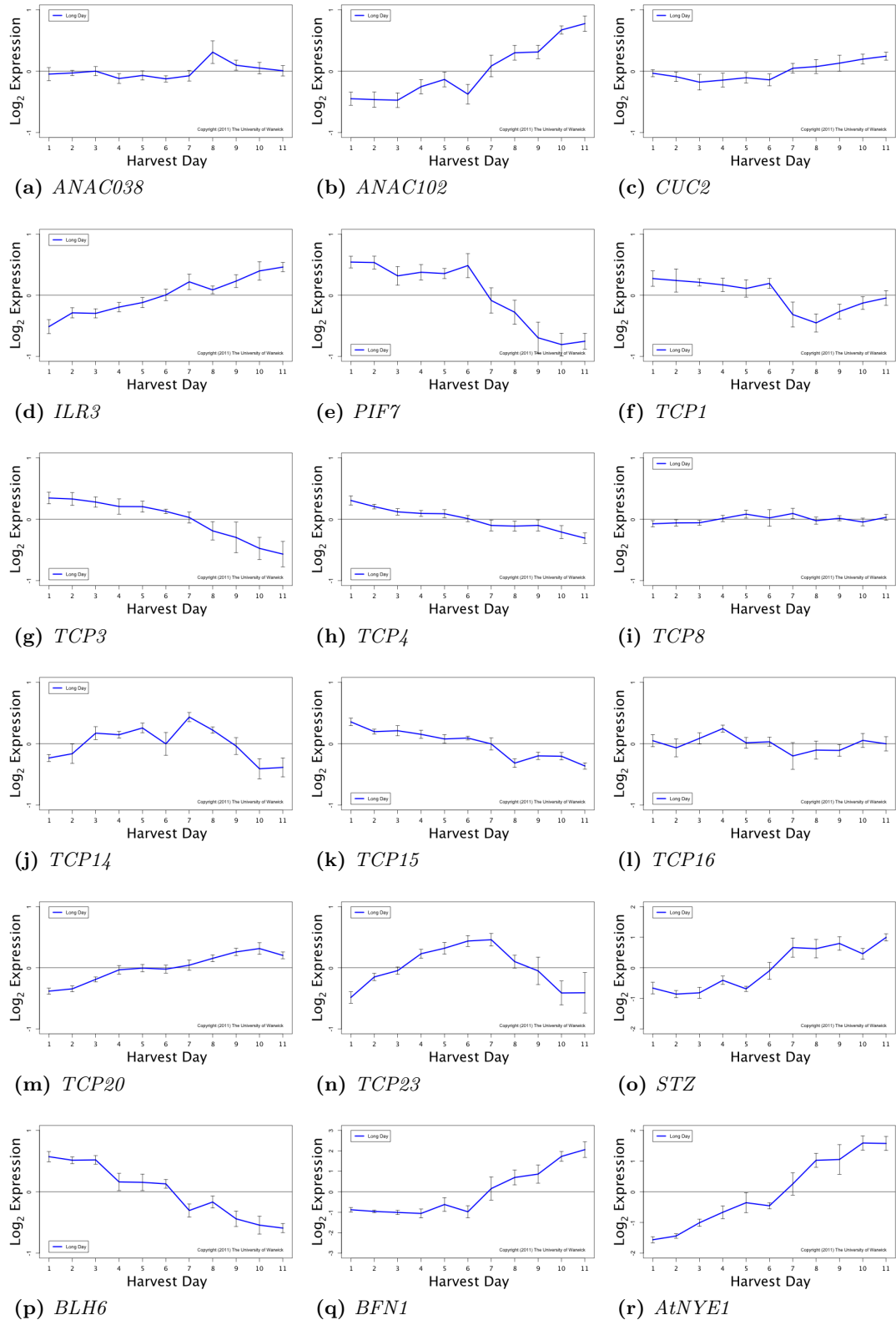


Figure 5.5 – Gene expression profiles for transcription factors identified by yeast-1-hybrid and their targets

Plots are of relative changes in gene expression over the full time course of the original microarray experiment on a log_2 scale. Plots (a) to (n) are of the transcription factors identified by yeast-1-hybrid, whilst plots (o) to (r) are of the genes capable of being bound by those transcription factors.

level.

Recent studies have shown that TCP1 negatively regulates the expression of *DWF4* which is part of the brassinosteroid synthesis pathway which also then regulates the expression of *TCP1* (Guo *et al.*, 2010). An excess of brassinosteroids are known to trigger leaf senescence (Clouse, 1996; Clouse & Sasse, 1998) and so it might be implicated that TCP1 therefore does play a role in leaf senescence although the results shown here would imply that this is not via the regulation of *STZ*. A further study of TCP1 has shown further evidence for its activity in plant development by demonstrating that the reduced expression of *TCP1* in *Arabidopsis* leads to abnormal expansion of cells (Koyama *et al.*, 2010). Very little is known about the function of TCP16 although research has shown that it is involved in the cell cycle process (Takeda *et al.*, 2006).

TCP23 (At1g35560) has been shown to be able to bind to the promoter region of both *STZ* and *BLH6*. Although the expression profile of *TCP23*, shown in Figure 5.5(n), does show an increased expression level in the lead up to senescence followed by a reduction in expression immediately after the peak change of senescence, there is very little known about the function of this transcription factor. The rising expression of *STZ* coupled with the similar but inverse falling expression of *BLH6* does suggest that they may be regulated by the same transcription factors. If TCP23 were one of these, it would therefore positive regulate *STZ* and negatively regulate *BLH6*.

TCP14 (At3g47620) has been shown to bind to every fragment screened here, which may suggest that its binding motif is very simple and that it must couple with other proteins to actually cause altered transcription in those promoters that it binds to. Its expression profile, shown in Figure 5.5(j), does show a similar response to that of *TCP23* described above with almost identical expression changes in the last part of the time course. All of the downstream targets screened here demonstrate a large alteration to their expression immediately after the 5th or 6th observation and this correlates with the sudden dip in expression of *TCP14*, which may be significant. It was shown by Tatematsu *et al.* (2008) that *tcp14* mutants were hypersensitive to exogenously applied abscisic acid which might indicate that TCP14 blocks the response to ABA, but that the reduction in expression of *TCP14* at observation six allows *STZ*, which has been associated with ABA response, to begin to rise in expression.

All of the remaining transcription factors are neither members of the TCP family nor do they show binding affinity for the promoter of any gene other than *STZ*, with the exception of CUC2 (At5g53950) which can also bind to the promoter region of *AtNYE1*. The expression levels of *CUC2* are shown in Figure 5.5(c) where it can be seen that it changes only very gradually throughout senescence and also demonstrates initial reduction in expression followed by a steady state and a final gradual increase in expression. It is hard to find an association between these gradual changes and the expression levels of *STZ* and *AtNYE1* since both of those genes only show steady periods and rising periods of expression. It is known that CUC2 is regulated by the presence of miR164 (Nikovics *et al.*, 2006) much like ANAC092 was shown to be by Kim

et al. (2009). It has also been shown that miR164 and CUC2 can directly cause altered serration during leaf development and alter the phyllotaxy of the leaves (Hasson *et al.*, 2011; Larue *et al.*, 2009a,b; Peaucelle *et al.*, 2007; Nikovics *et al.*, 2006) but despite obviously being related to leaf development, no research could be found to link CUC2 to senescence.

CUC2 is one of the initial transcription factors which founded the definition of the NAC domain (Ooka *et al.*, 2003). Amongst the other transcription factors found to be able to bind to the promoter region of *STZ* were ANAC038 (At2g24430) and ANAC102 (At5g63790). The expression profiles of these genes, found in Figures 5.5(a) and 5.5(b), indicate that although *ANAC038* does demonstrate a sudden change of expression, after a long period of no change, at the eighth observation, this is unlikely to be early enough to be the cause of any changes seen in the expression levels of *STZ*. However, the changes in the expression of *ANAC102* are a very close mimic of those found in *STZ* but would not be sufficient to entirely explain the interaction between them. *ANAC038* has been implicated as being part of pathways leading to altered sensitivity to gibberellic acid and auxin levels (Folta *et al.*, 2003) and therefore may not be relevant to senescence.

ANAC102 has been predicted, in a study of its closest homologue ATAF2 with which it shares 92% of its amino acid sequence, to be involved in the response to biotic stresses and the application of jasmonic acid and salicylic acid which are involved in wounding responses (Delessert *et al.*, 2005). Whilst the senescence response can be induced by biotic stresses and wounding, these stresses were not present in the natural senescence experiment and so may mean that ANAC102 is also responsive to other causes of senescence. A more recent study by Christianson *et al.* (2009) has shown that ANAC102 is likely to be the direct response to low-oxygen in roots during water-logging and that mutants of *ANAC102* do not show a phenotypical difference from the wild type unless stressed. Therefore *ANAC102* is most likely a stress response gene and may be one of the primary causes of the senescence response under those stresses but because of its homology with ATAF2, the natural senescence response is not altered in a single-gene mutation.

The specificity of NAC domain transcription factors to *STZ* in this analysis is interesting but explainable by looking for known NAC-domain binding site motifs in the sequence. Whilst the strict NAC binding motif, TTNCGTA, cannot be found in the sequence of *STZ*, a more general sequence, WKNCGTR (Olsen *et al.*, 2005), can be found directly in the centre of the conserved region of fragment 5. The identified sequence is TTGCGTG and so is very close to the strict NAC binding motif shown above. Furthermore, the binding motif for ANAC078, THNCKTGKG (Yabuta *et al.*, 2010), can be found closer to the transcription start site in fragment 5. These findings begin to explain the number of NAC domain transcription factors binding to the promoter of *STZ* and also suggests that others beyond those found by yeast-1-hybrid have the potential to bind to the promoter of *STZ*.

Of the two remaining transcription factors, *ILR3* (At5g54680) is shown to rise in expression steadily throughout the entire senescence process (Figure 5.5(d)) whilst the expression level of *PIF7* (At5g61270) is seen to remain steady for the first 6 observations before rapidly falling to a lower level over the remaining period (Figure 5.5(e)). Given that *STZ* begins to rise in expression after the 5th observation, before the point at which *PIF7* expression is lowered, it is difficult to identify how PIF7 could be regulating *STZ* expression, but if it were it would certainly be a repressor of *STZ*. However, if *ILR3* is a trigger for the increased expression of *STZ*, it can be seen that *ILR3* increases in expression monotonically and so, once at a given threshold, *STZ* would not lower in expression again as was observed in Figure 5.5(o).

ILR3 has been found to be capable of regulating several metal transporter genes which infers that it is involved in metal homeostasis in cells (Rampey *et al.*, 2006; Long *et al.*, 2010). However, this does not explain the relationship with *STZ* which is not yet known to be involved in metal homeostasis. As its name suggests, PIF7 (phytochrome interacting factor 7) is known to interact with phyB and its over-expression has been found to reduce sensitivity to red and far-red light in *Arabidopsis* (Leivar *et al.*, 2008). Whilst this does not provide evidence for any relationship with *STZ*, another study shows that PIF7 also negatively regulates the expression of DREB1B and DREB1C, drought response elements, but only when the plant was subject to light (Kidokoro *et al.*, 2009). These findings explain the relationship of PIF7 with light response proteins as well as possibly justifying the association with *STZ* which is known to respond to drought conditions and therefore may contain similar promoter elements to DREB1B and DREB1C.

The findings of the yeast-1-hybrid screen are informative, but cannot be used alone to determine the structure of the true interactions around genes in the theoretical model. Further experiments, presented throughout the remainder of this chapter, were also performed and the results combined with those shown here to develop a better understanding of probable interactions between senescence related genes.

5.3 Reduced-Expression Microarray Experiment

A microarray experiment was used to assess the dynamics of the model by comparing the expression of genes between the wild type and a series of reduced-expression mutant lines. By doing this, it was possible to identify which downstream genes exhibit altered expression in response to reduced expression of specific genes and hence which must be either directly or indirectly transcriptionally regulated by that gene. For this analysis, eleven SALK reduced-expression mutants were considered for the experiment including four mutants for *STZ*, four mutants for *SCL3*, two mutants for *BLH6* and one mutant for *ANAC092*. The SALK lines are shown in Table 5.4 where the IM line identifiers are listed and these will be used to refer to the lines in this chapter.

It should be noted that IM327 and IM328 were allocated SALK IDs which are

Gene	Line Identifier	SALK ID
<i>stz</i>	IM137	SALK_054092
<i>stz</i>	IM324	SALK_004580
<i>stz</i>	IM325	SALK_016589
<i>stz</i>	IM326	SALK_016596
<i>anac092</i>	IM188	SALK_090154
<i>scl3</i>	IM244	SALK_099576
<i>scl3</i>	IM323	SALK_002516
<i>scl3</i>	IM329	SALK_023428
<i>scl3</i>	IM330	SALK_107509
<i>blh6</i>	IM327	SALK_018707
<i>blh6</i>	IM328	SALK_018708

Table 5.4 – SALK lines considered for the model validation microarray experiment

Eleven reduced-expression mutant plant lines were acquired for a microarray experiment to identify the downstream effects of such a knockout when compared with the wild type. Lines IM137 and IM188 were already available from those screened in Chapter 4 whilst all others were ordered from the Nottingham *Arabidopsis* Stock Centre.

sequential. The SALK T-DNA FAQ website (http://signal.salk.edu/tdna_FAQs.html) states that sequentially numbered lines for insertion in the same gene can be because a contamination occurred between the samples during sequencing and hence only one should have the appropriate insertion.

5.3.1 Sample Harvesting

Given that the intention of the experiment was to identify changes in the dynamics of the senescence process given the reduced expression of a known gene, a novel method was devised to allow samples to be collected at the same relative stage of senescence rather than at the same absolute time by using the photograph analysis technique employed during quantitative phenotype screening described in Section 4.4.1. This prevented the observation of lines which exhibited delayed senescence before senescence had begun, or the observation of lines which exhibited early senescence after critical stages had completed, ensuring a more informative comparison with the wild type. Following the methods of Section 2.9.1, 24 biological replicates were sown, grown and harvested onto observation plates. Senescence was induced by light starvation and, on each day, those biological replicates which had reached their randomly assigned target senescence value, calculated as the normalised ratio of red and green in the leaf, were harvested for microarrays. For the twelve lines, including the wild type, four biological replicates were harvested at five senescence values: 0.7, 0.8, 0.9, 1.0 and 1.1, each replicate providing leaves 3, 4 and 5 for analysis. The harvesting senescence values were chosen because the value was 0.7 before the plants were subjected to light starvation and after 1.1 the leaf was so senescent that much of the mRNA in the leaf had already been degraded.

The target and actual senescence values of harvested leaves are shown in Table 5.5 where it can be seen that most harvests were very close to their intended level of senescence with the exception of a few replicates which are highlighted. As senescence reaches its peak change at a value of around 1.0, it becomes difficult to harvest at a pre-defined value because the leaf changes so rapidly from green to yellow in only a few hours.

Since the day of harvest was also recorded for each line and for each senescence value during the collection of samples, a comparison of the mean day of harvest for each line to the mean day of harvest for the wild type gives an approximation of the delay or advancement of senescence which has been caused by the reduced-expression in the mutant. The harvest day of the initial 0.7 replicates were not included in this mean as they were all harvested on the same day at the beginning of the experiment. The harvest days for each line were compared with the harvest days of Col-0 in a t-test which indicates whether the observed change in mean harvest day is significant. The final columns of Table 5.5 shows these statistics where it can be seen that IM326 shows significant difference from the wild type and IM188 shows very significant difference from the wild type.

blh6 mutants, IM327 and IM328, were allocated sequential SALK IDs with the possible explanation that one would contain the insert whilst the other would not. Although neither shows a significantly altered mean harvesting day, it would be likely that IM328, with a higher *p*-value of 0.7270 indicating greater similarity to the wild type, is the line without an insert.

5.3.2 Identifying Homozygous T-DNA Insertions

Using the same techniques as were used in Section 4.4.2, DNA was extracted from a mixed-leaf sample of the eleven lines, gathered from unneeded replicates during sample harvesting. These extractions were performed by following the methods of Section 2.6.1. The lines were screened by PCR amplification for a correctly located T-DNA insertion using primers specific to the gene and to the border of the T-DNA insert, following the methods of Section 2.6.2. The sequences of the primers were obtained directly from SALK and are shown in Figure 5.6.

Three PCR reactions per line were prepared to establish whether the insert was present and whether it was homozygous or heterozygous. Gene-specific primers were used to amplify genomic DNA; the same primers were used to amplify DNA extracted from the reduced-expression mutant and; the LBb1.3 T-DNA specific primer was used alongside the gene-specific reverse primer to amplify the reduced-expression mutant DNA. The results of these reactions are shown in Table 5.6 alongside conclusions which can be drawn from the results.

It can be seen that most of the lines screened were heterozygous. Therefore, if this generation of the plant was selfed and assuming that the reduced-expression was not detrimental to the growth of the plant, one quarter of the offspring would

	5'	Forward	3'	5'	Reverse	3'
IM137		TATTTTGTAAAGCGGCATCAG		AAGTCAAACCGAGGCTTCTTC		
IM324-326		GAGGGAATTTTCAAGGAAACG		CTAGTAGCGTGTTCCAACCTCCG		
IM244		TGATCATGAGCTGCATTCTTG		AAGGTCA TGGTGGTCACTGAG		
IM323		TTCCCTCTGTTCTTTAACCC		AGCGCAGTTCTTTCTCATGAG		
IM329		TGAAGATCATGGTGGAAAGCTC		CACGAGAAGCTTGAGAAATGG		
IM330		TGTAGAACTCTCTGGTCAATGG		TCCACGTCTCTGGTGGTTATC		
IM327-328		TGGTTACCTTTTATAGGCCTTTTG		TAAATA TCCACCCCGAAAAAG		
LBb1.3		ATTTTGCCGATTTTCGGAAC				

Figure 5.6 – Primers used to screen for T-DNA insertions in reduced-expression mutants

The primers shown were used to screen for T-DNA insertions in the reduced-expression mutants. Where a T-DNA insert exists in the correct location, the forward and reverse primers is able to amplify from the extracted DNA of the mutant whilst the LBb1.3 forward primer and the mutant reverse primer is able to to amplify from the same DNA.

Primers: Template:	GSP gDNA	GSP IM DNA	LBb1.3 IM DNA	Conclusions
IM137	✓	✗	✓	Homozygous
IM188	✓	✗	✓	Homozygous
IM244	✓	Faint ✓	✓	Heterozygous
IM323	✓	✓	✓	Heterozygous
IM324	✓	✓	✓	Heterozygous
IM325	✓	✓	✗	Wild Type
IM326	✓	✓	✓	Heterozygous
IM327	✓	✓	✓	Heterozygous
IM328	✓	✓	✗	Wild Type
IM329	✓	✓	✓	Heterozygous
IM330	✓	✗	✓	Homozygous

Table 5.6 – Results of a T-DNA insertion screen performed on eleven reduced-expression mutants

Where the primers are listed as GSP, these were the gene-specific primers for a given plant line. LBb1.3 indicates that the LBb1.3 primer was used in conjunction with the reverse GSP primer. In the template header, gDNA refers to genomic DNA whilst IM DNA refers to DNA extracted from the reduced-expression mutant. Where a tick is shown, this indicates that a product was able to amplify from the template whilst a cross indicates that no product could be amplified. For a line to be homozygous, the gene-specific primers must be able to generate a product in genomic DNA, but not in DNA extracted from the mutant, and there must be a product from the LBb1.3 primer combination. In the case of IM244, the product from the reaction amplifying mutant DNA with gene-specific primers was very feint, but it was identifiable and so taken as a positive result.

be homozygous, one quarter would be wild type and one half would be heterozygous. However, this would require at least two generations to obtain homozygous seed and was therefore unfeasible at this stage of the PhD. However, because the product was so faint in the amplification of IM244 DNA using gene specific primers, this line was retained for further analysis as RT-PCR and microarray analysis would be able to confirm whether there was a change in expression of the gene.

As predicted by the t-test performed on the mean harvest day data, IM328 is a wild type line, whilst IM327 contains the T-DNA insertion, even though it is heterozygous. However, the failure to find a homozygous line for reduced *BLH6* expression meant that, regrettably, this part of the predicted model could not be verified by microarray analysis in this study. The lines which were selected for continued study were, therefore, IM137 (*stz*), IM188 (*anac092*), IM244 (*scl3*) and IM330 (*scl3*).

5.3.3 Screening Expression Levels

It appeared appropriate to save leaf 4 for analysis by microarray since it was the leaf used to determine the stage of senescence whilst harvesting samples in Section 5.3.1. Therefore, leaf 5 was chosen for RT-PCR analysis to determine the expression level of the gene containing a T-DNA insert. The leaves screened by this process were those harvested at the third time point, when senescence had reached a value of 0.9. This time point was selected because, whilst senescence had visibly begun as indicated by the degradation of chlorophyll in the leaves, much of the leaf was still green and would provide sufficient mRNA for the screen.

Leaf 5 of all four replicates of each of the lines were combined for this screen by grinding with a pestle and mortar. Total RNA was extracted from the resulting tissue by the Trizol (Invitrogen) methods shown in Section 2.6.3.1. However, after achieving disappointing mRNA concentrations measured by NanoDrop, further total RNA was extracted from another aliquot of the same tissue samples using a *mirVana* RNA extraction kit (Ambion) as shown in Section 2.6.3.2. The concentrations and purities obtained from both extraction methods are shown in Table 5.7 whilst results from the Bioanalyzer 2100 Expert (Agilent) can be seen in Figure 5.7. IM188 was not screened because it had been previously shown to demonstrate a lowered expression by the same method in Section 4.4.3.

The results show that *mirVana* has provided an overall higher yield of total RNA, similar level of purity according to the 260/280 nm absorbance ratio and lower level of RNA degradation. For these reasons, the *mirVana* kit was used in all further RNA extractions.

Since neither of the genes being screened by RT-PCR (*STZ* or *SCL3*) contain an intron, conversion of RNA to cDNA by reverse transcriptase would have resulted in a DNA sequence identical to that of the genomic DNA. This would mean that PCR amplification would result in excessive product, regardless of the quantity of transcribed mRNA for the gene. Therefore, before transcribing RNA back to cDNA, all traces

	Trizol Extraction		<i>mir</i> Vana Extraction	
	ng/ μ L	260/280 nm	ng/ μ L	260/280 nm
IM137	17.8	2.20	20.7	2.30
IM244	16.5	2.15	25.5	2.14
IM330	18.7	2.26	32.3	2.04
Col-0	24.2	2.11	40.2	2.01

Table 5.7 – Total RNA concentrations obtained by extraction from harvested leaf tissue

Two methods of total RNA extraction were employed and compared for aliquots of leaf tissue from senescing leaves of reduced-expression lines. It was found that the *mir*Vana extraction kit produced higher yields of RNA than Trizol across all samples whilst retaining adequate purity.

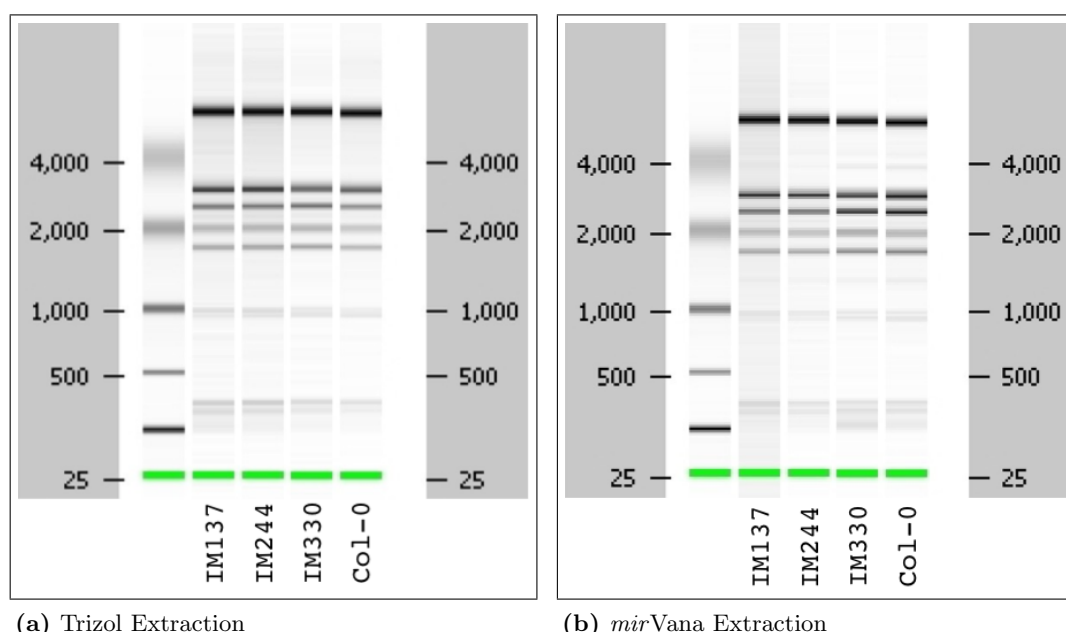


Figure 5.7 – Bioanalyser results for total RNA samples extracted using two alternative methods

It can be seen that, in those samples extracted using the Trizol method, a greater amount of degradation has occurred, as indicated by the pale grey smears along the length of the fragment sizes. In comparison, this grey smear is much less prominent in the samples extracted by the *mir*Vana method. The only exception is IM137 where the level of degradation is similar in both cases.

of DNA were removed by treatment with Turbo DNase (Ambion) by following the manufacturers instructions. The success of this treatment was determined by PCR of a small aliquot of each RNA sample with primers specific to an actin gene. No products were amplified during this reaction, indicating that DNA removal was a success.

RT-PCR analysis was performed in a two stage process, first using Superscript[™] II Reverse Transcriptase (Invitrogen) to produce cDNA from mRNA according to the methods described in Section 2.6.7.2. The cDNA could then be used to confirm the concentration of mRNA for specific genes by normal PCR amplification.

After the initial reverse transcription to produce cDNA for each of the four

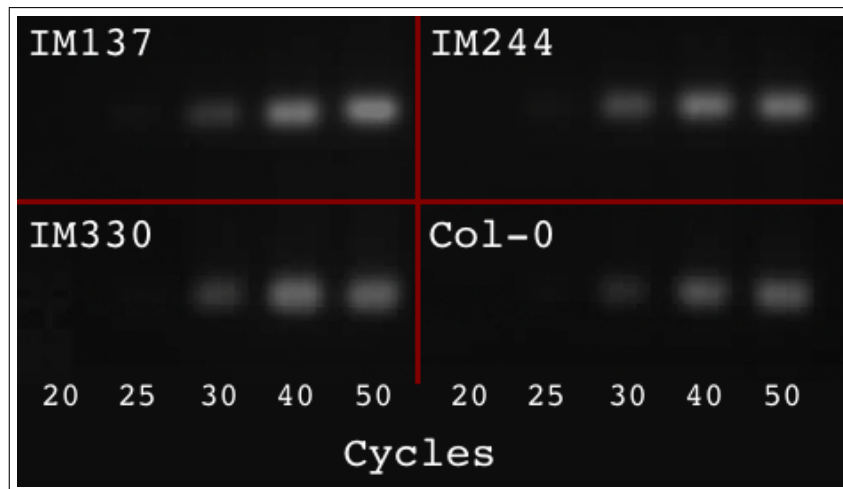


Figure 5.8 – Comparison of cDNA concentrations by PCR amplification of a highly expressed gene

By amplifying cDNA of a highly expressed gene in all samples, it is possible to determine the relative concentrations of each. Number of cycles refers to the number of PCR cycles before the aliquot shown was removed from the reaction. It can be seen, by comparing the band after 30 cycles, that all four samples were of a similar concentration with the possible exception of the wild type, Col-0, which is very slightly weaker. Given that these samples will be used to determine whether the mutants show reduced-expression of their respective gene relative to the wild type, this discrepancy should not present any problems in the case of a positive result.

RNA samples, the concentrations of the cDNA were compared by amplifying each with primers which were specific to an actin gene known to be highly expressed. By removing aliquots of the reaction mixture at 20, 25, 30, 40 and 50 cycles and comparing the brightness of bands on an agarose electrophoresis gel, the relative concentrations could be determined. The resulting gel image can be found in Figure 5.8 where it can be seen that the brightness of the bands across all three reduced-expression lines is very similar at all stages of the reaction, indicating similar starting concentrations. The wild type, Col-0, is very slightly lower in concentration, but this would not affect the outcome of a positive result where the mutant would show a reduced-expression in the respective gene.

Comparison of the transcriptional levels of specific genes was performed using four primer pairs: two pairs for IM137, one pair for IM244 and another pair for IM330. In the case of IM244 and IM330, the T-DNA insertion is in the 3' UTR and 5' UTR region respectively, hence why only a single pair of primers was necessary, whilst IM137 contained the T-DNA insert 278 bases downstream of the 5' UTR and hence primers were designed to amplify a region either side of this site. Figure 5.9 shows the results of all four comparisons with the wild type for each primer pair.

Conclusions which can be drawn from this result are that *STZ* is reduced in expression in IM137 as indicated by both pairs of primers. Whilst IM330 shows a similar result for *SCL3*, IM244 shows no change in expression. This may be because of the position of the T-DNA insert since in this line, the insert is in the 3' UTR

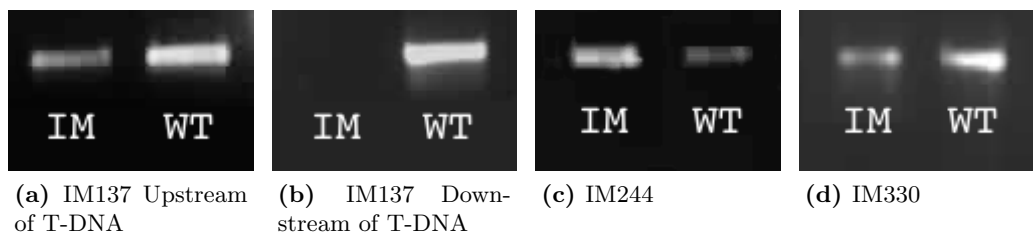


Figure 5.9 – Relative expression levels of genes containing T-DNA in reduced-expression lines

The results are representative of the expression levels of genes containing T-DNA inserts during RT-PCR amplification. IM refers to the insertion mutant line whilst WT refers to the wild type. Amplification for both bands were performed under the same conditions for the same number of cycles with the same volume of cDNA template. It can be seen that IM137, (a) and (b), demonstrates reduced expression of *STZ*, as determined by primers on both sides of the T-DNA insert. IM244 (c) demonstrates no change in expression of *SCL3*, since it is known that the wild type cDNA is slightly less concentrated (see Figure 5.8). The results for IM330 (d) indicate that *SCL3* has been reduced in expression in this line.

region which may result in either no transcriptional change, or partial transcription. In this case, RT-PCR would not accurately determine the abundance of functional *SCL3* protein in IM244.

5.3.4 Preparing Samples for Hybridisation

Although IM244 was suspected to be a heterozygous T-DNA insertion and also showed no apparent change in expression by RT-PCR, it was considered that it should still be included in the microarray experiment as the probes of the microarray would determine whether *SCL3* was reduced in expression or not. Therefore, four mutant lines, IM137, IM188, IM244 and IM330 were to be hybridised to CATMA microarrays alongside Col-0 to identify which genes exhibited altered expression and would therefore likely be downstream of the genes containing T-DNA insertions in the transcriptional network.

Leaf 4 of the harvested samples was selected to provide 4 biological replicates for each line since this leaf was the one which had been observed during the analysis of the progress of senescence whilst harvesting. Similarly to RT-PCR, leaves collected at the third time point when senescence had reached a value of 0.9 were selected for the microarrays. This therefore provided 5 lines and 4 biological replicates for each. Total RNA was extracted from these 20 leaves by individual grinding using a Dremel drill bit and then subsequent extraction by use of a *mirVana* RNA extraction kit (Ambion) following the methods described in Section 2.6.3.2.

Once extracted, the Bioanalyzer Expert 2100 (Agilent) was able to determine the condition of the extracted total RNA. The results of this analysis are shown in Figure 5.10 where it can be seen that all the samples show little degradation with the possible exception of the first replicate of IM244 which demonstrates one smaller fragment not found in the other samples. However, the concentration was good in this

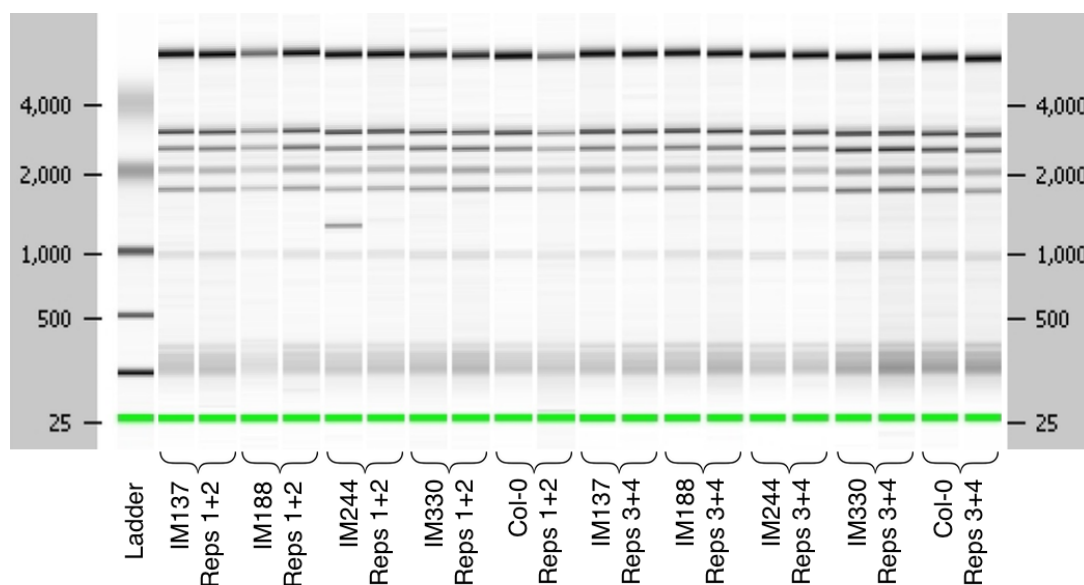


Figure 5.10 – Pseudo-electrophoresis gel produced during bioanalysis of total RNA samples for microarrays

The pseudo-gel shows the fragments of total RNA extracted from all 20 leaf samples to be analysed by microarray hybridisation. It can be seen that the samples exhibit little degradation with the exception of replicate 1 of IM244 which demonstrates a single smaller fragment than the other samples. It can also be seen that the lower concentrations of replicate 1 of IM188 and replicate 2 of Col-0 are visible as duller bands than the other samples.

sample and so it was not repeated.

The samples were processed using an Ambion MessageAmp II kit according to the methods shown in Section 2.9.2 to provide high concentrations of aRNA for labelling. After these steps, the samples were placed on the NanoDrop to determine the concentration of aRNA. Some of the samples failed to amplify sufficiently and were reattempted. After a second attempt, some insufficiently amplified samples still remained. As at least three biological replicates for each line had provided sufficient aRNA, a decision was made to reduce each line to three biological replicates, and the NanoDrop results for these are shown in Table 5.8. For each sample, an aRNA concentration of at least 400 ng/ μ L was required to be sufficient for fluorescent dye labelling and it can be seen that each of the samples exceeds this threshold and that each demonstrates a very high purity in the 260/280 nm ratio. The sample numbers shown in the final column are used throughout the rest of the experiment and particularly in the hybridisation design.

To confirm that the aRNA samples demonstrate a range of alternative fragment sizes, the Agilent Bioanalyzer Expert 2100 was used to analyse each sample. The pseudo-electrophoresis gel for these fragment sizes for each of the 15 samples can be found in Figure 5.11 where it can be seen that each sample shows a smooth gradient of change between larger and smaller fragments. A possible explanation for the diminishing mean fragment size towards the right of the gel may be that the samples were degrading during

Line & Rep	Concentration (ng/ μ L)	260/280 nm	Sample Number
IM137 - 2	616.3	2.19	1
IM137 - 3	665.0	2.19	2
IM137 - 4	1011.9	2.23	3
IM188 - 2	435.6	2.22	4
IM188 - 3	1313.6	2.21	5
IM188 - 4	1188.9	2.23	6
IM244 - 1	851.8	2.20	7
IM244 - 3	1526.6	2.24	8
IM244 - 4	1796.4	2.21	9
IM330 - 1	435.1	2.24	10
IM330 - 3	1136.7	2.23	11
IM330 - 4	618.7	2.24	12
Col-0 - 1	904.3	2.22	13
Col-0 - 3	594.2	2.26	14
Col-0 - 4	741.8	2.29	15

Table 5.8 – Concentrations of aRNA samples to be labelled with fluorescent dyes determined by NanoDrop

The concentrations of the 15 samples of aRNA to be labelled for microarray hybridisation are shown. The replicate number in the first column relates to the replicates as shown in Figure 5.10 whilst the sample number in the final column indicates the sample number as used in all further references to samples. As all samples were required to have a concentration of 400 ng/ μ L or more for labelling with fluorescent dyes, all are sufficiently concentrated to continue and all have a very high purity as indicated by the 260/280 nm ratios.

the time in which the bioanalyzer was processing prior samples. However, all samples are of sufficient quality for microarray analysis and could be labelled with fluorescent dye.

Fluorescent dye labelling was performed according to the methods of Section 2.9.3. Each aRNA sample was labelled in two separate reactions. One for labelling with Cy3 dye and one for labelling with Cy5 dye. Each labelled reaction had to be sufficient for two hybridisations since there were four technical replicates of each sample, those being equally split between the dyes so that dye-swapping had occurred. Where initial attempts to label samples provided fewer than 30 pmol of labelled sample per hybridisation, further attempts to label were performed until as much labelled sample as possible was available. Some samples were unable to reach a total of 60 pmols to achieve two hybridisations, but both time and budget did not allow repetition of aRNA synthesis and so all samples were used and split equally, where necessary, to maximise the concentration of dye in each hybridisation.

Microarrays were hybridised according to the design shown in Section 2.9.4. Hybridisation occurred overnight and microarrays were scanned the following day by following the methods of Section 2.9.7 during which time they were also analysed to produce numeric data describing the intensity of individual spots in each channel.

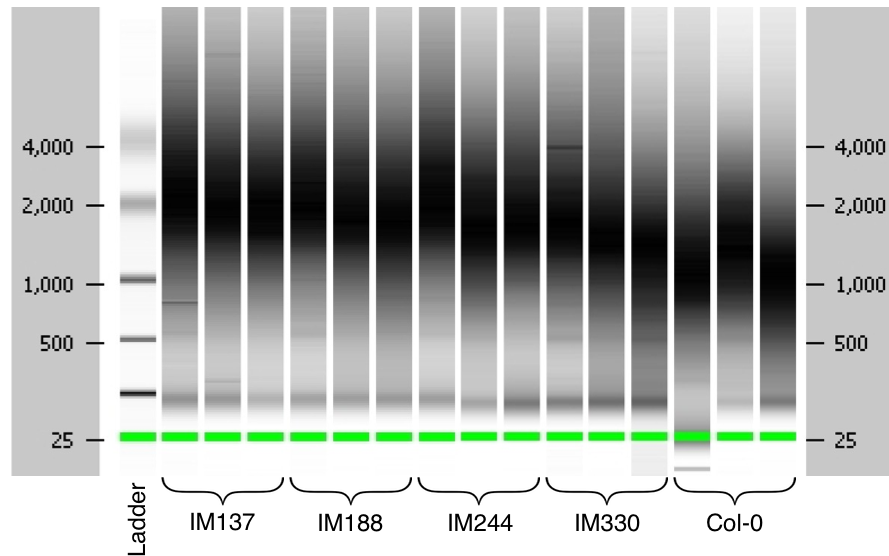


Figure 5.11 – Pseudo-electrophoresis gel produced during bioanalysis of aRNA samples for microarrays

The pseudo-gel shows that, after production of aRNA, all samples show a clean band of various sizes as would be expected. The slight reduction in mean size shown here in samples to the right of the gel can be attributed to degradation that may have occurred during the time that the bioanalyzer was analysing prior samples. Each of the samples shown can be labelled with fluorescent dyes.

5.3.5 Microarray Data Analysis

Microarray data was analysed by using the same MAANOVA package as was originally developed and used in Chapter 3. In the same way as had been used on the original senescence experiment, background corrected data was selected because this allowed the removal of background noise present on the microarray. The corrective procedures applied to the data were in the same order as that of the original senescence experiment analysis. It was noticed and expected that those samples which were labelled with lower dye concentrations provided a weaker signal overall, but because of the analysis by MAANOVA which does not use ratios of raw data during model fitting and is able to pool biological replicate information when establishing the effects for each line, the imbalance would not present a problem.

Because four technical replicates had been implemented in the microarray experimental design, it was possible to use the new method of analysis developed previously, **techrepcheck**, to test whether any technical replicates are outlying on a probe-by-probe basis. As was seen in the original senescence microarray experiment, a number of probes on each microarray demonstrated statistically significant between technical replicate variability, but as was explained during that original analysis, high levels of technical replicate variability will always exist for some probes in high-throughput technology such as microarrays and the benefits of reducing this must outweigh the costs and time required to do so. In the case of the microarrays being analysed here, the probes exhibiting this characteristic were not isolated to any particular region of the

microarrays and so it was concluded that no major issues existed prior to expression data modelling.

An ANOVA model was fitted to the observed microarray expression data using the model definition:

$$\text{Expression} \sim \text{Dye} + \text{Array} + \text{Line}/\text{BioRep}$$

in which both **Dye** and **Array** were specified as being random terms. It was then possible to analyse the expression data of each line and produce t-tests using the **matest** function of MAANOVA with a contrast matrix of the levels within the **Line** term above. The contrast matrix ensured that the only comparisons made were between the expression levels of the wild type and each of the mutant lines. To improve accuracy, the test was performed with a computationally expensive 1,000 permutations to refine the p -values. Returned p -values were adjusted for multiple testing by using the **stepdown** method previously applied to the F-tests of the original senescence experiment data. Genes showing statistically significant ($p \leq 0.05$) altered expression after correction in each mutant line were extracted and modelled expression data used to identify whether the gene is more highly expressed than wild type or vice versa.

Figure 5.12 illustrates the number of genes which were significantly altered in expression in each plant line when compared with the expression of the Col-0 wild type. It can be seen, by summing the number of genes in each lozenge, that there were 199 genes associated with IM137, 414 associated with IM188, 138 associated with IM244 and 97 associated with IM330. Of those genes, 7 were associated with all lines and may represent those genes which are very low in the hierarchy of the network.

Before drawing any conclusions from these figures, it was important to assess whether the lines were demonstrating reduced expression of the expected gene and so, for each line, the target gene's p -value for differential expression was assessed. These figures are shown in Table 5.9 where it can be seen that IM188 does show a very significant reduction in expression of *ANAC092* and that IM244 shows a borderline-significant reduction in expression of *SCL3*. Each of IM137 and IM330 showed no significant change in expression of their respective genes.

Although this might be considered a poor result, it may be partly explained by the position of the CATMA probe binding location, as shown in Figure 5.13. Here it can be seen that, in the case of *STZ* (IM137) and *ANAC092* (IM188), the binding site lies upstream of the T-DNA insertion. The T-DNA insertion, being in the coding region of the gene, will almost certainly prevent complete transcription and/or translation of the gene in this location, but may result in a partial and stable mRNA fragment being transcribed and therefore being detected by a CATMA probe upstream of the T-DNA insertion. If this is the case, this might explain why non-significant changes in the expression of *STZ* can be found in IM137. RT-PCR in Section 5.3.3 proved that the expression of *STZ* and *ANAC092* is definitely lowered in IM137 and IM188 respectively.

In the case of *SCL3*, the lowered expression observed by the CATMA probe in

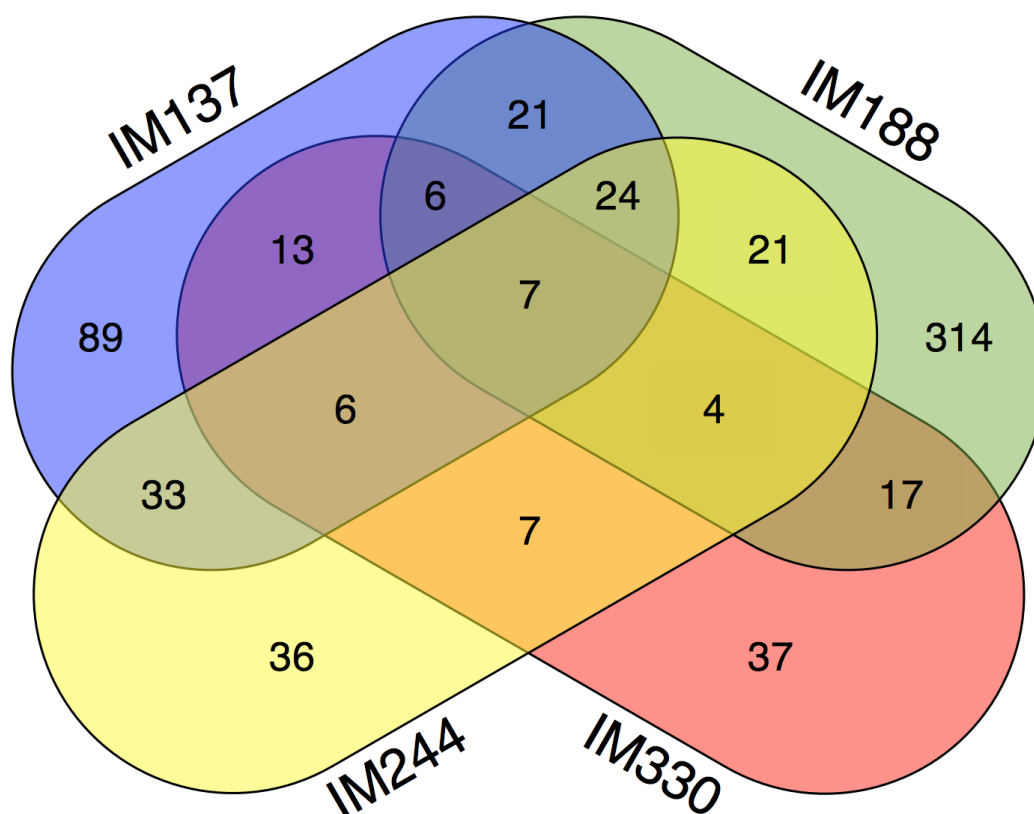


Figure 5.12 – A 4-way Venn diagram showing numbers of genes with altered expression for each line

Those genes which were significantly differentially expressed for each line when compared with the Col-0 wild type were analysed to find which genes could be associated with more than one line. The numbers shown indicate association with the four lines where it can be seen that 89, 314, 36 and 37 genes were unique to lines IM137, IM188, IM244 and IM330 respectively.

Line	Gene	Col-0 Expression	Mutant Expression	Corrected <i>p</i> -value
IM137	<i>stz</i>	8.673	8.285	0.9736
IM188	<i>anac092</i>	13.280	12.093	0.0000
IM244	<i>scl3</i>	13.678	13.201	0.0505
IM330	<i>scl3</i>	13.678	13.561	0.9812

Table 5.9 – A comparison of expression levels for reduced-expression genes in T-DNA insertion mutants

Each of the reduced-expression mutants were compared with the Col-0 wild type for the reported expression of the gene containing the T-DNA insertion. It was expected that these results would show highly reduced expression of that gene in the mutant line, but, as can be seen, only IM188 showed significantly reduced expression of *ANAC092*. IM244 was very close to showing significant reduction in expression of *SCL3* after correction of *p*-values but IM137 and IM330 did not show a significant change in the expression of *STZ* and *SCL3* respectively.

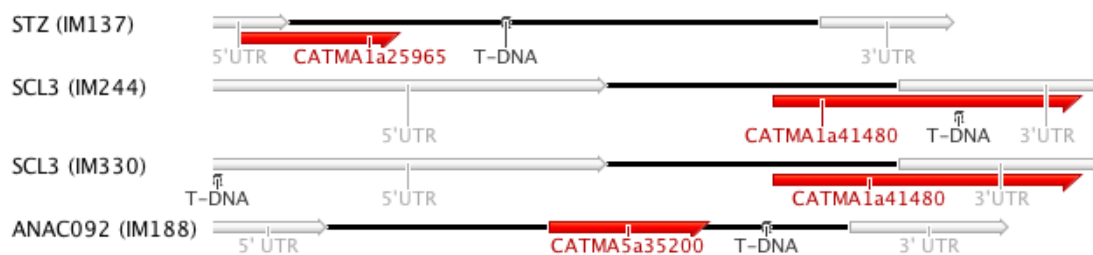


Figure 5.13 – Binding sites for CATMA probes specific to reduced-expression genes

Sequences lengths shown are not to scale. The figure shows where the CATMA probes used to identify the expression levels of reduced-expression genes bind in relation to the T-DNA insertion used to disrupt the transcription and/or translation of the gene. White arrows represent untranslated regions of the gene whilst red arrows represent the CATMA binding sites and the small black blocks indicate the expected insertion point of the transfer-DNA sequence. Because CATMA probes bind to mRNA before translation has occurred, it could be that partially transcribed sequences that maintain integrity can be shown to have high expression levels, even though the gene product may not be functional. This is a particular risk when the T-DNA sequence lies after (to the right/3'-end of) the binding site of the probe as can be seen in IM137 and IM188. The CATMA probe for *SCL3* binds to the right of the T-DNA sequence in IM330 and so should give a true impression of the abundance of gene product. The same CATMA probe straddles the T-DNA sequence in IM244 and so may give artificially reduced expression levels by only partial binding of the probe to the cDNA sequence.

IM244 may be because the probe is not able to completely bind to the sequence and, instead, only partially binds either upstream or downstream of the T-DNA insert. In the case of IM330, it should be possible for the CATMA probe to accurately determine the expression of *SCL3* and, therefore, because no significant change from the wild type is observed on the microarrays, this line should be disregarded as it would seem the T-DNA insertion is too far upstream of the coding region and having little to no effect on the expression of *SCL3*. Due to the previous uncertainty of IM244 as a homozygous line and the negative result from RT-PCR as well as the uncertainty that the CATMA probe is accurately determining the expression of *SCL3* with the position of the T-DNA insert as it is, results from this line can not be trusted and will also be disregarded.

Concentrating on the results of IM137 and IM188, the Venn diagram can be reconstructed to show only genes significantly differentially expressed in those two lines. This diagram can be found in Figure 5.14.

Despite *STZ* being downstream of *ANAC092* in the putative network model, no significant altered expression of *STZ* can be found in the microarray data for IM188, which might mean that the expression of *STZ* is influenced by multiple regulators and the reduction in expression of *ANAC092* can be compensated for by one of these. Knowing that *STZ* is not significantly altered in expression in IM188 allows the conclusion that the 58 genes shown in the centre of Figure 5.14 are likely to be independently regulated by *ANAC092* and *STZ*. The other two figures in the diagram represent those genes which are uniquely downstream of the respective mutated gene.

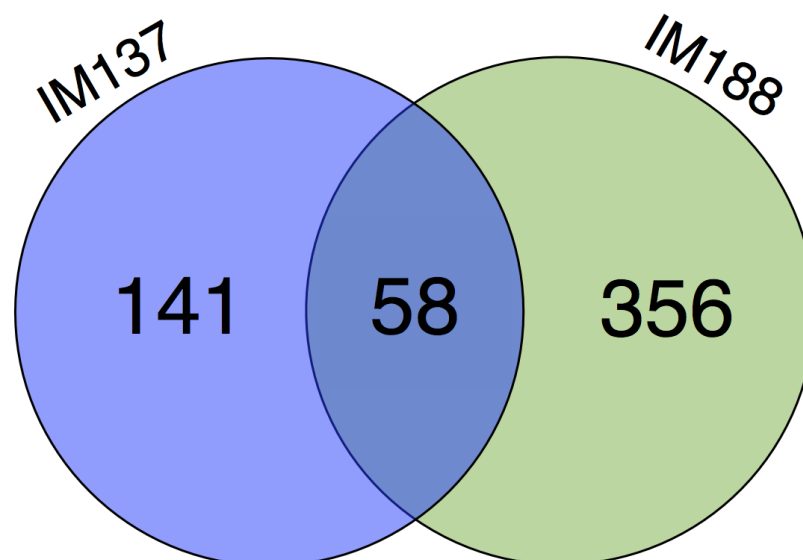


Figure 5.14 – A 2-way Venn diagram showing numbers of genes with altered expression for IM137 and IM188 only

Since IM224 and IM330 provide doubtful results regarding the reduced-expression of *SCL3*, they were removed from the Venn diagram to provide the one shown here.

Of particular interest is the list of genes regulated only by *ANAC092* as these can be cross referenced with results obtained about transcription factors able to bind to the sequence of *STZ* during yeast-1-hybrid. It was previously identified that potential regulators of *STZ* include *ANAC038*, *ANAC102*, *CUC2*, *ILR3*, and *PIF7* amongst a number of TCP transcription factors (see Table 5.3). However, none of the genes encoding those transcription factors are shown to exhibit altered-expression during reduced-expression of *ANAC092*. When accompanied with the information that IM188 does not show reduced expression of *STZ* either, this provides substantial evidence that *STZ* may not be downstream of *ANAC092* at all, although their functions are undoubtedly similar if they affect the expression of a common subset of genes and also have similar senescence phenotypes.

The putative network model shown in Figure 5.1 suggests that three genes are regulated by *STZ*, *SINA* (At3g13672), *BFN1* (At1g11190) and *AtNYE1* (At4g22920). *SINA* was represented by two probes on the microarrays. Whilst one probe found it not to be differentially expressed in either of the lines, the other probe found significant down-regulation of *SINA* in both the *stz* and *anac092* mutants before, but not after multiple testing correction. *BFN1* demonstrated significantly reduced expression in the *anac092* mutant, but no significant change in the *stz* mutant although this was predicted by the putative network model. *AtNYE1* was significantly reduced in expression in the *anac092* mutant prior to multiple testing correction, but not after correction. The *stz* mutant did not demonstrate altered expression of *AtNYE1*.

Past research has suggested that *STZ* may be a transcriptional repressor of *LOX3* (At1g17420), a jasmonic acid biosynthesis gene which may be involved in the senescence response (Pauwels & Goossens, 2008). However, whilst the *stz* mutant does show a

Description	# of Genes	<i>p</i> -value	Corrected <i>p</i> -value
Response to cadmium ion	8	3.24×10^{-7}	9.21×10^{-5}
Response to metal ion	8	1.52×10^{-6}	2.16×10^{-4}
Response to inorganic substance	8	7.48×10^{-6}	7.08×10^{-4}
Response to abiotic stimulus	12	1.49×10^{-5}	8.75×10^{-4}
Response to stimulus	20	1.54×10^{-5}	8.75×10^{-4}
Response to temperature stimulus	7	2.33×10^{-5}	1.10×10^{-3}
Response to stress	14	5.83×10^{-5}	2.37×10^{-3}
Response to chemical stimulus	12	5.20×10^{-4}	1.85×10^{-2}
Glycerol metabolic process	2	1.30×10^{-3}	3.68×10^{-2}
Folic acid and derivative metabolic process	2	1.30×10^{-3}	3.68×10^{-2}
Alditol metabolic process	2	1.42×10^{-3}	3.68×10^{-2}
Reactive oxygen species metabolic process	2	2.11×10^{-3}	5.00×10^{-2}

Table 5.10 – Results of a BiNGO analysis of the 58 genes differentially expressed in both IM137 and IM188

The gene ontologies shown are those biological processes which are over-represented in the 58 genes found to be differentially expressed during dark-induced senescence of IM137 and IM188. The full CATMAv4 probe set was used as a reference set for the analysis and only those with a *p*-value corrected by the Benjamini & Hochberg False Discovery Rate of less than 0.05 are shown.

higher expression of *LOX3* than the others, the change in expression was not large enough to be identified as significant in the t-tests here.

By analysing each of the three classifications of differentially expressed genes presented in Figure 5.14 with BiNGO, it is possible to identify biological process gene ontologies which are over-represented in those sets relative to the abundance of those ontologies across the genes analysed on the CATMA microarrays. This allows the processes relevant to each of STZ and ANAC092 to be identified.

In the set of 141 genes uniquely altered in expression by the *stz* mutant, 91 gene ontology terms are significantly ($p \leq 0.05$) over-represented after false discovery rate correction. Amongst those most highly over-represented are carboxylic acid metabolic processes, responses to jasmonic acid and jasmonic acid biosynthetic processes.

In the set of 356 genes uniquely altered in expression by the *anac092* mutant, 26 gene ontology terms are significantly ($p \leq 0.05$) over-represented after false discovery rate correction. Amongst those most highly over-represented are translation and gene expression related processes, response to temperature stimulus and response to salt stress.

As it is known that both the *stz* and *anac092* mutants show a similar phenotypical senescence response, it would seem more likely that those 58 genes regulated by both are senescence specific whilst other genes regulated by one or the other are specific to other responses involving the respective mutated gene. For this reason, the focus of the rest of this analysis will be around those 58 genes.

In the set of 58 genes altered in expression by both the reduced-expression of *STZ* and *ANAC092*, 12 gene ontologies are significantly ($p \leq 0.05$) over-represented and are shown in Table 5.10. It can be seen that the most significant of these terms represent genes involved in the response to abiotic stimuli. It appears that most of the stimuli are metal ions, temperature and inorganic substances. This allows a number of genes to be selected from the list of 58 and researched further.

The 58 genes altered in expression by both IM137 and IM188 are shown in Table 5.11. Observation of the direction of altered expression for senescence and stress related genes in the list as well as a literature review of their known functions and relationship to other genes of similar function were used to identify new elements in the transcriptional networks surrounding STZ and ANAC092.

It is important to remember that the samples were collected for each line at a time determined by the level of senescence in the leaf, rather than by the time that had elapsed since treatment with reduced light conditions. This means that rather than synchronising the treatment time between samples, as would be traditional, the output response has been synchronised instead. This should effectively isolate the local network around the reduced-expression gene and prevents those genes which provide the senescence response from showing as differentially expressed. In theory, the only genes which could be differentially expressed are those which are regulated by the reduced-expression gene or are altered in expression to compensate for the loss of its function.

Whilst 55 of the genes in the list show the same direction of altered expression, up-regulated or down-regulated in both lines, three genes were curiously up-regulated in IM137 but down-regulated in IM188. These genes were *AtGDPD2* (At5g41080), an unknown function gene (At2g18690) and *AOX1A* (At3g22370). It could be expected that any genes associated specifically with causing the senescence response would also show the same change in expression across both lines. Therefore, these three genes may be those which are found downstream of both STZ and ANAC092 but with alternative regulatory patterns whereby STZ represses their expression and ANAC092 promotes them.

One of the genes which has been shown to be up-regulated in both mutants is *SRG3* (At3g02040) which has been found to accumulate after the growth-arrest phase of *Arabidopsis* cell-suspension cultures (Callard *et al.*, 1996). This behaviour is also observed in the original senescence microarray data where *SRG3* is seen to accumulate steadily after the third observation and reaches a level 5-fold greater by the end of the experiment. Subsequent study of *SRG3* has elucidated that it contains a single GDPD domain, demonstrating conservation of critical residues of the active site along with five other proteins (Cheng *et al.*, 2011). *SRG3* has therefore been given a second alias, *AtGDPD1*, and is likely to be functionally related to *AtGDPD2* as mentioned above.

These genes appear to play an important role in the metabolism of phospholipids into inorganic phosphate required for growth and development. Interestingly, *AtGDPD1*'s closest homologue is *AtGDPD2* (Cheng *et al.*, 2011) and so their similar up-regulation in the *stz* mutant might be anticipated, but their responses in the *anac092* mutant are opposite with *AtGDPD1* being up-regulated and *AtGDPD2* being down-regulated. It was noted by Cheng *et al.* (2011) that *AtGDPD1* is highly induced by inorganic phosphate deficiency and that if it has been muted and cannot be expressed under these conditions, this alone is enough to prevent normal root and rosette growth.

As this family of proteins has only recently been established, no research has yet

AGI	Gene Description	Basic Level	IM137 Change	IM188 Change
At3g01345	Unknown protein	10.3373	÷4.613	÷4.121
At1g67870	Glycine-rich protein	11.9025	÷1.649	÷2.129
At1g67870	Glycine-rich protein	12.5436	÷1.792	÷2.084
At1g21010	Unknown protein	13.0213	÷1.335	÷2.025
At1g20450	Early Response to Dehydration (ERD10)	12.4483	÷1.748	÷1.801
At5g41080	PLC-like phosphodiesterases protein (AtGDPD2)	13.1464	×1.767	÷1.784
At2g23120	Late embryogenesis abundant protein, group 6	13.1193	÷1.594	÷1.773
At3g28100	Nodulin MtN21-like transporter family protein	13.4205	÷1.671	÷1.734
At5g18600	Thioredoxin superfamily protein	12.8716	÷1.673	÷1.723
At5g62680	Major facilitator superfamily protein	13.1166	÷1.508	÷1.696
At2g18690	Unknown protein	13.0544	×1.477	÷1.617
At4g38470	ACT-like protein tyrosine kinase family protein	13.6864	÷1.541	÷1.545
At3g22370	Alternative oxidase 1A (AOX1A)	12.1523	×1.374	÷1.514
At1g30490	Phavoluta (PHV)	13.4753	÷1.383	÷1.485
At1g71950	Proteinase inhibitor, propeptide	13.3395	÷1.352	÷1.446
At5g62760	P-loop containing protein	10.5934	÷1.546	÷1.443
At3g47950	H(+)-ATPase 4 (HA4)	12.8593	÷1.939	÷1.329
At3g06350	Maternal Effect Embryo Arrest 32 (MEE32)	11.5393	×1.279	×1.228
At1g78660	Gamma-glutamyl hydrolase 1 (GGH1)	11.3889	×1.586	×1.326
At5g55970	RING/U-box superfamily protein	11.4159	×1.428	×1.376
At4g32020	Unknown protein	11.4827	×1.862	×1.511
At1g03900	Non-intrinsic ABC protein 4 (NAP4)	11.9434	×1.599	×1.544
At5g67300	MYB domain protein R1 (MYBR1)	11.9911	×1.658	×1.566
At3g22890	ATP sulfurylase 1 (APS1)	12.1270	×1.639	×1.593
At3g17800	Unknown protein	10.4928	×1.591	×1.605
At5g46250	RNA-binding protein	11.9501	×1.617	×1.607
At1g53320	Tubby like protein 7 (TLP7)	11.1119	×1.588	×1.640
At3g56310	Melibiose family protein	11.4049	×1.643	×1.700
At2g41100	Touch 3 (TCH3)	12.7647	×1.927	×1.709
At3g12050	AHA1 domain-containing protein	9.3608	×1.779	×1.718
At4g32300	S-domain-2 5 (SD2-5)	9.5664	×1.700	×1.719
At4g23700	Cation/H ⁺ exchanger 17 (CHX17)	10.5125	×1.919	×1.752
At1g50480	10-formyltetrahydrofolate synthetase (THFS)	11.3216	×1.509	×1.770
At5g19120	Eukaryotic aspartyl protease family protein	11.5251	×1.526	×1.785
At1g31130	Unknown protein	9.4244	×1.957	×1.818
At5g42900	Cold regulated gene 27 (COR27)	11.9780	×1.453	×1.823
At5g23140	Nuclear-encoded CLP protease P7 (NCLPP7)	9.5468	×1.603	×1.828
At1g33610	Leucine-rich repeat (LRR) family protein	9.5847	×2.054	×1.829
At2g30490	Cinnamate-4-hydroxylase (C4H)	11.1215	×1.479	×1.879
At1g67810	Sulfur E2 (SUF2)	10.0035	×1.760	×1.890
At1g56220	Dormancy/auxin associated family protein	11.3248	×1.662	×1.902
At3g02040	Senescence-related gene 3 (SRG3)	11.2894	×2.170	×1.911
At2g25490	EIN3-binding F box protein 1 (EBF1)	12.4339	×1.979	×1.925
At4g34460	GTP binding protein beta 1 (AGB1)	10.1661	×1.901	×1.946
At2g30870	Glutathione S-transferase PHI 10 (GSTF10)	10.1927	×2.098	×1.957
At2g30860	Glutathione S-transferase PHI 9 (GSTF9)	11.5819	×2.217	×1.967
At1g07890	Ascorbate peroxidase 1 (APX1)	10.8858	×2.593	×2.009
At3g02360	6-phosphogluconate dehydrogenase family protein	9.6839	×1.825	×2.046
At1g10070	Branched-chain amino acid transaminase 2 (BCAT-2)	11.0475	×1.961	×2.130
At4g17040	CLP protease R subunit 4 (CLPR4)	12.0460	×1.807	×2.317
At5g26920	Cam-binding protein 60-like G (CBP60G)	10.3443	×1.809	×2.399
At5g51440	HSP20-like chaperones superfamily protein	9.6403	×2.326	×2.568
At3g09440	Heat shock protein 70 (Hsp 70) family protein	10.2657	×2.338	×2.595
At2g39030	Acyl-CoA N-acyltransferases superfamily protein	9.7003	×2.767	×2.599
At4g02715	Unknown protein	9.7656	×1.570	×2.631
At3g49620	Dark Inducible 11 (DIN11)	10.2378	×2.953	×2.643
At5g02500	Heat shock cognate protein 70-1 (HSC70-1)	9.9876	×1.802	×2.753
At5g14170	CHC1	9.9878	×1.927	×3.025
At2g41410	Calcium-binding EF-hand family protein	11.5021	×2.120	×3.314

Table 5.11 – Genes found to be altered in expression during reduced expression of *STZ* and *ANAC092*

The 58 genes shown were found to be altered in expression in both IM137 and IM188 mutants. 59 entries exist in the table, but At1g67870 (lines 2 and 3) was identified by two independent CATMA probes to exhibit significant altered expression. Basic Level represents the log₂ level of expression in the Col-0 wild type whilst IM137 Change represents the relative change to the expression level on the absolute scale invoked by the IM137 mutant at the same point in senescence. IM188 Change shows this same information for the IM188 mutant.

been published about the mechanics of *AtGDPD2*. The coupling of the findings found here with those of Cheng *et al.* (2011) do seem to insinuate, however, that both *STZ* and *ANAC092* may be involved in the response to changing levels of inorganic phosphate and that *ANAC092* promotes the expression of *AtGDPD2* whilst *STZ* represses it. Whether *AtGDPD2* then has an affect on the rate of senescence is uncertain, but these results do show how complex the system can be in its response to many different stress types.

It does seem unlikely that *SRG3* is negatively regulated by *ANAC092* or *STZ* as this microarray data suggests because it was seen to increase in expression alongside *STZ* and *ANAC092* during natural senescence. It could be suggested, therefore, that its expression is regulated by another transcription factor which is compensating for the loss of *STZ* and/or *ANAC092* function.

Those genes which are down-regulated are likely to either be positively regulated by both *STZ* and *ANAC092* or are part of an alternative pathway in which their negative regulation causes a senescence response which compensates the loss of *STZ* and/or *ANAC092*. The gene with the largest change in expression in both lines encodes an unknown protein (*At3g01345*) which maintains a constant expression during the entire period of the original senescence microarray experiment at a level which is low in comparison to active genes in that experiment. It only has one inferred function based on a recognisable glycoside hydrolase domain. It would appear unlikely that both *ANAC092* and *STZ* could positively regulate a gene but that its expression would still only remain low and constant and so its altered expression observed here may indicate that it is a member of an alternative pathway to the induction of senescence.

The second most down-regulated gene was identified by two independent probes on the CATMA microarrays with very similar reported changes in expression in both probes for each line. This glycine-rich protein (*At1g67870*) is more strongly down-regulated by *IM188*, which conforms with the longer period of delay in senescence seen in *IM188*, perhaps suggesting that its expression had been constantly reduced during that longer period until it reached the observed lower level. Like the unknown protein above, the changes in expression of this gene were minimal during the initial senescence microarray experiment but at a much higher level. A study of the cabbage leaf curl virus (*CaLCuV*) (Ascencio-Ibanez *et al.*, 2008), which causes an early senescence response, also found this gene to be significantly down-regulated by the infection and so might imply that it plays a role in the repression of leaf senescence. The exact function of the gene has, however, not yet been identified.

ERD10 (*At1g20450*) is part of the dehydrin family of proteins which are more highly expressed in times of abiotic stress as have been associated with *STZ* and *ANAC092* (Kovacs *et al.*, 2008; Kim & Nam, 2010). Rossel *et al.* (2007) observed that in times of high-light exposure or under oxidative stress, both *STZ* and *ERD10* are up-regulated which is the same but inverse behaviour observed here. During natural senescence, *ERD10*'s expression was observed to fall for a single day around the time that leaf senescence began to occur, before steadily rising to its maximum expression.

It would seem feasible that, given that STZ and ANAC092 are known to be involved in responses to abiotic stresses (Sakamoto *et al.*, 2004; He *et al.*, 2005; Balazadeh *et al.*, 2010b), they would cause up-regulation of *ERD10*, but that in their absence the expression levels of *ERD10* would be reduced as is being observed here.

Only one gene, *AHA4* (At3g47950) demonstrates a much larger significant change in expression in IM137 than that of IM188. *AHA4* has been found to play a role in the salt resistance of *Arabidopsis* (Vitart *et al.*, 2001) as have STZ and ANAC092, although the former is more well characterised in this response and may account for the difference in altered expression of *AHA4*. If *AHA4* were up-regulated by both STZ and ANAC092, this would explain the observed reduction in expression of *AHA4* when those transcription factors are not expressed.

Other abiotic stress related genes in the list include *COR27* (At5g42900), *HSP70* (At3g09440) and *HSC70-1* (At5g02500), all of which are responsive to temperature change. It has been shown that *STZ* is induced by cold stress (Sakamoto *et al.*, 2004) but ANAC092 is not known to be regulated during this type of stress. It might be assumed, then, that these genes are involved in a separate pathway that is able to compensate for the loss of function from STZ and ANAC092. *COR27* has only recently been found to be induced in response to cold (Mikkelsen & Thomashow, 2009) and only then when in the correct phase of the Circadian clock. *COR27*'s expression profile bears some resemblance to *STZ*'s during natural senescence and so perhaps they share a common regulator or stimulus.

As was expected due to the induction method of senescence in this experiment, a number of dark inducible genes were up-regulated in the samples, but not in a way that would cause the mutant lines to express differently compared with the wild type unless those genes have a secondary function related to the reduced-expression gene. One of these genes, *DIN11* (At3g49620), is shown to be induced at a level 16-fold greater than was observed during the natural senescence microarray experiment. This is in agreement with Fujiki *et al.* (2001) where observable changes in expression of *DIN11* can be seen after 24 hours of darkness. It was also shown that this accumulation of *DIN11* can be suppressed by exogenous application of sucrose on the leaves, however, suggesting that it is the reducing sugar levels of the leaf caused by cessation of photosynthesis that induces *DIN11*. This then provides a possible explanation for why *DIN11* is highly expressed in the wild type, but even more highly expressed in the two mutants. Given that the samples from the mutants were harvested at a later time due to the slower senescence response in these lines, it would seem likely that the remaining sugars of the leaf would be lower at that point in time since the leaves had been exposed to a greater length of darkness. This would then suggest that *DIN11* is not necessarily related to STZ and ANAC092 or senescence specifically, but that the delayed senescence caused by the silencing of those transcription factors is merely allowing the levels of sugars in the leaf to fall.

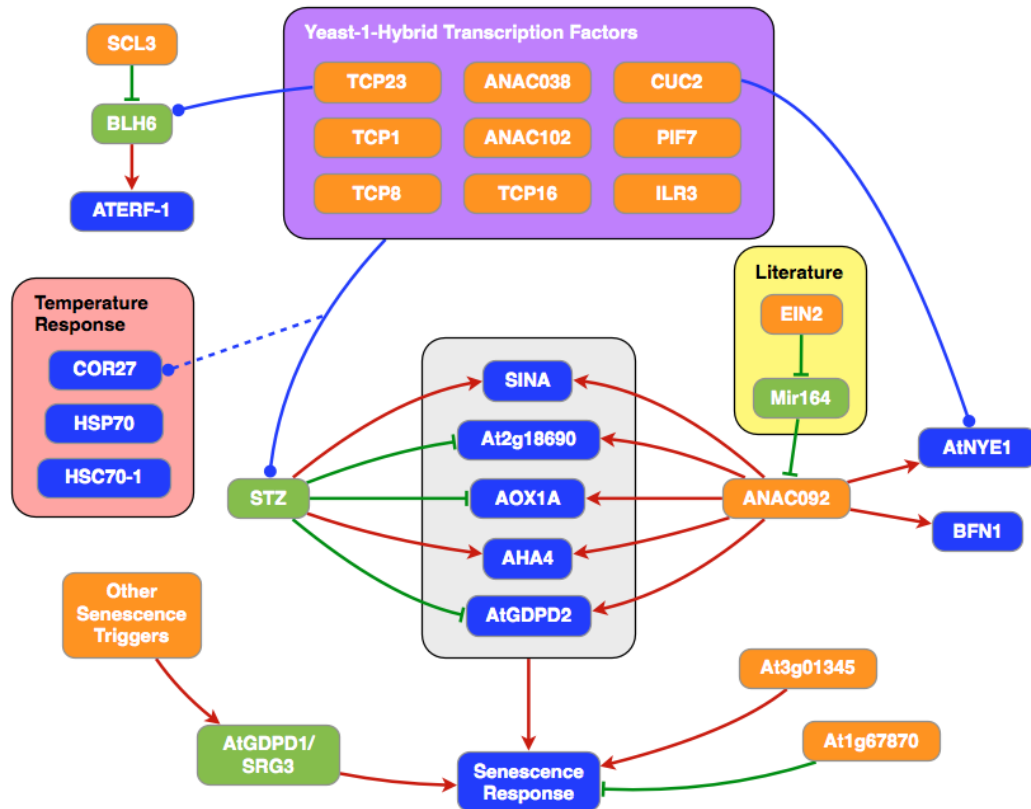


Figure 5.15 – An updated putative transcriptional network model based on results of yeast-1-hybrid screens, literature resources and microarray expression analysis

Using the information obtained by yeast-1-hybrid and microarray analysis of key genes in the transcriptional network discovered in Chapter 4, a newly updated putative transcriptional network model was produced. Red arrows indicate up-regulation whilst green arrows indicate down-regulation. Both red and green arrows may be indicative of an indirect interaction. Orange genes are thought to be hub genes high in the network hierarchy. Green genes act as intermediary genes to the senescence response. Blue genes lead to networks which cause a senescence response. Blue arrows are putative direct regulatory relationships of unknown kinetics which may cause up or down-regulation. Solid blue lines were defined by yeast-1-hybrid results whilst the dotted blue line represents a hypothesis that COR27 and STZ are regulated by a common transcription factor. The purple box groups transcription factors discovered by yeast-1-hybrid screen. The pale red box groups temperature response genes thought to form a separate network leading to senescence. The yellow box groups interactions known from the study by Kim *et al.* (2009).

5.4 Updating the Putative Transcriptional Network Model

The information gathered from the yeast-1-hybrid screening, microarray experiment and literary sources can be used to adapt the structure of the putative theoretical network model introduced in Figure 5.1. The new model can be found in Figure 5.15 where the discovered interactions link elements of the network in a way which could be tested in future progression of this work.

It can be seen that the yeast-1-hybrid data has suggested that ANAC092 does

not cause direct transcriptional regulation of *STZ* and this has also been verified by the microarray data in which the expression of *STZ* was not altered in response for the *anac092* mutant. Therefore, these elements have been separated, but are considered as hubs which act in union as controllers of the senescence response.

A number of potential regulators of *STZ* have been identified by the yeast-1-hybrid process. These were all found to bind to sequences close to the start site of *STZ* and have been included in the purple box. One of these transcription factors, CUC2, was also found to bind upstream of *AtNYE1* and this is shown. TCP23 was found to bind upstream of *BLH6*. Other TCP transcription factors were identified by yeast-1-hybrid but bound to most, if not all, of the fragments being screened which promotes doubt as to whether they would have a transcriptional effect on so many genes or whether, instead, their preferred binding site is relatively simple and found in most upstream regions of genes.

No evidence was given by the yeast-1-hybrid screen that *BLH6* has a regulatory effect on *STZ*, nor that *SCL3* has a regulatory effect on *BLH6*. As the lines with reduced-expression of *SCL3* proved unreliable in the microarray analysis, no evidence was available to prove these interactions. A negative result in yeast-1-hybrid cannot be considered proof of absence of an interaction and so these interactions are present in the updated model, to be tested in future work.

Other transcription factors which were proven to be able to bind to the upstream region of genes screened by yeast-1-hybrid include CUC2 binding to *AtNYE1* and TCP23 binding to *BLH6* and so these interactions are shown.

Kim *et al.* (2009) identified that *ANAC092*'s expression was negatively regulated by Mir164 and that this was negatively regulated by EIN2. These have been included in a yellow box to differentiate them from the interactions hypothesised by the results of this analysis.

Of the three putatively up-regulated genes downstream of *STZ* in the original model, *SINA* was found to be up-regulated by both *STZ* and *ANAC092* only before multiple testing correction was applied to the t-tests. Given that the StepDown method of false discovery rate was shown to be very stringent in Section 3.4.5 and that previous microarray data has shown this gene to be up-regulated by *ANAC092* (Balazadeh *et al.*, 2008, 2010a), these regulatory interactions have been included regardless. *BFN1* was found to be up-regulated by *ANAC092* but unchanged by *STZ* and so this interaction is shown. *AtNYE1*, similarly to *SINA*, was only significantly reduced in expression in the *anac092* mutant before correction, and this interaction is shown.

Senescence related genes with high levels of significant differential expression in both the *stz* mutant and the *anac092 mutant* have been included. Amongst these are *AHA4*, which is proposed to be up-regulated by both *STZ* and *ANAC092*, whilst *At2g18690* (Unknown), *AOX1A* and *AtGDPD2* are proposed to be up-regulated by *ANAC092* but down-regulated by *STZ*. Collectively, changes in the expression of these genes are thought to lead to the senescence response. Further evidence suggesting the

direct regulation of *SINA*, *AtNYE1*, *AHA4* and *AtGDPD2* can be found by analysing their promoter regions (1,000 bp upstream of the transcription start site) in which multiple copies of the known ANAC092 binding site motif, TDRCGTRHD (Olsen *et al.*, 2005), can be found.

Comparing these results with the findings of the *ANAC092* enhanced-expression microarray dataset previously mentioned in Section 4.5 (Balazadeh *et al.*, 2008, 2010a), it can be seen that *SRG3* is 24th most reduced in expression during enhanced-expression of *ANAC092*. This complements the results identified here. *AtGDPD2* and *AOX1A* are both similarly reduced in expression during enhanced-expression of *ANAC092*, but this is the same response as is seen here in the reduced-expression of *ANAC092* and may indicate that their regulation is more dominant by STZ than it is by ANAC092.

A further comparison with an independently hypothesised transcriptional network model can be made by considering the model inferred from microarray data of Breeze *et al.* (2011). In that model, *BFN1* was thought to be regulated by STZ, but not by ANAC092 - this was, however, not seen in the microarrays performed here. In that same model, ILR3 was thought to be a regulator of *STZ* which has not been shown here. STZ was also predicted to be a regulator of *ANAC102* in that model, whilst the evidence shown here is that ANAC102 is probably a regulator of *STZ*. It may be that both are correct and that a feedback loop exists between the two transcription factors.

Three temperature response genes, *COR27*, *HSP70* and *HSC70-1*, are thought to form a separate network similar to that shown in this model. These are therefore grouped in the pale red box in the updated transcriptional network model. The expression profile of *COR27* during the natural senescence experiment bears similarity to that of *STZ* and so it was hypothesised that they may share a common regulator: this is indicated by the dotted blue connection.

Senescence triggers other than STZ and ANAC092 were thought to cause up-regulation of *AtGRPD1/SRG3* and so these are shown at the bottom of the theoretical network. Other genes, At3g01345 and At1g67870, also thought to cause or repress the senescence response but not form part of the network with STZ and ANAC092 are also shown.

It is obvious, when considering the structure of this new network that much more work is required to reach a conclusive network, but also that this is an enormous achievement in a relatively short amount of time. It is worth noting that although any regulations suggested by blue lines would be direct as they were determined by yeast-1-hybrid which assesses the ability of one gene to bind to the upstream region of another, all other regulations should be considered as indicating that an interaction exists but that it might pass through other entities.

The discovery methods of the structure of this network make it hard to envisage how such a complex system might have evolved, but that is because this thesis has focussed on identifying key causes of senescence and trying to identify how they lead to the senescence response. From an evolutionary perspective, some of the first elements of

the network would likely have been those at the bottom of the network with a senescence response leading to the survival of fitter plants, after which the separate pathways leading to senescence may have been constructed as these would allow the timing of senescence to be more controlled and beneficial to the development of the plant. What appears to be shown in this updated network are two branches causing senescence, one from STZ and one from ANAC092 which appear to have somewhat independent upstream regulation but that both lead to a similarly regulated core senescence network leading to a senescence response. The diversity of the senescence network has been shown to be large and the complexity of interactions within it is vast. However, by following a process similar to that shown in this thesis, the network can be elucidated and will eventually be determined as high-throughput technologies underlying these techniques are improved.

5.5 Conclusions

This chapter has introduced an approach which has lead to the verification of a number of putative interactions established in a transcriptional network model formed at the end of Chapter 4. Yeast-1-hybrid screens have been successfully used to verify whether the anticipated transcription factors regulating four different genes in the model are able to bind to the sequence upstream of those genes and, therefore, potentially cause transcriptional changes in them. Similarly, microarrays have been used to determine which genes in the genome exhibit significantly altered expression when hubs STZ or ANAC092 are silenced.

The yeast-1-hybrid analysis has identified 9 potential transcription factors as regulators of STZ whilst two of them are also seen to regulate other members of the network. The microarray analysis has identified a number of new genes which may be downstream of both STZ and ANAC092 which are then thought to cause a senescence response. At least four further genes have been identified as further potential sources of a senescence response which may work in small networks which are adjacent to both STZ and ANAC092.

5.5.1 Future Development

If time had permitted, it would have been valuable to perform the microarray experiment testing the altered expression of genes between mutant lines across several time points. Such a change would have provided more accurate results by introducing temporal data.

The new network model presents a second iteration of the systems biology process to elucidating a definitive regulatory network of genes which lead to a senescence response. If further time were available, the newly predicted downstream interactions of ANAC092 and STZ could be confirmed by using alternative approaches. These may include producing yeast-1-hybrid clones for these genes to find out whether STZ and ANAC092 were able to bind to their upstream sequences. Those genes which have been

identified as being part of adjacent networks to that which has been focussed on here could be analysed using similar processes of theoretical modelling as used in Chapter 4 to try to find which genes they are likely to interact with both up-stream and down-stream in the transcriptional network. These smaller networks could then be verified using similar biological techniques to those shown here which might help to explain how their behaviour links with that of STZ and ANAC092.

Eventually, after a number of iterations, a bigger picture would form and highly interconnected transcriptional networks would result. These networks would have large quantities of evidence supporting their structure and would form one of the first large scale discoveries of a biologically validated transcriptional network of genes associated with a specific function in *Arabidopsis thaliana*.

Chapter 6

General Discussion

This thesis reports the systems biology approaches used to begin the identification of neighbouring transcriptional networks of genes found to alter the phenotype of the senescence response in *Arabidopsis thaliana* leaves. A high-resolution time series profile of the expression of every gene found on CATMA microarrays has been produced using techniques which minimise the technical and experimental variance associated with two-channel microarrays. Those genes which demonstrate an altered expression throughout the senescence time series have been isolated and grouped into functional clusters so that they may be included in a large pool of genes for theoretical network modelling. By selecting subsets of genes from large pools, those genes demonstrating extensive downstream transcriptional regulation were selected for further study. By investigating the phenotypical effect of reduced-expression in hub genes, those with statistically significant altered phenotypes during dark-induced senescence were selected for local network modelling. By modelling genes with altered phenotypes, local putative networks were identified. Sets of genes likely to be both down and upstream of those hubs were combined to provide a final model for biological validation. Tests of key interactions with established experimental techniques identified a large number of new interactions which had not yet been predicted. These findings have allowed a new network to be proposed which could lead to the identification of further interactions if used as prior information in a second round of modelling.

6.1 Microarray Data Analysis

Development and customisation of the MAANOVA microarray analysis package has allowed the design of the senescence microarray experiment to be used in the analysis of the microarrays and for normalised single value expression data to be obtained for each of the genes at each of the time points. Improving the quality control techniques of MAANOVA by providing semi-automated statistical analysis of probe intensities has allowed artefacts amongst the 11,421,696 probe measurements to be easily identified and for informed decisions to be made about the best mode of action to avoid those artefacts.

Additional information is also provided by the new **TechRepCheck** function which compares the intensities of probes between technical replicates to identify those probes which are significantly outlying on one or more replicates. By integrating this information with the existing **ArrayView** function, the location of those outlying probes can now be plotted to quickly illustrate problem areas of specific microarrays. It has been demonstrated that this analysis provides information about misaligned grids during microarray scan analysis, amongst other flaws, that are otherwise invisible to existing analysis techniques. If left untreated, these misaligned grids would influence the final expression data of specific time points for those probes, but this can now be avoided.

The adaptations made to MAANOVA have allowed complex designs in microarray experiments to be analysed with the same simplicity as single-channel highly replicable systems such as AffyMetrix GeneChips at a fraction of the cost. This allows extensive replication to be used which benefits both this analysis and also those of further applications such as transcriptional network modelling.

6.1.1 Future Developments

Removal of Time of Day Variation Prior to the start of this project, microarrays were hybridised with cDNA obtained from senescing leaf samples that were collected at two times of day. The reason samples were collected in this way was to allow the study of other processes such as the circadian clock genes as well as those involved in senescence. Whilst it has been possible to negate the effects of sampling at two times of day whilst analysing the microarrays with MAANOVA, a better solution would have been to collect the samples at one time of day only.

Extension of Tech Rep Check TechRepCheck, which analyses the variation between technical replicates to identify probes which are outliers in one replicate, is only capable of analysing four technical replicates. Whilst this was ideal for the analysis of the senescence microarrays presented in this thesis, this could be extended for datasets with greater numbers of replicates to find outliers even more accurately. This would require extensive consideration of possible formations of the technical replicates as was performed in Table 3.3, from which alternative rules for different numbers of replicates could be produced.

Introduction of BioRepCheck Another potential analysis which could be performed in a similar manner to TechRepCheck was considered in which the biological replicates would be analysed by BioRepCheck. In this case, each observation is represented by four biological replicates and each of those are represented by four technical replicates providing sixteen comparable biological replicates. The technical replicates could be averaged for the purpose of this analysis to reduce the number of comparisons to four again. The outcome of such an analysis would be that specific biological replicates could be identified as a poor representation of a time point if an unidenti-

fied bias has been imposed upon some samples. This information may not be useful unless the sample can be easily replaced, but this would result in many microarray re-hybridisations.

6.2 Analysis of Gene Expression Data

Application of theoretical network modelling techniques to gene expression data has allowed the simultaneous inference of whole networks of genes rather than just single interactions between genes as would normally be discovered by a non-systems biology approach. Although theoretical network modelling is becoming a popular technique amongst geneticists, the selection of genes for modelling are still heavily driven by biological knowledge, inhibiting the discovery of new elements. Two methods have been developed to enhance the discovery of new network elements.

Multi-modelling has allowed vast numbers of genes to be rapidly screened for their importance in theoretical networks. By ensuring that all genes from a large pool are given multiple opportunities to demonstrate their influence on every other gene in the pool, those genes with transcriptional influences on many downstream genes can be isolated. Whilst biologists have taken this approach to identifying hub genes in the past, an automated approach like that demonstrated here improves the accuracy of the analysis by being completely unbiased in gene selection and allowing many more models to be completed in the minimum time possible.

To demonstrate that genes identified as hubs were important to the senescence process, a new phenotype screening process has been developed here. This process provides a statistical analysis for the yellowing of senescing *Arabidopsis* leaves by observing the temporal changes in the ratio of the red and green components of digital photographs taken during the senescent process. By using a logistic curve fitted to those ratios, ANOVA statistics can be used to identify significant differences in the shape of the curves for a mutant plant line versus the wild type control. This removes inaccuracies introduced by subjective assessment of photographs as was previously used to identify altered phenotypes and therefore demonstrates a vast improvement in the standardisation of screens performed by separate individuals. The application of this screen on mutants of genes predicted to be hubs of senescence by multi-modelling demonstrated that greater numbers of altered phenotypes can be identified using that method when compared with random mutant screens.

A method for the identification of networks local to those hub genes demonstrating an altered senescence phenotype has been given the name Metropolis-like VBSSM. This method has allowed the identification of an optimised set of genes from a pool of genes thought to be related to the hub gene, chosen because of their similar profiles. A number of genes were determined to be putatively interacting with hub genes by using this method. Whilst finding directly interacting factors with key genes is undoubtedly the main aim of all studies involving modelling, this method provides a fully automated

approach to finding those genes which form the most robust network. The method is expensive in both computer processing and time, but these quickly pale into insignificance when compared with costly laboratory work discovering dead ends that would otherwise be avoided by extensive preliminary theoretical modelling.

6.2.1 Future Developments

Alternative Clustering The multi-modelling approach to identify hub genes is very reliant on initially identifying a large pool of genes and this was reliant on a good clustering method for the differentially-expressed genes. SplineCluster, as was used here, has undergone several changes since the beginning of the project, and some of these have been incorporated into the analyses performed here such as sweeping for misallocated genes after each merger. Whilst this has allowed SplineCluster to become a hybrid between a hierarchical clustering method and a partitional clustering method, a more biologically relevant method, temporal clustering by affinity propagation (TCAP), has also been developed by Kiddle *et al.* (2010).

TCAP scores every gene to be clustered using the similarity score by Qian *et al.* (2001) which allows directly correlating expression profiles to score highly as well as those which are inversely correlated, correlated after a time-shift and correlated in only part of the profile. These types of matches are all biologically relevant when looking for functional sets of genes, which adds power to the clustering method. The scores between genes are used to identify which are most related as a cluster using affinity propagation. For very large numbers of genes, this can be a time-consuming process, but the end result would most likely have provided even more biological relevance than SplineCluster's hybrid clustering method.

Improved Logistic Curve Analysis The analysis of altered phenotypes during screening of potential hub genes was improved a number of times throughout the project and the results shown in this thesis are based on the latest improvements which test two parameters of a logistic curve mapping the progression of senescence: the x-shift and the y-shift of the curve. However, two other parameters exist in the logistic curve, the x-amplitude and the steepness of the curve, and these could have been used to improve the analysis further. These were not used here because the analysis of only the first two parameters provided enough information to define which were statistically significant in their altered expression.

Choice of Theoretical Network Modelling Methods VBSSM has been used throughout the methods of Chapter 4 for theoretical network modelling and has been shown to be successful in improving the reliability of hub gene selection when used in multi-modelling. However, a number of limitations were identified when using VBSSM to identify local networks surrounding specific hub genes when applied using the Metropolis-like VBSSM approach and these were discussed in Section 4.7.2.

There are many alternative network inference methods available, recent examples being BANJO (Yu *et al.*, 2004), NIR (Gardner *et al.*, 2003), TSNI (Gatta *et al.*, 2008) and ARACNE (Basso *et al.*, 2005; Margolin *et al.*, 2006). These four methods have been compared for accuracy in determining the underlying network of an isolated synthetic *in vivo* network of five genes in *Saccharomyces cerevisiae* (Cantone *et al.*, 2009) where it was found that TSNI was the most reliable method for identifying the network, successfully identifying 4 of the 6 true interactions whilst only reporting 1 false interaction. In a comparison to this result, Steven Kiddle of the University of Warwick obtained the expression data used in this analysis and applied VBSSM to reverse engineer the same network. VBSSM was also able to establish 4 true interactions whilst only reporting 1 false interaction, demonstrating that it is equally as capable as the best method tested by Cantone *et al.* (2009). This demonstrates that, despite some drawbacks associated with the reliability of network models identified by VBSSM, it is still one of the best systems available in identifying underlying transcriptional networks for gene expression data.

6.3 Biological Validation of Theoretical Models

In vivo testing of predictions made by VBSSM involved the analysis of transcription factors able to bind to downstream targets by yeast-1-hybrid and the analysis of downstream genome wide expression changes during the mid-point of senescence in reduced-expression mutants for key regulators of the putative network. Yeast-1-hybrid analysis has shown that the predicted regulators of *STZ*, *BLH6*, *BFN1* and *AtNYE1* are unlikely to be correct, but yeast-1-hybrid is thought to miss many true interactions and also only identifies direct interactions, which the predictions may not be due to the hidden states of VBSSM models. The analysis of downstream expression changes in reduced-expression mutants of *stz* and *anac092* identified a large number of genes which are altered in expression as a result of the mutation. Given that the VBSSM network predicts a close relationship between *STZ* and *ANAC092*, those 58 genes with altered expression in both mutants were carefully assessed and a number of potential relationships with the mutated genes identified.

A combination of both the yeast-1-hybrid results and the results of the reduced-expression mutant analysis allowed a new putative network to be formed which shows both up and downstream elements neighbouring *ANAC092* and *STZ*. This model would make an ideal continuation point for the ongoing analysis of these genes in the senescence context and represent the start of the second iteration of the systems biology approach to elucidating senescence related transcriptional networks.

6.3.1 Future Developments

Choice of Altered Expression Mutants With limited time and resources available, the only plant lines to be tested by microarray analysis were reduced-expression lines

available from SALK. Regretfully, many of the seed stock ordered did not demonstrate a homozygous knockout of the intended gene and were therefore wasted. Others were homozygous, but did not demonstrate a reduced-expression due to the location of the T-DNA insertion. If more time had been available, it would have been beneficial to the interpretation of the results shown in this thesis to have also tested enhanced-expression lines for key genes and also analyse double-knockouts for genes thought to be functionally-redundant. Despite these possible improvements, the results which have been obtained for *stz* and *anac092* have proven valuable in the progression of the work shown here.

Introduction of ChIP-seq and DNase-seq Although the methods shown here have focussed on yeast-1-hybrid and microarray analysis, other methods which would complement these methods include Chromatin immunoprecipitation sequencing (ChIP-seq) (O'Neill & Turner, 1995, 1996) and DNase-seq (Boyle *et al.*, 2008). ChIP-seq involves cross-linking transcription factors to the DNA they bind through application of formaldehyde. This DNA is then sonicated to break it into small fragments where breaks are less likely to occur at the point of bound transcription factors. Antibodies attached to small beads are used to immunoprecipitate those fragments bound by a specific transcription factor and enzymes used to digest those transcription factors. The DNA is then sequenced to identify which DNA sequences can be bound by that transcription factor.

DNase-seq is a complementary procedure in which DNase I is used to digest DNA and high-throughput sequencing used to identify the resulting fragments. Where the DNA is found to be hyper-sensitive to DNase treatment, indicated by the fragment ends, this indicates that nucleosomes have been displaced and is most likely due to the binding of transcription factors to promoter regions causing chromatin to remain open and become more susceptible to DNase activity. By using this information to specify cloned regions for yeast-1-hybrid analysis, the likelihood of finding true positives in those clones would be increased.

6.4 Closing Remarks

Overall, this thesis demonstrates that a systems biology approach to the identification of senescence associated transcriptional networks using theoretical network models complemented by experimental validation techniques can provide vast quantities of valuable information about genes involved in the senescence response and their involvement in those networks. Whereas previous studies, as described in Section 1.3, have lead to the discovery of small networks associated with one gene at a time, the methods implemented here have provided extensive networks associated with several senescence related genes and have also shown crosstalk known to exist between some of those networks. A large number of hub genes remain that have not been studied in detail

here and would provide an excellent progression of this research as well as the iteration of theoretical modelling and experimental study to refine the understanding of these senescence-related genes and their neighbouring transcriptional networks further.

Appendix A

VBSSM Controlling Script

The MATLAB script shown below is that which was used to manage the VBSSM modelling. A pseudo-code interpretation of its functionality can be found in Procedure 2.2 on page 47 of this thesis.

```
1 mkdir('F_KK');
2 mkdir('log');
3 mkdir('vbnet');
4 load data.txt;
5 [Obs_seq, Input_seq] = norm_genes(data, 11);
6 save Obs_seq.mat Obs_seq;
7 save Input_seq.mat Input_seq;
8 multipth = repmat({pwd}, 1, 10);
9 s = dfeval(@FSvskk,num2cell(1:10),multipth,'PathDependencies',{pwd},'
    Configuration',defaultParallelConfig);
10 [MaxF MaxK] = F_vs_k_plot(10,'./F_KK');
11 cd ..;
12 save MaxF.txt MaxF -ascii;
13 print('-dpasc','F_vs_K_Plot.eps');
14 [M] = makeCytoscape(0,10,0,'./vbnet/VB_base_net_seed',['_kk_' num2str(MaxK) '.mat
    ']);
15 cd ./vbnet;
16 for i=1:10;
17     for j=1:20;
18         if j ~= MaxK;
19             delete(['VB_base_net_seed' num2str(i) '_kk_' num2str(j) '.mat']);
20         end;
21     end;
22 end;
23 exit;
```

Appendix B

WHRIMAANOVA Function

Definitions

This appendix lists the functions used during the analysis of the senescence data along with all their parameters such that they may be better understood if being applied to alternative datasets. Return values and descriptions have not been included for sakes of space, but can be identified within the text of Chapter 3 and also in the help files of WHRIMAANOVA.

B.1 read.madata

B.1.1 Usage

```
read.madata(datafile=datafile, designfile=designfile, covM = covM,  
            arrayType=c("oneColor", "twoColor"), header=TRUE, spotflag=FALSE,  
            n.rep=1, avgreps=0, log.trans=FALSE, metarow, metacol, row, col,  
            probeid, intensity, ...)
```

B.1.2 Parameters

datafile	The name of a file, relative to the current working directory, containing the microarray data.
designfile	The name of a file, relative to the current working directory, containing the experimental design.
covM	Gene specific covariate matrix (optional).
arrayType	Defines whether the microarrays are one or two colour.
header	A logical value defining whether the input files have headers.
spotflag	A logical value specifying whether spot data exists.
n.rep	An integer representing the number of replicates.
avgreps	An integer representing whether to collapse replicates or not. 0 means no collapsing, 1 means take the mean of replicates, 2 means take the median of replicates.
log.trans	A logical value defining whether to log ₂ transform the raw data.
metarow	The index of the column in the data file containing meta-row information.
metacol	The index of the column in the data file containing meta-column information.
row	The index of the column in the data file containing row information.
col	The index of the column in the data file containing column information.
probeid	The index of the column in the data file containing probe identification information.
intensity	The index of the first column in the data file containing intensity information.
...	The index of the column in the data file containing any other gene information.

B.2 gridcheck

B.2.1 Usage

```
gridcheck(object, array1, array2, num.split = 2, highlight.flag = TRUE,
          flag.color = "Orange", margin = c(4.6, 5.1, 3.1, 1.1))
```

B.2.2 Parameters

object	The microarray data object loaded with <code>read.madata()</code> .
array1	An integer of the first array number to plot. (optional)
array2	An integer of the second array number to plot (for cross-array comparisons).
num.split	An integer representing the number of pages to split the plots over.
highlight.flag	A logical value defining whether to plot flagged probes in an alternate colour.
flag.color	A string defining the colour to plot flagged probes with.
margin	A vector of margin widths for plots. (bottom, left, top, right)

B.3 riplot

B.3.1 Usage

```
riplot(object, title, array, color = "blue", highlight.flag = TRUE,
        flag.color = "Red", idx.highlight, highlight.color = "Green",
        rep.connect = FALSE, onScreen=TRUE)
```

B.3.2 Parameters

object	The microarray data object loaded with <code>read.madata()</code> .
title	A string array of plot titles which must be of length 'number of arrays to plot'.
array	A vector of array numbers to plot.
color	A string defining the colour to use for plotting the probes.
highlight.flag	A logical value specifying whether flagged probes should be highlighted.
flag.color	A string defining the colour to use for plotting highlighted probes.
idx.highlight	A vector of probe indexes to highlight separately.
highlight.color	A string defining the colour to use for plotting indexed probes for highlighting.
rep.connect	A logical value specifying whether to connect replicate points or not.
onScreen	A logical value specifying whether to display plots on screen.

B.4 arrayview

B.4.1 Usage

```
arrayview(object, array, colormap, num.split = 2, per.page = c(2, 4, 8),
          onScreen=TRUE, margin=c(4.6,5.1,3.1,1.1), ...)
```

B.4.2 Parameters

object	The microarray data object loaded with <code>read.madata()</code> .
array	A vector of array numbers to plot.
colormap	A colour-map to be used as an alternative to Red → Black → Green
num.split	An integer specifying the number of pages to plot each array across (cannot be used with <code>per.page</code> below).
per.page	An integer specifying the number of arrays to plot per page.
onScreen	A logical value representing whether to plot to the screen.
margin	A vector representing the margins of each plot (bottom, left, top, right).

B.5 techrepcheck

B.5.1 Usage

```
techrepcheck(object, array, sample, high.flag = TRUE,
              flagBothColor = "Purple", flag1stColor = "Green",
              flag2ndColor = "Orange", margin = c(6.1, 5.6, 4.1, 2.6))
```

B.5.2 Parameters

object	The microarray data object loaded with <code>read.madata()</code> .
array	A vector of integers specifying which arrays to check the technical replicates of. (optional).
sample	A vector of integers specifying which samples to check the technical replicates of (not to be specified at the same time as <code>array</code>).
high.flag	A logical value indicating whether to highlight flagged probes.
flagBothColor	A string specifying the colour to use for plotting probes flagged in both technical replicates.
flag1stColor	A string specifying the colour to use for plotting probes flagged in the first technical replicate only.
flag2ndColor	A string specifying the colour to use for plotting probes flagged in the second technical replicate only.
margin	A vector representing the margins of each plot (bottom, left, top, right).

B.6 transform.madata

B.6.1 Usage

```
transform.madata(_data, method=c("shift", "glowess", "rLOWESS", "mgLOWESS",  
                                "linlog", "linlogshift"), lolim, uplim, f=0.1, iter=3,  
                degree=1, cg=0.3, cr=0.3, n.bin=10,  
                draw=c("screen", "dev", "off"), ...)
```

B.6.2 Parameters

_data The microarray data object loaded with `read.madata()`.

method The method to apply for transformation.

lolim Low shift limit.

uplim High shift limit.

f The span parameter for a LOWESS transformation.

iter Number of iterations to make the LOWESS fit more robust.

degree The degree of polynomials used in LOWESS fits, up to 2.

cg The percentage of genes to be transformed linearly for the green channel.

cr The percentage of genes to be transformed linearly for the red channel.

n.bin The number of bins for calculating the variance after linlogshift.

draw Where to plot the transformation RIPlots.

B.7 fitmaanova

B.7.1 Usage

```
fitmaanova(madata, formula, random= ~1, covariate = ~1,  
           method=c("REML", "ML", "MINQE-I", "MINQE-UI", "noest"),  
           verbose=TRUE, subCol=FALSE)
```

B.7.2 Parameters

madata The microarray data object loaded with `read.madata()`.

formula A model formula defining the expected relationship between experimental design terms.

random A formula indicating which terms are randomly effects.

covariate Array specific covariates.

method The method used for solving the mixed-model equation. Default is REML.

verbose A logical value specifying whether to display messages about calculation progress.

subCol A logical value specifying whether to subtract column mean. Default is FALSE but two-colour arrays are always TRUE.

B.8 matest

B.8.1 Usage

```
matest(data, anovaobj, term, Contrast, n.perm=1000, nnodes=1,
        critical=.9, test.type = c("ttest", "ftest"),
        shuffle.method=c("sample", "resid"),
        MME.method=c("REML", "noest", "ML"),
        test.method=c(1,1), pval.pool=TRUE, verbose=TRUE)
```

B.8.2 Parameters

<code>data</code>	The microarray data object loaded with <code>read.madata()</code> .
<code>anovaobj</code>	A MAANOVA data object obtained from <code>fitmaanova()</code> .
<code>term</code>	The term to be tested with an F-test.
<code>Contrast</code>	A contrast matrix specifying levels of the term to test, each row is a separate test, whilst columns represent the levels of the term.
<code>n.perm</code>	An integer specifying the number of permutations to perform.
<code>nnodes</code>	The number of nodes in the MPI cluster for permutation tests.
<code>critical</code>	The percentile of the F-distribution to use for conducting permutation tests.
<code>test.type</code>	The type of test to perform. Default is <code>ftest</code> where <code>Contrast</code> is missing and <code>ttest</code> where <code>Contrast</code> has been provided.
<code>shuffle.method</code>	The method used for data shuffling.
<code>MME.method</code>	The method used for solving mixed-model equations. The default is <code>REML</code> .
<code>test.method</code>	A vector of integers indicating whether to (1) or not to (0) perform the F1 and Fs tests, respectively.
<code>pval.pool</code>	A logical value specifying whether to use pooled permutation F values to calculate P values.
<code>verbose</code>	A logical value specifying whether to display messages indicating calculation progress.

B.9 ConvertMATest

B.9.1 Usage

```
convertmatest(bioreptest, alttest)
```

B.9.2 Parameters

<code>bioreptest</code>	An object returned by <code>matest()</code> representing the F-test results of a biological replicate term.
<code>alttest</code>	An object returned by <code>matest()</code> representing the F-test results of a non-biological replicate term.

B.10 adjPval

B.10.1 Usage

```
adjPval(matestobj, method=c("stepup", "adaptive", "stepdown", "jsFDR"))
```

B.10.2 Parameters

matestobj An object returned by either `matest()` or `ConvertMATEst()`.
method A string specifying the false discovery rate method to apply.

B.11 AnalyseMATEst

B.11.1 Usage

```
analysematest <- (fstat=c("Fs", "F1"), test1, test2, test3, sig1 = 0.05,  
                  sig2 = 0.05, sig3 = 0.05, useAdjPVals = F, prefix = "CATMA")
```

B.11.2 Parameters

fstat A string specifying which F-statistic to use for plotting a Venn diagram.

test1 An object representing the first F-test results to be plotted. It must have been returned by `ConvertMATEst()` and then optionally processed by `adjPval()`.

test2 An object representing the second F-test results to be plotted. It must have been returned by `ConvertMATEst()` and then optionally processed by `adjPval()`.

test3 An object representing the third F-test results to be plotted. It must have been returned by `ConvertMATEst()` and then optionally processed by `adjPval()`.

sig1 The significance level to use in deciding which genes are significantly differentially expressed in **test1**.

sig2 The significance level to use in deciding which genes are significantly differentially expressed in **test2**.

sig3 The significance level to use in deciding which genes are significantly differentially expressed in **test3**.

useAdjPVals A logical value specifying whether adjusted *p*-values should be used.

prefix A string indicating the prefix of probe IDs to be included in the Venn diagram. Can be set as an empty string when not required.

Appendix C

MAANOVA Analysis Script

This R script is the top-level script for the microarray analysis of the senescence data. It calls a number of functions which operate MAANOVA's pre-ANOVA quality control, ANOVA model fitting and post-ANOVA identification of significantly differentially expressed genes.

```
1 rm(list=ls())
2 library(whrimaanova)
3
4 # Change to the folder containing data from the experiment
5 setwd("~/Documents/MAANOVA/Data")
6
7 # Read in the data
8 catma <- read.madata("catma_bc_medians.txt", designfile="design.txt",
9   metarow=1, metacol=2, row=3, col=4, probeid=5,
10   intensity=6, arrayType="twoColor", log.trans=T, spotflag=T)
11
12 # Change to a folder where output can be placed
13 setwd("~/Documents/MAANOVA/")
14
15 # GridCheck, RIPlot and TechRepCheck the data
16 gridcheck(catma)
17 riplot(catma)
18 catma <- techrepcheck(catma)
19
20 # Save the output before attempting transformation
21 save(catma, file="./After Estimation.RData")
22
23 # Transform the data to remove anomalies
24 hri <- transform.madata(catma, method="mglowess", f=0.1, draw="pdf")
25 file.rename("./Output/Data Transformation RIPlots.pdf",
26   "./Output/MGLowess Transformation RIPlots.pdf")
27 hri <- transform.madata(catma, method="rlowess", f=0.1, draw="pdf")
28 file.rename("./Output/Data Transformation RIPlots.pdf",
29   "./Output/RLowess Transformation RIPlots.pdf")
30
31 # Check the final result in graphical form
32 arrayview(catma, per.page="2")
33
34 # Save the data ready for model fitting
35 save(catma, file="./Ready for Model Fitting.RData")
```

```

36
37 # Fit a model to the data based on the terms of variation given
38 anova <- fitmaanova(catma, ~Dye+Array+(Day*ToD)/BioRep, ~Dye+Array)
39
40 # Save the ANOVA output
41 save(anova, file="./After Anova.RData")
42
43 # Do F-tests on the terms to find how much variation they provide to the model
44 # Remove the ftests as they are complete because they take up a lot of RAM
45 ftest4 <- matest(catma, anova, "Day:ToD:BioRep", n.perm=1)
46 save(ftest4, file="./Output/F-Test Day x ToD x BioRep.RData")
47
48 ftest1 <- matest(catma, anova, "Day", n.perm=1)
49 save(ftest1, file="./Output/F-Test Day.RData")
50 test1 <- convertmatest(ftest4, ftest1)
51 rm(ftest1)
52
53 ftest2 <- matest(catma, anova, "ToD", n.perm=1)
54 save(ftest2, file="./Output/F-Test ToD.RData")
55 test2 <- convertmatest(ftest4, ftest2)
56 rm(ftest2)
57
58 ftest3 <- matest(catma, anova, "Day:ToD", n.perm=1)
59 save(ftest3, file="./Output/F-Test Day x ToD.RData")
60 test3 <- convertmatest(ftest4, ftest3)
61 rm(ftest3, ftest4)
62
63 # Adjust P Values for false discovery rate
64 test1 <- adjPval(test1, "stepdown")
65 test2 <- adjPval(test2, "stepdown")
66 test3 <- adjPval(test3, "stepdown")
67
68 # Test the terms and draw a Venn Diagram
69 analysematest("Fs", test1, test2, test3, useAdjPVals=T)

```

Appendix D

DrawGOGraphs() Source Code

This R script accepts output BGO files from BiNGO and returns a PDF of informative plots about the GO terms which were identified as over-represented.

```
1 DrawGOGraphs <- function(bgoFolder, csvOutFile = "BiNGOSpreadsheet.csv",
2   pdfOutFile = "BiNGOPlots.pdf", useCorrectedPvalue = F, plotLabelInCentre = F)
3   {
4     # A function to pick black or white as a contrasting text colour
5     SetTextContrastColor <- function(color)
6     {
7       ifelse( mean(col2rgb(color)) > 127, "black", "white")
8     }
9
10    # Get the lists of files from the BINGO output
11    bgoList <- dir(bgoFolder, ".bgo", full.names=T)
12
13    # Setup a legend table of colors
14    colorList <- colors()[c(555,29,144,258,99,588,153,182,231,79,41,20)]
15    contrastTextColor <- unlist(lapply(colorList, SetTextContrastColor))
16    colorList <- data.frame(density = rep(c(100, 40, 40, 40, 40), each=12), angle
17      = rep(c(0, 22, 157, 112, 67), each = 12), color=rep(colorList, 5),
18      contrast=c(contrastTextColor, rep("black",48)), stringsAsFactors=F)
19
20    # Prepare to produce the ontList from the bgo files
21    ontList <- NULL
22
23    # Read in the data from the BGO files
24    clusters <- NULL
25    for (i in 1:length(bgoList)) {
26      # Read the data in to a messy table with header info
27      curCluster <- read.table(bgoList[i], fill=T, sep="\t", quote="", comment.
28        char="", blank.lines.skip=F, as.is=T, col.names=c("GO-ID", "p-value",
29        "corr p-value", "x in cluster", "x in universe", "cluster size", "
30        universe size", "Description", "Genes in test set"))
31
32      # Find the end of the header and cut it off
33      lineBeforeData <- which(curCluster[,1] == "GO-ID")
34      curCluster <- curCluster[-(1:lineBeforeData),]
35
36      # Add the ontologies to the ontList
37      if (nrow(curCluster) > 0) {
38        ontList <- rbind(ontList, cbind(i, curCluster))
39      }
40    }
41  }
```

```

32     }
33
34     # Store this cluster's data in the clusters variable
35     clusters[[i]] <- curCluster
36 }
37
38 colnames(ontList) <- c("Cluster No.", "GO-ID", "p-value", "corr p-value", "x
    in cluster", "x in universe", "cluster size", "universe size", "
    Description", "Genes in test set")
39 write.csv(ontList, csvOutFile, row.names=F)
40
41 # Open a PDF file
42 pdf(pdfOutFile, height=7.5, width=11)
43 layout(matrix(1:2,1), widths=c(2,3))
44
45 # Go through the clusters one at a time
46 i <- 0
47 for (curData in clusters) {
48     i <- i + 1
49     # Check that the cluster has any GO terms
50     if (nrow(curData) > 0) {
51         # Sort the data into order of GO ontologies instead of P-value
52         curData <- curData[order(as.numeric(curData$p.value)),]
53
54         # Find the values which will be used to plot
55         pvalCol <- 1 * useCorrectedPvalue + 2
56         plotValues <- -log(as.numeric(curData[,pvalCol]))
57
58         # Get the labels to be plotted above the bars
59         labelValues <- as.numeric(curData[, "x.in.cluster"]) / as.numeric(
            curData[, "cluster.size"]) * 100
60         labelValues <- sprintf("%0.2f%s", labelValues, "%")
61
62         # Make a plot title
63         plotTitle <- sprintf("Cluster %d (%d members)", i, as.numeric(curData
            [1,6]))
64         par(mar=c(1,4.5,2.5,0))
65
66         # Draw a barplot and put labels on the bars
67         xvals <- barplot(plotValues, density=colorList$density, angle=
            colorList$angle, col=colorList$color, main=plotTitle, ylab="-log
            (p-value)", cex.lab=1.2, col.main="blue2", col.lab="red2", ylim=c
            (0, ceiling(max(plotValues))))
68         nBars <- length(xvals)
69         cexFactor <- min(12 / nBars, 1)
70         if (plotLabelInCentre) {
71             text(xvals, plotValues/2, labelValues, srt=90, col=colorList$
                contrast, cex=cexFactor)
72         } else {
73             text(xvals, plotValues, labelValues, pos=3)
74         }
75
76         par(mar=c(0,1,0,0))
77
78         # Draw histogram legend
79         legendSize <- min(25 / nBars, 1)
80         plot.new()
81         legend("left", legend=curData$Description, fill=colorList$color,
            density=colorList$density, angle=colorList$angle, bty="n", cex=

```

```
            legendSize)
82         }
83     }
84
85     dev.off()
86 }
```

Appendix E

Supplementary GO term results

After analysis, by BiNGO, of clusters returned by SplineCluster, a number of Gene Ontologies were identified as over-represented. This table lists those terms and shows the expected abundance of the term given the number of genes with that ontology in the *Arabidopsis* genome and also the actual observed abundance in that cluster. The discrepancy between these values allows the calculation of a p -value following a hypergeometric test. This p -value was then corrected using a Bonferroni multiple testing correction to eliminate most falsely identified significant terms. Those terms which were still below the ($p < 0.05$) significance threshold are the only ones to be shown.

Cluster Number	GO Term Description	Term Abundance	Observed in Cluster	Corrected p -value
1	cell wall organization	0.83%	7.21%	1.17×10^{-3}
1	cell wall modification	0.63%	6.31%	1.87×10^{-3}
1	cell wall organization or biogenesis	1.37%	7.21%	4.83×10^{-2}
5	response to jasmonic acid stimulus	0.74%	5.88%	3.11×10^{-2}
6	cellular respiration	0.35%	5.71%	3.41×10^{-2}
6	energy derivation by oxidation of organic compounds	0.35%	5.71%	3.41×10^{-2}
10	defense response	3.08%	15.09%	4.17×10^{-2}
12	response to starvation	0.52%	9.52%	1.82×10^{-2}
19	response to water	1.43%	14.49%	1.06×10^{-5}
19	response to water deprivation	1.40%	13.04%	1.12×10^{-4}
19	response to abscisic acid stimulus	1.76%	13.04%	7.62×10^{-4}
19	response to organic substance	5.89%	20.29%	1.05×10^{-2}
19	response to desiccation	0.11%	4.35%	1.09×10^{-2}
19	response to hormone stimulus	4.42%	17.39%	1.14×10^{-2}
19	response to endogenous stimulus	4.76%	17.39%	2.35×10^{-2}
19	cold acclimation	0.14%	4.35%	2.59×10^{-2}
25	transport	8.72%	20.66%	1.57×10^{-2}
25	establishment of localization	8.80%	20.66%	1.82×10^{-2}
25	localization	8.97%	20.66%	2.51×10^{-2}
25	response to osmotic stress	2.57%	9.92%	2.72×10^{-2}
29	pyrimidine nucleotide biosynthetic process	0.06%	5.00%	7.65×10^{-4}

Continued on next page

Cluster Number	GO Term Description	Term Abundance	Observed in Cluster	Corrected p -value
29	pyrimidine nucleotide metabolic process	0.08%	5.00%	1.90×10^{-3}
37	cellular process	37.68%	59.74%	3.92×10^{-2}
39	post-translational protein modification	5.38%	28.57%	2.37×10^{-2}
39	response to fungus	0.96%	14.29%	3.97×10^{-2}
44	response to wounding	0.86%	22.73%	1.67×10^{-4}
44	response to stress	10.68%	45.45%	6.29×10^{-3}
45	response to ozone	0.14%	5.88%	1.33×10^{-2}
46	regulation of vegetative phase change	0.03%	3.77%	1.85×10^{-2}
50	translation	2.74%	11.90%	2.65×10^{-2}
50	cellular biosynthetic process	11.33%	26.19%	3.34×10^{-2}
51	DNA metabolic process	1.48%	11.58%	5.83×10^{-5}
51	DNA replication	0.55%	7.37%	3.40×10^{-4}
51	cellular nitrogen compound metabolic process	8.32%	22.11%	1.25×10^{-2}
51	nitrogen compound metabolic process	8.65%	22.11%	2.22×10^{-2}
51	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	5.16%	15.79%	4.58×10^{-2}
55	translation	2.74%	39.74%	1.16×10^{-26}
55	gene expression	4.48%	42.31%	2.34×10^{-22}
55	cellular macromolecule biosynthetic process	4.87%	41.03%	5.40×10^{-20}
55	macromolecule biosynthetic process	4.93%	41.03%	8.01×10^{-20}
55	cellular biosynthetic process	11.33%	46.15%	1.98×10^{-12}
55	biosynthetic process	12.27%	46.15%	2.43×10^{-11}
55	cellular protein metabolic process	11.10%	43.59%	4.87×10^{-11}
55	cellular macromolecule metabolic process	15.12%	47.44%	2.66×10^{-9}
55	protein metabolic process	13.04%	43.59%	5.17×10^{-9}
55	macromolecule metabolic process	17.26%	47.44%	1.42×10^{-7}
55	primary metabolic process	27.23%	52.56%	4.05×10^{-4}
55	cellular metabolic process	27.29%	52.56%	4.31×10^{-4}
55	cellular process	37.68%	61.54%	3.60×10^{-3}
55	metabolic process	34.58%	57.69%	5.48×10^{-3}
55	nucleosome organization	0.32%	5.13%	2.40×10^{-2}
55	nucleosome assembly	0.32%	5.13%	2.40×10^{-2}
55	chromatin assembly	0.34%	5.13%	2.90×10^{-2}
55	DNA packaging	0.34%	5.13%	2.90×10^{-2}
55	protein-DNA complex assembly	0.35%	5.13%	3.48×10^{-2}
58	photosynthesis	1.51%	15.00%	5.71×10^{-5}
60	anatomical structure development	6.92%	20.48%	1.82×10^{-2}
66	cellular nitrogen compound metabolic process	8.32%	22.58%	4.29×10^{-4}
66	chloroplast organization	0.65%	6.45%	5.81×10^{-4}
66	nitrogen compound metabolic process	8.65%	22.58%	9.28×10^{-4}
66	plastid organization	1.08%	7.26%	3.70×10^{-3}
66	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	5.16%	14.52%	3.49×10^{-2}
67	photosynthesis	1.51%	9.38%	1.54×10^{-4}
67	photosynthesis, light harvesting	0.26%	4.69%	2.09×10^{-4}

Continued on next page

Cluster Number	GO Term Description	Term Abundance	Observed in Cluster	Corrected p -value
67	photosynthesis, light reaction	0.86%	6.25%	4.56×10^{-3}
67	photosynthesis, light harvesting in photosystem I	0.06%	2.34%	1.14×10^{-2}
68	photosynthesis	1.51%	8.42%	3.65×10^{-2}
68	pseudouridine synthesis	0.11%	3.16%	4.45×10^{-2}
73	photosynthesis	1.51%	20.75%	3.06×10^{-17}
73	photosynthesis, light reaction	0.86%	12.26%	1.08×10^{-9}
73	photosynthetic electron transport chain	0.35%	8.49%	1.46×10^{-8}
73	generation of precursor metabolites and energy	1.68%	14.15%	6.04×10^{-8}
73	electron transport chain	0.57%	9.43%	7.81×10^{-8}
73	oxidation reduction	1.06%	10.38%	3.98×10^{-6}
73	photosynthetic electron transport in photosystem I	0.22%	5.66%	1.64×10^{-5}
73	cellular metabolic process	27.29%	45.28%	1.76×10^{-2}
73	defense response to bacterium	1.29%	7.55%	2.23×10^{-2}
74	small molecule catabolic process	1.34%	8.25%	1.98×10^{-2}
74	photosynthesis	1.51%	8.25%	4.66×10^{-2}
74	cellular carbohydrate metabolic process	2.90%	11.34%	4.90×10^{-2}
75	embryonic development	2.54%	9.40%	1.23×10^{-2}
75	embryonic development ending in seed dormancy	2.22%	8.72%	1.27×10^{-2}
75	seed development	2.48%	8.72%	4.11×10^{-2}
76	photosynthesis	1.51%	15.29%	1.04×10^{-7}
76	translation	2.74%	18.82%	2.87×10^{-7}
76	gene expression	4.48%	21.18%	9.16×10^{-6}
76	cellular macromolecule biosynthetic process	4.87%	20.00%	1.79×10^{-4}
76	photosynthesis, light reaction	0.86%	9.41%	1.92×10^{-4}
76	macromolecule biosynthetic process	4.93%	20.00%	2.14×10^{-4}
76	cellular biosynthetic process	11.33%	30.59%	4.36×10^{-4}
76	cellular metabolic process	27.29%	51.76%	4.73×10^{-4}
76	biosynthetic process	12.27%	30.59%	1.98×10^{-3}
76	electron transport chain	0.57%	7.06%	2.58×10^{-3}
76	photosynthetic electron transport chain	0.35%	5.88%	3.48×10^{-3}
76	metabolic process	34.58%	57.65%	3.69×10^{-3}
76	generation of precursor metabolites and energy	1.68%	10.59%	3.94×10^{-3}
76	photosynthetic electron transport in photosystem I	0.22%	4.71%	9.02×10^{-3}
76	oxidation reduction	1.06%	8.24%	1.05×10^{-2}
76	cellular process	37.68%	58.82%	2.03×10^{-2}
77	pigment biosynthetic process	0.71%	5.41%	4.50×10^{-3}
77	pigment metabolic process	0.82%	5.41%	1.34×10^{-2}
77	quinone cofactor metabolic process	0.14%	2.70%	1.78×10^{-2}
77	quinone cofactor biosynthetic process	0.14%	2.70%	1.78×10^{-2}
77	metabolic process	34.58%	50.00%	4.04×10^{-2}

Appendix F

Multimodelling Source Code

This R script is the co-ordinating script for the multi-modelling process. It assembles lists of genes appropriate for VBSSM, and then initiates an instance of MATLAB which is tasked to provide the modelling results of VBSSM. Once the model is complete, another is initiated until all pairs have been present at least 5 times.

```
1 bigGeneList <- as.matrix(read.table("GeneList.txt")) # Load in the gene list
2 nGenes <- nrow(bigGeneList) # Count the number of rows
3
4 # Check the two data files and if they are missing, create new empty ones
5 if (!file.exists("Pairs.RData")) {
6   warning("Pairs.RData doesn't exist; creating a new one")
7   pairs <- matrix(0, nrow=nGenes, ncol=nGenes)
8   save(pairs, file = "Pairs.RData")
9 }
10
11 if (!file.exists("Counts.RData")) {
12   warning("Counts.RData doesn't exist; creating a new one")
13   counts <- rep(0, nGenes)
14   save(counts, file = "Counts.RData")
15 }
16
17 load("Pairs.RData") # Load the pairs file
18 load("Counts.RData") # Load the counts file
19
20 threshold <- 1 # Only pairs picked less than this number of times are selected
   for modelling
21 checkcapacity <- FALSE # Don't check capacity the first time
22
23 while (threshold <= 5) { # Continue until the threshold is greater than
24   repeat {
25     repeat {
26       if (!checkcapacity) { # If check isn't required, skip to workings
27         break
28       }
29       system("sleep 270") # Wait to check for more capacity on the cluster
30       maxpercent <- read.table("/Users/sdjmhchattie/Model Parameters/MaxPercent.
   txt")[[1]] # Read in the max cluster should be used
31       allprocesses <- length(system("qstat", intern=T)) - 2 # Find out how many
   things are running
32       if ((allprocesses + 10) / 56 * 100 <= maxpercent) { # If new processes won
```

```

        't push over the limit
33     break
34 }
35 }
36
37 idx <- matrix(F, ncol=nGenes, nrow=nGenes) # Start an index which will
        diagonally cut the pairs matrix
38 pair <- matrix(0, ncol=2, nrow=nGenes * (nGenes - 1) / 2) # Start a matrix
        of gene pairs
39 numPairs <- nGenes * (nGenes - 1) / 2 # Calculate total number of pairs
40 for (i in 1:(nGenes-1)) { # Loop rows of the table
41     for (j in (i + 1):nGenes) { # Loop cols of the table
42         idx[i,j] <- T # Set a filter index for the pairs table
43         pairCount <- ((j - 1) * (j - 2) / 2) + i
44         pair[pairCount,] <- c(i, j) # Make an entry in pair storing which genes
            are in each pair
45     }
46 }
47 lowPairs <- which(pairs[idx] < threshold) # Find out which pairs are
        represented less than 4 times
48 if (length(unique(as.vector(pair[lowPairs,]))) < 88) { # Less than 88 genes
        left at threshold
49     threshold <- threshold + 1 # Raise threshold
50     cat("\nThreshold is being raised to", threshold, "\n\n") # Log the change
51     checkcapacity <- FALSE # Don't check capacity next loop
52     break # Exit repeat loop (drop back to while loop)
53 }
54 geneidx <- NULL # Start the geneidx variable
55 while(length(geneidx) < 88) { # Keep going until the geneidx has at least 88
        entries
56     geneidx <- unique(c(geneidx, pair[sample(lowPairs, 1),])) # Pull pairs at
        random and make sure genes remain unique in the index
57 }
58
59 folders <- dir(pattern="^[0-9]+$") # Get a list of files/folders with
        numbers only
60 folders <- as.numeric(folders) # Change the list into integers
61 newfolder <- sprintf("%04d", setdiff(1:9999, folders)[1]) # Get next
        available number
62 system(paste("cp -Rp Template", newfolder)) # Make an exact copy of the
        Template folder
63
64 curdir <- getwd() # Record the current directory
65 setwd(paste("./", newfolder, sep="")) # Change to the new folder
66
67 smallGeneList <- bigGeneList[geneidx] # Subset the larger list
68 write.table(smallGeneList, "Genes.txt", row.names=F, col.names=F, quote=F)
69
70 cat("\nStarting list number", newfolder, "at", format(Sys.time(), "%H:%M on %
        d-%m-%Y"), "\n") # Log to the screen
71 cat("Threshold is currently", threshold, "\n") # Log to the screen
72
73 source("extract.R") # Extract the data from the ANOVA object
74 system("matlab -nodisplay -nodesktop -nosplash -r script > output.txt 2>
        errors.txt", wait=F) # Initiate the SSM Model
75
76 checkcapacity <- TRUE # Check capacity next time now that new processes are
        running
77

```

```

78     setwd(curdir) # Change back to the original folder
79
80     # Set values in Pairs and Counts
81     counts[geneidx] <- counts[geneidx] + 1
82     for (i in geneidx) {
83         for (j in geneidx) {
84             if (i != j) {
85                 pairs[i,j] <- pairs[i,j] + 1
86             }
87         }
88     }
89
90     # Backup RData files in case of cluster crash
91     file.copy("Pairs.RData", paste("./RData Backups/Pairs before ", newfolder, ".
      RData", sep=""))
92     file.copy("Counts.RData", paste("./RData Backups/Counts before ", newfolder,
      ".RData", sep=""))
93
94     # Save RData files with new data
95     save(pairs, file="Pairs.RData")
96     save(counts, file="Counts.RData")
97 }
98 }
99
100 cat("\nAll models complete!\n")

```

Appendix G

Metropolis-like VBSSM Source Code

G.1 Altered VBSSM Process

The VBSSM processing script was split into four files so that the second and third could be iterated until a maximum mean F could be found. The contents of each of these files are as shown.

G.1.1 VBSSMPart1.m

```
1 scriptPath='/Volumes/cluster/sdjmachattie/Work/LocalisedModels/VBSSMScripts/';
2 addpath(scriptPath);
3 mkdir('F_KK');
4 mkdir('log');
5 mkdir('vbnet');
6 load data.txt;
7 [Obs_seq, Input_seq] = norm_genes(data, 11);
8 save Obs_seq.mat Obs_seq;
9 save Input_seq.mat Input_seq;
10 delete('VBSSMPart1.m');
11 delete('VBSSMPart1_errors.txt');
12 delete('VBSSMPart1_output.txt');
13 exit;
```

G.1.2 VBSSMPart2.m

```
1 scriptPath='/Volumes/cluster/sdjmachattie/Work/LocalisedModels/VBSSMScripts/';
2 addpath(scriptPath);
3 inp1 = num2cell(1:5);
4 inp2 = repmat({pwd}, 1, 5);
5 inp3 = repmat({lowkk}, 1, 5);
6 inp4 = repmat({highkk}, 1, 5);
7 copyfile([scriptPath, 'FSvskk.m'], 'FSvskk.m');
8 s = dfeval(@FSvskk, inp1, inp2, inp3, inp4, 'PathDependencies', {pwd}, 'Configuration',
9         defaultParallelConfig);
10 delete('FSvskk.m');
11 delete('VBSSMPart2.m');
```

```

11 delete('VBSSMPart2_errors.txt');
12 delete('VBSSMPart2_output.txt');
13 exit;

```

G.1.3 VBSSMPart3.m

```

1 scriptPath='/Volumes/cluster/sdjmchattie/Work/LocalisedModels/VBSSMScripts/';
2 addpath(scriptPath);
3 [MaxF MaxK] = F_vs_k_plot(5, './F_KK');
4 cd ..;
5 save MaxF.txt MaxF -ascii;
6 save MaxK.txt MaxK -ascii;
7 print('-dpsc', 'F_vs_K_Plot.eps');
8 delete('VBSSMPart3.m');
9 delete('VBSSMPart3_errors.txt');
10 delete('VBSSMPart3_output.txt');
11 exit;

```

G.1.4 VBSSMPart4.m

```

1 scriptPath='/Volumes/cluster/sdjmchattie/Work/LocalisedModels/VBSSMScripts/';
2 addpath(scriptPath);
3 [M] = makeCytoscape(0,5,0, './vbnet/VB_base_net_seed', ['_kk_' num2str(MaxK) '.mat'
4   ]);
5 cd ./vbnet;
6 for i=1:5;
7     for j=lowkk:highkk;
8         if j ~= MaxK;
9             delete(['VB_base_net_seed' num2str(i) '_kk_' num2str(j) '.mat']);
10        end;
11    end;
12 end;
13 cd ..;
14 delete('VBSSMPart4.m');
15 delete('VBSSMPart4_errors.txt');
16 delete('VBSSMPart4_output.txt');
17 delete('Genes.txt');
18 exit;

```

G.2 Coordinating R Script

This R script is the co-ordinating script for the Metropolis-like VBSSM process. It assembles lists of genes appropriate for VBSSM, and then initiates an instance of MATLAB which is tasked to provide the modelling results of VBSSM. After each model, the gene list is updated based on the fit of the previous model and this is iterated until no better fitting model can be identified.

```

1 specificGene <- read.table("../GeneCATMA.txt", sep="\t", quote="", as.is=T,
2   check.names=F)[[1]] # Load in the specific gene for modelling

```

```

2 sigGenes <- read.table("CloseMatches.txt", sep="\t", quote="", as.is=T, check.
  names=F)[[1]] # Load in the list of significant genes from F-test
3
4 modFactor <- 1.05 # Set the factor to use for modifying swapNumber
5 modelNum <- 1 # Set the model number to be done next
6 targetListSize <- 88 # Set the target number of genes per model
7 swapNumber <- 20 # Set the number of genes to swap before the next model
8 Fgold <- -100000 # Set the F-score to beat
9 goldGeneList <- c(specificGene, sample(sigGenes, (targetListSize - 1))) # Make a
  random list of genes
10 acceptance <- c(1, 0, 0, 0, 1, 0, 0, 0) # Make a vector of 0 for rejected and 1
  for accepted
11
12 folders <- dir(pattern="^Run[0-9]+$") # Get a list of files/folders with runs
13
14 unfinished <- NULL # Set a new variable to store the unfinished folder name
15 write.table(NULL, "TimeCheck") # Create a file for checking the time since last
  model
16 timeNow <- file.info("TimeCheck")["mtime"][[1]] # Find the time right now
17 for (folder in folders) {
18   # Go through the folders
19   modHours <- difftime(timeNow, file.info(file.path(folder, "RunData.RData"))["
    mtime"][[1]], unit="hours") # Find the number of hours since the run was
    modified
20   if (modHours > 2) {
21     # If run hasn't been updated in 2 hours, assume not running
22     test <- new.env(parent = baseenv()) # Start a new environment
23     load(file.path(folder, "RunData.RData"), test) # Load the RunData into the
      environment
24     if (round(test$swapNumber) > 0) {
25       # If the run still has more to do
26       rm(test) # Remove the test environment
27       unfinished <- folder # Set the folder to continue into unfinished
28       break # Break the for loop
29     }
30     rm(test) # Remove the test environment
31   }
32 }
33
34 if (is.null(unfinished)) {
35   # If there is no unfinished run
36   folders <- as.numeric(substr(folders, 4, 6)) # Change the list into integers
37   newfolder <- sprintf("Run%03d", setdiff(1:999, folders)[1]) # Get next
    available number
38   dir.create(newfolder) # Make the new folder
39   setwd(newfolder) # Change to the new folder
40 } else {
41   setwd(unfinished) # Set the working directory to the unfinished run
42   load("RunData.RData") # Load the data for that run, ready to continue
43 }
44
45 logFile <- file("RunLog.txt", "a") # Append the runlog
46 if (file.exists("RunData.RData")) {
47   # If this is a continued run
48   cat("\n\nRecovering and continuing run from previous attempt...\n\n\n", file=
    logFile) # Log this fact to the log file
49 }
50
51 # Keep going until the swapNumber reaches zero

```

```

52 while(round(swapNumber) > 0) {
53   newfolder <- sprintf("%04d", modelNum) # Create a folder for the model
54   if (file.exists(newfolder)) {
55     suppressValue <- unlink(newfolder, T) # Delete it as it's incomplete
56   }
57
58   curdir <- getwd() # Record the current directory
59   dir.create(newfolder) # Make the new folder
60   setwd(newfolder) # Change to the new folder
61
62   numToKeep <- (targetListSize - 1) - round(swapNumber) # Find number of genes
                    to keep in goldGeneList
63   geneList <- c(specificGene, sample(goldGeneList[-1], numToKeep)) # Keep that
                    number of genes at random
64   while (length(geneList) < targetListSize) {
65     # Continue to add genes until it's full sized
66     numToAdd <- targetListSize - length(geneList) # Calculate how many genes to
                    add to the list
67     geneList <- c(geneList, sample(sigGenes, numToAdd)) # Add the genes
68     geneList <- unique(geneList) # Make sure no duplicates exist in the list
69   }
70   write.table(geneList, "Genes.txt", row.names=F, col.names=F, quote=F) # Write
                    the list to the folder
71
72   write.table(NULL, "../TimeCheck") # Create a file for checking the time
                    since last model
73   timeNow <- file.info("../TimeCheck")["mtime"][[1]] # Find the time right
                    now
74   cat("Beginning model", newfolder, "at", format(timeNow, "%H:%M on %d-%m-%Y"), "
...\\n", file=logFile) # Log information
75
76   source("~/Work/LocalisedModels/extract.R") # Extract the data from the ANOVA
                    object
77   supressValue <- file.copy("/Volumes/cluster/sdjmhchattie/Work/LocalisedModels/
                    VBSSMScripts/VBSSMPart1.m", "VBSSMPart1.m") # Copy normalisation script in
78   system("matlab -nodisplay -nodesktop -nosplash -r VBSSMPart1 > VBSSMPart1_
                    output.txt 2> VBSSMPart1_errors.txt", wait=T) # Normalise the data
79
80   low <- lowkk <- 8 # Set initial lowkk to 8
81   high <- highkk <- 9 # Set initial highkk to 9
82   loopflag <- TRUE # Set the flag to allow loop
83   while (loopflag) {
84     doModelAgain <- TRUE # Set a flag to redo a model if FSvskk plot fails
85     while (doModelAgain) {
86       doModelAgain <- FALSE # Set it to false as we'll assume it went OK
87       # While the flag is set
88       newFile <- file("VBSSMPart2.m", "w") # Open a new file for the middle
                    script
89       cat("lowkk = ", low, ";\n", sep="", file=newFile) # Put the lowkk into the
                    script
90       cat("highkk = ", high, ";\n", sep="", file=newFile) # Put the highkk into
                    the script
91       close(newFile) # Close the new file
92       system("cat /Volumes/cluster/sdjmhchattie/Work/LocalisedModels/VBSSMScripts/
                    VBSSMPart2.m >> VBSSMPart2.m", wait=T) # Append the contents of the
                    middle script
93
94       # Sleep for a while if the qstat list is full or another model was recently
                    run

```

```

95   looped <- FALSE # Set the loop test to FALSE
96   while (TRUE) {
97     if (looped) {
98       system("sleep 50") # Wait for 45 seconds before checking capacity
99     }
100    looped <- TRUE # Indicate that the loop has happened at least once
101    write.table(NULL, "~/ModelParameters/TimeCheck") # Create a file for
      checking the time since last model
102    timeNow <- file.info("~/ModelParameters/TimeCheck")["mtime"][[1]] # Find
      the time right now
103    lastModelTime <- file.info("~/ModelParameters/LastModelStarted")["mtime"
      ][[1]] # Find the time of the last model to start
104    secsSinceLastModel <- difftime(timeNow, lastModelTime, unit="secs") #
      Calculate the minutes between the two times
105    if (secsSinceLastModel > 45) {
106      # If more than 45 seconds have passed since the last model started then
      consider starting another
107      qstat <- system("qstat", intern=T) # Get the qstat list
108      qstat <- qstat[-(1:2)] # Remove header lines
109      qstat <- qstat[substr(qstat, 41, 41) == "r" | substr(qstat, 41, 41) ==
        "q"] # Only look at running and waiting jobs
110      runningJobs <- sum(as.numeric(substr(qstat, 98, 102))) # Get the
        column for slots and sum it
111      maxPercent <- read.table("~/ModelParameters/MaxCores.txt", sep="\t",
        quote="", as.is=T, check.names=F)[[1]] # Load in the max
        percentage of cluster to use
112      if ((runningJobs + 5) < maxPercent) {
113        break # Get out of the loop
114      }
115    }
116  }
117
118  write.table(NULL, "~/ModelParameters/LastModelStarted") # Generate a file
      to time check the last model run
119  system("matlab -nodisplay -nodesktop -nosplash -r VBSSMPart2 > VBSSMPart2_
      output.txt 2> VBSSMPart2_errors.txt", wait=T) # Initiate the SSM Model
120
121  suppressValue <- file.copy("/Volumes/cluster/sdjmchattie/Work/
      LocalisedModels/VBSSMScripts/VBSSMPart3.m", "VBSSMPart3.m") # Copy
      normalisation script in
122
123  waitBeforeCheck <- FALSE
124  countLoops <- 0 # Set the number of attempts to plot FSvskk to 0
125  while (!file.exists("MaxK.txt")) {
126    if (waitBeforeCheck) {
127      system("sleep 15") # Wait for 15 seconds before checking again
128    }
129    system("matlab -nodisplay -nodesktop -nosplash -r VBSSMPart3 > VBSSMPart3
      _output.txt 2> VBSSMPart3_errors.txt", wait=T) # Draw the F vs K
      plot
130    countLoops <- countLoops + 1 # Add one to the loop counter
131    if (countLoops >= 40) {
132      # If been trying for around 10 minutes
133      unlink("./F_KK/*") # Delete old attempt at model
134      unlink("FS_on.mat") # Delete old FS_on file
135      low <- lowkk # Set the lowk value to the lowest used
136      high <- highkk # Set the highk value to the highest used
137      doModelAgain <- TRUE # Get the model started again
138      break

```

```

139     }
140     waitBeforeCheck <- TRUE
141   }
142 }
143
144 maxK <- read.table("MaxK.txt")[[1]] # Find the MaxK identified so far
145 if (maxK <= lowkk) {
146   lowkk <- lowkk - 1
147   low <- high <- lowkk
148   unlink("MaxK.txt")
149 } else if (maxK >= highkk) {
150   highkk <- highkk + 1
151   low <- high <- highkk
152   unlink("MaxK.txt")
153 } else {
154   loopflag <- FALSE
155 }
156 }
157
158 newFile <- file("VBSSMPart4.m", "w") # Create a new finishing script
159 cat("lowkk = ", lowkk, ";\n", sep="", file=newFile) # Set the lowkk value
160 cat("highkk = ", highkk, ";\n", sep="", file=newFile) # Set the highkk value
161 cat("MaxK = ", maxK, ";\n", sep="", file=newFile) # Set the MaxK variable
162 close(newFile) # Close the new file
163 system("cat /Volumes/cluster/sdjmachattie/Work/LocalisedModels/VBSSMScripts/
VBSSMPart4.m >> VBSSMPart4.m", wait=T) # Append the generic finishing
script
164 system("matlab -nodisplay -nodesktop -nosplash -r VBSSMPart4 > VBSSMPart4_
output.txt 2> VBSSMPart4_errors.txt", wait=T) # Run the finishing script
165
166
167 Fnew <- read.table("MaxF.txt")[[1]] # Find the maximum F-score of the VBSSM
168 cat("\nModel", newfolder, "completed...\n", file=logFile) # Log the end of the
modelling
169 cat("MaxF for model:", Fnew, "\n", file=logFile) # Log the F-score of the
model
170 if (Fnew > Fgold || runif(1) < exp(-abs(Fnew - Fgold))) {
171   # If the new model is better
172   goldGeneList <- geneList # Set the goldGeneList
173   Fgold <- Fnew # Set the gold F-score
174   acceptance <- c(acceptance, 1) # Set the last model as accepted
175   if (swapNumber > (targetListSize - 1)) {
176     swapNumber <- (targetListSize - 1) # If it has, set it back to the
maximum swapNumber
177   }
178   cat("Model was kept\n", file=logFile) # Log that the model is the new gold
179   cat("New Fgold value:", Fgold, "\n", file=logFile) # Log the new gold F-
score
180   cat("New goldGeneList:", paste(goldGeneList, collapse=", "), "\n\n", file=
logFile) # Log the new list of genes
181 } else {
182   acceptance <- c(acceptance, 0) # Set that the model was rejected
183   cat("Model was discarded\n\n", file=logFile) # Log that it was rejected
184 }
185
186
187 acceptedRatio <- sum(acceptance[-(1:modelNum)]) / 8 # Find the ratio of
accepted models
188 modRatio <- acceptedRatio - 0.25 # Subtract the target ratio

```

```

189   cat("Percentage of accepted models is now:", sprintf("%.2f%%", (acceptedRatio
      * 100)), "\n", file=logFile) # Log the ratio
190
191   if (acceptance[modelNum + 8] == 1) {
192     # If the last model was accepted
193     swapNumber <- swapNumber * modFactor ^ 3 # Increase the swapNumber three
      times
194   } else if (modRatio > 0.05) {
195     # If the ratio is too high
196     swapNumber <- swapNumber * modFactor # Make the swapNumber bigger to walk
      parameter space faster
197   } else if (modRatio < -0.05) {
198     # If the ratio is too low
199     swapNumber <- swapNumber / modFactor # Reduce swapNumber to reduce speed of
      random walk
200   }
201
202   if (swapNumber > (targetListSize - 1)) {
203     # Check that swapNumber hasn't exceeded the maximum
204     swapNumber <- (targetListSize - 1) # If it has, set it to the maximum
205   }
206
207   if (round(swapNumber) < 1) {
208     cat("\n\n*****\n** SwapNumber reached zero and has been set back to 5\n
      *****\n\n", file=logFile)
209     swapNumber <- 5
210   }
211
212   cat("SwapNumber is now:", round(swapNumber), "\n\n\n", file=logFile) # Log the
      new swapNumber
213
214   setwd(curdir) # Change back to the original folder
215   modelNum <- modelNum + 1 # Get ready for next model number
216
217   save(list=c("modelNum", "swapNumber", "Fgold", "acceptance", "goldGeneList"),
      file="RunData.RData") # Save the RunData out
218 }
219
220 cat("\nHalting!", file=logFile) # Log end of run
221
222 close(logFile) # Finish with the log file
223 q("no") # Exit R

```

Appendix H

White Balance Transformation Script

This R script is used in the white balance of images used in quantitative phenotype screening. Once the white balance has been applied, the newly normalised values are then used to calculate the level of senescence in the leaf area which has been supplied.

```
1 gatherData <- function(path) {
2   modalvalue <- function(x, na.rm=FALSE) {
3     x = unlist(x);
4     if(na.rm) x = x[!is.na(x)]
5     u = unique(x);
6     n = length(u);
7     frequencies = rep(0, n);
8     for(i in 1:n)
9     {
10       if(is.na(u[i]))
11       {
12         frequencies[i] = sum(is.na(x))
13       } else
14       {
15         frequencies[i] = sum(x==u[i], na.rm=TRUE)
16       }
17     }
18     u[which.max(frequencies)]
19   }
20
21   inputFiles <- dir(path)
22   if (length(inputFiles) == 0) {
23     stop("No files found in that path")
24   }
25   lineNames <- unique(sub("[^-]*$", "", inputFiles))
26   repNames <- unique(sub("^.*-([^-]*)\\.txt$", "\\1", inputFiles))
27   temp <- as.matrix(read.table(paste(path, inputFiles[1], sep=""), quote="",
28     header=F))
29   nTimes <- nrow(temp) / 256
30   dataArray <- array(0, c(length(lineNames), 13, length(repNames), nTimes), c("
    Line", "RGB Data", "Replicate", "Time"))
31 }
```

```

32   for (lineNum in 1:length(lineNames)) {
33     lineName <- lineNames[lineNum]
34     cat("Processing ", lineName, "...\\n", sep="")
35
36     for (repNum in 1:length(repNames)) {
37       repName <- repNames[repNum]
38
39       fileToOpen <- paste(path, lineName, "-", repName, ".txt", sep="")
40       rawRGB <- as.matrix(read.table(fileToOpen, quote="", header=F))
41
42       for (i in 1:nTimes) {
43         from <- 256 * i - 255
44         to <- 256 * i
45
46         rData <- NULL
47         gData <- NULL
48         bData <- NULL
49
50         for (j in from:to) {
51           rData <- c(rData, rep(rawRGB[j,1], rawRGB[j,2]))
52           gData <- c(gData, rep(rawRGB[j,1], rawRGB[j,3]))
53           bData <- c(bData, rep(rawRGB[j,1], rawRGB[j,4]))
54         }
55
56         if (repNum > 1) {
57           rTrans <- 200 / dataArray[lineNum, 1, 1, i]
58           gTrans <- 200 / dataArray[lineNum, 2, 1, i]
59           bTrans <- 200 / dataArray[lineNum, 3, 1, i]
60           rData <- rData * rTrans
61           gData <- gData * gTrans
62           bData <- bData * bTrans
63         }
64
65         dataArray[lineNum, , repNum, i] <-
66           c(mean(rData), mean(gData), mean(bData),
67             median(rData), median(gData), median(bData),
68             modalvalue(rData), modalvalue(gData), modalvalue(bData),
69             sd(rData), sd(gData), sd(bData),
70             length(rData))
71       }
72     }
73   }
74
75   rownames(dataArray) <- c(lineNames)
76   colnames(dataArray) <- c("rMean", "gMean", "bMean", "rMedian", "gMedian", "
       bMedian", "rMode", "gMode", "bMode", "rSD", "gSD", "bSD", "count")
77
78   lineNames <- rownames(dataArray)
79   nLines <- length(lineNames)
80   nReps <- dim(dataArray)[3] - 1
81   nTimes <- dim(dataArray)[4]
82
83   scoreCard <- matrix(0, nrow=nLines*nReps, ncol=nTimes)
84   rownames(scoreCard) <- paste(rep(lineNames, each=nReps), "- Rep", 1:nReps)
85
86   for (l in 1:nLines) {
87     for (r in 2:(nReps + 1)) {
88       rownum <- (l - 1) * nReps + r - 1
89       for (t in 1:nTimes) {

```

```
90         scoreCard[rownum, t] <- dataArray[1,1,r,t] / dataArray[1,2,r,t]
91     }
92 }
93 }
94
95 write.table(scoreCard, file=outfile, quote=F, sep="\t", row.names=T, col.
96             names=F)
```

Appendix I

Logistic Curve Model Fitting Script

This R script is used to fit a logistic curve to the phenotype screen data and then process the resulting curves to identify whether differences exist between the wild type response and the altered expression lines.

```
1 plotLogistic <- function(y, x, xshift = 10, xcoef = 1, yshift = 0.7, ycoef = 1,
2   plot = T, add = F, pch = 4, cex = 0.5, ...) {
3   plotdata <- data.frame(x, y)
4   if (plot) {
5     if (!add) {
6       plot(plotdata, pch = pch, cex = cex, ...)
7     } else {
8       points(plotdata, pch = pch, cex = cex, ...)
9     }
10  }
11  try(model <- nls(y ~ m * (1/(1 + exp(n * (-x + a)))) + b, plotdata, list(m =
12    ycoef, n = xcoef, a = xshift, b = yshift)))
13  if (!exists("model") || model$convInfo$stopMessage != "converged") {
14    stop("The model failed, try different starting parameters")
15  }
16  modelpars <- model$m$getAllPars()
17
18  curveFunc <- function(x, pars = modelpars) {
19    y <- pars["m"] * (1/(1 + exp(pars["n"] * (-x + pars["a"])))) + pars["b"]
20  }
21
22  if (plot) {
23    curve(curveFunc, add=T, ...)
24  }
25
26  invisible(list(ycoef=modelpars["m"],yshift=modelpars["b"],xcoef=modelpars["n"]
27    ,xshift=modelpars["a"]))
28 }
29 processSenData <- function(filename) {
30   senescence <- read.table(filename,quote="",sep="\t",fill=T)
```

```

31   timepoints <- as.matrix(senescence[1,])
32   senescence <- as.matrix(senescence [-1,])
33
34   y <- NULL
35   x <- NULL
36
37   for (i in 1:nrow(senescence)) {
38     for (j in 1:ncol(senescence)) {
39       curdata <- senescence[i,j]
40       if (!is.na(curdata)) {
41         x <- c(x, timepoints[j])
42         y <- c(y, curdata)
43       }
44     }
45   }
46
47   returnList <- NULL
48   returnList$x <- x
49   returnList$y <- y
50   returnList$tps <- timepoints
51   returnList$repdata <- senescence
52
53   invisible(returnList)
54 }
55
56 interpolate <- function(target, axis, data) {
57   returnvals <- NULL
58   for (repro in 1:nrow(data$repdata)) {
59     pars <- NULL
60     try (pars <- plotLogistic(data$repdata[repro,], t(data$tps), plot=F))
61     if (is.null(pars)) {
62       warning("Unable to evaluate logistic curve for replicate ", repro)
63     }
64     else if (axis == 1 || axis == "x") {
65       returnvals <- c(returnvals, pars$ycoef * (1/(1 + exp(pars$xcoef * (-
66         target + pars$xshift))))) + pars$yshift)
67     }
68     else if (axis == 2 || axis == "y") {
69       returnvals <- c(returnvals, -1 * ((log((pars$ycoef / (target - pars$
70         yshift)) - 1) / pars$xcoef) - pars$xshift))
71     }
72     else {
73       stop("Invalid option specified for 'axis': ", axis)
74     }
75   }
76   invisible(returnvals)
77 }
78 makePlot <- function(mutant, wt, title, subtitle1="", subtitle2="", seprep=F, cex
79   =0) {
80   plotErrorBars <- function(x, y, yplus, yminus, widthCoef = 0.008, col="black"
81     , ...) {
82     nbars <- length(x)
83     xmax <- max(x)
84     xmin <- min(x)
85     ebWidth <- (xmax-xmin) * widthCoef
86     for (bar in 1:nbars) {
87       xbar <- x[bar]

```

```

86     upperbar <- y[bar] + yplus[bar]
87     lowerbar <- y[bar] - yminus[bar]
88     plotx <- c(xbar, xbar)
89     ploty <- c(upperbar, lowerbar)
90     lines(plotx, ploty, col=col, ...)
91     plotx <- c(xbar-ebWidth, xbar+ebWidth)
92     ploty <- c(upperbar, upperbar)
93     lines(plotx, ploty, col=col, ...)
94     ploty <- c(lowerbar, lowerbar)
95     lines(plotx, ploty, col=col, ...)
96   }
97 }
98
99 par(mar = c(5, 4, 6, 4) + 0.1)
100 xlim <- range(wt$x)
101 ylim <- range(mutant$y, wt$y)
102 ylim[1] <- floor(ylim[1] * 10) / 10
103 ylim[2] <- ceiling(ylim[2] * 10) / 10
104 plot(NULL, xlim=xlim, ylim=ylim, xaxt="n", yaxt="n", xlab="Days after
    Darkness", ylab="Yellowing", main=title)
105 posx <- xlim[1] + (xlim[2]-xlim[1]) * 0.2
106 posy <- ylim[1] + (ylim[2]-ylim[1]) * 0.95
107 text(posx, posy, subtitle1, cex=0.7)
108 text(posx, posy, pos=1, offset=0.7, subtitle2, cex=0.7)
109 xticks <- xlim[1]:xlim[2]
110 yticks <- seq(ylim[1],ylim[2],0.1)
111 axis(1, xticks, cex.axis=0.7)
112 axis(2, yticks, cex.axis=0.7, las=2)
113 for (i in xticks) {
114   abline(v=i, col="lightgrey", lty="dashed", lwd=0.5)
115 }
116 for (i in yticks) {
117   abline(h=i, col="lightgrey", lty="dashed", lwd=0.5)
118 }
119
120 if (seprep) {
121   mainlwd <- 2
122 }
123 else {
124   mainlwd <- 1
125 }
126
127
128 for (repro in 1:nrow(wt$repdata)) {
129   wtpars <- NULL
130   try (wtpars <- plotLogistic(wt$repdata[repro,], t(wt$tps), col="
    palegreen4", cex=cex, add=T, plot=seprep))
131   if (is.null(wtpars)) {
132     warning("Unable to evaluate logistic curve for wildtype replicate ",
        repro)
133   }
134   else {
135
136   }
137 }
138 for (repro in 1:nrow(mutant$repdata)) {
139   mutantpars <- NULL
140   try (mutantpars <- plotLogistic(mutant$repdata[repro,], t(mutant$tps),
        col="pink", cex=cex, add=T, plot=seprep))

```

```

141     if (is.null(mutantpars)) {
142         warning("Unable to evaluate logistic curve for mutant replicate ",
143             reprow)
144     }
145     else {
146     }
147 }
148
149 wtpars <- plotLogistic(wt$y, wt$x, col="green", cex=cex, add=T, lwd=mainlwd)
150 if (!seprep) {
151     ebx <- sort(unique(wt$x))
152     eby <- wtpars$ycoef * (1/(1 + exp(wtpars$xcoef * (-ebx + wtpars$xshift))))
153         ) + wtpars$yshift
154     ebse <- NULL
155     for (x in ebx) {
156         yvals <- wt$y[wt$x == x]
157         ebse <- c(ebse, sd(yvals)/sqrt(length(yvals)))
158     }
159     plotErrorBars(ebx, eby, ebse, ebse, col="green")
160 }
161
162 mutantpars <- plotLogistic(mutant$y, mutant$x, col="red", cex=cex, add=T, lwd
163     =mainlwd)
164 if (!seprep) {
165     ebx <- sort(unique(mutant$x))
166     eby <- mutantpars$ycoef * (1/(1 + exp(mutantpars$xcoef * (-ebx +
167         mutantpars$xshift)))) + mutantpars$yshift
168     ebse <- NULL
169     for (x in ebx) {
170         yvals <- mutant$y[mutant$x == x]
171         ebse <- c(ebse, sd(yvals)/sqrt(length(yvals)))
172     }
173     plotErrorBars(ebx, eby, ebse, ebse, col="red")
174 }
175
176 maxwtx <- wtpars$xshift
177 maxwty <- wtpars$ycoef / 2 + wtpars$yshift
178 mutantx <- -1 * ((log((mutantpars$ycoef / (maxwty - mutantpars$yshift)) - 1)
179     / mutantpars$xcoef) - mutantpars$xshift)
180 mutanty <- mutantpars$ycoef * (1/(1 + exp(mutantpars$xcoef * (-maxwtx +
181     mutantpars$xshift)))) + mutantpars$yshift
182 abline(v=maxwtx, col="green")
183 abline(h=maxwty, col="green")
184 abline(v=mutantx, col="red")
185 abline(h=mutanty, col="red")
186 maxwtx <- round(maxwtx, 2)
187 maxwty <- round(maxwty, 2)
188 mutantx <- round(mutantx, 2)
189 mutanty <- round(mutanty, 2)
190 axis(3, c(maxwtx, mutantx), cex.axis=0.7, las=2, lwd=0, lwd.ticks=1)
191 axis(4, c(maxwty, mutanty), cex.axis=0.7, las=2, lwd=0, lwd.ticks=1)
192 }
193
194 wt1 <- processSenData("1-Col10.txt")
195 im244 <- processSenData("1-IM244.txt")
196
197 xlevels <- t1 <- t2 <- p1 <- p2 <- NULL
198 wtpars <- plotLogistic(wt1$y, wt1$x, plot=F)

```

```

194 maxwt1x <- wtipars$xshift
195 maxwt1y <- wtipars$ycoef / 2 + wtipars$yshift
196
197 anovaTable <- data.frame(NULL)
198 anovaTable <- rbind(anovaTable, data.frame(Line = "WT", Yellowing = interpolate(
199   maxwt1x, 1, wt1)))
200 anovaTable <- rbind(anovaTable, data.frame(Line = "IM244", Yellowing =
201   interpolate(maxwt1x, 1, im244)))
202 test <- aov(Yellowing ~ Line, anovaTable)
203 df <- test$df.residual
204 rms <- sum(test$residuals^2)/df
205 meanstable <- model.tables(test, "means")
206 linereps <- meanstable$n$Line
207 linemeans <- meanstable$tables$Line
208 n1 <- linereps[1]
209 xbar1 <- linemeans[1]
210 for (line in 2:length(linereps)) {
211   n2 <- linereps[line]
212   xbar2 <- linemeans[line]
213   sed <- sqrt(rms*(1/n1 + 1/n2))
214   newt <- (xbar1 - xbar2) / sed
215   newp <- 1-pt(abs(newt), df)
216   t1 <- c(t1, newt)
217   p1 <- c(p1, newp)
218 }
219 xlevels <- c(xlevels, test$xlevels[[1]][-1])
220
221 anovaTable <- data.frame(NULL)
222 anovaTable <- rbind(anovaTable, data.frame(Line = "WT", Time = interpolate(
223   maxwt1y, 2, wt1)))
224 anovaTable <- rbind(anovaTable, data.frame(Line = "IM244", Time = interpolate(
225   maxwt1y, 2, im244)))
226 test <- aov(Time ~ Line, anovaTable)
227 df <- test$df.residual
228 rms <- sum(test$residuals^2)/df
229 meanstable <- model.tables(test, "means")
230 linereps <- meanstable$n$Line
231 linemeans <- meanstable$tables$Line
232 n1 <- linereps[1]
233 xbar1 <- linemeans[1]
234 for (line in 2:length(linereps)) {
235   n2 <- linereps[line]
236   xbar2 <- linemeans[line]
237   sed <- sqrt(rms*(1/n1 + 1/n2))
238   newt <- (xbar1 - xbar2) / sed
239   newp <- 1-pt(abs(newt), df)
240   t2 <- c(t2, newt)
241   p2 <- c(p2, newp)
242 }
243
244 anovareresults <- data.frame(Line = xlevels, SameTimeTStat = t1, SameTimePvalue =
245   p1,
246   SameResponseTStat = t2, SameResponsePvalue = p2)
247
248 sub1 <- paste("p-value for vertical comparison =", signif(p1, 4))
249 sub2 <- paste("p-value for horizontal comparison =", signif(p2, 4))

```

```
248 pdf("Phenotype Means.pdf", width=8, height=11)
249 layout(matrix(1:2,ncol=1))
250
251 makePlot(im244, wt1, "IM244 - SCL3", sub1[3], sub2[3])
252
253 dev.off()
254
255
256 pdf("Phenotype Replicates.pdf", width=8, height=11)
257 layout(matrix(1:2,ncol=1))
258
259 makePlot(im244, wt1, "IM244 - SCL3", sub1[3], sub2[3], T)
260
261 dev.off()
```

Appendix J

Hub Genes Identified by Multi-modelling

The tables in this Appendix show the list of genes identified as hubs by multi-modelling. Criteria for entry in this list is that the gene must have been seen to have an influence on the expression of at least 2 other genes on more than 50% of the occasions when those genes were presented in the same model. The interaction was only considered to be true if at least one of the modelling seeds was more than 95% confident of the interaction. The entries of the table are sorted into order of the number of genes downstream of each hub (labelled Upstream Of) and their TAIR description is presented, where available, in adjacent columns.

AGI No.	Upstream of	Description	Functions In	Involved In
At5g05410	366	DRE-binding protein 2A (DREB2A)		
At3g05690	292	nuclear factor Y, subunit A2 (NF-YA2)		
At1g27730	266	salt tolerance zinc finger (STZ)		
At1g68990	262	male gametophyte defective 3 (MGP3)	DNA-directed RNA polymerase activity, DNA binding	pollen tube growth, transcription, ovule development, embryo development ending in seed dormancy
At2g47040	249	VANGUARD1 (VGD1)	enzyme inhibitor activity, pectinesterase activity	pollen tube growth
At2g38880	238	nuclear factor Y, subunit B1 (NF-YB1)	sequence-specific DNA binding transcription factor activity	response to water deprivation, regulation of transcription, DNA-dependent
At4g25490	192	C-repeat/DRE binding factor 1 (CBF1)		
At2g25900	151	ATCTH		
At5g06510	140	nuclear factor Y, subunit A10 (NF-YA10)	sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent
At3g15210	136	ethylene responsive element binding factor 4 (ERF4)		
At5g62040	135	PEBP (phosphatidylethanolamine-binding protein) family protein		
At3g14020	132	nuclear factor Y, subunit A6 (NF-YA6)	sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent
At2g34710	117	PHABULOSA (PHB)		
At5g18140	101	Chaperone DnaJ-domain superfamily protein	unfolded protein binding, heat shock protein binding	protein folding
At1g63850	96	BTB/POZ domain-containing protein		
At3g25540	95	LONGEVITY ASSURANCE GENE 1 (LAG1)		N-terminal protein myristoylation
At5g39610	95	NAC domain containing protein 6 (NAC6)		
At5g13750	87	zinc induced facilitator-like 1 (ZIFL1)	tetracycline:hydrogen antiporter activity	response to karrikin
At5g57565	82	Protein kinase superfamily protein	protein serine/threonine kinase activity, protein kinase activity, ATP binding	protein amino acid phosphorylation

Table J.1 – Hub genes identified by multi-modelling for the gene pool generated using SplineCluster

Once multi-modelling was completed for the gene list produced using SplineCluster as a guide, 118 genes were identified as potential hubs on the basis that, at a 95% confidence level, they were observed to regulate two or more genes on more than 50% of the occasions that those genes were present in the same model as the hub gene. The number of genes from the pool of 580 found to be regulated by them is listed and this is used as the sorting criteria for the table. Where an entry existed in the TAIR10 database the description, function and process they are involved in is shown.

AGI No.	Upstream of	Description	Functions In	Involved In
At2g40340	79	DREB2C		
At3g15510	79	NAC domain containing protein 2 (NAC2)	sequence-specific DNA binding transcription factor activity	multicellular organismal development, regulation of transcription
At1g05055	77	general transcription factor II H2 (GTF2H2)		
At5g38140	72	nuclear factor Y, subunit C12 (NF-YC12)	DNA binding, sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-binding transcription factor activity
At2g22760	69	basic helix-loop-helix (bHLH) DNA-binding superfamily protein	DNA binding, sequence-specific DNA binding transcription factor activity	regulation of transcription
At3g12250	58	TGACG motif-binding factor 6 (TGA6)		
At5g37200	57	RING/U-box superfamily protein		
At4g35985	54	Senescence/dehydration-associated protein-related		
At1g51500	51	ECERIFERUM 5 (CER5)	ATPase activity, coupled to transmembrane movement of substances	wax biosynthetic process
At3g60010	47	SKP1-like 13 (SK13)	ubiquitin-protein ligase activity, protein binding	ubiquitin-dependent protein catabolic process
At1g53170	46	ethylene response factor 8 (ERF8)		
At4g13100	45	RING/U-box superfamily protein		
At3g16370	40	GDSL-like Lipase/Acylhydrolase superfamily protein	hydrolase activity, acting on ester bonds, carboxylesterase activity	lipid metabolic process
At2g16480	38	Unknown		
At1g20440	37	cold-regulated 47 (COR47)		
At1g76720	37	eukaryotic translation initiation factor 2 (eIF-2) family protein	GTP binding, GTPase activity, translation initiation factor activity	translational initiation
At2g42880	34	MAP kinase 20 (MPK20)		
At1g64060	33	respiratory burst oxidase protein F (RBOH F)	NAD(P)H oxidase activity	in 10 processes
At1g11850	30	Unknown		
At5g13920	28	GRF zinc finger / Zinc knuckle protein	zinc ion binding, nucleic acid binding	biological process unknown
At3g24730	27	mRNA splicing factor, thioredoxin-like U5 snRNP	catalytic activity	mitosis
At3g19820	26	DWARF 1 (DWF1)	calmodulin binding, catalytic activity	response to light stimulus, steroid biosynthetic process, unidimensional cell growth, brassinosteroid biosynthetic process

Table J.1 – Continued from previous page

AGI No.	Upstream of	Description	Functions In	Involved In
At2g01110	24	ALBINO AND PALE GREEN 2 (APG2)	proton motive force dependent protein transmembrane transporter activity	thylakoid membrane organization, double fertilization forming a zygote and endosperm
At2g22540	24	SHORT VEGETATIVE PHASE (SVP)		
At2g40940	24	ethylene response sensor 1 (ERS1)	ethylene binding, protein histidine kinase activity, receptor activity	negative regulation of ethylene mediated signaling pathway
At4g23750	24	cytokinin response factor 2 (CRF2)		
At4g24540	21	AGAMOUS-like 24 (AGL24)	sequence-specific DNA binding, protein homodimerization activity, protein heterodimerization activity, sequence-specific DNA binding transcription factor activity	floral whorl development, maintenance of floral meristem identity, maintenance of inflorescence meristem identity, floral meristem determinacy, regulation of timing of transition from vegetative to reproductive phase
At5g40020	19	Pathogenesis-related thaumatin superfamily protein		response to other organisms
At3g29035	15	NAC domain containing protein 3 (NAC3)		
At4g14560	15	indole-3-acetic acid inducible (IAA1)		
At5g47910	15	respiratory burst oxidase homologue D (RBOHD)	NAD(P)H oxidase activity	oxygen and reactive oxygen species metabolic process, defense response to fungus, response to heat, defense response, negative regulation of programmed cell death
At3g04420	13	NAC domain containing protein 48 (NAC048)	sequence-specific DNA binding transcription factor activity	multicellular organismal development, regulation of transcription
At4g13670	13	plastid transcriptionally active 5 (PTAC5)	unfolded protein binding, heat shock protein binding	protein folding, metabolic process
At4g17500	13	ethylene responsive element binding factor 1 (ERF-1)		
At2g25930	12	EARLY FLOWERING 3 (ELF3)		
At4g32280	12	indole-3-acetic acid inducible 29 (IAA29)	sequence-specific DNA binding transcription factor activity	response to auxin stimulus, response to cyclopentenone, response to red light, response to far red light
At1g17420	11	lipoxygenase 3 (LOX3)	oxidoreductase activity, acting on single donors with incorporation of molecular oxygen, incorporation of two atoms of oxygen, lipoxygenase activity, iron ion binding, metal ion binding	growth, response to fungus, jasmonic acid biosynthetic process, response to wounding, defense response

Table J.1 – Continued from previous page

AGI No.	Upstream of	Description	Functions In	Involved In
At3g60390	11	homeobox-leucine zipper protein 3 (HAT3)	sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent, transcription, regulation of transcription
At2g40740	10	WRKY DNA-binding protein 55 (WRKY55)		
At1g71190	9	senescence associated gene 18 (SAG18)	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amides	ceramide metabolic process
At3g01470	9	homeobox 1 (HB-1)	sequence-specific DNA binding, DNA binding, transcription activator activity, protein homodimerization activity, sequence-specific DNA binding transcription factor activity	response to salt stress, response to blue light, positive regulation of transcription, regulation of transcription, DNA-dependent, leaf morphogenesis
At3g24440	9	VERNALIZATION 5 (VRN5)		histone methylation, chromatin modification, vernalization response, regulation of flower development
At5g25190	9	Integrase-type DNA-binding superfamily protein		
At5g45040	9	Cytochrome c	electron carrier activity, iron ion binding, heme binding	
At1g53160	8	squamosa promoter binding protein-like 4 (SPL4)	DNA binding, sequence-specific DNA binding transcription factor activity	regulation of vegetative phase change, regulation of transcription
At4g23810	8	WRKY53	protein binding, DNA binding, transcription activator activity, sequence-specific DNA binding transcription factor activity	in 6 processes
At4g36540	8	BR enhanced expression 2 (BEE2)	DNA binding, sequence-specific DNA binding transcription factor activity	regulation of transcription
At1g22190	7	Integrase-type DNA-binding superfamily protein	DNA binding, sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent
At2g28820	7	Integrase-type DNA-binding superfamily protein	sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent
At3g11650	7	NDR1/HIN1-like 2 (NHL2)		
At5g24800	7	basic leucine zipper 9 (BZIP9)	DNA binding, protein heterodimerization activity, sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent

Table J.1 – Continued from previous page

AGI No.	Upstream of	Description	Functions In	Involved In
At1g04430	6	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein		
At2g47450	6	CHAOS (CAO)	chromatin binding	response to high light intensity, protein import into chloroplast thylakoid membran
At3g04260	6	plastid transcriptionally active 3 (PTAC3)	DNA binding, nucleic acid binding	
At3g50970	6	LOW TEMPERATURE-INDUCED 30 (LTI30)		
At3g51910	6	heat shock transcription factor A7A (HSA7A)	DNA binding, sequence-specific DNA binding transcription factor activity	response to high light intensity, response to heat, regulation of cellular protein metabolic proces
At4g24120	6	YELLOW STRIPE like 1 (YSL1)		
At5g03730	6	CONSTITUTIVE TRIPLE RESPONSE 1 (CTR1)		
At5g57390	6	AINTEGUMENTA-like 5 (AIL5)		
At5g64340	6	SUPPRESSOR OF ACAULIS 51 (SAC51)		
At1g02130	5	RAS 5 (RA-5)	GTP binding	response to cadmium ion, ER to Golgi vesicle-mediated transport
At1g74840	5	Homeodomain-like superfamily protein	DNA binding, sequence-specific DNA binding transcription factor activity	
At1g07050	4	CCT motif family protein		
At1g09530	4	phytochrome interacting factor 3 (PIF3)		
At1g15690	4	AVP1	hydrogen-translocating pyrophosphatase activity	response to water deprivation, response to salt stress, establishment or maintenance of transmembrane electrochemical gradient, leaf development, auxin polar transport
At1g49010	4	Duplicated homeodomain-like superfamily protein		
At1g66390	4	myb domain protein 90 (MYB90)		
At1g73730	4	ETHYLENE-INSENSITIVE3-like (EIL3)	3 DNA binding, sequence-specific DNA binding transcription factor activity	ethylene mediated signaling pathway, cellular response to sulfate starvation, regulation of sulfur metabolic process, regulation of transcriptio
At1g80920	4	J8	unfolded protein binding, heat shock protein binding	protein folding, response to stress
At2g30360	4	SOS3-interacting protein 4 (SIP4)		

Table J.1 – Continued from previous page

AGI No.	Upstream of	Description	Functions In	Involved In
At3g20910	4	nuclear factor Y, subunit A9 (NF-YA9)	sequence-specific DNA binding transcription factor activity, specific transcriptional repressor activity	negative regulation of gene-specific transcription, regulation of DNA-dependent
At3g52150	4	RNA-binding (RRM/RBD/RNP motifs) family protein	RNA binding, nucleotide binding, nucleic acid binding	
At4g10710	4	global transcription factor C (SPT16)		vegetative to reproductive phase transition of meriste
At5g36940	4	cationic amino acid transporter 3 (CAT3)	cationic amino acid transmembrane transporter activity	transport, amino acid transport, N-terminal protein myristoylation, transmembrane transport
At1g29280	3	WRKY DNA-binding protein 65 (WRKY65)	sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent, regulation of transcription
At1g49130	3	B-box type zinc finger protein with CCT domain	sequence-specific DNA binding transcription factor activity, zinc ion binding	regulation of transcription
At1g65480	3	FLOWERING LOCUS T (FT)	phosphatidylethanolamine binding, protein binding	photoperiodism, flowering, positive regulation of flower development, regulation of flower development
At1g78080	3	related to AP2 4 (RAP2.4)		
At4g22130	3	STRUBBELIG-receptor family 8 (SRF8)	protein serine/threonine kinase activity, protein kinase activity, kinase activity, ATP binding	protein amino acid phosphorylation
At5g39830	3	DEG8	serine-type peptidase activity, peptidase activity	photosystem II repair, proteolysis
At1g01200	2	RAB GTPase homolog A3 (RABA3)	GTP binding	protein transport, small GTPase mediated signal transduction
At1g19350	2	BRH-EMS-SUPPRESSOR 1 (BES1)		
At1g22640	2	myb domain protein 3 (MYB3)		
At1g66340	2	ETHYLENE RESPONSE 1 (ETR1)		
At1g67260	2	TCPI		
At1g74430	2	myb domain protein 95 (MYB95)		
At2g34720	2	nuclear factor Y, subunit A4 (NF-YA4)		
At2g42520	2	P-loop containing nucleoside triphosphate hydrolases superfamily protein	helicase activity, nucleic acid binding, ATP binding, ATP-dependent helicase activity	
At2g47090	2	zinc ion binding		

Table J.1 – Continued from previous page

AGI No.	Upstream of	Description	Functions In	Involved In
At3g50260	2	cooperatively regulated by ethylene and jasmonate 1 (CEJ1)		
At3g58120	2	BZIP61	DNA binding, transcription activator activity, sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-
At4g29100	2	basic helix-loop-helix (bHLH) DNA-binding superfamily protein	sequence-specific DNA binding transcription factor activity	regulation of transcriptio
At5g20010	2	RAS-related nuclear protein-1 (RAN-1)	protein binding, GTP binding, GTPase activity	response to cadmium ion, response to salt stress, protein import into nucleu
At5g38660	2	ACCLIMATION OF PHOTOSYNTHESIS TO ENVIRONMENT (APE1)		
At5g39660	2	cycling DOF factor 2 (CDF2)		
At5g50580	2	SUMO-activating enzyme 1B (SAE1B)		
At5g57900	2	SKP1 interacting partner 1 (SKIPI)		
At5g61380	2	TIMING OF CAB EXPRESSION 1 (TOC1)		
At5g61850	2	LEAFY (LFY)		

Table J.1 – Continued from previous page

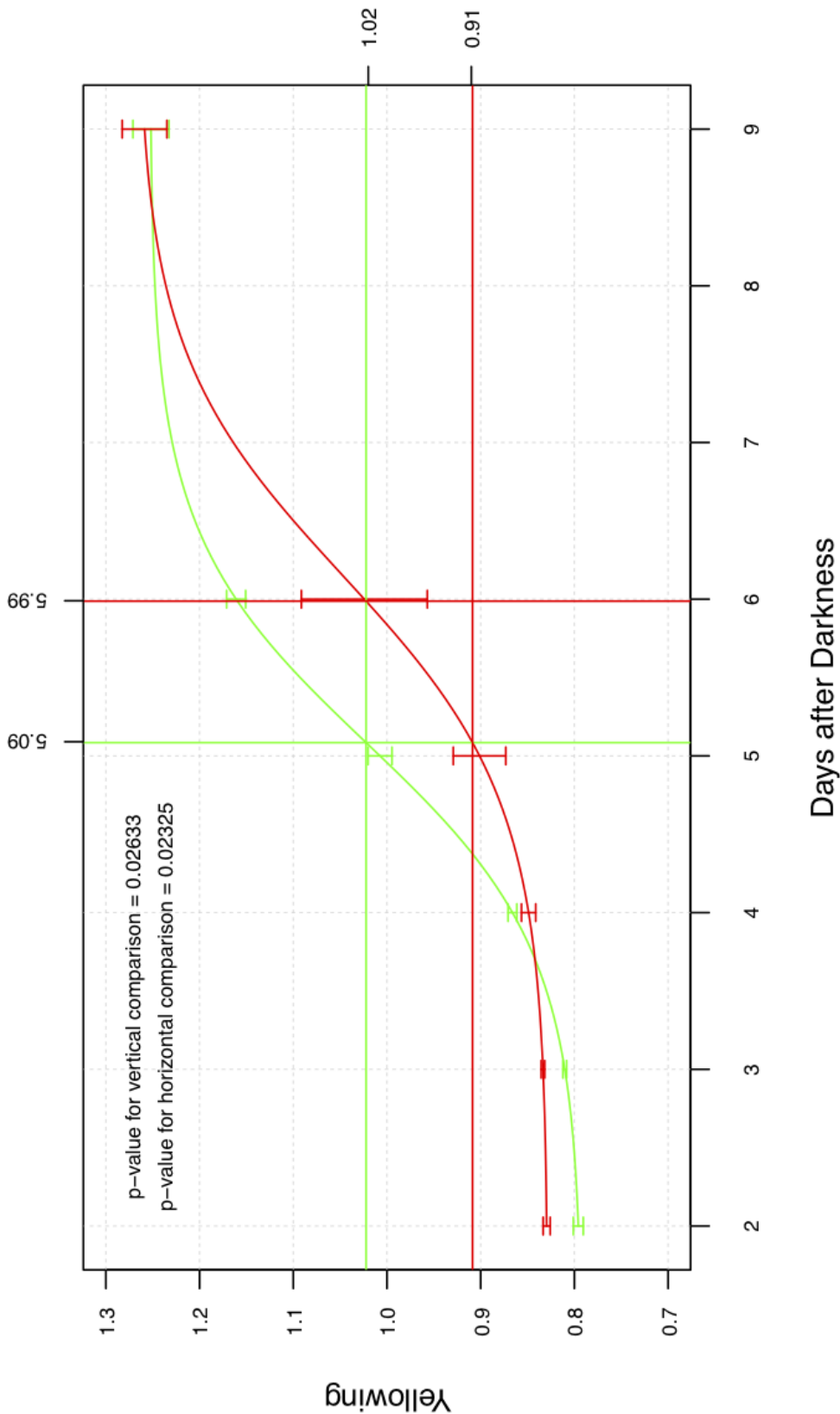
Appendix K

Logistic Curve Plots for Quantitative Phenotype Screens

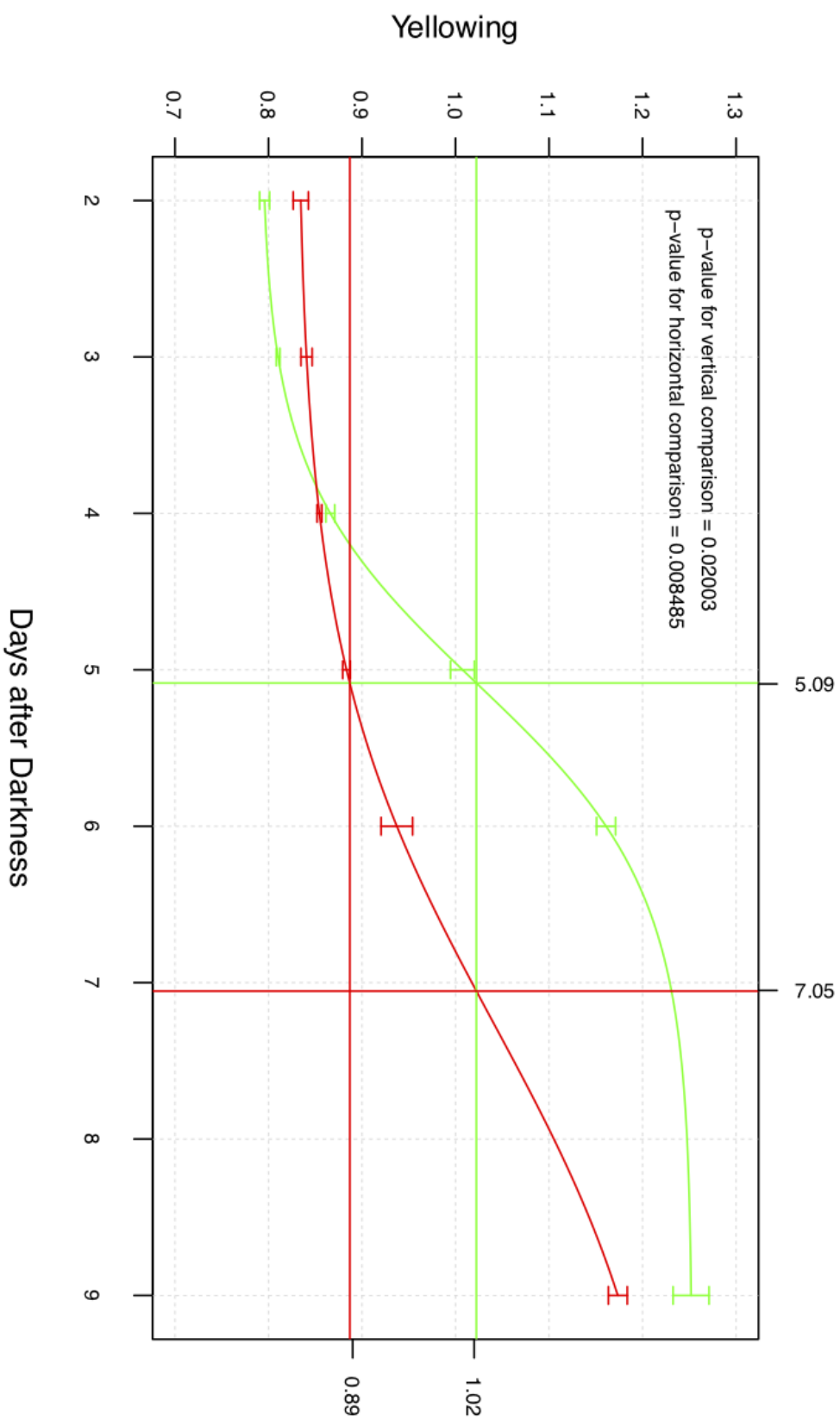
The plots shown on the following pages are the logistic curves fit to data collected during the phenotype screening of altered expression mutants. The horizontal axis of each plot represents the time at which observations were made whilst the vertical axis of each plot represents the normalised level of yellowing in leaf 4 of the rosette. The green line defines the logistic curve which best fits the yellowing data of the wild type whilst the red line defines the logistic curve which best fits the yellowing data of the altered expression mutant. Error bars are based on one standard error of the three biological replicates. The horizontal and vertical green lines indicates the point of maximum rate of change (the mid-point) in the wild type curve. The horizontal red line indicates the level of yellowing in the altered expression mutant at the time the wild type reaches maximum rate of change (i.e. the point at which the vertical green line intersects the red curve). The vertical red line indicates the time at which the altered expression mutant is as yellow as the wild type was at its maximum rate of change (i.e. the point at which the horizontal green line intersects the red curve).

The p -value for vertical comparison indicates the significance of the difference between the two levels of yellowing as indicated by the horizontal lines. The p -value for horizontal comparison indicates the significance of the difference between the two time stages indicated by the vertical lines. Where both these values satisfy $p \leq 0.05$, the phenotypical response is considered to be significantly altered in the altered expression mutant. Only those mutants which demonstrated a significantly altered phenotypical response are shown.

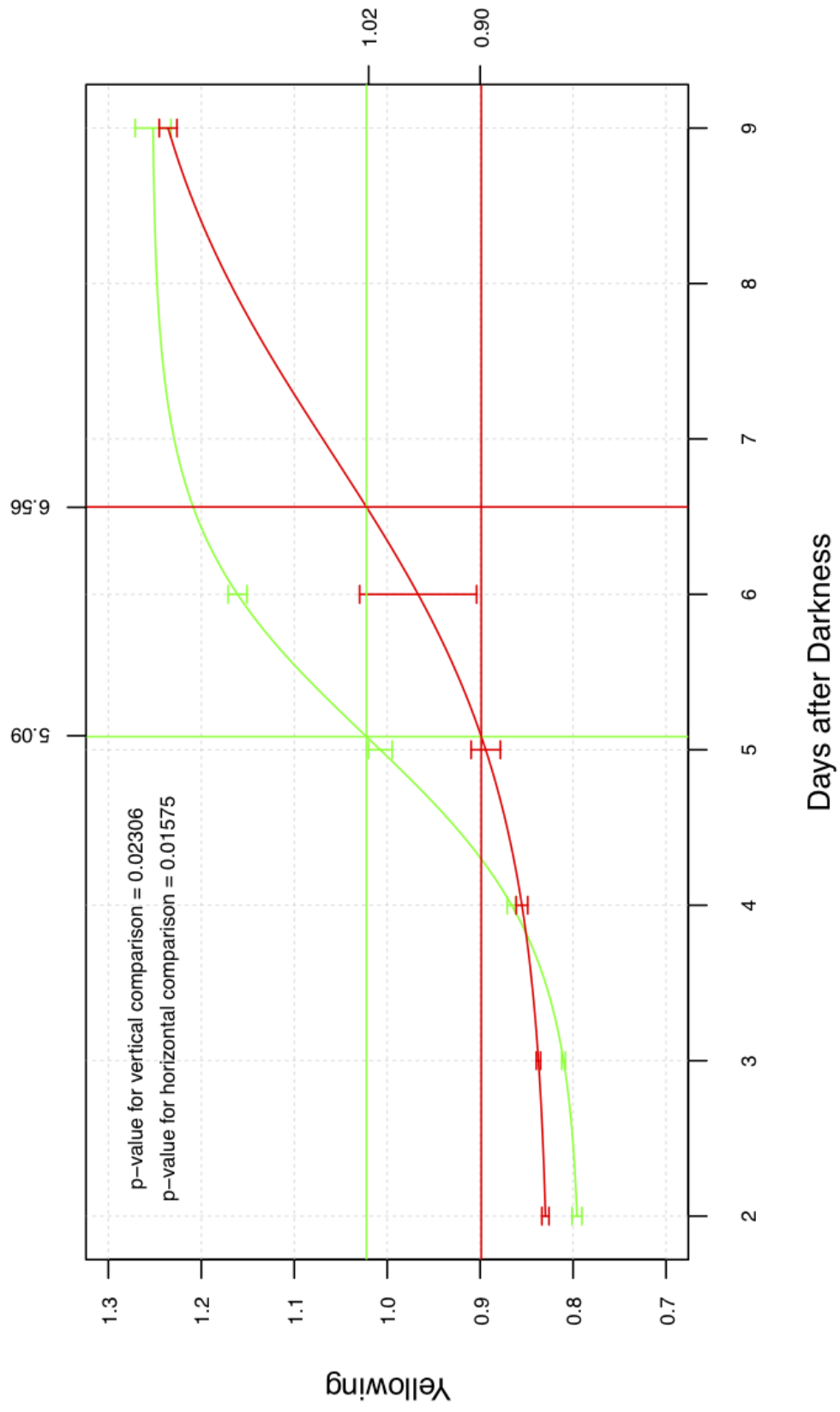
IM137 - At1g27730 - STZ



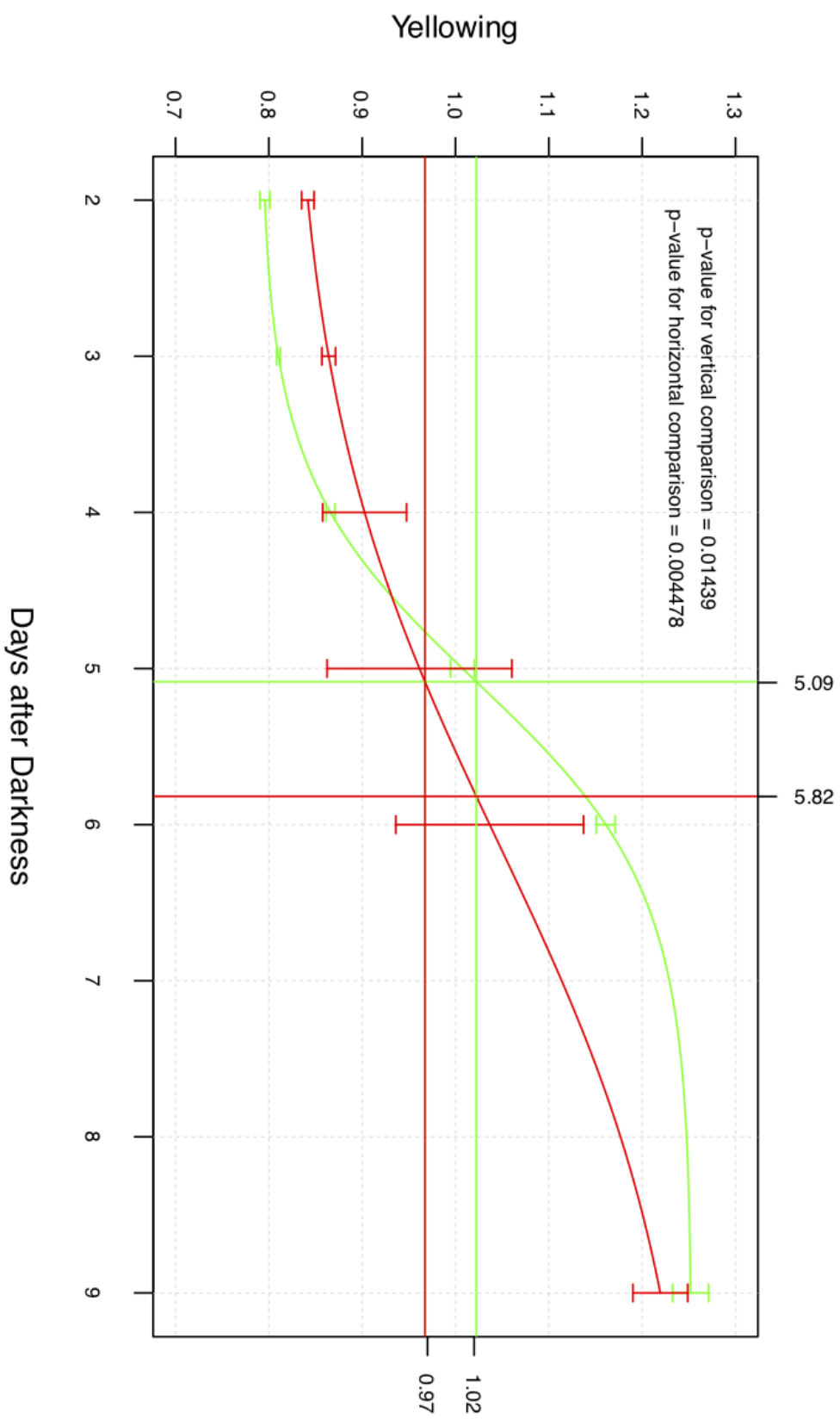
IM147 - At1g69690 - TCP family transcription factor



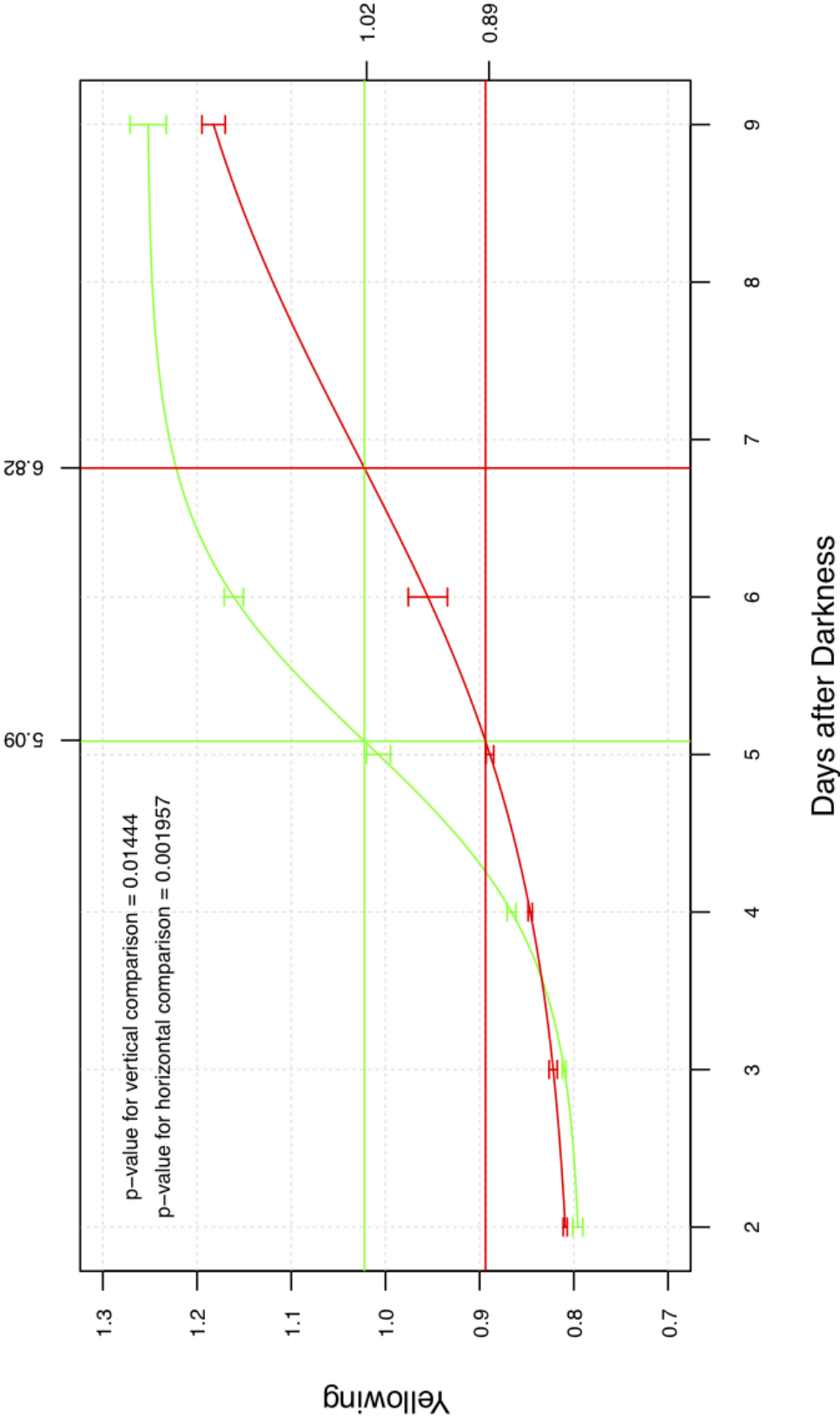
IM160 - At2g38880 - HAP3A/AtNF-YB1



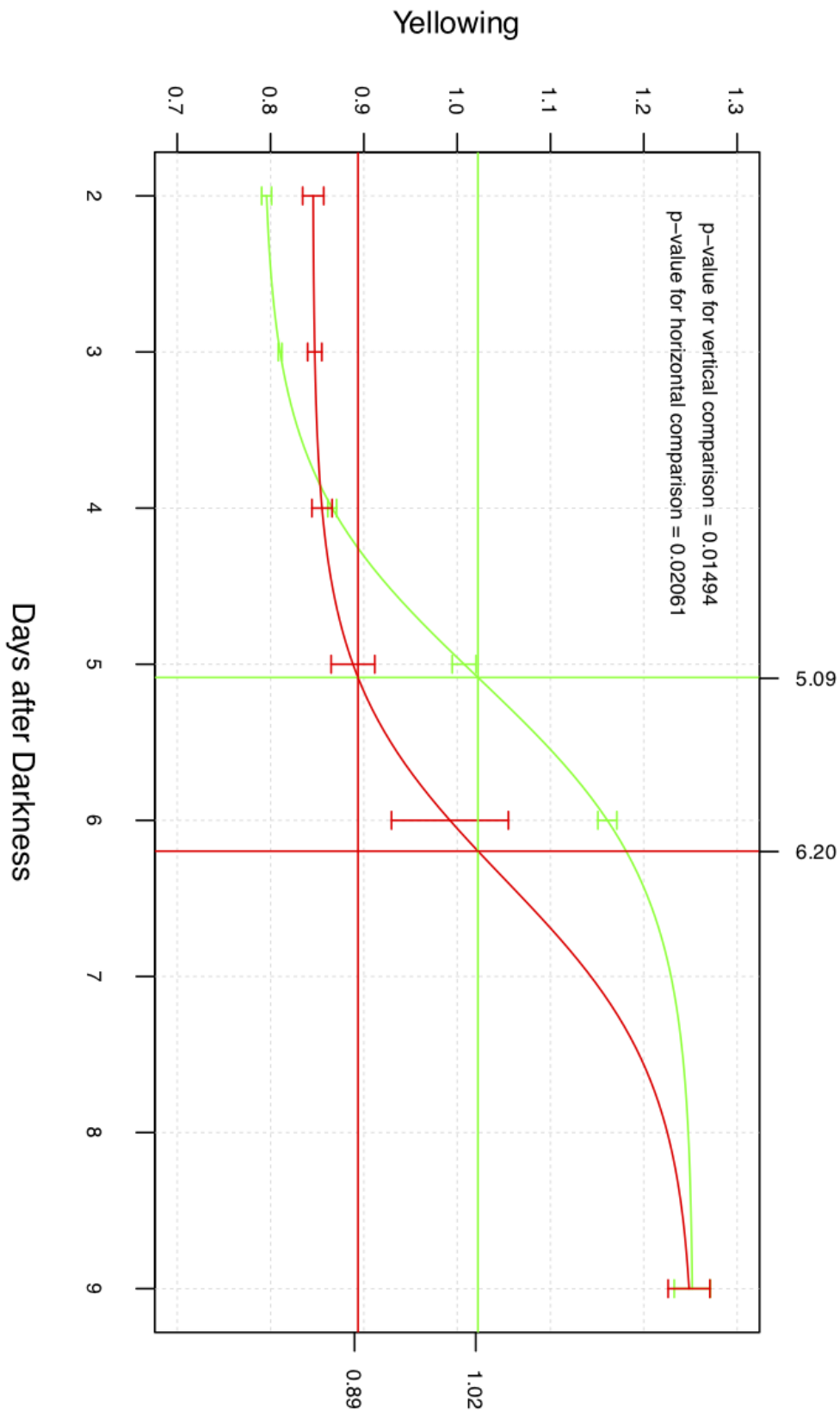
IM175 - At3g60390 - HAT3



IM193 - At5g62040 - Brother of FT and TFL1



TP010 - At4g05100 - MYB74



Appendix L

Yeast-1-Hybrid Sequencing Results

The following table shows which colonies were able to grow on a selective media of SD-Leu-Trp-His during yeast-1-hybrid screening. The transcription factor library was arranged over four plates and so the library number is given to show which of the selective plates provided a positive result and the position of the colony on that plate is also given. Library numbers 001 and 002 represent the transcription factor library in one orientation whilst 101 and 102 represent the same library in an alternative orientation and so where results are duplicated between both libraries, this may indicate a more positive result. The best BLAST result was identified by BLASTing the sequence obtained from BigDye sequencing of the transcription factor found in the pDEST22 plasmid. Only entries that gave a good sequencing result are shown: results with many ambiguous bases or no results at all were removed. Where multiple lines show the same information, this indicates that more than one colony was found to be growing and were found to contain the same transcription factor.

Fragment Number	Library Number	Plate Position	Best BLAST Result
1	001	E03	TCP3 mRNA, complete cds
1	001	F03	AT-TCP20
1	001	F10	AtTCP14
1	001	F10	AtTCP14
1	001	F10	AtTCP14
1	001	H03	TCP8
1	002	G03	TCP15
1	101	B04	AT-TCP20
1	101	G02	AtTCP14
1	102	B08	TCP15
1	102	B08	TCP15
1	102	E12	TCP4 (TCP family TF 4)
2	001	E03	TCP3

Continued on next page...

Table L.1 – continued from previous page

Fragment Number	Library Number	Plate Position	Best BLAST Result
2	001	F10	TCP14
2	001	F10	TCP14
2	001	F10	TCP14
2	002	G03	TCP14
2	002	G03	TCP15
2	101	B03	TCP3 mRNA
2	101	B03	TCP3 mRNA
2	101	B03	TCP3 mRNA
2	101	B04	AT-TCP20
2	101	B05	TCP8
2	101	B05	TCP8
2	101	G02	TCP14
2	101	G02	TCP14
2	101	G02	TCP14
2	101	G02	TCP14
2	102	B09	TCP14
3	001	E03	TCP3
3	001	E03	TCP3
3	001	F03	AT-TCP20
3	001	F03	AT-TCP20
3	001	F03	AT-TCP20
3	001	F10	TCP14
3	001	F10	TCP14
3	001	F10	TCP14
3	002	G03	TCP15
3	002	G03	TCP15
3	002	G03	TCP15
3	002	G03	TCP15
3	101	B03	TCP3
3	101	B03	TCP3
3	101	B03	TCP3
3	101	B03	TCP3
3	101	B03	TCP3
3	101	B03	TCP3
3	101	B03	TCP3
3	101	B03	TCP3
3	101	B03	TCP3
3	101	B04	AT-TCP20
3	101	B04	AT-TCP20
3	101	B12	TCP4 (TCP family TF 4)
3	101	B12	TCP4 (TCP family TF 4)
3	102	B08	TCP15
3	102	B08	TCP15
3	102	B08	TCP15
4	001	D02	PIF7 (PHYTOCHROME-INTERACTING FACTOR7)
4	001	D02	PIF7 (PHYTOCHROME-INTERACTING FACTOR7)
4	001	E03	TCP3

Continued on next page...

Table L.1 – continued from previous page

Fragment Number	Library Number	Plate Position	Best BLAST Result
4	001	E03	TCP3
4	001	F03	TCP16 (TCP domain protein 16)
4	001	F03	AT-TCP20
4	001	F10	TCP14
4	001	F10	TCP14
4	001	F10	TCP14
4	001	F10	TCP14
4	002	G03	TCP15
4	002	G03	TCP15
4	002	G03	TCP15
4	002	G07	ANAC102
4	101	A11	PIF7 (PHYTOCHROME-INTERACTING FACTOR7)
4	101	B01	TCP1
4	101	B03	TCP3
4	101	B04	AT-TCP20
4	101	B04	AT-TCP20
4	101	B04	AT-TCP20
4	101	B04	AT-TCP20
4	101	B05	TCP8
4	101	B05	TCP8
4	101	G02	TCP14
4	102	B05	ILR3 (iaa-leucine resistant3)
4	102	B05	ILR3 (iaa-leucine resistant3)
4	102	B05	ILR3 (iaa-leucine resistant3)
4	102	B05	ILR3 (iaa-leucine resistant3)
4	102	B05	ILR3 (iaa-leucine resistant3)
4	102	B08	TCP15
4	102	B08	TCP15
4	102	B08	TCP15
4	102	B08	TCP15
4	102	B09	TCP14
4	102	B09	TCP14
4	102	B09	TCP14
5	001	A12	ANAC038
5	001	A12	ANAC038
5	001	E02	CUC2 (CUP-SHAPED COTYLEDON 2)
5	001	E03	TCP3
5	001	E03	TCP3
5	001	F03	AT-TCP20
5	001	F03	AT-TCP20
5	001	F03	AT-TCP20
5	001	F03	AT-TCP20
5	001	F03	AT-TCP20
5	001	F10	TCP14
5	001	F10	TCP14
5	001	F10	TCP14
5	001	F10	TCP14

Continued on next page...

Table L.1 – continued from previous page

Fragment Number	Library Number	Plate Position	Best BLAST Result
5	101	B02	TCP23
5	101	B02	TCP23
5	101	B03	TCP3
5	101	B03	TCP3
5	101	B04	AT-TCP20
5	101	B04	AT-TCP20
5	101	B04	AT-TCP20
5	101	B04	AT-TCP20
5	101	B04	AT-TCP20
5	101	B05	TCP8
5	101	B05	TCP8
5	101	B05	TCP8
5	101	G02	TCP14
5	102	B08	TCP15
5	102	B08	TCP15
5	102	B08	TCP15
5	102	B08	TCP15
5	102	B08	TCP15
5	102	E12	TCP4 (TCP family TF 4)
5	102	E12	TCP4 (TCP family TF 4)
6	001	F03	AT-TCP20
6	001	F03	AT-TCP20
6	001	F03	AT-TCP20
6	002	G03	TCP14
6	101	B02	TCP23
6	101	B03	TCP3
6	101	B03	TCP3
6	101	B04	AT-TCP20
6	101	B04	AT-TCP20
6	101	B04	AT-TCP20
6	101	B04	AT-TCP20
6	102	B08	TCP15
6	102	B08	TCP15
6	102	B08	TCP15
6	102	B08	TCP15
6	102	B09	TCP14
7	001	F10	TCP14
7	001	F10	TCP14
7	002	G03	TCP15
7	002	G03	TCP14
7	101	B03	TCP14
7	101	B03	TCP3
7	101	B03	TCP3
7	101	B03	TCP3
7	101	B03	TCP3
7	101	B03	TCP3
7	101	B03	TCP3

Continued on next page...

Table L.1 – continued from previous page

Fragment Number	Library Number	Plate Position	Best BLAST Result
7	101	B04	AT-TCP20
7	101	B04	AT-TCP20
7	101	B12	TCP4 (TCP family TF 4)
7	101	B12	TCP4 (TCP family TF 4)
7	101	G02	TCP14
7	101	G02	TCP14
7	101	G02	TCP14
7	101	G02	TCP14
7	101	G02	TCP14
7	101	G02	TCP14
7	102	B08	TCP15
7	102	B08	TCP15
7	102	B08	TCP15
7	102	B09	TCP14
7	102	E12	TCP4 (TCP family TF 4)
7	102	E12	TCP4 (TCP family TF 4)
7	102	E12	TCP4 (TCP family TF 4)
8	001	E02	CUC2 (CUP-SHAPED COTYLEDON 2)
8	001	F10	TCP14
8	001	F10	TCP14
8	001	F10	TCP14
8	101	B03	TCP3
8	101	B03	TCP3
8	101	B03	TCP3
8	101	G02	TCP14
8	101	G02	TCP14
8	101	G02	TCP14
8	102	B09	TCP14
8	102	E12	TCP4 (TCP family TF 4)

Appendix M

Balanced Design Source Code

The source code shown is written in R and produces every possible permutation of 15 samples in a 30 microarray design and tests each to find which are the most balanced.

```
1 require(gtools)
2 skeleton <- read.table("5x3sampleSkeleton.txt", header=T, sep="\t")
3
4 testSampleBalance <- function(cy3Samples, cy5Samples) {
5   # Check that each sample exists equally on both channels
6   dyeTable <- table(c(rep(c("Cy3", "Cy5"), each=length(cy3Samples))), c(
7     cy3Samples, cy5Samples))
8   if (any(dyeTable != 2)) {
9     return(FALSE)
10  }
11
12  # Check that the samples are connected only once to samples of other lines
13  lineTable <- table(c(skeleton$Cy3Line, skeleton$Cy5Line), c(cy5Samples,
14    cy3Samples))
15  if (max(lineTable) > 1) {
16    cat(".")
17    return(FALSE)
18  }
19
20  return(TRUE)
21 }
22
23 setwd("./BalancedSampleDesigns")
24
25 # Produce sample layout
26 design <- 1
27 cy3Samples <- cy5Samples <- rep(0, 30)
28 cy3Samples[1:6] <- cy5Samples[7:12] <- rep(1:3, 2)
29
30 permsOf3 <- permutations(3,3)
31 nPerms3 <- nrow(permsOf3)
32 permsOfPerms <- permutations(nPerms3, 4, , , T)
33 first12 <- t(matrix(t(permsOf3[t(permsOfPerms),]), nrow=ncol(permsOfPerms) * ncol(
34   permsOf3), ncol=nrow(permsOfPerms)))
35 for (i in seq(3, 12, 3)) {
36   first12[, (i-2):i] <- first12[, (i-2):i] + i
37 }
38 first12 <- first12[, c(1,2,4,7,8,10,5,6,3,11,12,9)]
39
```

```

37 for (i in sample(1:nrow(first12), 10)) {
38   cat("Starting", i, "of", nrow(first12), "at", format(Sys.time(), "%H:%M:%S on
      %d-%m-%Y"), "...\\n")
39
40   # Arrange first 12 comparisons
41   cy5Samples[1:6] <- first12[i,1:6]
42   cy3Samples[7:12] <- first12[i,7:12]
43
44   # Update left
45   cy3Samples[13:30] <- cy5Samples[13:30] <- rep(0, 18)
46   cy3Left <- 2 - summary(factor(cy3Samples, levels=1:15))[1:15]
47   cy5Left <- 2 - summary(factor(cy5Samples, levels=1:15))[1:15]
48
49   # Prepare remaining B samples
50   cy3B <- rep(4:6, cy3Left[4:6])
51   cy5B <- rep(4:6, cy5Left[4:6])
52   cy3Perms <- permutations(length(cy3B), length(cy3B), cy3B, F)
53   cy5Perms <- permutations(length(cy5B), length(cy5B), cy5B, F)
54   x <- c(skeleton$Cy5Line[13:17], skeleton$Cy3Line[18:21])
55   bPerms <- NULL
56   for (v in 1:nrow(cy3Perms)) {
57     for (w in 1:nrow(cy5Perms)) {
58       y <- c(cy3Perms[v,], cy5Perms[w,])
59       if (var(as.vector(table(x,y))) == 0) {
60         bPerms <- rbind(bPerms, y)
61       }
62     }
63   }
64   for (iBPerm in sample(1:nrow(bPerms), 10)) {
65     cy3Samples[13:17] <- bPerms[iBPerm,1:5]
66     cy5Samples[18:21] <- bPerms[iBPerm,6:9]
67     permsOfPerms <- permutations(nPerms3, 3, , , T)
68     next9 <- t(matrix(t(permsOf3[t(permsOfPerms),]), nrow=ncol(permsOfPerms)
      * ncol(permsOf3), ncol=nrow(permsOfPerms)))
69     for (i in seq(3, 9, 3)) {
70       next9[, (i-2):i] <- next9[, (i-2):i] + i + 3
71     }
72     next9 <- next9[, c(1,2,7,8,4,3,5,6,9)]
73     for (j in sample(1:nrow(next9), 10)) {
74       cy5Samples[13:17] <- next9[j,1:5]
75       cy3Samples[18:21] <- next9[j,6:9]
76
77     # Update left
78     cy3Samples[22:30] <- cy5Samples[22:30] <- rep(0, 9)
79     cy3Left <- 2 - summary(factor(cy3Samples, levels=1:15))[1:15]
80     cy5Left <- 2 - summary(factor(cy5Samples, levels=1:15))[1:15]
81     if(any(c(cy3Left, cy5Left) < 0)) next
82
83     # Prepare remaining C samples
84     cy3C <- rep(7:9, cy3Left[7:9])
85     cy5C <- rep(7:9, cy5Left[7:9])
86     cy3Perms <- permutations(length(cy3C), length(cy3C), cy3C, F)
87     cy5Perms <- permutations(length(cy5C), length(cy5C), cy5C, F)
88     x <- c(skeleton$Cy5Line[22:24], skeleton$Cy3Line[25:27])
89     cPerms <- NULL
90     for (v in 1:nrow(cy3Perms)) {
91       for (w in 1:nrow(cy5Perms)) {
92         y <- c(cy3Perms[v,], cy5Perms[w,])
93         if (var(as.vector(table(x,y))) == 0) {

```

```

94         cPerms <- rbind(cPerms, y)
95     }
96 }
97 }
98 for (iCPerm in sample(1:nrow(cPerms), 10)) {
99     cy3Samples[22:24] <- cPerms[iCPerm,1:3]
100     cy5Samples[25:27] <- cPerms[iCPerm,4:6]
101
102     permsOfPerms <- permutations(nPerms3, 2, , , T)
103     next6 <- t(matrix(t(permsOf3[t(permsOfPerms),]), nrow=ncol(
104         permsOfPerms) * ncol(permsOf3), ncol=nrow(permsOfPerms)))
105     for (i in seq(3, 6, 3)) {
106         next6[, (i-2):i] <- next6[, (i-2):i] + i + 6
107     }
108     next6 <- next6[, c(1,2,4,5,6,3)]
109     last3 <- next6[, c(4:6,1:3)]
110
111     for (k in sample(1:nrow(next6), 10)) {
112         cy5Samples[22:24] <- next6[k,1:3]
113         cy3Samples[25:27] <- next6[k,4:6]
114
115         # Update left
116         cy3Samples[28:30] <- cy5Samples[28:30] <- rep(0, 3)
117         cy3Left <- 2 - summary(factor(cy3Samples, levels=1:15))[1:15]
118         cy5Left <- 2 - summary(factor(cy5Samples, levels=1:15))[1:15]
119         if(any(c(cy3Left, cy5Left) < 0) | max(cy3Left) > 1 | max(
120             cy5Left) > 1) next
121
122         for (l in 1:nrow(last3)) {
123             cy5Samples[28:30] <- last3[l,1:3]
124             cy3Samples[28:30] <- last3[l,4:6]
125
126             # Update left
127             cy3Left <- 2 - summary(factor(cy3Samples, levels=1:15))
128                 [1:15]
129             cy5Left <- 2 - summary(factor(cy5Samples, levels=1:15))
130                 [1:15]
131             if(any(c(cy3Left, cy5Left) != 0)) next
132             if (testSampleBalance(cy3Samples, cy5Samples)) {
133                 write.table(data.frame(Cy3Sample=cy3Samples,
134                     Cy5Sample=cy5Samples), file=sprintf("%08d.txt",
135                     design), sep="\t", row.names=F, quote=F)
136                 cat("Design number", design, "written out\n")
137                 design <- design + 1
138             }
139         }
140     }
141 }
142 }
143 }
144 }
145 }
146 }
147 }
148 }

```

The results of the above script can then be used to find which assignment of biological replicates to the samples produces the most balanced comparison between replicates.

```

1 require(gtools)
2
3 testBioReps <- function(cy3Samples, cy3BioReps, cy5Samples, cy5BioReps) {
4   # Check that all bioreps are linked equally to each other overall
5   globalScore <- var(as.vector(table(cy3BioReps, cy5BioReps)))
6
7   # Check for minimisation of variance in bioreps connected to each sample
8   sampleTable <- table(c(cy3Samples, cy5Samples), c(cy5BioReps, cy3BioReps))
9   sampleScore <- sum(apply(sampleTable, 1, var))
10
11   return(c(globalScore, sampleScore))
12 }
13
14 fileSelection <- sprintf("./BalancedSampleDesigns/%08d.txt", 1:120716)
15 designFiles <- sample(fileSelection, 1000)
16
17 biorepOptions <- permutations(3,3,toupper(letters[1:3]))
18 nBR0pts <- nrow(biorepOptions)
19 biorepSelections <- permutations(nBR0pts, 5, repeats.allowed=T)
20 biorepTable <- t(matrix(t(biorepOptions[t(biorepSelections),]), nrow=ncol(
21   biorepSelections) * ncol(biorepOptions), ncol=nrow(biorepSelections)))
22 nPerms <- nrow(biorepTable)
23
24 globalScore <- sampleScore <- matrix(0, ncol=1000, nrow=nrow(biorepTable))
25 colnames(globalScore) <- colnames(sampleScore) <- sub("./BalancedSampleDesigns/",
26   "", sub(".txt", "", designFiles))
27
28 for (file in designFiles) {
29   skeleton <- read.table(file, sep="\t", header=T)
30   designNumber <- sub("./BalancedSampleDesigns/", "", sub(".txt", "", file))
31
32   cy3Samples <- as.numeric(as.vector(skeleton$Cy3Sample))
33   cy5Samples <- as.numeric(as.vector(skeleton$Cy5Sample))
34
35   for (i in 1:nPerms) {
36     cy3BioReps <- biorepTable[i,cy3Samples]
37     cy5BioReps <- biorepTable[i,cy5Samples]
38     scores <- testBioReps(cy3Samples, cy3BioReps, cy5Samples, cy5BioReps)
39     globalScore[i,designNumber] <- scores[1]
40     sampleScore[i,designNumber] <- scores[2]
41     if (i %% 1000 == 0) {
42       cat("Completed", i, "of", nPerms, "options for design", designNumber,
43         "at", format(Sys.time(), "%H:%M:%S on %d-%m-%Y"), "...\\n")
44     }
45   }
46   save(list = c("globalScore", "sampleScore"), file="Scores.RData")
47 }

```


Bibliography

- Abramoff, M. D., Magelhaes, P. J., & Ram, S. J. (2004). Image Processing with ImageJ. *Biophotonics International*, 11(7), 36–42.
- Addicott, F. T., Lynch, R. S., & Carns, H. R. (1955). Auxin gradient theory of abscission regulation. *Science*, 121(3148), 644–5.
- Äijö, T., & Lähdesmäki, H. (2009). Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics. *Bioinformatics*, 25(22), 2937–44.
- Allemeersch, J., Durinck, S., Vanderhaeghen, R., Alard, P., Maes, R., Seeuws, K., Bogaert, T., Coddens, K., Deschouwer, K., van Hummelen, P., Vuylsteke, M., Moreau, Y., Kwekkeboom, J., Wijfjes, A. H. M., May, S., Beynon, J., Hilson, P., & Kuiper, M. T. R. (2005). Benchmarking the CATMA microarray. A novel tool for Arabidopsis transcriptome analysis. *Plant Physiol*, 137(2), 588–601.
- Aloni, R., Aloni, E., Langhans, M., & Ullrich, C. I. (2006). Role of auxin in regulating Arabidopsis flower development. *Planta*, 223(2), 315–28.
- Alonso, J. M., Hirayama, T., Roman, G., Nourizadeh, S., & Ecker, J. R. (1999). EIN2, a bifunctional transducer of ethylene and stress responses in Arabidopsis. *Science*, 284(5423), 2148–52.
- Alonso, J. M., Stepanova, A. N., Leisse, T. J., Kim, C. J., Chen, H., Shinn, P., Stevenson, D. K., Zimmerman, J., Barajas, P., Cheuk, R., Gadrinab, C., Heller, C., Jeske, A., Koesema, E., Meyers, C. C., Parker, H., Prednis, L., Ansari, Y., Choy, N., Deen, H., Geralt, M., Hazari, N., Hom, E., Karnes, M., Mulholland, C., Ndubaku, R., Schmidt, I., Guzman, P., Aguilar-Henonin, L., Schmid, M., Weigel, D., Carter, D. E., Marchand, T., Risseuw, E., Brogden, D., Zeko, A., Crosby, W. L., Berry, C. C., & Ecker, J. R. (2003). Genome-wide insertional mutagenesis of Arabidopsis thaliana. *Science*, 301(5633), 653–7.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3), 403–10.
- Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408(6814), 796–815.

- Ascencio-Ibanez, J., Sozzani, R., Lee, T., Chu, T., Wolfinger, R., Cella, R., & Hanley-Bowdoin, L. (2008). Global analysis of Arabidopsis gene expression uncovers a complex array of changes impacting pathogen response and cell cycle during geminivirus infection. *Plant Physiol*, 148(1), 436.
- Balazadeh, S., Kwasniewski, M., Caldana, C., Mehrnia, M., Zanol, M. I., Xue, G.-P., & Mueller-Roeber, B. (2011). ORS1, an H₂O₂-responsive NAC transcription factor, controls senescence in Arabidopsis thaliana. *Mol Plant*, 4(2), 346–60.
- Balazadeh, S., Riaño-Pachón, D. M., & Mueller-Roeber, B. (2008). Transcription factors regulating leaf senescence in Arabidopsis thaliana. *Plant Biol (Stuttg)*, 10 Suppl 1, 63–75.
- Balazadeh, S., Siddiqui, H., Allu, A. D., Matallana-Ramirez, L. P., Caldana, C., Mehrnia, M., Zanol, M.-I., Köhler, B., & Mueller-Roeber, B. (2010a). A gene regulatory network controlled by the NAC transcription factor ANAC092/AtNAC2/ORE1 during salt-promoted senescence. *Plant J*.
- Balazadeh, S., Wu, A., & Mueller-Roeber, B. (2010b). Salt-triggered expression of the ANAC092-dependent senescence regulon in Arabidopsis thaliana. *Plant Signaling & Behavior*, 5(6), 733–5.
- Bansal, M., Gatta, G. D., & di Bernardo, D. (2006). Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, 22(7), 815–22.
- Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R., & Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. *Nat Genet*, 37(4), 382–90.
- Baxter, L., & Ott, S. (unpublished). Analysis of Plant Promoter-Linked Elements.
- Beal, M. J. (2003). Variational Algorithms for Approximate Bayesian Inference. *Doctorate Thesis*, (p. 281).
- Beal, M. J., Falciani, F., Ghahramani, Z., Rangel, C., & Wild, D. L. (2005). A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics*, 21(3), 349–56.
- Becker, W., & Apel, K. (1993). Difference in gene expression between natural and artificially induced leaf senescence. *Planta*, 189, 74–9.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, 57(1), 289–300.

- Benjamini, Y., & Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, 25(1), 60.
- Benjamini, Y., & Liu, W. (1999). A step-down multiple hypotheses testing procedure that control the false discovery rate under independence. *Journal of Statistical Planning and Inference*, 82, 163–170.
- Benková, E., Michniewicz, M., Sauer, M., Teichmann, T., Seifertová, D., Jürgens, G., & Friml, J. (2003). Local, efflux-dependent auxin gradients as a common module for plant organ formation. *Cell*, 115(5), 591–602.
- Berger, S. (2002). Jasmonate-related mutants of Arabidopsis as tools for studying stress signaling. *Planta*, 214(4), 497–504.
- Boyes, D. C., Zayed, A. M., Ascenzi, R., McCaskill, A. J., Hoffman, N. E., Davis, K. R., & Görlach, J. (2001). Growth stage-based phenotypic analysis of Arabidopsis: a model for high throughput functional genomics in plants. *Plant Cell*, 13(7), 1499–510.
- Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., & Crawford, G. E. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2), 311–22.
- 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., & 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), 873–94.
- Breeze, E., Harrison, E., Page, T., Warner, N., Shen, C., Zhang, C., & Buchanan-Wollaston, V. (2008). Transcriptional regulation of plant senescence: from functional genomics to systems biology. *Plant Biol (Stuttg)*, 10 Suppl 1, 99–109.
- Brody, J. P., Williams, B. A., Wold, B. J., & Quake, S. R. (2002). Significance and statistical errors in the analysis of DNA microarray data. *Proc Natl Acad Sci USA*, 99(20), 12975–8.
- Brown, R. G., & Hwang, P. Y. (1997). *Introduction to Random Signals and Applied Kalman Filtering*. John Wiley & Sons, NY.
- Buchanan-Wollaston, V. (1997). The molecular biology of leaf senescence. *J Exp Bot*.
- Buchanan-Wollaston, V., Earl, S., Harrison, E., Mathas, E., Navabpour, S., Page, T., & Pink, D. (2003). The molecular analysis of leaf senescence—a genomics approach. *Plant Biotechnology Journal*, 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., & Leaver, C. J. (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), 567–85.
- Burg, S. P. (1968). Ethylene, plant senescence and abscission. *Plant Physiol*, 43(9 Pt B), 1503–11.
- Callard, D., Axelos, M., & Mazzolini, L. (1996). Novel molecular markers for late phases of the growth cycle of Arabidopsis thaliana cell-suspension cultures are expressed during organ senescence. *Plant Physiology*, 112(2), 705.
- Cantone, I., Marucci, L., Iorio, F., Ricci, M. A., Belcastro, V., Bansal, M., Santini, S., di Bernardo, M., di Bernardo, D., & Cosma, M. P. (2009). A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell*, 137(1), 172–81.
- Cao, H., Bowling, S. A., Gordon, A. S., & Dong, X. (1994). Characterization of an Arabidopsis Mutant That Is Nonresponsive to Inducers of Systemic Acquired Resistance. *Plant Cell*, 6(11), 1583–1592.
- Carabelli, M., Morelli, G., Whitelam, G., & Ruberti, I. (1996). Twilight-zone and canopy shade induction of the Athb-2 homeobox gene in green plants. *Proc Natl Acad Sci USA*, 93(8), 3530–5.
- Carabelli, M., Sessa, G., Baima, S., Morelli, G., & Ruberti, I. (1993). The Arabidopsis Athb-2 and -4 genes are strongly induced by far-red-rich light. *Plant J*, 4(3), 469–79.
- Carmona, M. J., Calonje, M., & Martínez-Zapater, J. M. (2007). The FT/TFL1 gene family in grapevine. *Plant Mol Biol*, 63(5), 637–50.
- Cheng, Y., Zhou, W., Sheery, N. I. E., Peters, C., Li, M., Wang, X., & Huang, J. (2011). Characterization of the Arabidopsis glycerophosphodiester phosphodiesterase (GDPD) family reveals a role of the plastid-localized AtGDPD1 in maintaining cellular phosphate homeostasis under phosphate starvation. *Plant J*, 66(5), 781–95.
- Christianson, J. A., Wilson, I. W., Llewellyn, D. J., & Dennis, E. S. (2009). The low-oxygen-induced NAC domain transcription factor ANAC102 affects viability of Arabidopsis seeds following low-oxygen treatment. *Plant Physiol*, 149(4), 1724–38.
- Chung, K. S., Yoo, S. Y., Yoo, S. J., Lee, J. S., & Ahn, J. H. (2010). BROTHER OF FT AND TFL1 (BFT), a member of the FT/TFL1 family, shows distinct pattern of expression during the vegetative growth of Arabidopsis. *Plant Signaling & Behavior*, 5(9), 1102–4.

- Churchill, G. A. (2004). Using ANOVA to analyze microarray data. *BioTechniques*, 37(2), 173–5, 177.
- Ciarbelli, A. R., Ciolfi, A., Salvucci, S., Ruzza, V., Possenti, M., Carabelli, M., Fruscalzo, A., Sessa, G., Morelli, G., & Ruberti, I. (2008). The Arabidopsis homeodomain-leucine zipper II gene family: diversity and redundancy. *Plant Mol Biol*, 68(4-5), 465–78.
- Clouse, S. D. (1996). Molecular genetic studies confirm the role of brassinosteroids in plant growth and development. *Plant J*, 10(1), 1–8.
- Clouse, S. D., & Sasse, J. M. (1998). BRASSINOSTEROIDS: Essential Regulators of Plant Growth and Development. *Annu Rev Plant Physiol Plant Mol Biol*, 49, 427–451.
- Cominelli, E., & Tonelli, C. (2009). A new role for plant R2R3-MYB transcription factors in cell cycle regulation. *Cell Res*, 19(11), 1231–2.
- Crowe, M. L., Serizet, C., Thareau, V., Aubourg, S., Rouzé, P., Hilson, P., Beynon, J., Weisbeek, P., van Hummelen, P., Reymond, P., Paz-Ares, J., Nietfeld, W., & Trick, M. (2003). CATMA: a complete Arabidopsis GST database. *Nucleic Acids Res*, 31(1), 156–8.
- Cubas, P., Lauter, N., Doebley, J., & Coen, E. (1999). The TCP domain: a motif found in proteins regulating plant growth and development. *Plant J*, 18(2), 215–22.
- Cui, X., & Churchill, G. A. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*, 4(4), 210.
- Cui, X., Kerr, M. K., & Churchill, G. A. (2003). Transformations for cDNA microarray data. *Statistical Applications in Genetics and Molecular Biology*, 2(1), Article 4.
- Curtis, I. S., Hanada, A., Yamaguchi, S., & Kamiya, Y. (2005). Modification of plant architecture through the expression of GA 2-oxidase under the control of an estrogen inducible promoter in Arabidopsis thaliana L. *Planta*, 222(6), 957–67.
- Dai, N., Schaffer, A., Petreikov, M., Shahak, Y., Giller, Y., Ratner, K., Levine, A., & Granot, D. (1999). Overexpression of Arabidopsis hexokinase in tomato plants inhibits growth, reduces photosynthesis, and induces rapid senescence. *Plant Cell*, 11(7), 1253–66.
- Delessert, C., Kazan, K., Wilson, I. W., Straeten, D. V. D., Manners, J., Dennis, E. S., & Dolferus, R. (2005). The transcription factor ATAF2 represses the expression of pathogenesis-related genes in Arabidopsis. *Plant J*, 43(5), 745–57.
- Drummond, A., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Heled, J., Kearse, M., Moir, R., Stones-Havas, S., Sturrock, S., Thierer, T., & Wilson, A. (2010). Geneious

v5.3.

URL <http://www.geneious.com>

- Dudoit, S., Yang, Y., Callow, M., & Speed, T. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12(1), 111–140.
- Edwards, D., Murray, J. A., & Smith, A. G. (1998). Multiple genes encoding the conserved CCAAT-box transcription factor complex are expressed in Arabidopsis. *Plant Physiology*, 117(3), 1015–22.
- Edwards, K., Anderson, P., Hall, A., Salathia, N., Locke, J., Lynn, J., Straume, M., Smith, J., & Millar, A. (2006). FLOWERING LOCUS C mediates natural variation in the high-temperature response of the Arabidopsis circadian clock. *The Plant Cell Online*, 18(3), 639.
- Egawa, C., Kobayashi, F., & Ishibashi..., M. (2006). Differential regulation of transcript accumulation and alternative splicing of a DREB2 homolog under abiotic stress conditions in common wheat. *Genes & genetic ...*
- Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA*, 95(25), 14863–8.
- Falcon, S., & Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2), 257–8.
- Feldmann, K. (1991). T-DNA insertion mutagenesis in Arabidopsis: mutational spectrum. *The Plant Journal*.
- Finnemann, J., & Schjoerring, J. K. (2000). Post-translational regulation of cytosolic glutamine synthetase by reversible phosphorylation and 14-3-3 protein interaction. *Plant J*, 24(2), 171–81.
- Flanders Interuniversity Institute for Biotechnology (2010). *BiNGO User Guide*.
- Folta, K. M., Pontin, M. A., Karlin-Neumann, G., Bottini, R., & Spalding, E. P. (2003). Genomic and physiological studies of early cryptochrome 1 action demonstrate roles for auxin and gibberellin in the control of hypocotyl growth by blue light. *Plant J*, 36(2), 203–14.
- Fraley, C., & Raftery, A. E. (1998). How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis. *Computer Journal*, 41, 578–588.
- Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science*, 315(5814), 972–6.

- Fujiki, Y., Yoshikawa, Y., Sato, T., Inada, N., Ito, M., Nishida, I., & Watanabe, A. (2001). Dark-inducible genes from *Arabidopsis thaliana* are associated with leaf senescence and repressed by sugars. *Physiol Plant*, 111(3), 345–352.
- Gaffney, T., Friedrich, L., Vernooij, B., Negrotto, D., Nye, G., Uknes, S., Ward, E., Kessmann, H., & Ryals, J. (1993). Requirement of salicylic Acid for the induction of systemic acquired resistance. *Science*, 261(5122), 754–6.
- Gan, S., & Amasino, R. M. (1995). Inhibition of leaf senescence by autoregulated production of cytokinin. *Science*, 270(5244), 1986–8.
- Gardner, T. S., di Bernardo, D., Lorenz, D., & Collins, J. J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301(5629), 102–5.
- Gatta, G. D., Bansal, M., Ambesi-Impiombato, A., Antonini, D., Missero, C., & di Bernardo, D. (2008). Direct targets of the TRP63 transcription factor revealed by a combination of gene expression profiling and reverse engineering. *Genome Res*, 18(6), 939–48.
- Gay, A., Thomas, H., Roca, M., James, C., Taylor, J., Rowland, J., & Ougham, H. (2008). Nondestructive analysis of senescence in mesophyll cells by spectral resolution of protein synthesis-dependent pigment metabolism. *New Phytol*, 179(3), 663–74.
- Gepstein, S., & Thimann, K. V. (1980). Changes in the abscisic acid content of oat leaves during senescence. *Proceedings of the National Academy of Sciences of the United States of America*, 77(4), 2050–3.
- Giraud, E., Ng, S., Carrie, C., Duncan, O., Low, J., Lee, C. P., Aken, O. V., Millar, A. H., Murcha, M., & Whelan, J. (2010). TCP transcription factors link the regulation of genes encoding mitochondrial proteins with the circadian clock in *Arabidopsis thaliana*. *Plant Cell*, 22(12), 3921–34.
- Glauser, G., Grata, E., Dubugnon, L., Rudaz, S., Farmer, E. E., & Wolfender, J.-L. (2008). Spatial and temporal dynamics of jasmonate synthesis and accumulation in *Arabidopsis* in response to wounding. *J Biol Chem*, 283(24), 16400–7.
- Glazebrook, J., Zook, M., Mert, F., Kagan, I., Rogers, E. E., Crute, I. R., Holub, E. B., Hammerschmidt, R., & Ausubel, F. M. (1997). Phytoalexin-deficient mutants of *Arabidopsis* reveal that PAD4 encodes a regulatory factor and that four PAD genes contribute to downy mildew resistance. *Genetics*, 146(1), 381–92.
- Gong, Z., Koiwa, H., Cushman, M. A., Ray, A., Bufford, D., Kore-eda, S., Matsumoto, T. K., Zhu, J., Cushman, J. C., Bressan, R. A., & Hasegawa, P. M. (2001). Genes that are uniquely stress regulated in salt overly sensitive (sos) mutants. *Plant Physiol*, 126(1), 363–75.

- Grbić, V., & Bleecker, A. (1995). Ethylene regulates the timing of leaf senescence in Arabidopsis. *The Plant Journal*, 8(4), 595–602.
- Green, P. (1994). The ribonucleases of higher plants. *Annual Review of Plant Physiology and Plant Molecular Biology*, 45, 421–45.
- Gresham, D., Dunham, M. J., & Botstein, D. (2008). Comparing whole genomes using DNA microarrays. *Nat Rev Genet*, 9(4), 291–302.
- Guo, Y., & Gan, S. (2006). AtNAP, a NAC family transcription factor, has an important role in leaf senescence. *Plant J*, 46(4), 601–12.
- Guo, Z., Fujioka, S., Blancaflor, E. B., Miao, S., Gou, X., & Li, J. (2010). TCP1 modulates brassinosteroid biosynthesis by regulating the expression of the key biosynthetic gene DWARF4 in Arabidopsis thaliana. *Plant Cell*, 22(4), 1161–73.
- Hanzawa, Y., Money, T., & Bradley, D. (2005). A single amino acid converts a repressor to an activator of flowering. *Proc Natl Acad Sci USA*, 102(21), 7748–53.
- Harper, J. W., Adami, G. R., Wei, N., Keyomarsi, K., & Elledge, S. J. (1993). The p21 Cdk-interacting protein Cip1 is a potent inhibitor of G1 cyclin-dependent kinases. *Cell*, 75(4), 805–16.
- Hasson, A., Plessis, A., Blein, T., Adroher, B., Grigg, S., Tsiantis, M., Boudaoud, A., Damerval, C., & Laufs, P. (2011). Evolution and diverse roles of the CUP-SHAPED COTYLEDON genes in Arabidopsis leaf development. *Plant Cell*, 23(1), 54–68.
- He, X.-J., Mu, R.-L., Cao, W.-H., Zhang, Z.-G., Zhang, J.-S., & Chen, S.-Y. (2005). AtNAC2, a transcription factor downstream of ethylene and auxin signaling pathways, is involved in salt stress response and lateral root development. *Plant J*, 44(6), 903–16.
- He, Y., & Gan, S. (2002). A gene encoding an acyl hydrolase is involved in leaf senescence in Arabidopsis. *Plant Cell*, 14(4), 805–15.
- He, Y., Tang, W., Swain, J. D., Green, A. L., Jack, T. P., & Gan, S. (2001). Networking senescence-regulating pathways by using Arabidopsis enhancer trap lines. *Plant Physiology*, 126(2), 707–16.
- Heard, N. A. (2011). Iterative Reclassification in Agglomerative Clustering. *Journal of Computational and Graphical Statistics*.
- Heard, N. A., Holmes, C. C., & Stephens, D. A. (2006). A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes: An Application of Bayesian Hierarchical Clustering of Curves. *Journal of the American Statistical Association*, 101(473), 18–29.

- Heard, N. A., Holmes, C. C., Stephens, D. A., Hand, D. J., & Dimopoulos, G. (2005). Bayesian coclustering of *Anopheles* gene expression time series: study of immune defense response to multiple experimental challenges. *Proc Natl Acad Sci USA*, 102(47), 16939–44.
- Hervé, C., Dabos, P., Bardet, C., Jauneau, A., Auriac, M. C., Ramboer, A., Lacout, F., & Tremousaygue, D. (2009). In vivo interference with AtTCP20 function induces severe plant growth alterations and deregulates the expression of many genes important for development. *Plant Physiol*, 149(3), 1462–77.
- Hilson, P., Allemeersch, J., Altmann, T., Aubourg, S., Avon, A., Beynon, J., Bhalerao, R. P., Bitton, F., Caboche, M., Cannoot, B., Chardakov, V., Cognet-Holliger, C., Colot, V., Crowe, M., Darimont, C., Durinck, S., Eickhoff, H., de Longevialle, A. F., Farmer, E. E., Grant, M., Kuiper, M. T. R., Lehrach, H., Léon, C., Leyva, A., Lundeberg, J., Lurin, C., Moreau, Y., Nietfeld, W., Paz-Ares, J., Reymond, P., Rouzé, P., Sandberg, G., Segura, M. D., Serizet, C., Tabrett, A., Taconnat, L., Thareau, V., van Hummelen, P., Vercruysse, S., Vuylsteke, M., Weingartner, M., Weisbeek, P. J., Wirta, V., Wittink, F. R. A., Zabeau, M., & Small, I. (2004). Versatile gene-specific sequence tags for Arabidopsis functional genomics: transcript profiling and reverse genetics applications. *Genome Res*, 14(10B), 2176–89.
- Himelblau, E., & Amasino, R. (2001). Nutrients mobilized from leaves of *Arabidopsis thaliana* during leaf senescence. *J. Plant Physiol.*, 158(10), 1317–1323.
- Hinder, B., Schellenberg, M., Rodoni, S., Ginsburg, S., Vogt, E., Martinoia, E., Matile, P., & Hörtensteiner, S. (1996). How plants dispose of chlorophyll catabolites. Directly energized uptake of tetrapyrrolic breakdown products into isolated vacuoles. *J Biol Chem*, 271(44), 27233–6.
- Hong, S. B., Sexton, R., & Tucker, M. L. (2000). Analysis of gene promoters for two tomato polygalacturonases expressed in abscission zones and the stigma. *Plant Physiology*, 123(3), 869–81.
- Hörtensteiner, S., & Feller, U. (2002). Nitrogen metabolism and remobilization during senescence. *J Exp Bot*, 53(370), 927–37.
- Hörtensteiner, S., Wüthrich, K. L., Matile, P., Ongania, K. H., & Kräutler, B. (1998). The key step in chlorophyll breakdown in higher plants. Cleavage of pheophorbide a macrocycle by a monooxygenase. *J Biol Chem*, 273(25), 15335–9.
- Hung, K. T., & Kao, C. H. (2003). Nitric oxide counteracts the senescence of rice leaves induced by abscisic acid. *J Plant Physiol*, 160(8), 871–9.
- Hung, K. T., & Kao, C. H. (2004). Hydrogen peroxide is necessary for abscisic acid-induced senescence of rice leaves. *J Plant Physiol*, 161(12), 1347–57.

- James, P., Halladay, J., & Craig, E. A. (1996). Genomic libraries and a host strain designed for highly efficient two-hybrid selection in yeast. *Genetics*, 144(4), 1425–36.
- James, W., & Stein, C. (1961). Estimation with quadratic loss. In *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, vol. 1, (pp. 361–379).
- Jenssen, T.-K., Langaas, M., Kuo, W. P., Smith-Sørensen, B., Myklebost, O., & Hovig, E. (2002). Analysis of repeatability in spotted cDNA microarrays. *Nucleic Acids Res*, 30(14), 3235–44.
- Jing, H.-C., Schippers, J. H. M., Hille, J., & Dijkwel, P. P. (2005). Ethylene-induced leaf senescence depends on age-related changes and OLD genes in Arabidopsis. *J Exp Bot*, 56(421), 2915–23.
- Jing, H.-C., Sturre, M. J. G., Hille, J., & Dijkwel, P. P. (2002). Arabidopsis onset of leaf death mutants identify a regulatory pathway controlling leaf senescence. *Plant J*, 32(1), 51–63.
- Jung, S. (2004). Effect of chlorophyll reduction in Arabidopsis thaliana by methyl jasmonate or norflurazon on antioxidant systems. *Plant Physiology and Biochemistry*, 42(3), 225–231.
- Kamachi, K., Yamaya, T., Hayakawa, T., Mae, T., & Ojima, K. (1992). Changes in Cytosolic Glutamine Synthetase Polypeptide and its mRNA in a Leaf Blade of Rice Plants during Natural Senescence. *Plant Physiol*, 98(4), 1323–1329.
- Kerr, M., & Churchill, G. (2001). Experimental design for gene expression microarrays. *Biostatistics*, 2(2), 183.
- Kiddle, S. J., Windram, O. P. F., McHattie, S., Mead, A., Beynon, J., Buchanan-Wollaston, V., Denby, K. J., & Mukherjee, S. (2010). Temporal clustering by affinity propagation reveals transcriptional modules in Arabidopsis thaliana. *Bioinformatics*, 26(3), 355–62.
- Kidokoro, S., Maruyama, K., Nakashima, K., Imura, Y., Narusaka, Y., Shinwari, Z. K., Osakabe, Y., Fujita, Y., Mizoi, J., Shinozaki, K., & Yamaguchi-Shinozaki, K. (2009). The phytochrome-interacting factor PIF7 negatively regulates DREB1 expression under circadian control in Arabidopsis. *Plant Physiol*, 151(4), 2046–57.
- Kim, H. J., Ryu, H., Hong, S. H., Woo, H. R., Lim, P. O., Lee, I. C., Sheen, J., Nam, H. G., & Hwang, I. (2006). Cytokinin-mediated control of leaf longevity by AHK3 through phosphorylation of ARR2 in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, 103(3), 814–9.
- Kim, J. H., Woo, H. R., Kim, J., Lim, P. O., Lee, I. C., Choi, S. H., Hwang, D., & Nam, H. G. (2009). Trifurcate feed-forward regulation of age-dependent cell death involving miR164 in Arabidopsis. *Science*, 323(5917), 1053–7.

- Kim, S. Y., & Nam, K. H. (2010). Physiological roles of ERD10 in abiotic stresses and seed germination of Arabidopsis. *Plant Cell Rep*, 29(2), 203–9.
- Kinoshita, T., Yamada, K., Hiraiwa, N., Kondo, M., Nishimura, M., & Hara-Nishimura, I. (1999). Vacuolar processing enzyme is up-regulated in the lytic vacuoles of vegetative tissues during senescence and under various stressed conditions. *Plant J*, 19(1), 43–53.
- Klemm, S. L. (2008). Causal structure identification in non-linear dynamical systems. *Cambridge, UK: University of Cambridge, UK*.
- Klevebring, D., Gry, M., Lindberg, J., Eidefors, A., & Lundeberg, J. (2009). Automation of cDNA synthesis and labelling improves reproducibility. *J Biomed Biotechnol*, 2009, 396808.
- Kong, Z., Li, M., Yang, W., Xu, W., & Xue, Y. (2006). A novel nuclear-localized CCCH-type zinc finger protein, OsDOS, is involved in delaying leaf senescence in rice. *Plant Physiology*, 141(4), 1376–88.
- Kovacs, D., Kalmar, E., Torok, Z., & Tompa, P. (2008). Chaperone activity of ERD10 and ERD14, two disordered stress-related plant proteins. *Plant Physiol*, 147(1), 381.
- Koyama, T., Sato, F., & Ohme-Takagi, M. (2010). A role of TCP1 in the longitudinal elongation of leaves in Arabidopsis. *Biosci. Biotechnol. Biochem.*, 74(10), 2145–7.
- Larue, C. T., Wen, J., & Walker, J. C. (2009a). Genetic interactions between the miRNA164-CUC2 regulatory module and BREVIPEDICELLUS in Arabidopsis developmental patterning. *Plant Signaling & Behavior*, 4(7), 666–8.
- Larue, C. T., Wen, J., & Walker, J. C. (2009b). A microRNA-transcription factor module regulates lateral organ size and patterning in Arabidopsis. *Plant J*, 58(3), 450–63.
- Lèbre, S. (2009). Inferring dynamic genetic networks with low order independencies. *Statistical Applications in Genetics and Molecular Biology*, 8(1), Article 9.
- Lee, I. C., Hong, S. W., Whang, S. S., Lim, P. O., Nam, H. G., & Koo, J. C. (2011). Age-dependent action of an ABA-inducible receptor kinase, RPK1, as a positive regulator of senescence in Arabidopsis leaves. *Plant Cell Physiol*, 52(4), 651–62.
- Legaie, R., & McHattie, S. D. J. (2010). PRESTA Senescence Timeseries Data.
URL <http://www2.warwick.ac.uk/fac/sci/lifesci/research/presta/data/senescence>
- Leivar, P., Monte, E., Al-Sady, B., Carle, C., Storer, A., Alonso, J. M., Ecker, J. R., & Quail, P. H. (2008). The Arabidopsis phytochrome-interacting factor PIF7, together

- with PIF3 and PIF4, regulates responses to prolonged red light by modulating phyB levels. *Plant Cell*, 20(2), 337–52.
- Leung, J., & Giraudat, J. (1998). ABSCISIC ACID SIGNAL TRANSDUCTION. *Annu Rev Plant Physiol Plant Mol Biol*, 49, 199–222.
- Li, C., Potuschak, T., Colón-Carmona, A., Gutiérrez, R., & Doerner, P. (2005). Arabidopsis TCP20 links regulation of growth and cell division control pathways. *Proceedings of the National Academy of Sciences of the United States of America*, 102(36), 12978.
- Li, J., Li, X., Guo, L., Lu, F., Feng, X., He, K., Wei, L., Chen, Z., Qu, L.-J., & Gu, H. (2006). A subgroup of MYB transcription factor genes undergoes highly conserved alternative splicing in Arabidopsis and rice. *J Exp Bot*, 57(6), 1263–73.
- Li, J. J., & Herskowitz, I. (1993). Isolation of ORC6, a component of the yeast origin recognition complex by a one-hybrid system. *Science*, 262(5141), 1870–4.
- Lim, P. O., Kim, H. J., & Nam, H. G. (2007). Leaf senescence. *Annual review of plant biology*, 58, 115–36.
- Lin, X., Kaul, S., Rounsley, S., Shea, T. P., Benito, M. I., Town, C. D., Fujii, C. Y., Mason, T., Bowman, C. L., Barnstead, M., Feldblyum, T. V., Buell, C. R., Ketchum, K. A., Lee, J., Ronning, C. M., Koo, H. L., Moffat, K. S., Cronin, L. A., Shen, M., Pai, G., Aken, S. V., Umayam, L., Tallon, L. J., Gill, J. E., Adams, M. D., Carrera, A. J., Creasy, T. H., Goodman, H. M., Somerville, C. R., Copenhaver, G. P., Preuss, D., Nierman, W. C., White, O., Eisen, J. A., Salzberg, S. L., Fraser, C. M., & Venter, J. C. (1999). Sequence and analysis of chromosome 2 of the plant Arabidopsis thaliana. *Nature*, 402(6763), 761–8.
- Lloyd, A. M., Barnason, A. R., Rogers, S. G., Byrne, M. C., Fraley, R. T., & Horsch, R. B. (1986). Transformation of Arabidopsis thaliana with Agrobacterium tumefaciens. *Science*, 234(4775), 464–6.
- Long, T. A., Tsukagoshi, H., Busch, W., Lahner, B., Salt, D. E., & Benfey, P. N. (2010). The bHLH transcription factor POPEYE regulates response to iron deficiency in Arabidopsis roots. *Plant Cell*, 22(7), 2219–36.
- Mach, J. M., Castillo, A. R., Hoogstraten, R., & Greenberg, J. T. (2001). The Arabidopsis-accelerated cell death gene ACD2 encodes red chlorophyll catabolite reductase and suppresses the spread of disease symptoms. *Proc Natl Acad Sci USA*, 98(2), 771–6.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium of Mathematics, Statistics and Probability*, (pp. 281–297). University of California.

- Maere, S., Heymans, K., & Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16), 3448–9.
- Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., & Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7 Suppl 1, S7.
- MATLAB (2010). *version 7.11.0 (R2010b)*. Natick, Massachusetts: The Mathworks Inc.
- Mayer, K., Schüller, C., Wambutt, R., Murphy, G., Volckaert, G., Pohl, T., Düsterhöft, A., Stiekema, W., Entian, K. D., Terryn, N., Harris, B., Ansorge, W., Brandt, P., Grivell, L., Rieger, M., Weichselgartner, M., Simone, V. D., Obermaier, B., Mache, R., Müller, M., Kreis, M., Delseny, M., Puigdomenech, P., Watson, M., Schmidtheini, T., Reichert, B., Portatelle, D., Perez-Alonso, M., Boutry, M., Bancroft, I., Vos, P., Hoheisel, J., Zimmermann, W., Wedler, H., Ridley, P., Langham, S. A., McCullagh, B., Bilham, L., Robben, J., der Schueren, J. V., Grymonprez, B., Chuang, Y. J., Vandenbussche, F., Braeken, M., Weltjens, I., Voet, M., Bastiaens, I., Aert, R., Defoor, E., Weitzenegger, T., Bothe, G., Ramsperger, U., Hilbert, H., Braun, M., Holzer, E., Brandt, A., Peters, S., van Staveren, M., Dirske, W., Mooijman, P., Lankhorst, R. K., Rose, M., Hauf, J., Kötter, P., Berneiser, S., Hempel, S., Feldpausch, M., Lamberth, S., den Daele, H. V., Keyser, A. D., Buysschaert, C., Gielen, J., Villarroel, R., Clercq, R. D., Montagu, M. V., Rogers, J., Cronin, A., Quail, M., Bray-Allen, S., Clark, L., Doggett, J., Hall, S., Kay, M., Lennard, N., McLay, K., Mayes, R., Pettett, A., Rajandream, M. A., Lyne, M., Benes, V., Rechmann, S., Borkova, D., Blöcker, H., Scharfe, M., Grimm, M., Löhnert, T. H., Dose, S., de Haan, M., Maarse, A., Schäfer, M., Müller-Auer, S., Gabel, C., Fuchs, M., Fartmann, B., Granderath, K., Dauner, D., Herzl, A., Neumann, S., Argiriou, A., Vitale, D., Liguori, R., Piravandi, E., Massenet, O., Quigley, F., Clabaud, G., Mündlein, A., Felber, R., Schnabl, S., Hiller, R., Schmidt, W., Lecharny, A., Aubourg, S., Chedfor, F., Cooke, R., Berger, C., Montfort, A., Casacuberta, E., Gibbons, T., Weber, N., Vandenbol, M., Barges, M., Terol, J., Torres, A., Perez-Perez, A., Purnelle, B., Bent, E., Johnson, S., Tacon, D., Jesse, T., Heijnen, L., Schwarz, S., Scholler, P., Heber, S., Francs, P., Bielke, C., Frishman, D., Haase, D., Lemcke, K., Mewes, H. W., Stocker, S., Zaccaria, P., Bevan, M., Wilson, R. K., de la Bastide, M., Habermann, K., Parnell, L., Dedhia, N., Gnoj, L., Schutz, K., Huang, E., Spiegel, L., Sehkun, M., Murray, J., Sheet, P., Cordes, M., Abu-Threideh, J., Stoneking, T., Kalicki, J., Graves, T., Harmon, G., Edwards, J., Latreille, P., Courtney, L., Cloud, J., Abbott, A., Scott, K., Johnson, D., Minx, P., Bentley, D., Fulton, B., Miller, N., Greco, T., Kemp, K., Kramer, J., Fulton, L., Mardis, E., Dante, M., Pepin, K., Hillier, L., Nelson, J., Spieth, J., Ryan, E., Andrews, S., Geisel, C., Layman, D., Du, H., Ali, J., Berghoff, A., Jones, K., Drone, K., Cotton, M., Joshu, C., Antonoiu, B., Zidanic, M., Strong, C., Sun, H.,

- Lamar, B., Yordan, C., Ma, P., Zhong, J., Preston, R., Vil, D., Shekher, M., Matero, A., Shah, R., Swaby, I. K., O'Shaughnessy, A., Rodriguez, M., Hoffmann, J., Till, S., Granat, S., Shohdy, N., Hasegawa, A., Hameed, A., Lodhi, M., Johnson, A., Chen, E., Marra, M., Martienssen, R., & McCombie, W. R. (1999). Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature*, 402(6763), 769–77.
- Mead, A. (**in preparation**). Experimental designs for unbiased, highly connected, two-channel microarray experiments.
- Merlot, S., Gosti, F., Guerrier, D., Vavasseur, A., & Giraudat, J. (2001). The ABI1 and ABI2 protein phosphatases 2C act in a negative feedback regulatory loop of the abscisic acid signalling pathway. *Plant J*, 25(3), 295–303.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., & Teller, A. H. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6), 1088–92.
- Miao, Y., Laun, T., Zimmermann, P., & Zentgraf, U. (2004). Targets of the WRKY53 transcription factor and its role during leaf senescence in *Arabidopsis*. *Plant Molecular Biology*, 55(6), 853–67.
- Miao, Y., & Zentgraf, U. (2007). The antagonist function of *Arabidopsis* WRKY53 and ESR/ESP in leaf senescence is modulated by the jasmonic and salicylic acid equilibrium. *Plant Cell*, 19(3), 819–30.
- Mikkelsen, M., & Thomashow, M. (2009). A role for circadian evening elements in cold-regulated gene expression in *Arabidopsis*. *The Plant Journal*, 60(2), 328–339.
- Miller, C., Skoog, F., Saltza, M., & Strong, F. M. (1955). Kinetin, a cell division factor from deoxyribonucleic acid. *Journal of the American Chemical Society*, 77, 1392.
- Monnier, A., Liverani, S., Bouvet, R., Jesson, B., Smith, J., Mosser, J., Corellou, F., & Bouget, F. (2010). Orchestrated transcription of biological processes in the marine picoeukaryote *Ostreococcus* exposed to light/dark cycles. *BMC Genomics*, 11(1), 192.
- Morris, K., MacKerness, S. A., Page, T., John, C. F., Murphy, A. M., Carr, J. P., & Buchanan-Wollaston, V. (2000). Salicylic acid has a role in regulating gene expression during leaf senescence. *Plant J*, 23(5), 677–85.
- Mu, R.-L., Cao, Y.-R., Liu, Y.-F., Lei, G., Zou, H.-F., Liao, Y., Wang, H.-W., Zhang, W.-K., Ma, B., Du, J.-Z., Yuan, M., Zhang, J.-S., & Chen, S.-Y. (2009). An R2R3-type transcription factor gene *AtMYB59* regulates root growth and cell cycle progression in *Arabidopsis*. *Cell Res*, 19(11), 1291–1304.
- Munné-Bosch, S., & Alegre, L. (2004). Die and let live: leaf senescence contributes to plant survival under drought stress. *Functional Plant Biology*, 31(3), 203–216.

- Mutarelli, M., Cicatiello, L., Ferraro, L., Grober, O., Ravo, M., Facchiano, A., Angelini, C., & Weisz, A. (2008). Time-course analysis of genome-wide gene expression data from hormone-responsive human breast cancer cells. *BMC Bioinformatics*, 9(Suppl 2), S12.
- Nelson, D. E., Repetti, P. P., Adams, T. R., Creelman, R. A., Wu, J., Warner, D. C., Anstrom, D. C., Bensen, R. J., Castiglioni, P. P., Donnarummo, M. G., Hinchey, B. S., Kumimoto, R. W., Maszle, D. R., Canales, R. D., Krolikowski, K. A., Dotson, S. B., Gutterson, N., Ratcliffe, O. J., & Heard, J. E. (2007). Plant nuclear factor Y (NF-Y) B subunits confer drought tolerance and lead to improved corn yields on water-limited acres. *Proc Natl Acad Sci USA*, 104(42), 16450–5.
- Nikovics, K., Blein, T., Peaucelle, A., Ishida, T., Morin, H., Aida, M., & Laufs, P. (2006). The balance between the MIR164A and CUC2 genes controls leaf margin serration in Arabidopsis. *Plant Cell*, 18(11), 2929–45.
- Noh, Y. S., & Amasino, R. M. (1999). Identification of a promoter region responsible for the senescence-specific expression of SAG12. *Plant Molecular Biology*, 41(2), 181–94.
- Olsen, A., Ernst, H., Leggio, L., & Skriver, K. (2005). DNA-binding specificity and molecular functions of NAC transcription factors. *Plant Science*, 169(4), 785–797.
- O'Neill, L. P., & Turner, B. M. (1995). Histone H4 acetylation distinguishes coding regions of the human genome from heterochromatin in a differentiation-dependent but transcription-independent manner. *EMBO J*, 14(16), 3946–57.
- O'Neill, L. P., & Turner, B. M. (1996). Immunoprecipitation of chromatin. *Meth Enzymol*, 274, 189–97.
- Ooka, H., Satoh, K., Doi, K., Nagata, T., Otomo, Y., Murakami, K., Matsubara, K., Osato, N., Kawai, J., Carninci, P., Hayashizaki, Y., Suzuki, K., Kojima, K., Takahara, Y., Yamamoto, K., & Kikuchi, S. (2003). Comprehensive analysis of NAC family genes in *Oryza sativa* and *Arabidopsis thaliana*. *DNA Res*, 10(6), 239–47.
- Park, J. H., Oh, S. A., Kim, Y. H., Woo, H. R., & Nam, H. G. (1998). Differential expression of senescence-associated mRNAs during leaf senescence induced by different senescence-inducing factors in Arabidopsis. *Plant Molecular Biology*, 37(3), 445–54.
- Park, P. J., Cao, Y. A., Lee, S. Y., Kim, J.-W., Chang, M. S., Hart, R., & Choi, S. (2004). Current issues for DNA microarrays: platform comparison, double linear amplification, and universal RNA reference. *J Biotechnol*, 112(3), 225–45.
- Pauwels, L., & Goossens, A. (2008). Fine-tuning of early events in the jasmonate response. *Plant Signaling & Behavior*, 3(10), 846–7.

- Pauwels, L., Morreel, K., Witte, E. D., Lammertyn, F., Montagu, M. V., Boerjan, W., Inzé, D., & Goossens, A. (2008). Mapping methyl jasmonate-mediated transcriptional reprogramming of metabolism and cell cycle progression in cultured Arabidopsis cells. *Proceedings of the National Academy of Sciences*, 105(4), 1380–5.
- Peaucelle, A., Morin, H., Traas, J., & Laufs, P. (2007). Plants expressing a miR164-resistant CUC2 gene reveal the importance of post-meristematic maintenance of phyllotaxy in Arabidopsis. *Development*, 134(6), 1045–50.
- Penfold, C., & Wild, D. (2011). How to infer gene networks from expression profiles, revisited. *Journal of the Royal Society Interface Focus*.
- Pérez-Amador, M. A., Abler, M. L., Rocher, E. J. D., Thompson, D. M., van Hoof, A., LeBrasseur, N. D., Lers, A., & Green, P. J. (2000). Identification of BFN1, a bifunctional nuclease induced during leaf and stem senescence in Arabidopsis. *Plant Physiology*, 122(1), 169–80.
- Pyke, K. A., & Leech, R. M. (1994). A Genetic Analysis of Chloroplast Division and Expansion in Arabidopsis thaliana. *Plant Physiology*, 104(1), 201–207.
- Qian, J., Dolled-Filhart, M., Lin, J., Yu, H., & Gerstein, M. (2001). Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *J Mol Biol*, 314(5), 1053–66.
- Quirino, B. F., Noh, Y. S., Himelblau, E., & Amasino, R. M. (2000). Molecular aspects of leaf senescence. *Trends Plant Sci*, 5(7), 278–82.
- Quirino, B. F., Normanly, J., & Amasino, R. M. (1999). Diverse range of gene activity during Arabidopsis thaliana leaf senescence includes pathogen-independent induction of defense-related genes. *Plant Molecular Biology*, 40(2), 267–78.
- Quirino, B. F., Reiter, W. D., & Amasino, R. D. (2001). One of two tandem Arabidopsis genes homologous to monosaccharide transporters is senescence-associated. *Plant Mol Biol*, 46(4), 447–57.
- R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Rampey, R. A., Woodward, A. W., Hobbs, B. N., Tierney, M. P., Lahner, B., Salt, D. E., & Bartel, B. (2006). An Arabidopsis basic helix-loop-helix leucine zipper protein modulates metal homeostasis and auxin conjugate responsiveness. *Genetics*, 174(4), 1841–57.
- Rao, M. V., & Davis, K. R. (1999). Ozone-induced cell death occurs via two distinct mechanisms in Arabidopsis: the role of salicylic acid. *Plant J*, 17(6), 603–14.

- Rhee, S. Y., Beavis, W., Berardini, T. Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., Miller, N., Mueller, L. A., Mundodi, S., Reiser, L., Tacklind, J., Weems, D. C., Wu, Y., Xu, I., Yoo, D., Yoon, J., & Zhang, P. (2003). The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res*, 31(1), 224–8.
- Richmond, A. E., & Lang, A. (1957). Effect of kinetin on protein content and survival of detached Xanthium leaves. *Science*, 125(3249), 650–651.
- Robertson, E. J., Rutherford, S. M., & Leech, R. M. (1996). Characterization of chloroplast division using the Arabidopsis mutant arc5. *Plant Physiology*, 112(1), 149–59.
- Rodoni, S., Mühlecker, W., Anderl, M., Kräutler, B., Moser, D., Thomas, H., Matile, P., & Hörtensteiner, S. (1997). Chlorophyll Breakdown in Senescent Chloroplasts (Cleavage of Pheophorbide a in Two Enzymic Steps). *Plant Physiol*, 115(2), 669–676.
- Rosenzweig, B. A., Pine, P. S., Domon, O. E., Morris, S. M., Chen, J. J., & Sistare, F. D. (2004). Dye bias correction in dual-labeled cDNA microarray gene expression measurements. *Environ Health Perspect*, 112(4), 480–7.
- Rossel, J. B., Wilson, P. B., Hussain, D., Woo, N. S., Gordon, M. J., Mewett, O. P., Howell, K. A., Whelan, J., Kazan, K., & Pogson, B. J. (2007). Systemic and intracellular responses to photooxidative stress in Arabidopsis. *The Plant Cell*, 19(12), 4091–110.
- Roweis, S., & Ghahramani, Z. (1999). A unifying review of linear gaussian models. *Neural Comput*, 11(2), 305–45.
- Ryu, S. B., & Wang, X. (1995). Expression of Phospholipase D during Castor Bean Leaf Senescence. *Plant Physiology*, 108(2), 713–719.
- Sakamoto, H., Maruyama, K., Sakuma, Y., Meshi, T., Iwabuchi, M., Shinozaki, K., & Yamaguchi-Shinozaki, K. (2004). Arabidopsis Cys2/His2-type zinc-finger proteins function as transcription repressors under drought, cold, and high-salinity stress conditions. *Plant Physiology*, 136(1), 2734–46.
- Salanoubat, M., Lemcke, K., Rieger, M., Ansorge, W., Unseld, M., Fartmann, B., Valle, G., Blöcker, H., Perez-Alonso, M., Obermaier, B., Delseny, M., Boutry, M., Grivell, L. A., Mache, R., Puigdomènech, P., Simone, V. D., Choisne, N., Artiguenave, F., Robert, C., Brottier, P., Wincker, P., Cattolico, L., Weissenbach, J., Saurin, W., Quétier, F., Schäfer, M., Müller-Auer, S., Gabel, C., Fuchs, M., Benes, V., Wurmbach, E., Drzonek, H., Erfle, H., Jordan, N., Bangert, S., Wiedelmann, R., Kranz, H., Voss, H., Holland, R., Brandt, P., Nyakatura, G., Vezzi, A., D'Angelo,

- M., Pallavicini, A., Toppo, S., Simionati, B., Conrad, A., Hornischer, K., Kauer, G., Löhnert, T. H., Nordsiek, G., Reichelt, J., Scharfe, M., Schön, O., Bargues, M., Terol, J., Climent, J., Navarro, P., Collado, C., Perez-Perez, A., Ottenwälder, B., Duchemin, D., Cooke, R., Laudie, M., Berger-Llauro, C., Purnelle, B., Masuy, D., de Haan, M., Maarse, A. C., Alcaraz, J. P., Cottet, A., Casacuberta, E., Monfort, A., Argiriou, A., flores, M., Liguori, R., Vitale, D., Mannhaupt, G., Haase, D., Schoof, H., Rudd, S., Zaccaria, P., Mewes, H. W., Mayer, K. F., Kaul, S., Town, C. D., Koo, H. L., Tallon, L. J., Jenkins, J., Rooney, T., Rizzo, M., Walts, A., Utterback, T., Fujii, C. Y., Shea, T. P., Creasy, T. H., Haas, B., Maiti, R., Wu, D., Peterson, J., Aken, S. V., Pai, G., Militscher, J., Sellers, P., Gill, J. E., Feldblyum, T. V., Preuss, D., Lin, X., Nierman, W. C., Salzberg, S. L., White, O., Venter, J. C., Fraser, C. M., Kaneko, T., Nakamura, Y., Sato, S., Kato, T., Asamizu, E., Sasamoto, S., Kimura, T., Idesawa, K., Kawashima, K., Kishida, Y., Kiyokawa, C., Kohara, M., Matsumoto, M., Matsuno, A., Muraki, A., Nakayama, S., Nakazaki, N., Shinpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M., Tabata, S., Consortium, E. U. C. . A. S., for Genomic Research, I., & Institute, K. D. R. (2000). Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana*. *Nature*, 408(6814), 820–2.
- Sasaki, Y., Asamizu, E., Shibata, D., Nakamura, Y., Kaneko, T., Awai, K., Amagai, M., Kuwata, C., Tsugane, T., Masuda, T., Shimada, H., Takamiya, K., Ohta, H., & Tabata, S. (2001). Monitoring of methyl jasmonate-responsive genes in *Arabidopsis* by cDNA macroarray: self-activation of jasmonic acid biosynthesis and crosstalk with other phytohormone signaling pathways. *DNA Res*, 8(4), 153–61.
- Sawa, S., Ohgishi, M., Goda, H., Higuchi, K., Shimada, Y., Yoshida, S., & Koshiba, T. (2002). The HAT2 gene, a member of the HD-Zip gene family, isolated as an auxin inducible gene by DNA microarray screening, affects auxin response in *Arabidopsis*. *Plant J*, 32(6), 1011–22.
- Schaller, F. (2001). Enzymes of the biosynthesis of octadecanoid-derived signalling molecules. *J Exp Bot*, 52(354), 11–23.
- Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235), 467–70.
- Schenk, P. M., Kazan, K., Wilson, I., Anderson, J. P., Richmond, T., Somerville, S. C., & Manners, J. M. (2000). Coordinated plant defense responses in *Arabidopsis* revealed by microarray analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 97(21), 11655–60.
- Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., Minx, P., Reily, A. D., Courtney, L., Kruchowski,

- S. S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S. M., Belter, E., Du, F., Kim, K., Abbott, R. M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S. M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M. J., McMahan, L., Buren, P. V., Vaughn, M. W., Ying, K., Yeh, C.-T., Emrich, S. J., Jia, Y., Kalyanaraman, A., Hsia, A.-P., Barbazuk, W. B., Baucom, R. S., Brutnell, T. P., Carpita, N. C., Chaparro, C., Chia, J.-M., Deragon, J.-M., Estill, J. C., Fu, Y., Jeddelloh, J. A., Han, Y., Lee, H., Li, P., Lisch, D. R., Liu, S., Liu, Z., Nagel, D. H., McCann, M. C., SanMiguel, P., Myers, A. M., Nettleton, D., Nguyen, J., Penning, B. W., Ponnala, L., Schneider, K. L., Schwartz, D. C., Sharma, A., Soderlund, C., Springer, N. M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T. K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J. L., Dawe, R. K., Jiang, J., Jiang, N., Presting, G. G., Wessler, S. R., Aluru, S., Martienssen, R. A., Clifton, S. W., McCombie, W. R., Wing, R. A., & Wilson, R. K. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science*, 326(5956), 1112–5.
- Schommer, C., Palatnik, J. F., Aggarwal, P., Chételat, A., Cubas, P., Farmer, E. E., Nath, U., & Weigel, D. (2008). Control of jasmonate biosynthesis and senescence by miR319 targets. *PLoS Biol*, 6(9), e230.
- Schopfer, C. R., Nasrallah, M. E., & Nasrallah, J. B. (1999). The male determinant of self-incompatibility in Brassica. *Science*, 286(5445), 1697–700.
- Sclep, G., Allemeersch, J., Liechti, R., Meyer, B. D., Beynon, J., Bhalerao, R., Moreau, Y., Nietfeld, W., Renou, J.-P., Reymond, P., Kuiper, M. T., & Hilson, P. (2007). CATMA, a comprehensive genome-scale resource for silencing and transcript profiling of Arabidopsis genes. *BMC Bioinformatics*, 8, 400.
- Seo, M., Koiwai, H., Akaba, S., Komano, T., Oritani, T., Kamiya, Y., & Koshida, T. (2000). Abscissic aldehyde oxidase in leaves of Arabidopsis thaliana. *Plant J*, 23(4), 481–8.
- Seth, A. K. (2010). A MATLAB toolbox for Granger causal connectivity analysis. *J Neurosci Methods*, 186(2), 262–73.

- Shalon, D., Smith, S. J., & Brown, P. O. (1996). A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res*, 6(7), 639–45.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11), 2498–504.
- Skowrya, D., Craig, K. L., Tyers, M., Elledge, S. J., & Harper, J. W. (1997). F-box proteins are receptors that recruit phosphorylated substrates to the SCF ubiquitin-ligase complex. *Cell*, 91(2), 209–19.
- Steindler, C., Carabelli, M., Borello, U., Morelli, G., & Ruberti, I. (1997). Phytochrome A, phytochrome B and other phytochrome(s) regulate ATHB-2 gene expression in etiolated and green Arabidopsis plants. *Plant*, 20, 759–763.
- Steinhaus, H. (1957). Sur la division des corps materiels en parties. *Bulletin of the Polish Academy of Sciences*, 4, 801–804.
- Stekel, D. (2003). *Microarray Bioinformatics*. Cambridge University Press.
- Stephenson, A. G., Doughty, J., Dixon, S., Elleman, C., Hiscock, S., & Dickinson, H. G. (1997). The male determinant of self-incompatibility in *Brassica oleracea* is located in the pollen coating. *The Plant Journal*, 12(6), 1351–1359.
- Storey, J. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3), 479–498.
- Surplus, S. L., Jordan, B. R., Murphy, A. M., Carr, J. P., Thomas, B., & A-H-Mackerness, S. (1998). Ultraviolet-B-induced responses in *Arabidopsis thaliana*: role of salicylic acid and reactive oxygen species in the regulation of transcripts encoding photosynthetic and acidic pathogenesis-related proteins. *Plant, Cell and Environment*, 21, 685–694.
- Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A., Singh, S., Swing, V., Tissier, C., Zhang, P., & Huala, E. (2008). The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res*, 36, D1009–14.
- Tabata, S., Kaneko, T., Nakamura, Y., Kotani, H., Kato, T., Asamizu, E., Miyajima, N., Sasamoto, S., Kimura, T., Hosouchi, T., Kawashima, K., Kohara, M., Matsumoto, M., Matsuno, A., Muraki, A., Nakayama, S., Nakazaki, N., Naruo, K., Okumura, S., Shinpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M., Sato, S., de la Bastide, M., Huang, E., Spiegel, L., Gnoj, L., O’Shaughnessy, A., Preston, R., Habermann, K., Murray, J., Johnson, D., Rohlffing, T., Nelson, J., Stoneking, T.,

- Pepin, K., Spieth, J., Sekhon, M., Armstrong, J., Becker, M., Belter, E., Cordum, H., Cordes, M., Courtney, L., Courtney, W., Dante, M., Du, H., Edwards, J., Fryman, J., Haakensen, B., Lamar, E., Latreille, P., Leonard, S., Meyer, R., Mulvaney, E., Ozersky, P., Riley, A., Strowmatt, C., Wagner-McPherson, C., Wollam, A., Yoakum, M., Bell, M., Dedhia, N., Parnell, L., Shah, R., Rodriguez, M., See, L. H., Vil, D., Baker, J., Kirchoff, K., Toth, K., King, L., Bahret, A., Miller, B., Marra, M., Martienssen, R., McCombie, W. R., Wilson, R. K., Murphy, G., Bancroft, I., Volckaert, G., Wambutt, R., Düsterhöft, A., Stiekema, W., Pohl, T., Entian, K. D., Terryn, N., Hartley, N., Bent, E., Johnson, S., Langham, S. A., McCullagh, B., Robben, J., Grymonprez, B., Zimmermann, W., Ramsperger, U., Wedler, H., Balke, K., Wedler, E., Peters, S., van Staveren, M., Dirkse, W., Mooijman, P., Lankhorst, R. K., Weitzenegger, T., Bothe, G., Rose, M., Hauf, J., Berneiser, S., Hempel, S., Feldpausch, M., Lamberth, S., Villarroel, R., Gielen, J., Ardiles, W., Bents, O., Lemcke, K., Kolesov, G., Mayer, K., Rudd, S., Schoof, H., Schueller, C., Zaccaria, P., Mewes, H. W., Bevan, M., Fransz, P., Institute, K. D. R., Harbor, C. S., in St Louis Sequencing Consortium, W. U., & Consortium, E. U. A. G. S. (2000). Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*. *Nature*, 408(6814), 823–6.
- Takamiya, K. I., Tsuchiya, T., & Ohta, H. (2000). Degradation pathway(s) of chlorophyll: what has gene cloning revealed? *Trends Plant Sci*, 5(10), 426–31.
- Takasaki, T., Hatakeyama, K., Suzuki, G., Watanabe, M., Isogai, A., & Hinata, K. (2000). The S receptor kinase determines self-incompatibility in *Brassica stigma*. *Nature*, 403(6772), 913–6.
- Takeda, T., Amano, K., Ohto, M.-A., Nakamura, K., Sato, S., Kato, T., Tabata, S., & Ueguchi, C. (2006). RNA interference of the *Arabidopsis* putative transcription factor TCP16 gene results in abortion of early pollen development. *Plant Mol Biol*, 61(1-2), 165–77.
- Tatematsu, K., Nakabayashi, K., Kamiya, Y., & Nambara, E. (2008). Transcription factor AtTCP14 regulates embryonic growth potential during seed germination in *Arabidopsis thaliana*. *Plant J*, 53(1), 42–52.
- Taylor, C. B., Bariola, P. A., delCardayré, S. B., Raines, R. T., & Green, P. J. (1993). RNS2: a senescence-associated RNase of *Arabidopsis* that diverged from the S-RNases before speciation. *Proceedings of the National Academy of Sciences of the United States of America*, 90(11), 5118–22.
- The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.*, 25(1), 25–9.
- Theologis, A., Ecker, J. R., Palm, C. J., Federspiel, N. A., Kaul, S., White, O., Alonso, J., Altafi, H., Araujo, R., Bowman, C. L., Brooks, S. Y., Buehler, E., Chan, A., Chao, Q., Chen, H., Cheuk, R. F., Chin, C. W., Chung, M. K., Conn, L., Conway, A. B.,

- Conway, A. R., Creasy, T. H., Dewar, K., Dunn, P., Etgu, P., Feldblyum, T. V., Feng, J., Fong, B., Fujii, C. Y., Gill, J. E., Goldsmith, A. D., Haas, B., Hansen, N. F., Hughes, B., Huizar, L., Hunter, J. L., Jenkins, J., Johnson-Hopson, C., Khan, S., Khaykin, E., Kim, C. J., Koo, H. L., Kremenetskaia, I., Kurtz, D. B., Kwan, A., Lam, B., Langin-Hooper, S., Lee, A., Lee, J. M., Lenz, C. A., Li, J. H., Li, Y., Lin, X., Liu, S. X., Liu, Z. A., Luros, J. S., Maiti, R., Marziali, A., Militscher, J., Miranda, M., Nguyen, M., Nierman, W. C., Osborne, B. I., Pai, G., Peterson, J., Pham, P. K., Rizzo, M., Rooney, T., Rowley, D., Sakano, H., Salzberg, S. L., Schwartz, J. R., Shinn, P., Southwick, A. M., Sun, H., Tallon, L. J., Tambunga, G., Toriumi, M. J., Town, C. D., Utterback, T., Aken, S. V., Vaysberg, M., Vysotskaia, V. S., Walker, M., Wu, D., Yu, G., Fraser, C. M., Venter, J. C., & Davis, R. W. (2000). Sequence and analysis of chromosome 1 of the plant *Arabidopsis thaliana*. *Nature*, 408(6814), 816–20.
- Thompson, J. E., Froese, C. D., Madey, E., Smith, M. D., & Hong, Y. (1998). Lipid metabolism during plant senescence. *Prog Lipid Res*, 37(2-3), 119–41.
- Tommasini, R., Vogt, E., Fromenteau, M., Hörtensteiner, S., Matile, P., Amrhein, N., & Martinoia, E. (1998). An ABC-transporter of *Arabidopsis thaliana* has both glutathione-conjugate and chlorophyll catabolite transport activity. *Plant J*, 13(6), 773–80.
- Tran, P. H., Peiffer, D. A., Shin, Y., Meek, L. M., Brody, J. P., & Cho, K. W. Y. (2002). Microarray optimizations: increasing spot accuracy and automated identification of true microarray signals. *Nucleic Acids Res*, 30(12), e54.
- Turner, J. G., Ellis, C., & Devoto, A. (2002). The jasmonate signal pathway. *Plant Cell*, 14 Suppl, S153–64.
- Ueda, J., & Kato, J. (1980). Isolation and Identification of a Senescence-promoting Substance from Wormwood (*Artemisia absinthium* L.). *Plant Physiol*, 66(2), 246–9.
- van der Graaff, E., Schwacke, R., Schneider, A., Desimone, M., Flügge, U.-I., & Kunze, R. (2006). Transcription analysis of arabidopsis membrane transporters and hormone pathways during developmental and induced leaf senescence. *Plant Physiol*, 141(2), 776–92.
- Vattani, A. (2009). k-means Requires Exponentially Many Iteractions Even in the Plane. In *Proceedings of the 25th annual symposium on Computational Geometry*, (pp. 324–332).
- Vijayan, P., Shockey, J., Lévesque, C. A., Cook, R. J., & Browse, J. (1998). A role for jasmonate in pathogen defense of *Arabidopsis*. *Proceedings of the National Academy of Sciences of the United States of America*, 95(12), 7209–14.

- Viola, I., Manassero, N. U., Ripoli, R., & Gonzalez, D. (2011). The Arabidopsis class I TCP transcription factor AtTCP11 is a developmental regulator with distinct DNA-binding properties due to the presence of a threonine residue at position 15 of the TCP domain. *Biochem J.*, 435(1), 143–55.
- Vitart, V., Baxter, I., Doerner, P., & Harper, J. (2001). Evidence for a role in growth and salt resistance of a plasma membrane H⁺-ATPase in the root endodermis. *The Plant Journal*, 27(3), 191–201.
- Wakefield, J., Zhou, C., & Self, S. (2003). Modeling Gene Expression Over Time: Curve Clustering with Informative Prior Distributions. In J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, & M. West (Eds.) *Bayesian Statistics 7*, (pp. 721–732). Oxford, U.K.: Clarendon Press.
- Wall, M. E., Rechtsteiner, A., & Rocha, L. M. (2003). Singular value decomposition and principal component analysis. In D. Berrar, W. Dubitzky, & M. Granzow (Eds.) *A Practical Approach to Microarray Data Analysis*, chap. 5, (pp. 91–109). Norwell, MA: Kluwer.
- Wang, X., Dong, J., Liu, Y., & Gao, H. (2010). A novel dehydration-responsive element-binding protein from *Caragana korshinskii* is involved in the response to multiple abiotic stresses and enhances stress tolerance in transgenic tobacco. *Plant Molecular Biology Reporter*, 28(4), 664–675.
- Wanner, L., Keller, F., & Matile, P. (1991). Metabolism of radiolabelled galactolipids in senescent barley leaves. *Plant Science*, 78(2), 199–206.
- Wasternack, C. (2007). Jasmonates: an update on biosynthesis, signal transduction and action in plant stress response, growth and development. *Annals of Botany*, 100(4), 681.
- Wasternack, C., Stenzel, I., Hause, B., Hause, G., Kutter, C., Maucher, H., Neumerkel, J., Feussner, I., & Miersch, O. (2006). The wound response in tomato—role of jasmonic acid. *J Plant Physiol*, 163(3), 297–306.
- Watanabe, C., Hachiya, T., Takahara, K., Kawai-Yamada, M., Uchimiya, H., Uesono, Y., Terashima, I., & Noguchi, K. (2010). Effects of AOX1a Deficiency on Plant Growth, Gene Expression of Respiratory Components and Metabolic Profile Under Low-Nitrogen Stress in *Arabidopsis thaliana*. *Plant and Cell Physiology*, 51(5), 810.
- Weaver, L. M., Gan, S., Quirino, B., & 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), 455–69.
- Wenkel, S., Turck, F., Singer, K., Gissot, L., Gourrierc, J. L., Samach, A., & Coupland, G. (2006). CONSTANS and the CCAAT box binding complex share a functionally

- important domain and interact to regulate flowering of Arabidopsis. *The Plant Cell*, 18(11), 2971–84.
- Wingender, E., Dietze, P., Karas, H., & Knüppel, R. (1996). TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res*, 24(1), 238–41.
- Woo, H. R., Chung, K. M., Park, J. H., Oh, S. A., Ahn, T., Hong, S. H., Jang, S. K., & Nam, H. G. (2001). ORE9, an F-box protein that regulates leaf senescence in Arabidopsis. *Plant Cell*, 13(8), 1779–90.
- Woo, H. R., Goh, C.-H., Park, J.-H., de la Serve, B. T., Kim, J.-H., Park, Y.-I., & Nam, H. G. (2002). Extended leaf longevity in the ore4-1 mutant of Arabidopsis with a reduced expression of a plastid ribosomal protein gene. *Plant J*, 31(3), 331–40.
- Woo, H. R., Kim, J. H., Nam, H. G., & Lim, P. O. (2004). The delayed leaf senescence mutants of Arabidopsis, ore1, ore3, and ore9 are tolerant to oxidative stress. *Plant Cell Physiol*, 45(7), 923–32.
- Wu, H., Kerr, K., Cui, X., & Churchill, G. A. (2003). MAANOVA: a software package for the analysis of spotted cDNA microarray experiments. In G. Parmigiani, E. S. Garret, R. A. Irizarry, & S. L. Zeger (Eds.) *The Analysis of Gene Expression Data: an Overview of Methods and Software*, chap. 14, (pp. 313–341). Springer, NY.
- Xie, D. X., Feys, B. F., James, S., Nieto-Rostro, M., & Turner, J. G. (1998). COI1: an Arabidopsis gene required for jasmonate-regulated defense and fertility. *Science*, 280(5366), 1091–4.
- Yabuta, Y., Morishita, T., Kojima, Y., Maruta, T., Nishizawa-Yokoi, A., & Shigeoka, S. (2010). Identification of recognition sequence of ANAC078 protein by the cyclic amplification and selection of targets technique. *Plant Signaling & Behavior*, 5(6), 695.
- Yamada, K., Matsushima, R., Nishimura, M., & Hara-Nishimura, I. (2001). A slow maturation of a cysteine protease with a granulin domain in the vacuoles of senescing Arabidopsis leaves. *Plant Physiology*, 127(4), 1626–34.
- Yanhui, C., Xiaoyuan, Y., Kun, H., Meihua, L., Jigang, L., Zhaofeng, G., Zhiqiang, L., Yunfei, Z., Xiaoxiao, W., Xiaoming, Q., Yunping, S., Li, Z., Xiaohui, D., Jingchu, L., Xing-Wang, D., Zhangliang, C., Hongya, G., & Li-Jia, Q. (2006). The MYB transcription factor superfamily of Arabidopsis: expression analysis and phylogenetic comparison with the rice MYB family. *Plant Mol Biol*, 60(1), 107–24.
- Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., & Ruzzo, W. L. (2001). Model-Based Clustering and Data Transformations for Gene Expression Data. *Bioinformatics*, 17, 977–987.

- Yoo, S. J., Chung, K. S., Jung, S. H., Yoo, S. Y., Lee, J. S., & Ahn, J. H. (2010). BROTHER OF FT AND TFL1 (BFT) has TFL1-like activity and functions redundantly with TFL1 in inflorescence meristem development in Arabidopsis. *Plant J*, 63(2), 241–53.
- Yu, J., Smith, V. A., Wang, P. P., Hartemink, A. J., & Jarvis, E. D. (2004). Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, 20(18), 3594–603.
- Zavaleta-Mancera, H. A., Franklin, K. A., Ougham, H. J., Thomas, H., & Scott, I. M. (1999). Regreening of senescent Nicotiana leaves. I. Reappearance of NADPH-protochlorophyllide oxidoreductase and light-harvesting chlorophyll a/b-binding protein. *J. Exp. Bot.*, 50(340), 1677–1682.
- Zavatele-Mancera, H. A., Thomas, B. J., Thomas, H., & Scott, I. M. (1999). Regreening of senescent Nicotiana leaves. II. Redifferentiation of plastids. *J. Exp. Bot.*, 50(340), 1683–1689.
- Zentgraf, U., Laun, T., & Miao, Y. (2010). The complex regulation of WRKY53 during leaf senescence of Arabidopsis thaliana. *Eur J Cell Biol*, 89(2-3), 133–7.
- Zhang, J., Schurr, U., & Davies, W. J. (1987). Control of Stomatal Behaviour by Absciscic Acid which Apparently Originates in the Roots. *J Exp Bot*, 38(192), 1174–1181.
- Zhou, M., Ma, J., Pang, J., Zhang, Z., Tang, Y., & Wu, Y. (2010). Regulation of plant stress response by dehydration responsive element binding (DREB) transcription factors. *African Journal of Biotechnology*, 9(54), 9255–9269.

Publications

The attached pages of this thesis present the publications in which I feature as an author during the course of the PhD.

Temporal clustering by affinity propagation reveals transcriptional modules in *Arabidopsis thaliana*

Steven J. Kiddle^{1,4,*}, Oliver P. F. Windram⁴, Stuart McHattie^{1,4}, Andrew Mead⁴, Jim Beynon^{1,4}, Vicky Buchanan-Wollaston^{1,4}, Katherine J. Denby^{1,4} and Sach Mukherjee^{2,3*}

¹Warwick Systems Biology Centre, Warwick University, CV4 7AL, UK

²Department of Statistics & ³Centre for Complexity Science, Warwick University, CV4 7AL, UK

⁴Warwick HRI, Warwick University, Wellesbourne, CV35 9EF, UK

Associate Editor: Dr. Joaquin Dopazo

ABSTRACT

Motivation: Identifying regulatory modules is an important task in the exploratory analysis of gene expression time series data. Clustering algorithms are often used for this purpose. However, gene regulatory events may induce complex temporal features in a gene expression profile, including time delays, inversions and transient correlations, which are not well accounted for by current clustering methods. As the cost of microarray experiments continues to fall, the temporal resolution of time course studies is increasing. This has led to a need to take account of detailed temporal features of this kind. Thus, while standard clustering methods are both widely used and much studied, their shared shortcomings with respect to such temporal features motivates the work presented here.

Results: Here, we introduce a temporal clustering approach for high-dimensional gene expression data which takes account of time delays, inversions and transient correlations. We do so by exploiting a recently introduced, message-passing-based algorithm called Affinity Propagation (AP). We take account of temporal features of interest following an approximate but efficient dynamic programming approach due to Qian *et al.* (2001). The resulting approach is demonstrably effective in its ability to discern non-obvious temporal features, yet efficient and robust enough for routine use as an exploratory tool. We show results on validated transcription factor-target pairs in yeast and on gene expression data from a study of *Arabidopsis thaliana* under pathogen infection. The latter reveals a number of biologically striking findings.

Availability: Matlab code for our method is available at <http://www.wsbc.warwick.ac.uk/stevenkiddle/tcap.html>.

Contact: {s.j.kiddle,s.n.mukherjee}@warwick.ac.uk

of the number of discrete time points sampled) has increased. Indeed, it is now common to see studies with 20 or more time points over timescales of hours to days. A central task in the exploratory analysis of these high-dimensional time series is that of identifying subsets of genes which are functionally related, for example transcription factors (TFs) and their targets, genes which share a regulatory program and so on. Following much of the recent literature we call such subsets modules (Bar-Joseph *et al.*, 2003; Segal *et al.*, 2003). Module identification plays a key role both in the generation of experimental hypotheses and in informing subsequent modelling. Microarray data which highlight a set of genes as possibly functionally related can suggest specific follow-up experiments, for example using interventions targeted at module members. Equally, module identification informs further computational work. The inference of gene regulatory networks (e.g. using Bayesian networks or Gaussian graphical models), for example, rapidly grows more challenging in higher dimensions. In the same way, mechanistic models of gene expression (ODE, PDE or statistical mechanical), become much more tractable for small sets of genes. Thus, identifying transcriptional modules can greatly aid downstream, detailed quantitative analysis.

Clustering algorithms are widely used for the purpose of identifying gene modules (e.g. Ghosh & Chinnaiyan, 2001; Heard *et al.*, 2005; Thalamuthu *et al.*, 2006). Such algorithms seek to partition the set of genes into subsets whose within-subset similarity is high relative to between-subset similarity. The most widely used notions of similarity are simple vector distances between temporal profiles, and include the Euclidean distance, Pearson's correlation coefficient (PCC) and Mahalanobis distance (used in Gaussian mixture models). Loosely speaking, these methods seek to find subsets of genes which *look* similar in the sense of having highly correlated expression profiles. This in turn means that these methods are well suited to detecting modules whose members are co-regulated (Yona *et al.*, 2006), for example by a shared TF, and where regulatory events are simultaneous, at least up to the temporal resolution of the dataset.

However, the general strategy of clustering by straightforward profile similarity suffers from a number of drawbacks. First, while

1 INTRODUCTION

Gene expression analysis by microarrays is now a well established approach in high-throughput biology. Time course studies are widely used to probe the dynamics of gene expression and uncover underlying regulatory programs. As costs per array have continued to fall, the temporal resolution of such studies (in the sense

*to whom correspondence should be addressed

it is arguably well suited to certain cases of simultaneous co-regulation, it is not as well suited to finding genes which regulate each other. In these settings there can be a time lag between a change in the profile of the regulator and the corresponding change in its target. At very low temporal resolutions, this may not be an issue, because the changes, if detected, may appear as *de facto* simultaneous. However, at higher temporal resolutions time lags become an important issue; we show experimental examples below.

Second, even when a set of putatively co-regulated genes can be identified, the task of identifying a shared TF remains a challenging one. A widespread approach is to use sequence analysis to discover upstream motifs, shared among module members, which may correspond to TF binding sites. However, even when upstream motifs can be found, TFs that bind to these sequences are often unknown, particularly in higher organisms. This motivates a need for module finding methods which can identify subsets including both regulator and targets directly from expression data.

Third, many existing approaches do not account for transient correlations, in which gene profiles are similar only within a certain time window, and not well correlated outside it. This can arise for example in longer time courses, where the underlying biological process driving profile similarity is itself transient, such that at its end, the genes revert almost to a background level of variation. Two-way clustering or biclustering (Hartigan, 1972; Lazzeroni & Owen, 2002; Balasubramaniyan *et al.*, 2004; Madeira & Oliveira, 2005; Meng *et al.*, 2009) has been used to address the issue of transient correlations. Here, clusters are sought which form subsets of both genes and (contiguous) time points. However, robust biclustering remains computationally challenging on account of the vast number of possible biclusters that can be formed. Finally, inversions in the sense of negative correlation/co-expression can be important when regulatory relationships are repressive, but are not always accounted for by clustering methods.

In order to account for these temporal features, a natural idea is to carry out cluster analysis using richer similarity measures in place of a simple vector distance; this idea appears several times in the literature (Qian *et al.*, 2001; Schmitt *et al.*, 2004; Balasubramaniyan *et al.*, 2004; Smith *et al.*, 2009). However, doing so brings with it a non-trivial computational burden, especially under conditions of high dimensionality and high temporal resolution (and resulting longer time lags). Under Euclidean distance and its variants clusters can be characterized by cluster-level statistics such as the mean; this in turn permits (relatively) fast iterative computation *via* algorithms such as K-means and Expectation-Maximization (EM). In contrast, temporally rich gene-gene similarity measures typically do not give an analogue to cluster mean. The standard approach then is to use an iterative algorithm known as K-centres (or K-medoids) (see e.g. Hastie *et al.*, 2001). However, K-centres is notoriously slow, requiring quadratic time in cluster size to find a cluster centre; it is also known to be highly sensitive to initialization. The resulting difficulty in clustering under rich gene-gene similarity measures has meant that existing work on such measures has not led to a widely applicable alternative to standard clustering.

We note that time delays are well accounted for in graphical model formulations (including dynamic Bayesian networks, state space models and hidden Markov models) where Markov assumptions are used to model these temporal effects. However, these approaches are computationally demanding and statistically challenging for high-dimensional data, and have for these reasons

not usually been exploited to provide practical alternatives to clustering for exploratory analysis. Hierarchical clustering (see e.g. Hastie *et al.*, 2001) and spectral clustering (Shi & Malik, 2000; Ng *et al.*, 2002) address the related but quite distinct problem of partitioning a dataset by recursively comparing pairs of observations. In particular, these methods do not ensure that all points within a cluster are similar to a cluster mean or centre and indeed quite often make splits which lead to clusters which do not have this property.

Here, we address these open issues by putting forward an approach for finding gene modules which incorporates these key temporal features — time lags, transient correlation and inversions — but is computationally efficient enough to provide a practical alternative to standard clustering. We do so by exploiting a recently proposed message-passing-based algorithm called Affinity Propagation (AP) (Frey & Dueck, 2007) which we show, using biological data, to be robust and efficient in this setting. As a similarity measure we choose a dynamic programming formulation due to Qian *et al.* (2001); this is fast but approximate, and we confirm empirically that it is sufficiently powerful to give good results in this setting.

Our work adds to the existing literature in two main ways. First, we put forward an approach for clustering microarray time series data which captures rich temporal features yet is robust, requires little or no user input and is fast enough for routine use in microarray data analysis. For example, in an analysis of real microarray data, this finds a substantially better value of the same objective function than any of 400 runs of K-centres, while requiring a fraction of the total compute time, and no user input whatsoever. Second, we show extensive results on experimental data, highlighting the biological relevance of richer temporal features and the importance of capturing such features during clustering. We are able to cluster together members of a recently identified gene regulatory network whose profiles would not have been clustered together by traditional clustering techniques. We also find several modules which suggest hypotheses to test experimentally.

The remainder of the paper is organized as follows. We begin by reviewing basic ideas and notation for clustering and then describe the methods used here. We show results on a validated set of TF-target pairs in yeast, and on experimental data from a study of *Botrytis cinerea* infection in *Arabidopsis thaliana*. We conclude with a discussion of the shortcomings of our work, possible extensions and its relationship to other methods.

2 BACKGROUND

2.1 Notation

Let X_{it} be the mRNA expression value of gene i at time t . A time series microarray dataset, \mathbf{X} , is a matrix containing the expression values of genes $i \in \mathcal{I} = \{1, 2, \dots, g\}$, for time points $t \in \mathcal{T} = \{1, 2, \dots, T\}$. The complete expression profile for gene i is denoted $X_i = [X_{i1}, X_{i2}, \dots, X_{iT}]^T$.

2.2 Clustering

Clustering is a form of unsupervised machine learning in which observations are partitioned into groups, called clusters, such that within-cluster similarity is large relative to between-cluster

similarity. In the present setting, observations correspond to gene expression profiles X_i .

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

$$J(\{c(i)\}, \{\mu_k\}) = \sum_{k \in \mathcal{K}} \sum_{i: c(i)=k} \|X_i - \mu_k\|^2 \quad (1)$$

where, $\|\cdot\|^2$ denotes (squared) Euclidean distance and $\{c(i)\}$ and $\{\mu_k\}$ are cluster assignments and cluster means respectively.

K-means minimizes this cost function by means of an iterative procedure in which the computation of cluster means alternates with cluster assignment. Mixture-model-based approaches can be viewed as a probabilistic generalization of K-means, in which observations are assigned to clusters in a “soft” manner, under a probability model in which cluster membership is treated as a latent variable. Model fitting is usually accomplished using the EM algorithm; as is well-known, K-means itself arises as a certain limiting case of EM applied to a Gaussian mixture model.

2.2.2 K-centres Cost function Eq. (1) directly uses cluster means $\{\mu_k\}$. In contrast, a matrix of similarities $\psi(i, j)$, $i, j \in \mathcal{I}$ between observations may not give an analogue to cluster mean. In this setting, a standard approach is to characterize a cluster by means of an observation within that cluster, referred to as the *centre* of the cluster. This formulation yields the following cost function:

$$J(\{e(i)\}) = - \sum_{i \in \mathcal{I} s.t. i \neq e(i)} \psi(i, e(i)) \quad (2)$$

where, $e : \mathcal{I} \mapsto \mathcal{E} \subset \mathcal{I}$, $|\mathcal{E}| = K$

is a cluster assignment function which in this case maps observations to the (index of) the corresponding cluster centre.

The K-centres algorithm is a K-means-like heuristic method for optimizing Eq. (2), in which a cluster characterization step is alternated with a cluster assignment step. Absent any notion of mean, the cluster characterization step involves searching over all members of each cluster to minimize within-cluster distance; this requires quadratic time in cluster size. Moreover, K-centres must be initialized, and the initialization can affect which local maximum the method will find.

Thus, while Eq. (2) provides a natural cost function for clustering under a similarity matrix ψ , it can be difficult to obtain good clusters in practice, and moreover to do so robustly and rapidly in applications with a large number of objects to be clustered.

3 METHODS

Here we describe the methods used in the remainder of the paper. We first discuss clustering by Affinity Propagation (AP) and then the similarity measure used here.

3.1 Affinity propagation

Affinity propagation (AP) is an algorithm by which to learn cluster assignments and cluster centres under the K-centres cost function Eq. (2). Like K-centres, AP uses observations themselves to characterize clusters;

however, unlike K-centres AP simultaneously considers *all* observations as candidate centres. Naïvely, this would be computationally intractable; in AP this is accomplished by an efficient message passing formulation (which can be derived as an instance of the max-sum algorithm for factor graphs). Two different kinds of messages are exchanged between observations: *responsibility* $r(i, j)$, which reflects point j ’s suitability as a centre for point i and *availability* $a(i, j)$, which reflects evidence in favour of i choosing j as its centre. Here we briefly describe the AP algorithm, as it used in the present application; for further details we refer the interested reader to Frey & Dueck (2007).

Update equations. AP is provided with a similarity matrix ψ^* , such as the one introduced in section 3.2.

Initially, availabilities $a(i, j)$ are set to zero; “self-similarities” $\psi^*(i, i)$ are given a user-set value s , this is discussed below. Then, responsibilities and availabilities are updated sequentially using the following:

$$r(i, j) \leftarrow \psi^*(i, j) - \max_{j' : j' \neq j} \{a(i, j') + \psi^*(i, j')\} \quad (3)$$

$$\forall i \neq j, \quad a(i, j) \leftarrow \min \left\{ 0, r(j, j) + \sum_{i' : i' \notin \{i, j\}} \max\{0, r(i', j)\} \right\} \quad (4)$$

$$a(j, j) \leftarrow \sum_{i' : i' \neq j} \max\{0, r(i', j)\} \quad (5)$$

A damping factor $\lambda \in [0, 1]$ is used to prevent numerical oscillations: each message is set to a weighted combination of its value from the previous iteration and its updated value, weighted by λ and $1 - \lambda$ respectively. In all our experiments we use a default value of $\lambda = 0.9$. Update equations are iterated until cluster centres remain unchanged for a user-set number of iterations (see below). Then, cluster centres $e(i)$ are given by maximizing over the sum of responsibility and availability:

$$e(i) = \underset{j \in \mathcal{I}}{\operatorname{argmax}} a(i, j) + r(i, j) \quad (6)$$

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

Algorithm parameters. The self-similarity value s influences the number of clusters discovered, higher values giving a greater number of clusters. However, in contrast to the parameter K in K-means and K-centres, this is not a hard specification; rather, the number of clusters found emerges from data, but is influenced by self-similarity s . In this sense, self-similarity is closer in spirit to a shrinkage/regularization strength or Bayesian hyperparameter than a pre-specified number of clusters. Importantly, this means that a default value for s can give good results for a wide range of problems; in all our experiments, we set s to the median of the (off-diagonal entries of) similarity matrix ψ^* . Finally, we call convergence if cluster centres remain unchanged for 100 iterations and further set the overall maximum number of iterations to 1000.

3.2 Similarity measure

As noted in the introduction, there are now a number of biologically plausible similarity measures for gene expression time series in the literature. We choose a similarity score due to Qian *et al.* (2001) which uses alignment to find time lags in gene expression time series, as outlined below. Although approximate, this approach is both efficient and rich enough to capture not only time lags but also inversions and transient correlations, and is therefore well suited to our goals.

Given time series data X_{it} for genes $i \in \mathcal{I}$ at times $t \in \mathcal{T}$, Algorithm 1 returns a matrix $\psi(i, j)$ of similarity scores for all gene pairs (i, j) . Data X_i for each gene profile are assumed to be normalized to mean zero and standard deviation one. For a given pair (i, j) dynamic programming is used to build up a matrix Ω^+ , which compares and scores each alignment between profiles X_i and X_j . Inversion or negative co-expression is captured in a second

matrix Ω^- , whose entries are obtained in a similar manner. Finally, transient correlations are captured by explicitly forcing each entry of Ω^+ and Ω^- to be non-negative. Then, similarity score ψ is simply the highest entry in Ω^+ or Ω^- . The alignment matrices Ω^+ or Ω^- further yield a “match type”, which may be positive/negative and simultaneous/delayed and describes the characteristics of the highest scoring alignment. Specifically, if $\omega^+ = \psi$ the profiles have a positive local correlation, whereas if $\omega^- = \psi$ the profiles have a negative local correlation. Likewise, if ψ is achieved at $\Omega_{t_1 t_2}^+$ or $\Omega_{t_1 t_2}^-$ with $t_1 = t_2$ then the local correlation is simultaneous, otherwise it is time delayed.

For AP a similarity matrix, where identical profiles have a score of zero, is constructed using the following transformation:

$$\psi^*(i, j) = \psi(i, j) - T + 1 \quad (7)$$

Algorithm 1 Computation of similarity measure ψ , following Qian *et al.* (2001).

(1) Initialise $\Omega_{t_0}^+$, Ω_{0t}^+ , $\Omega_{t_0}^-$ and Ω_{0t}^- equal to zero $\forall t \in \mathcal{T} \cup 0$.

(2) Initialise $t_1 = t_2 = 1$.

(3) Calculate $\Omega_{t_1 t_2}^+$ and $\Omega_{t_1 t_2}^-$:

$$\Omega_{t_1 t_2}^+ = \max(\Omega_{t_1-1 t_2-1}^+ + X_{it_1} X_{jt_2}, 0) \quad (8)$$

$$\Omega_{t_1 t_2}^- = \max(\Omega_{t_1-1 t_2-1}^- - X_{it_1} X_{jt_2}, 0) \quad (9)$$

(4) If $t_1 < T$ and $t_2 \leq T$ then set $t_1 = t_1 + 1$ and go to step 3.

(5) If $t_1 = T$ and $t_2 < T$ then set $t_1 = 1$ and $t_2 = t_2 + 1$ and go to step 3.

(6) Let $\omega^+ = \max_{t_1 t_2} \{\Omega_{t_1 t_2}^+\}$ and $\omega^- = \max_{t_1 t_2} \{\Omega_{t_1 t_2}^-\}$. Set: $\psi(i, j) = \max\{\omega^+, \omega^-\}$.

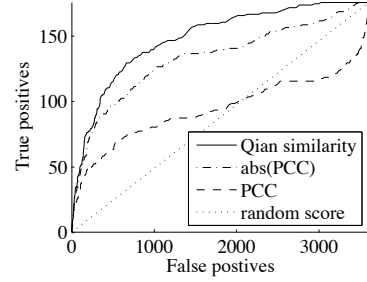
4 RESULTS

We first show results in which we investigate whether richer temporal features are indeed useful in uncovering biological relationships. We then compare the ability of K-centres and AP to cluster real microarray data under similarity matrix ψ . Finally, we present an analysis, using our temporal clustering approach, of a microarray time course experiment we carried out to better understand the response of *A. thaliana* to infection by the pathogen *B. cinerea* (Denby, manuscript in preparation).

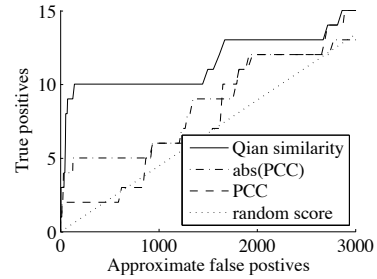
4.1 Validation of similarity measure ψ

We sought to investigate whether the similarity measure ψ does indeed capture biologically important relationships. To this end we used two biological examples, from yeast and Arabidopsis respectively, in which the underlying biology is relatively well understood.

TF-target pairs in yeast. The yeast genome has been well studied and provides a number of validated TF-target pairs. This makes yeast TF-target pairs well suited to a validation study. Here, we used published microarray data (Spellman *et al.*, 1998; Gasch *et al.*, 2000; Qian *et al.*, 2003) of such regulatory pairs, consisting of validated positive and negative examples. The positive examples were chosen from TRANSFAC and SCPD; negative examples were identified by finding genes without the known binding site of the transcription factor or permuting the gene (but not the transcription



(a) Yeast TF-target pairs



(b) Arabidopsis clock module

Fig. 1. Validation results using microarray data. (a) ROC plots obtained from microarray data for validated examples of TF-target pairs in yeast (data from Spellman *et al.* (1998); Gasch *et al.* (2000)). Similarity score ψ outperforms both Pearson’s correlation coefficient (PCC) and its absolute value. The dotted line corresponds to random guesswork. (b) ROC plots obtained from microarray data, comparing the expression profiles of genes from the *A. thaliana* circadian clock with that of random genes. Similarity score ψ outperforms the other measures of similarity, performing roughly twice as well as measures neglecting time lags.

factors) expression profile. The expression profiles cover a total of 79 time points, which gives a relatively high time resolution in line with the general motivation for our approach. We assessed the ability of the similarity score ψ to capture underlying biology by means of a Receiver Operator Characteristic (ROC) analysis. Similarity scores $\psi(i, j)$, for each TF-target pair (positive and negative), were thresholded to yield predictions of TF-target pairs. The predictions were then compared with the list of known positive and negative pairs to yield true positive and false positive rates as a function of threshold level. Varying the threshold gives a curve which is referred to as a ROC curve; this shows the sensitivity and specificity of the analysis across all possible thresholds on a single plot, giving a comprehensive view of the ability of the score to distinguish positive and negative examples. Fig 1(a) shows ROC curves obtained from these yeast data for the similarity score ψ , the widely-used Pearson’s correlation coefficient (PCC) and absolute PCC. The (expected) curve which would be obtained by random chance is also shown for comparison. Similarity score ψ performs better than both PCC and the absolute value of PCC in this instance, suggesting that the score is indeed able to detect instances of direct regulation.

Arabidopsis clock module. The results presented above pertain to direct regulatory relationships between TFs and validated targets. However, the complete set of pairwise relationships in a gene regulatory module naturally includes indirect as well as direct influences; e.g. within a module, if TF *A* has as its target gene *B*, which in turn has target *C*, the pair (*A*, *C*) is an example of an indirect relationship. We therefore sought to complement results from yeast TF-target pairs with a study of a well-studied gene regulatory network in *A. thaliana*. A small network of just six genes has been shown to jointly control the circadian clock in *A. thaliana* (Locke *et al.*, 2006). Microarray data for these six genes were supplemented with data for a further 560 genes, chosen at random from the *A. thaliana* genome. None of the 560 genes were annotated as belonging to the circadian clock (Swarbreck *et al.*, 2008). In the resulting set of pairs, those including only members of the known circadian clock module were treated as positive examples, while those with only one member of the circadian clock were considered to be false positives¹. As the similarity measure is symmetric, we have $\binom{6}{2} \times \frac{1}{2} = 15$ positive examples and $6 \times 560 = 3,360$ negative examples. Data were obtained from leaf samples taken every 2 hours for 48 hours. ROC curves were constructed in a similar manner to the TF-target case above.

Fig. 1(b) shows ROC curves obtained in this way: similarity score ψ very clearly outperforms PCC and its absolute value in this instance. For example 10 (out of 15) true positives are obtained at a cost of 141 false positives; in comparison, PCC requires 1649 and absolute PCC requires 1783 false positives. This suggests that ψ is indeed able to detect both direct and indirect regulation, even under highly sparse conditions, i.e. when true positives are scarce relative to false positives. We note also that the vast gains relative to random selection we see using all three similarity scores confirm that the data are indeed information rich.

4.2 Comparative results

The similarity measure ψ captures a quite different notion of closeness than a straightforward vector distance; we have shown biological evidence in Fig 1 above that in the context of regulatory relationships in time series data, ψ offers a superior ability to discern validated biology. Because of this underlying difference in the notion of closeness, clustering under ψ represents a fundamentally different formulation of the clustering problem than many widely-used methods (Hastie *et al.*, 2001; Ghosh & Chinnaiyan, 2001; Heard *et al.*, 2005; Thalamuthu *et al.*, 2006). In this sense, our approach and these widely used methods address different questions, which makes them difficult to compare directly. However, K-centres (Hastie *et al.*, 2001) represents a natural choice for clustering under the similarity measure ψ ; indeed, it has been used for this purpose in previous work (Qian *et al.*, 2001). We therefore compared our AP-based approach with K-centres, to investigate its ability to find clusters under similarity measure ψ . We used two microarray time series; 4,489 genes over 18 time points from a published study in yeast (Spellman *et al.*, 1998) and 6,000 genes over 24 time points from a study we have carried out on *A. thaliana*

leaves during infection by the necrotrophic fungal pathogen *B. cinerea*.

For each dataset we applied both methods to the full set of genes and also used smaller, randomly selected subsets, to investigate dependence on dimensionality. For each regime of dimensionality, 10 runs of K-centres and one run of AP (which is deterministic) was applied to the data. Since we use the same similarity measure in both cases, the underlying cost function Eq. (2) is identical. AP was applied using default parameters; AP is able to automatically learn a good number of clusters (Frey & Dueck, 2007). To ensure a fair comparison, we set the number *K* of clusters for K-centres to equal the number of clusters discovered by AP in each case. Fig 2(a) shows results obtained using the yeast dataset of Spellman *et al.* (1998), which is a time course of expression profiles of genes from cells synchronised by the addition of alpha pheromone. The *A. thaliana* dataset contains the expression profiles of 6,000 genes shown to be differentially expressed between infected and so-called “mock infected” leaves (i.e. a control set of leaves not inoculated with *B. cinerea* spores, but otherwise kept in identical experimental conditions). Figure 2(b) shows results on the *A. thaliana* data. In each case, boxplots show values of the objective function obtained using K-centres; AP is deterministic and gives a single result in each case.

Fig 2(c) shows an analysis in which we used 400 K-centres runs on the full *A. thaliana* dataset, with each run allowed the same compute time as a single run of our method. Our method is completely deterministic, and therefore not subject to variation due to initial conditions or stochastic steps. It is clear that K-centres is performing significantly worse than our method at producing clusters to minimize cost function (2).

4.3 Temporal clustering of *A. thaliana* time series data

Here, we apply our method to a microarray time series dataset of gene expression in *A. thaliana* leaves during infection by the necrotrophic fungal pathogen *B. cinerea*, as described in Section 4.2. We use the VirtualPlant software platform for GO term over-representation analysis, with p-values calculated using the hypergeometric distribution (G  tierrez *et al.*, 2005).

We first visually highlight the ability of our method to uncover non-obvious clusters by means of an illustrative example. Fig. 3(a) is an example of a cluster whose underlying temporal patterns are sufficiently complex as to make the cluster appear, at first glance, devoid of any coherent pattern. Fig. 3(b) shows the same cluster, adjusted for time lags and inversions: this is now highly coherent.

Application of our method produced 481 clusters; 143 of these were singleton clusters and so were ignored. In Fig. 4 we highlight several clusters which yielded modules with known interactions or novel modules which are biologically interesting.

Circadian clock. Fig. 4(a) shows a cluster which appears to have a 24 hour rhythm. The cluster contains two genes encoding known components of the circadian clock module. Gene *GI* is found to score highly with *LHY* with a delayed and inverted match. The delayed and inverted relationship between the two expression profiles fits extremely well with the known role of *LHY* as a transcriptional repressor of *GI* (Locke *et al.*, 2006). In addition, another member of the cluster, At1g56300, belongs to a class of genes known as Rapid Wounding Response (RWR) genes, which are known to be regulated by the circadian clock (Walley *et al.*, 2007).

¹ Despite these precautions, it is possible that some of the 560 genes are circadianly regulated, as their roles may not currently be fully known. However, it is highly unlikely that any more than a small minority are so regulated.

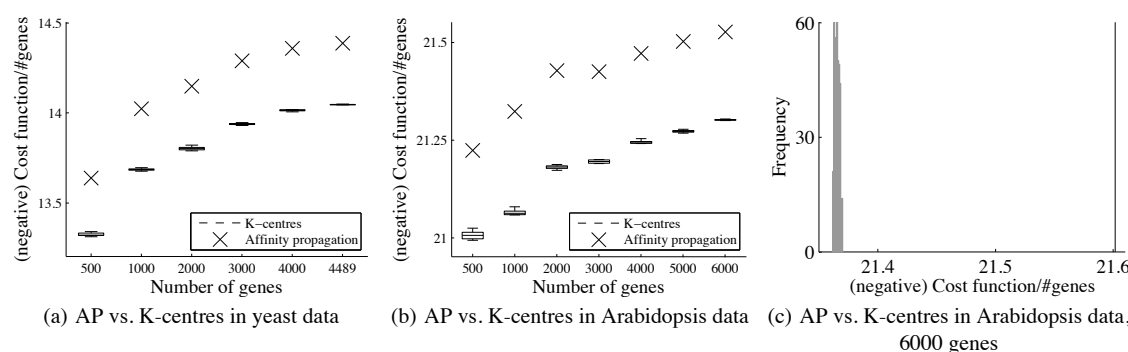
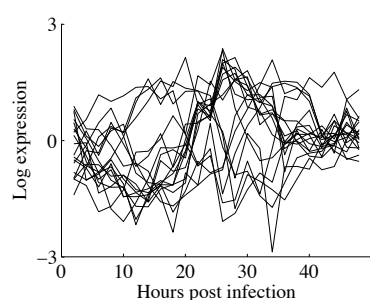
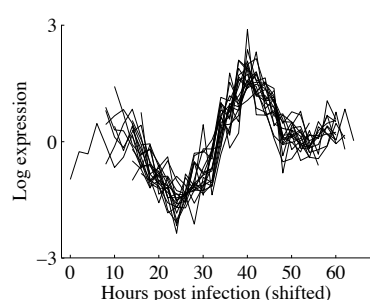


Fig. 2. Here the method proposed in Qian *et al.* (2001) is compared to our method. (a) They are both applied to data from Spellman *et al.* (1998), a time series consisting of 4,489 genes over 18 time points. Various subsets of this are clustered and the cost function, as given in Eq. (2) and then divided by the number of genes in the subset, is reported. 10 runs of K-centres each allowed to take as long as a single run of AP were applied to the data. (b) Both methods are applied to data from *A. thaliana* leaves during infection by the necrotrophic fungal pathogen *B. cinerea*. Various subsets of this are clustered and the cost function, as given in Eq. (2) and then divided by the number of genes in the subset, is reported. 10 runs of K-centres each allowed to take as long as a single run of AP were applied to the data. (c) Here the *A. thaliana* data is clustered again by both methods, but with 400 runs of K-centres (shown in the grey histogram) each allowed to take as long as a single run of AP (black line, representing the result of a single run of AP).



(a) Complex temporal cluster



(b) Complex temporal cluster, adjusted for delay and inversions

Fig. 3. (a) A cluster returned by our method. (b) The same cluster as in the previous figure, adjusted for time delays and anti-correlation. Some profiles in this plot have been shifted in time and/or vertically inverted according to their original match type.

The *de novo* discovery of a small cluster containing these genes is striking in light of the fact that the relationship between these genes took many years and much research effort to uncover. To the best of

our knowledge, the remaining cluster members have no known link to the circadian clock; however, given the highly validated nature of other cluster members, these further genes provide intriguing hypotheses for additional downstream targets.

Ethylene response. Fig. 4(b) shows a second cluster whose members form a striking and biologically coherent group. It is noteworthy that this cluster contains a regulator and known target genes of this regulator. The TF *ORA59* (At1g06160) is in this cluster, along with six genes (At1g59950, At2g43580, At3g23550, At3g56710, At4g11280, At4g24350) that have been previously found to be upregulated in an inducible overexpressor line of *ORA59* (Pré *et al.*, 2008). These genes are also upregulated in the present experiment. Moreover, *ORA59* and another TF, *ERF1* (At3g23240), are believed to jointly regulate *PDF1.2* (Pré *et al.*, 2008) and *ERF1* is also found in this cluster. *PDF1.2* itself is not in the dataset as there is no probe for it on the microarrays used. Both *ORA59* and *ERF1* are known to respond to the plant hormone ethylene; the cluster also has an over-representation, significant at 1%, of the GO term response to ethylene stimulus. Little is known in Arabidopsis about the relative timing of expression of TFs and their direct targets. However, in this case the time resolution of the dataset (2 hr) is apparently not sufficient to pick up a delay between the expression of the regulator *ORA59* and its targets.

Response to abscisic acid. The cluster of 13 genes shown in Fig. 4(c) highlights a novel putative transcriptional module. The only TF in this cluster, At1g71030 (*AtMYBL2*) scores highly for a match with the other genes with a time delay of 6 hours. This cluster has an over-representation, significant at 1%, of the GO term “response to abscisic acid (ABA)” and as such may represent a transcriptional module involved in signalling in response to this hormone. Intriguingly, ABA has been shown to play a role in the interaction between *B. cinerea* and plant hosts (Audenaert *et al.*, 2002; AbuQamar *et al.*, 2006).

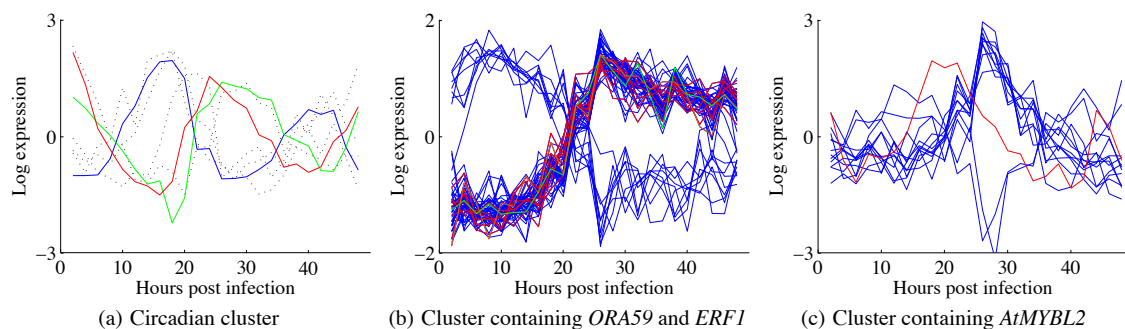


Fig. 4. Clusters found by applying our method to biological data, with default parameters. Data represents *Arabidopsis thaliana* gene expression levels following infection by *Botrytis cinerea*. (a) A circadian module. *LHY* (in blue) is known to be a transcriptional repressor of *GI* (in green). *Atlg56300* (in red) is a Rapid Wounding Response gene, which are known to be regulated by the circadian clock. Here black dotted lines represent the expression levels of four additional cluster members. (b) A cluster containing 6 genes co-regulated by *ORA59* (in red), *ORA59* (in orange) and gene *ERF1* (in green) that is believed to jointly regulate *PDF1.2* with *ORA59* (Pré *et al.*, 2008). (c) A putative transcriptional module. *AtMYBL2* (in red) is the only known transcription factor in this cluster, and peaks 3 time points before the rest of the genes.

5 DISCUSSION

In this paper we have introduced a clustering methodology that can reveal relatively complex temporal features in gene expression time series datasets. Our method is complementary to standard clustering approaches, but aimed specifically at high resolution time series and regulatory modules whose expression profiles have complex temporal relations. Here we discuss the shortcomings of our method, discuss possible extensions and the relationship of our method to others.

As transcriptional assays continue to mature higher resolution datasets are becoming more common; our method is best suited to data with (relatively) high temporal resolution, e.g. more than ten time points. Time series data with fewer time points will naturally give a higher chance of spurious correlations or missed time lags.

The method used here is able to detect transient co-expression, but is not as sensitive as biclustering methods to events occurring only within short windows of time. This is due to the conservative approximation strategy of Qian *et al.* (2001), that divides the overall score by the total number of time points rather than the number of time points where co-expression occurs. We could improve this by giving each ψ value a p-value using an empirical null distribution. For example, a local correlation across 5 time points could be compared to alignments of 5 time points in random expression profiles. A matrix could then be constructed from the p-values and clustered as described above. This would aid in identifying clusters that contain genes that are transiently co-expressed.

The deterministic approach we have used for alignment is effectively a (constrained) time-warping. An interesting extension would be to carry out alignment within a probabilistic framework using a Hidden Markov Model (Rabiner, 1989; Eddy, 1998). However, in such an approach the design of the state space would be crucial in capturing realistic gene expression time series using conventional i.i.d. Gaussian observation models. Moreover the resulting computational burden for all pairwise comparisons of $\sim 10^4$ genes would be considerably greater than the method used here, which is fast enough for interactive use as an exploratory tool.

As AP is an appropriate method to cluster arbitrary matrices of similarity, it provides a flexible framework in which to carry out further work in incorporating complementary information in the similarity measure, e.g. additional time series of the same genes under different environmental conditions, the identity of TFs, presence of known TF binding sites in a gene's promoter, protein-protein interactions, etc.

A recent paper by Smith *et al.* (2009) demonstrated a method called SCOW for aligning the profile of a gene with its profile in another time series. This is subtly different from clustering the profiles of different genes in the same time series, for example, shorting is not appropriate in this case. It also allows for unequal sampling. The problem of unequal sampling was partially treated in Qian *et al.* (2001), but could certainly be improved. One way that suggests itself is to record the spacing between time points, and on the basis of that allow skips in matrices Ω^+ and Ω^- that are acceptable given the spacing.

ACKNOWLEDGEMENT

We would like to thank anonymous referees for their constructive comments. S.K. and S.Mc. are supported by an Engineering and Physical Sciences Research Council/Biotechnology and Biological Sciences Research Council grant to Warwick Systems Biology Doctoral Training Centre. This work was supported by Biotechnology and Biological Sciences Research Council [PRESTA Project, grant number BB/F005806/1] to V.B.W., J.B. and K.D. SM is partially supported by the Engineering and Physical Sciences Research Council. We would like to acknowledge Edward Morrissey for inspiration. Dedicated to Carl Blakey.

REFERENCES

AbuQamar, S. *et al.* (2006) Expression profiling and mutant analysis reveals complex regulatory networks involved in *Arabidopsis* response to *Botrytis* infection. *The Plant Journal*, **48**, 28-44.

- Audenaert, K., DeMeyer, G.B., Höfte, M.M. (2002) Absciscic acid determines basal susceptibility of tomato to *Botrytis cinerea* and suppresses salicylic acid-dependent signaling mechanisms. *Plant Physiol.*, **128** (2), 491-501.
- Balasubramanian, R., Hüllermeier, E., Weskamp, N., Kämper, J. (2004) Clustering of gene expression data using a local shape-based similarity measure. *Bioinformatics*, **21** (7), 1069-1077.
- Bar-Joseph, Z. *et al.* (2003) Computational discovery of gene modules and regulatory networks. *Nature Biotechnology*, **22** (11), 1337-1342.
- Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14** (9), 755-63.
- Frey, B. and Dueck, D. (2007) Clustering by Passing Messages Between Data Points. *Science*, **315**, 972-976.
- Gasch, A.P. *et al.* (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell.*, **11**, 4241-4257.
- Ghosh, D. and Chinnaiyan, A.M. (2007) Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*, **18** (2), 275-286.
- Gutiérrez, R.A., Shasha, D.E., Coruzzi, G.M. (2005) Systems biology for the virtual plant. *Plant Physiology*, **138**, 550-554.
- Hartigan, J.A. (1972) Direct clustering of a data matrix. *Journal of the American Statistical Association*, **67** (337), 123-129.
- Hastie, T., Tibshirani, R. and Friedman, J.H. (2001) *The Elements of Statistical Learning*. Springer-Verlag.
- Heard, N.A. *et al.* (2005) Bayesian coclustering of *Anopheles* gene expression time series: Study of immune defense response to multiple experimental challenges. *Proc. Nat. Acad. Sci.*, **102** (47), 16939-16944.
- Lazzeroni, L. and Owen, A. (2002) Plaid models for gene expression data. *Statistica Sinica*, **12** (2002), 61-86.
- Locke, J.C.W. *et al.* (2006) Experimental validation of a predicted feedback loop in the multi-oscillator clock of *Arabidopsis thaliana*. *Molecular Systems Biology*, **2006**, 1-6.
- Madeira, S.C. and Oliveira, A.L. (2005) A linear time biclustering algorithm for time series gene expression data. *Lecture Notes In Computer Science*, **3692**, 3806-3807.
- Meng, J., Gao, S.J., Huang, Y. (2009) Enrichment constrained time-dependent clustering analysis for finding meaningful temporal transcription modules. *Bioinformatics*, **25** (12), 1521-1527.
- Ng, A., Jordan, M., Weiss, Y. (2002) On spectral clustering: analysis and an algorithm. In T. Dietterich, S. Becker, and Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems 14*. MIT Press.
- Pré, M. *et al.* (2008) The AP2/ERF domain transcription factor ORA59 integrates jasmonic acid and ethylene signals in plant defense. *Plant Physiology*, **147**, 1347-1357.
- Qian, J. *et al.* (2001) Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new biologically relevant interactions. *J. Mol. Biol.*, **314**, 1053-1056.
- Qian, J., Lin, J., Luscombe N.M., Yu, H., Gerstein, M. (2003) Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics*, **22** (13), 1917-1926.
- Rabiner, L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257-286 (1989).
- Schmitt Jr, W.A., Raab, R.M., Stephanopoulos, G. (2004) Elucidation of gene interaction networks through time-lagged correlation analysis of transcriptional data. *Genome Res.*, **2004** (14), 1654-1663.
- Segal, E. *et al.* (2003) Module networks: identifying regulatory networks and their condition specific regulators from gene expression data. *Nat. Genet.*, **34**, 166-176.
- Shi, J. and Malik, J. (2000) Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(8):888-905.
- Smith, A.S., Vollrath, A., Bradfield, C.A., Craven, M. (2009) Clustered alignments of gene-expression data. *Bioinformatics*, **25** (12), 1521-1527.
- Spellman, P.T. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell.*, **9**, 3273-3297.
- Swarbreck, D. *et al.* (2008) The Arabidopsis information resource (TAIR): gene structure and function annotation. *Nucleic Acids Research*, **36**, 1009-1014.
- Thalamuthu, A., Mukhopadhyay, I., Zheng, X., Tseng, G.C. (2006) Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, **22** (19), 2405-2412.
- Walley, J.W. *et al.* (2007) Mechanical stress induces biotic and abiotic stress responses via a novel cis-element. *PLoS Genet.*, **3** (10), 1800-1812.
- Yona, G., Dirks, W., Rahman, S., Lin, D.M. (2006) Effective similarity measures for expression profiles. *Bioinformatics*, **22** (13), 1616-1622.

LARGE-SCALE BIOLOGY ARTICLE

High-Resolution Temporal Profiling of Transcripts during *Arabidopsis* Leaf Senescence Reveals a Distinct Chronology of Processes and Regulation

Emily Breeze,^{a,1} Elizabeth Harrison,^{a,1} Stuart McHattie,^{a,b,1} Linda Hughes,^{a,2} Richard Hickman,^{a,b} Claire Hill,^a Steven Kiddle,^{a,b} Youn-sung Kim,^{a,3} Christopher A. Penfold,^b Dafyd Jenkins,^b Cunjin Zhang,^a Karl Morris,^a Carol Jenner,^a Stephen Jackson,^a Brian Thomas,^a Alexandra Tabrett,^a Roxane Legaie,^b Jonathan D. Moore,^b David L. Wild,^b Sascha Ott,^b David Rand,^b Jim Beynon,^{a,b} Katherine Denby,^{a,b} Andrew Mead,^a and Vicky Buchanan-Wollaston,^{a,b,4}

^aSchool of Life Sciences, University of Warwick, Wellesbourne, Warwick CV35 9EF, United Kingdom

^bWarwick Systems Biology, University of Warwick, Coventry CV4 7AL, United Kingdom

Leaf senescence is an essential developmental process that impacts dramatically on crop yields and involves altered regulation of thousands of genes and many metabolic and signaling pathways, resulting in major changes in the leaf. The regulation of senescence is complex, and although senescence regulatory genes have been characterized, there is little information on how these function in the global control of the process. We used microarray analysis to obtain a high-resolution time-course profile of gene expression during development of a single leaf over a 3-week period to senescence. A complex experimental design approach and a combination of methods were used to extract high-quality replicated data and to identify differentially expressed genes. The multiple time points enable the use of highly informative clustering to reveal distinct time points at which signaling and metabolic pathways change. Analysis of motif enrichment, as well as comparison of transcription factor (TF) families showing altered expression over the time course, identify clear groups of TFs active at different stages of leaf development and senescence. These data enable connection of metabolic processes, signaling pathways, and specific TF activity, which will underpin the development of network models to elucidate the process of senescence.

INTRODUCTION

During leaf senescence, the plant recovers and recycles valuable nutrient components that have been incorporated during growth that would otherwise be lost when the leaf dies or is shed. Efficient senescence is essential to maximize viability in the next season or generation, but premature senescence, a protective mechanism employed when plants are stressed, results in reduced yield and quality of crop plants. During the senescence

process, viability of cells within the leaf is actively maintained until maximum remobilization has occurred (Hörtensteiner and Feller, 2002). This requires meticulous control of each step of the process, regulated by internal and external signals via a series of interlinking signaling pathways involving gene expression changes and influenced by the balance of hormones and metabolites. Thus, senescence is a very complex process involving the expression of thousands of genes and many signaling pathways (Buchanan-Wollaston et al., 2005; van der Graaff et al., 2006). Elucidation of the relative influences of each pathway and the crosstalk between them is crucial in identifying the key regulatory genes that control senescence.

To date, genes with a role in leaf senescence have been identified either by forward genetic screening to find mutants with altered senescence rates followed by cloning of the genes involved or by using reverse genetics for functional analysis of genes that show differential expression during senescence (reviewed by Lim et al., 2007). Many of these altered senescence phenotypes occur as a result of altered hormone signaling, such as reduced ethylene signaling (Grbic and Bleeker, 1995) or increased cytokinin signaling (Kim et al., 2006), both of which result in delayed senescence. However, traditional molecular biology approaches in which one gene or mutant at a time is


¹ These authors contributed equally to this work.


² Current address: Wellcome Trust Centre for Human Genetics, Oxford OX3 7BN, United Kingdom.


³ Current address: Research and Development Center of GenDocs, 544-1 B-Station Bongmyung-Dong, Yusong-Gu, Daejeon, Republic of Korea, 305-301.

⁴ Address correspondence to vicky.b-wollaston@warwick.ac.uk.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: Vicky Buchanan-Wollaston (vicky.b-wollaston@warwick.ac.uk).

 Some figures in this article are displayed in color online but in black and white in the print edition.

 Online version contains Web-only data.

 Open Access articles can be viewed online without a subscription. www.plantcell.org/cgi/doi/10.1105/tpc.111.083345

identified and analyzed have resulted in interesting information but have generally failed to reveal a global picture of senescence regulatory networks, including likely feed-forward, feedback, and crosstalk mechanisms. To understand a system as complex as senescence, where the influence of many external and internal signals is balanced to allow controlled disassociation and dispersal of cellular components, it is essential to study the system in its entirety rather than focus on small parts. The first step in this global analysis is to identify the dynamic changes that are occurring in transcript levels as senescence progresses. Obviously, transcripts are only one part of the regulatory process; factors such as RNA stability, translation rates, protein processing and stability, metabolite concentrations, and many others will have essential roles in the fine-scale moderation of cellular activity. However, transcription plays a key role in regulating both senescence and hormone signaling; therefore, identification of regulatory networks based on transcript levels is an ideal starting point in identifying key switch points in senescence.

Here, we use high-resolution time series microarray data, collected over many time points during the development of the leaf, to identify and characterize the gene expression changes during the different steps that make up the senescence process. The resulting detailed measurement of transcript levels for 22 time points during the developmental process is highly valuable for the investigation of numerous complex processes, such as the discovery of metabolic pathway switches, the identification of key regulatory genes that are active at different time points, and the inference of gene regulatory networks. Analysis of the expression patterns has enabled us to propose a detailed chronology of transcriptional and functional changes during leaf senescence. Promoter motif and transcription factor (TF) analysis has revealed a progression of regulatory genes that influence gene expression at different times during development. Finally, a preliminary model, generated with selected genes from the array data, is presented to illustrate the value of this data set for future network inference.

RESULTS

Growth and Biochemical Changes during Senescence

All measurements in this study were made on samples collected from leaf 7, chosen because senescence and mobilization of constituents from this leaf occur concurrently with flower development and silique filling in our growth conditions and are thus likely to be controlled by developmental signals. Each sample was harvested from an independent *Arabidopsis thaliana* plant at each time point (Figure 1A), and samples were not pooled for any of the analyses. *Arabidopsis* Col-0 plants were grown in controlled conditions until leaf 7 was ~50% of its final size (19 d after sowing [DAS]). This leaf was harvested at defined time points until 39 DAS when it was visibly senescent (~50% of leaf area being yellow, Figure 1B). Samples were taken in the morning (7 h into the light period) and afternoon (14 h into the light period) every other day, resulting in 22 time points in total. Sampling was carried out at these two time points each day to allow us to distinguish genes that are altered in a diurnal rhythm, as well as

being differentially expressed over time; the times were selected based on likely maximum changes in expression. Plants started flowering from around 21 DAS. Leaf 7 started to show yellowing at the tip at around 31 DAS and was 25 to 50% yellow by 37 DAS. By the final sample time (39 DAS), the plants were fully flowering, and siliques were filling. Physiological parameters were measured in the morning samples only (i.e., 11 time points). Sampled leaves reached full expansion by 23 DAS (Figure 1C). However, leaf weight increased significantly between time points up to 25 DAS ($P < 0.01$) and continued to increase, reaching a maximum at 31 DAS when the first signs of yellowing were visible and then declined rapidly after 37 DAS ($P < 0.05$). Once the leaf is fully expanded, weight may continue to increase due to synthesis of macromolecules, expansion of organelles, and water uptake. Similarly, loss of fresh weight is primarily due to the decline in macromolecules and the loss of water as the leaf begins to dry.

Protein and chlorophyll levels are often used as markers for the progression of senescence since both are degraded during the senescence process. Levels of total chlorophyll and protein were measured (Figure 1D). Chlorophyll levels did not change significantly until after 31 DAS, when levels started to fall ($P < 0.001$ from maximum). However, relative protein levels started to drop considerably earlier at 23 DAS ($P < 0.05$ from maximal), which is before the time at which maximum leaf weight is reached. This implies that the leaf weight increase seen up to 31 DAS is not due to new protein synthesis but is probably due to increased water content and possibly continuing increases in cell wall density and membrane and other structural developments. Levels of the small and large subunit (SSU and LSU, respectively) of the photosynthetic carbon-fixation enzyme ribulose-1,5-bis-phosphate carboxylase/oxygenase (Rubisco) increased to maximum at 23 DAS (LSU) and 25 DAS (SSU) and then fell steadily during senescence ($P < 0.001$ from maximum; Figure 1E). Rubisco is abundant in a mature green leaf and has been suggested to have some role as a storage protein (Staswick, 1994). Early degradation of this protein may provide building blocks for synthesis of additional proteins required for senescence without affecting the rate of photosynthesis.

Senescence results in activation of signaling pathways involving the stress-related plant hormones salicylic acid (SA), jasmonic acid (JA), and ABA (Weaver et al., 1998; Morris et al., 2000; He et al., 2002). Levels of these three hormones were measured in the leaf 7 developmental time series and showed phased increases during senescence (Figure 2). SA levels were high in immature leaves, gradually decreased to minimal levels at 31 DAS ($P < 0.001$ from initial maximum) and then rose significantly ($P < 0.05$) from a relatively late stage (35 DAS). ABA levels significantly increased earlier at around 31 DAS ($P < 0.05$), with a subsequent increase to maximum at 39 DAS; JA levels showed a complex pattern with peaks at 25, 33, and 39 DAS.

Microarray Analysis over Multiple Time Points Identified Thousands of Differentially Expressed Genes

Four biological replicates for each time point were used for RNA isolation (88 samples in total), each hybridized as four technical replicates to the two channel microarrays. The resulting gene

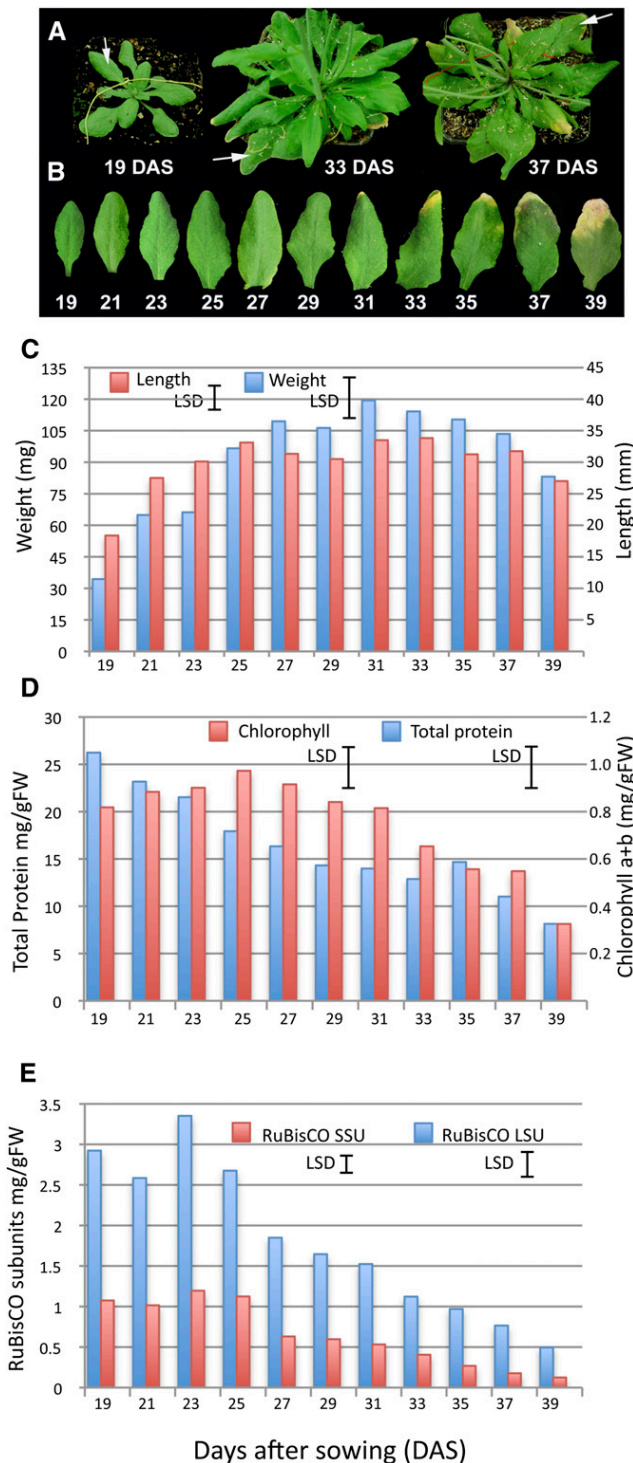


Figure 1. Plant Growth Parameters and Protein and Chlorophyll Measurements.

(A) Appearance of the *Arabidopsis* plants at three different stages of development, 19, 33, and 37 DAS. White arrows indicate leaf 7, the sampled leaf from each plant.

(B) An example of leaf 7 harvested from plants at 19 to 39 DAS (picture shows the morning sample only).

expression profiles were analyzed, and time point means were extracted using a local adaptation of the MAANOVA (MicroArray Analysis Of VAriance) package, which quality checks and normalizes the data and produces data files containing predicted means for each gene; in essence, a single normalized value for each gene for each biological replicate measured at each time point (Wu et al., 2003; Churchill, 2004). Two different data sets were obtained in this way following the completed analysis: the first contained predicted mean values for each of the four biological replicates at each of the 22 distinct time points, therefore including time-of-day variation, whereas the second contained predicted mean values for eight biological replicates for each of 11 d, the values calculated by omitting the time-of-day and day/time-of-day interaction effects from the fitted model. These data sets were both used in the further analysis described below.

F tests, constructed from the variance estimates obtained from the MAANOVA model-fitting process, were used to assess each gene for significant changes in gene expression between time points. The model fitting allowed separate assessments of the variation due to differences between days (averaged across time-of-day samples), differences between time of day (averaged across days), and the interaction between these terms. Significance levels for all tests were adjusted across genes for multiple testing using a step-down false discovery rate (FDR) controlling procedure (Westfall et al., 1998; Benjamini and Liu, 1999), resulting in 8878 genes showing significant ($P < 0.05$) variation due to day of sampling (19–39 DAS). Additional genes were identified as showing significant ($P < 0.05$) variation due to the time of day or the interaction between day of sampling and the time of day, and the numbers of genes having significant test results for combinations of these terms are summarized in a Venn diagram, together with sample expression profiles for each combination (see Supplemental Figure 1 online). The selection process that was used to identify the list of differentially expressed genes used in all further analyses is described in the Methods section and combined information about the adjusted significance levels of the statistical tests with visual examination of expression patterns. The final list of genes used for the analysis described below contains probes for 6323 genes (see Supplemental Data Set 1 online).

We have generated a web tool that illustrates the expression levels of each individual probe on the Complete *Arabidopsis thaliana* MicroArray (CATMA; Allemeersch et al., 2005) array

(C) Length (mm, red bars) and weight (mg, blue bars) of the sampled leaves over the time course. Least significant differences (LSD; 5%, 71 [length] and 99 [weight] *df*) calculated based on the minimum sample size of 6 (length) and consistent sample size of 10 (weight) for comparing pairs of means is shown for each variable, calculated from the ANOVA.

(D) Total protein (blue bars) and chlorophyll a+b (red bars) levels were measured in leaf samples at each stage of development. LSD (5%, 42 [protein] and 43 [chlorophyll] *df*) is shown for both variables.

(E) Levels of the large (LSU) and small (SSU) subunits of Rubisco were estimated from stained polyacrylamide gels. LSD (5%, 42 *df*) is shown for both variables. Values shown in **(D)** and **(E)** represent the means of five independent biological replicates per time point.

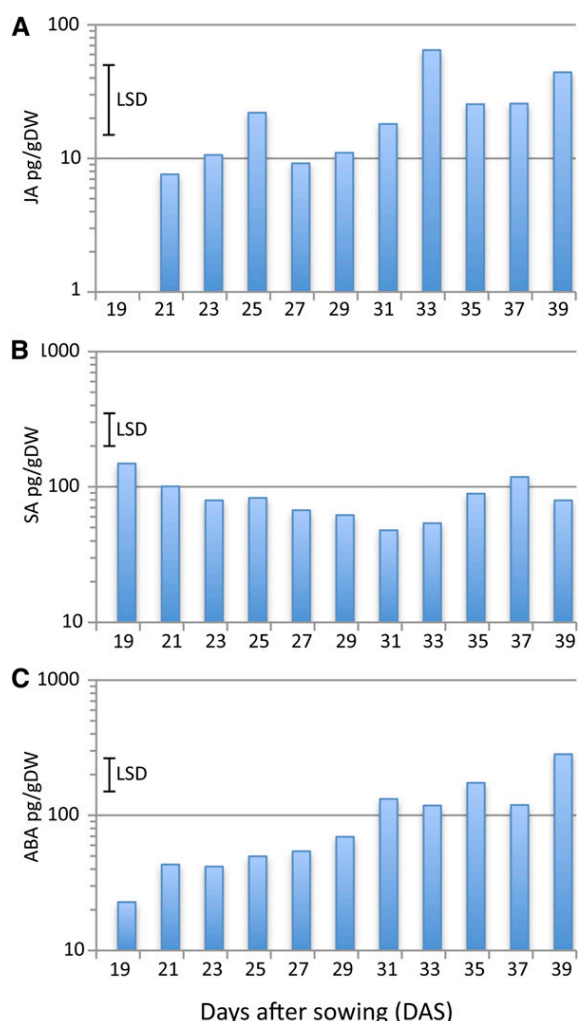


Figure 2. Hormone Levels during Leaf Development.

Levels of JA (A), SA (B), and ABA (C) were measured in leaf 7 harvested at different times during plant development. LSD (5%, 41 df) for comparisons between pairs of means are shown for each hormone, calculated from the ANOVA of log₁₀-transformed data. Values represent the means of five independent biological replicates per time point. DW, dry weight.

[See online article for color version of this figure.]

using the two alternative summaries of the senescence data. Expression patterns for each gene in the 22- and 11-time point data can be viewed in the Data section at <http://go.warwick.ac.uk/presta>.

Clustering Genes by Expression Pattern Illustrates the Extensive Metabolic Changes Occurring during Leaf Senescence

The 6323 differentially expressed genes were clustered using the time series-clustering software SplineCluster (Heard et al., 2006). Clustering analysis of both the 22- and 11-time point data was

performed, and 48 and 74 clusters were obtained respectively. Supplemental Data Set 1 online shows the cluster number for each differentially expressed gene in both the 11- and 22-time point clusters. For reasons of space, only the 22-time point clusters are analyzed in this article (Figure 3A). The heat map (Figure 3B) indicates that there are changes in gene expression at each time point but that there are several time points at which an obvious step change in the transcriptome occurs. Overall, the major switch in gene expression in leaf 7, both in genes up-regulated and downregulated, occurs between 29 and 33 DAS, and the genes identified as differentially expressed can be divided into two major groups, genes in clusters 1 through 24, which are downregulated during this period, and genes in clusters 27 through 48, which are upregulated. Some of the clusters in the center of the heat map show a more complex pattern; for example, cluster 26 genes are downregulated initially and then increase in expression, and genes in clusters 27, 28, and 29 show an initial increase followed by a decrease in expression (Figures 3A and 3B).

A clear diurnal variation in expression is seen with many of the differentially expressed genes, which show higher expression in either the morning or the afternoon samples from the same sample day. Other genes show a distinct morning to night rhythm that did not alter significantly over the 22 d of the experiment. These genes were not selected as being differentially expressed over time but were identified in the MAANOVA analysis as being significantly affected by the time of day term only (1086 genes; see Supplemental Figure 1 online). These genes show clearly that the clock does not decline as late senescence is reached. Morning genes include well-characterized clock genes such as *LATE ELONGATED HYPOCOTYL (LHY)*, *CIRCADIAN CLOCK ASSOCIATED1 (CCA1)*, and *PSEUDO-RESPONSE REGULATOR 7 (PRR7)*; Harmer, 2009) as well as genes involved in light signaling such as *PHYTOCHROME A (PHYA)*, *CRYPTOCHROME 1 (CRY1)*, and *PIF4*, a phytochrome-interacting factor (see Supplemental Figure 2A online). Afternoon genes include *EARLY FLOWERING 4 (ELF4)* and *PHYTOCLOCK1 (PCL1)*, both of which are negatively regulated by *CCA1* and *LHY* (Hazen et al., 2005; Kikis et al., 2005; see Supplemental Figure 2B online).

The 48 clusters of genes identified from SplineCluster analysis of the 22-time point data were analyzed using the gene ontology (GO) enrichment tool BiNGO (Maere et al., 2005). Initially, the two groups of genes showing either decreasing (clusters 1–24) or increasing (clusters 27–48) expression during leaf development were analyzed for overrepresented functions using the GoSlim Plants annotation. The BiNGO-derived graph (Figure 4) illustrates the most highly significant enrichment of specific functions. Downregulated genes are significantly enriched for genes linked to the plastid and thylakoid, and with functions in metabolic processes, particularly photosynthesis and carbohydrate and amino acid metabolism. A more detailed investigation using all GO terms (see Supplemental Table 1 online) shows overrepresentation for genes involved in chloroplast activity such as photosystem (PS) I and II, carbon fixation, chlorophyll (tetrapyrrole) biosynthesis, and amino acid metabolism. All these functions are essential for a growing and active leaf but are downregulated during senescence, when cellular structures are dismantled.

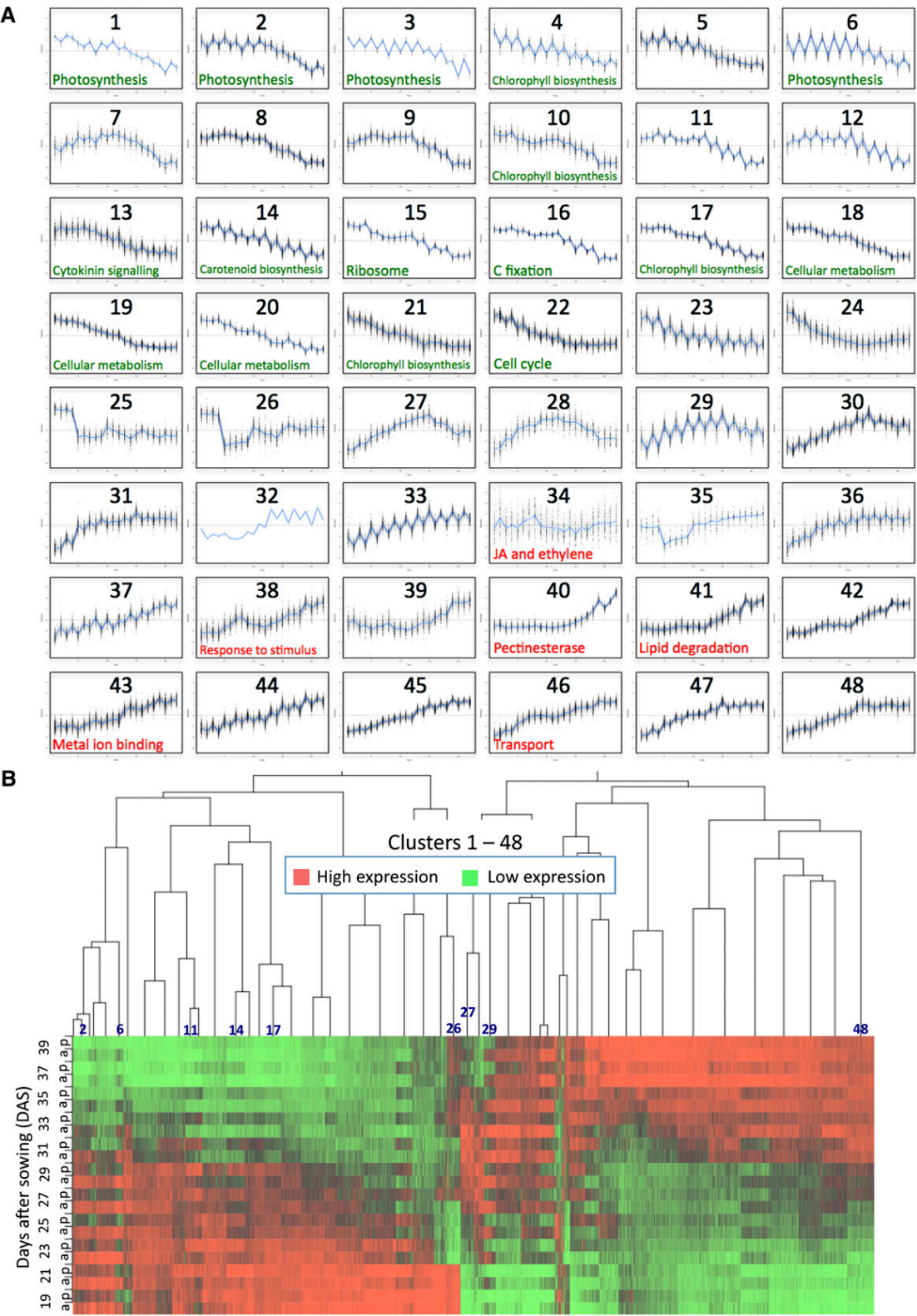


Figure 3. SplineCluster Analysis of Differentially Expressed Genes.

Upregulated genes show a very different picture, illustrating clearly the protective steps the plant takes to respond to the stress generated by the degradative and mobilization functions that occur during senescence (Figure 4). Only two cellular component terms are overrepresented, peroxisome and vacuole. Within the molecular function annotations, only transporter, protein binding, and transcription are overrepresented, whereas there is significant enrichment for stress response and catabolic processes in the biological process terms. In more detail (see Supplemental Table 1 online), enrichment is seen for genes involved in response to stimulus, particularly ABA and ethylene, and many stress responses such as osmotic, salt, and water stress. Enrichment of genes involved in metal ion binding is interesting. Many of these genes (64/222) encode zinc finger (C3H4-type RING) proteins, which may be involved in targeting specific proteins for ubiquitination and degradation. Other zinc binding proteins present have DNA binding activity and may act as TFs. Also, copper chaperones, metallothioneins, calcium binding proteins, and metal ion transporters are represented, which may illustrate the importance of the remobilization of valuable metal ions. Autophagy genes are a significant group; 15 *Arabidopsis* genes involved in autophagy are upregulated during senescence, showing the key role of autophagy in the controlled degradation of cellular components.

This global analysis is highly informative, as it shows broad classes of genes altered in expression during senescence and indicates the processes that are changing. Analysis of enriched GO terms in individual clusters should help to elucidate the chronology of gene expression and associated metabolic activities (Figure 3, see Supplemental Data Set 2 online). Not surprisingly, there was a very strong representation of photosynthesis-related genes in many of the clusters of downregulated genes. Clusters 2, 3, and 6, all of which show a strong diurnal variation with higher morning expression (Figure 3A), are highly enriched with photosynthesis genes, particularly those for the light reaction. Cluster 16, which shows less diurnal change, contains genes encoding enzymes such as Rubisco that are involved in carbon fixation. Clusters 4, 10, 17, and 21 are enriched for chlorophyll biosynthesis genes, and clusters 18 and 19 contain genes involved in cellular biosynthesis such as those for amino acid, polysaccharide, and lipid metabolism. Cluster 15 contains many genes encoding ribosomal proteins. Downregulation of these groups of genes reflects the shutdown of cellular biosynthetic activity as senescence occurs, and the coregulation is an indication of the organized control of this process. Cluster 13 is enriched for cytokinin signaling genes; a reduced level of

cytokinin is a key signal that initiates the senescence process (Noodén et al., 1990).

GO terms enriched in the clusters of genes showing increased expression during senescence are less informative than those for the downregulated gene clusters. Certain clusters are enriched for stress-related genes, e.g., genes involved in JA and ethylene signaling are overrepresented in cluster 34. Other clusters are enriched for genes involved with macromolecule degradation, such as clusters 40 and 41 containing genes involved in carbohydrate and lipid degradation, respectively. Metal ion binding, particularly calcium binding, is overrepresented in cluster 43 and transporter genes in cluster 46.

Distinct Pathways Become Active at Different Times during Senescence

Although SplineCluster is useful in identifying groups of genes that are coexpressed and hence may be coregulated across the entire time series, it is not easy to divide these clusters according to their time of differential expression because the overall pattern of expression is the driving factor for cluster membership. To identify an ordering of events, the rate of change of gene expression (gradient) was inferred using Gaussian process (GP) regression applied to the 11–time point data set (described in detail in Supplemental Methods 1 online). Where data are sufficiently time resolved, this method can be used to identify the time points at which the gradient of a gene's expression profile is significantly positive (increased), significantly negative (decreased), or not statistically different from zero (steady), whereas for less resolved data, it will identify times of significant change to the derivative of the global trend. The results are illustrated using the well-characterized, senescence-enhanced gene *SAG12* (Figure 5). Expression profiles (Figure 5A) are used to train a GP model of gene expression (Figure 5B), after which a GP model of the gradient is obtained (Figure 5C) and used to identify whether the gradient at any time is sufficiently far from zero at three different significance thresholds (Figure 5D). A numeric representation of Figure 5C is shown in Figure 5E and suggests that *SAG12* expression first becomes significantly enhanced around 31 DAS. This method can also be used to show when the gene expression gradient is maximal, i.e., the time of most rapid change. For example, the maximum change of the expression for *SAG12* occurs between 33 and 35 DAS (Figure 5E).

After examining the results for a number of genes, a significance stringency of two standard deviations was taken to

Figure 3. (continued).

SplineCluster analysis was performed on the 22–time point data using normalized data for the 6323 differentially expressed gene probes (average of the four biological replicates).

(A) Forty-eight clusters were obtained. The blue line on each plot represents the mean expression profile for the cluster. The genes present in each cluster may be viewed in Supplemental Data Set 1 online. Selected enriched GO terms (data shown in Supplemental Table 1 online) are indicated, in green for downregulated and red for upregulated genes.

(B) The heat map illustrates the expression profiles for genes in each cluster, with red representing high expression and green representing low expression. Morning (a) and afternoon (p) data are shown for 19 to 39 DAS. Darker shades show intermediate levels of expression. A few cluster positions are identified to compare with the cluster profiles shown in **(A)**.

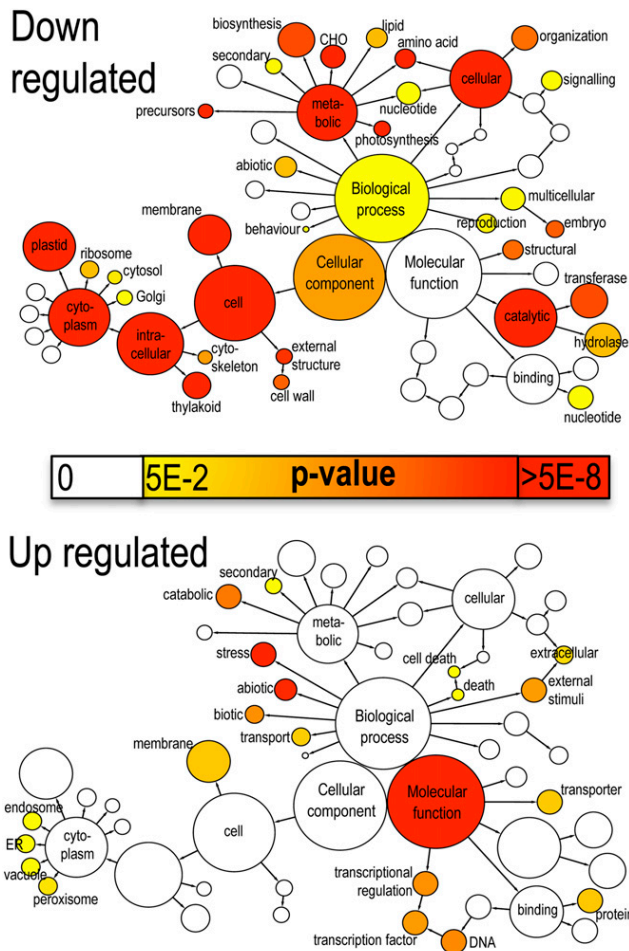


Figure 4. Enriched GO Terms in Genes Upregulated and Downregulated during Senescence.

The network graphs show BiNGO visualization of the overrepresented GO terms for the combined clusters of genes either downregulated (clusters 1–24, 2849 genes) or upregulated (clusters 27–48, 3292 genes) during senescence. Categories in GoSlimPlants (Maere et al., 2005) were used to simplify this analysis and the same nodes are shown on both graphs. Uncolored nodes are not overrepresented, but they may be the parents of overrepresented terms. Colored nodes represent GO terms that are significantly overrepresented (Benjamini and Hochberg corrected P value < 0.05), with the shade indicating significance as shown in the color bar. A more detailed analysis of the GO categories is shown in Supplemental Table 1 online. ER, endoplasmic reticulum.

represent a sufficient distance from zero and was used to generate discrete representations of the state of a gene (Figure 5D) for the 6323 differentially expressed genes. The resulting data were then sorted according to the time of first differential expression to identify 19 clusters (see Supplemental Data Set 3 online). GO term enrichment within the 19 clusters or subsets provided more clarity on the cellular and metabolic activities, showing step changes at each time point during the experiment than was gained from the cluster analysis described above (Figure 6, see Supplemental Data Set 4 online).

Downregulated Gene Clusters Show Step Changes in Cellular Dismantling

It is clear that there are progressive changes in genes being downregulated as senescence progresses, and these are highly informative in indicating changes in metabolic pathways. Genes downregulated from the first time point (19 DAS, cluster D1; see Supplemental Data Sets 3 and 4 online), before the leaf is fully expanded, are enriched for genes involved in amino acid metabolism, including those for biosynthesis of Arg, Trp, Lys, and Gln. Genes involved in tRNA aminoacylation and over 30 ribosomal protein genes are downregulated at 21 DAS (cluster D2), indicating that expansion of the ribosomal content of the cells has slowed down. This suggests that large-scale *de novo* protein synthesis has ceased and that leaf cells are fully developed and equipped for activity.

Many tetrapyrrole or chlorophyll biosynthesis genes are first downregulated at 23 DAS (cluster D3), including the two genes encoding HEMA (glutamyl-tRNA reductase), which catalyzes the rate-limiting and first committed step in tetrapyrrole biosynthesis, and two genes encoding the D subunit of Mg-chelatase, part of the enzyme that diverts the tetrapyrrole pathway toward chlorophyll biosynthesis (Tanaka and Tanaka, 2007). Thus, the requirement for *de novo* chlorophyll biosynthesis appears to cease at 23 DAS, indicating that all chloroplasts are fully developed. Three genes involved in a branch of the carotenoid biosynthesis pathway (*LUT1*, 2, and 5) show a correlated drop in expression at this stage. These genes encode enzymes in the pathway leading from trans-lycopene via α -carotene to lutein, the major carotenoid component in the leaf with an important role in light-harvesting complex-II structure and function and in photoprotection (Kim and DellaPenna, 2006). In addition, expression of three cytokinin-inducible transcription repressors (response regulators *ARR4*, 6, and 7) that mediate a negative feedback loop in cytokinin signaling (Hwang and Sheen, 2001) also drops at this time point.

At the next stage (25 DAS, cluster D4), there is significant overrepresentation of genes involved in fixation of carbon dioxide or the Calvin cycle, including two Rubisco small subunit genes and sedoheptulose biphosphatase, a key enzyme involved in the regeneration of the CO₂ acceptor molecule, ribulose-1,5-bisphosphate. The reduction in expression of the two Rubisco small subunit genes correlates with the reduction in protein levels shown in Figure 1E and indicates that photosynthetic activity probably starts to drop at this stage. At 27 DAS (cluster D5), expression of genes involved in Glycine metabolism declines, including Gly decarboxylase and Ser trans hydroxymethyl transferase 1, both involved in photorespiration, which presumably is also less important as photosynthesis becomes less active. Interestingly, five genes designated as *HIGH CHLOROPHYLL FLUORESCENCE PHENOTYPE* (*HCF101*, 109, 152, 173, and 208) are downregulated together at 27 DAS. Several such genes have been shown to have a role in maintaining the stability of chloroplast-encoded transcripts (Meurer et al., 1996; Meierhoff et al., 2003), and it may be that reduced expression of these genes enables enhanced degradation of photosynthetically related transcripts in the chloroplast. Finally, gene clusters that show expression that declines at 29, 31, and 33 DAS

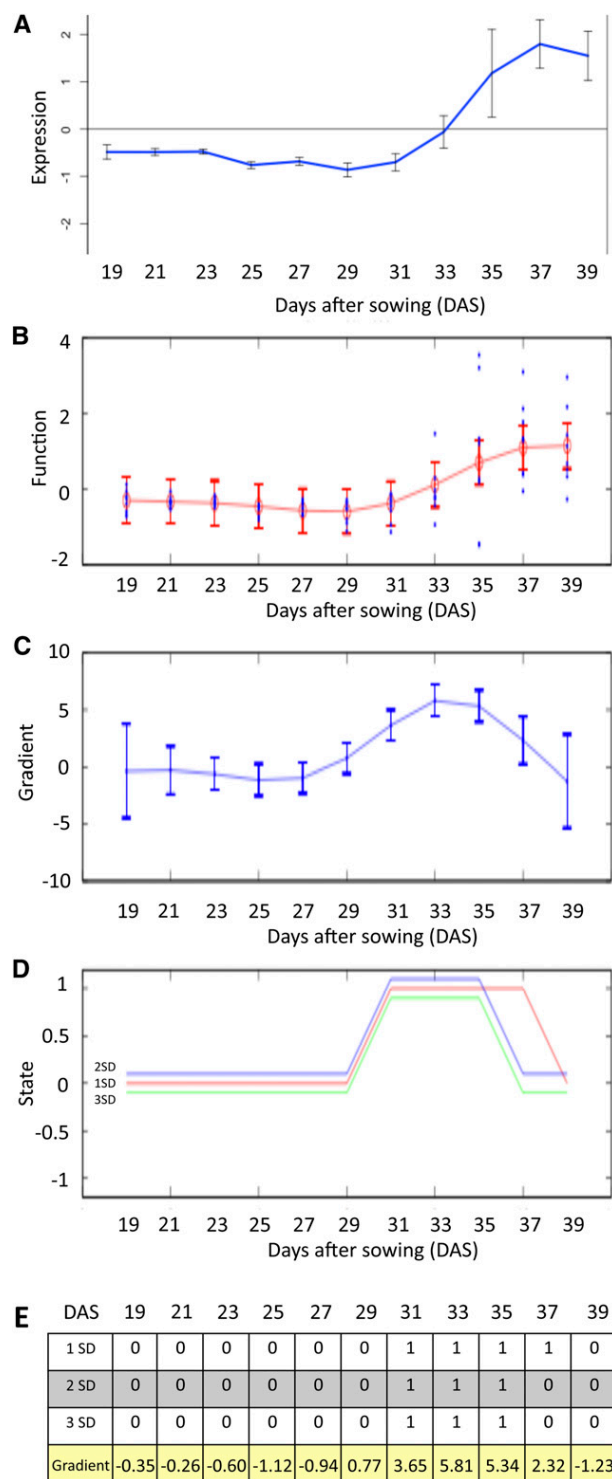


Figure 5. Gradient Analysis of *SAG12* to Identify the Time of First Significant Change in Expression.

(A) Expression levels from microarray data (output from Gene Viewer). The blue line shows the mean of the eight replicates ($n = 8$; error bars = standard deviation, SD).

(clusters D6, D7, and D8) are significantly overrepresented for photosynthesis-related genes. Cluster D7 includes the gene-encoding TF GOLDEN2-LIKE (GLK2) that, together with its functional homolog GLK1, has been shown to coordinate expression of the photosynthesis apparatus genes in *Arabidopsis* (Waters et al., 2009). Inducible expression of *GLK2* resulted in significantly increased expression of many photosynthesis-related genes (Waters et al., 2009), including those for the PSII chlorophyll binding proteins LHC2 2.2, 4.2, and 6 that are found in the same cluster of downregulated genes as *GLK2*, together with many others encoding subunits of the PSI and PSII complexes. The observation that expression of many photosynthesis-related genes is maintained until this late stage of development implies that there must be a continued requirement of these transcripts to retain sufficient energy production for the senescence process to occur.

Upregulated Gene Clusters Illustrate the Complexity of the Senescence Process and Reveal Novel Groups of Coregulated Genes

Genes that show increased expression at different time points during senescence were divided into clusters based on the time of first significant increase, but these clusters were also subdivided further depending on the subsequent expression patterns (see Supplemental Data Sets 3 and 4 online). This separation revealed additional enriched GO terms, as shown in Supplemental Data Set 4 online and Figure 6.

Many autophagy-related (*ATG*) genes are enhanced in expression from the start of the experiment (cluster U1), indicating that there may be a role for these proteins even before the leaf is fully expanded. Autophagy has a key role in the senescence process, and accelerated senescence has been observed in a number of autophagy-defective mutants (Doelling et al., 2002; Hanaoka et al., 2002; Yoshimoto et al., 2004). Nine of the 15 upregulated autophagy genes show increased expression from the first time point, with five others upregulated at 21 or 23 DAS and one, *ATG7*, being upregulated at 29 DAS. Investigation of the overall expression patterns of the autophagy genes shows four genes, *ATG7*, *ATG8H*, *ATG8A*, and *ATG8B* that show correlated and rapidly increased expression between 29 and 31 DAS (see Supplemental Figure 3A online). In yeast, *ATG7p* has been shown to be required for activation of *ATG8p* to allow conjugation with phosphatidylethanolamine (Ichimura et al., 2000), and the resulting *ATG8p*-phosphatidylethanolamine conjugates

(B) Expression levels used by the GP regression, showing the eight biological replicates and the 95% confidence interval.

(C) GP model of the gradient showing 95% confidence interval.

(D) Change in gradient measured at each time point, shown at three different significance values: 3 SD, 2 SD, and 1 SD. Positive value of 1 shows an increased expression, 0 shows no significant change in expression, and -1 shows a significant decrease in expression.

(E) Data output for each significance value, i.e., whether gradient is significantly positive or steady at each time point, with the actual gradient value shown below.

[See online article for color version of this figure.]

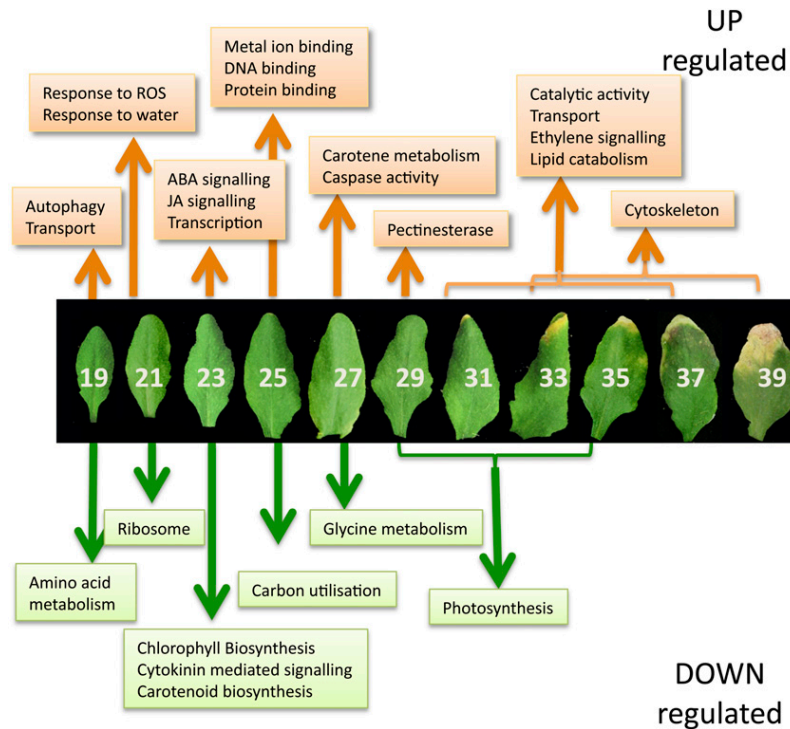


Figure 6. Metabolic Processes Initiated or Repressed at Different Time Points during Development.

Enriched GO terms were identified using BiNGO (Maere et al., 2005) in groups of genes that show first significant upregulation or downregulation at each time point during leaf development and senescence. ROS, reactive oxygen species; white numbers indicate DAS.

assist in the formation of the autophagosome. Thus, the timing of expression of *ATG7* (around 29 DAS) may be the key control point for autophagy activation in senescing leaf cells.

Genes induced at 21 DAS (cluster U2) are enriched for response to oxidative stress. These include TFs such as *DREB2A*, a key regulator in drought and heat stress responses (Sakuma et al., 2006), and *LSD1*, a zinc finger that monitors superoxide levels and regulates cell death (Epple et al., 2003). Increased expression of the mitogen-activated protein kinase MPK7, which is also induced by hydrogen peroxide treatment and enhances plant defense responses (Dóczi et al., 2007), and heat shock proteins such as HSP70, a stress-enhanced heat shock chaperone with a protective role, also indicates that the plant is protecting itself from the detrimental effects of oxidative stress caused by the early stages of senescence. Genes involved in response to water deprivation are also overrepresented; *DREB2A* described above, *Arabidopsis HISTIDINE KINASE3 (AHK3)*, a stress-responsive gene that has been shown to influence cytokinin control of leaf longevity (Kim et al., 2006), dehydration-responsive genes such as *ERD1* and *ERD14*, and *RAB18*, *ABF2*, and other ABA-responsive genes.

Genes involved in both JA and ABA responses are clearly overrepresented in those whose expression increases at 23 DAS (cluster U3). Many of these show a peak of activity at this time point followed by a drop, and this subset is highly enriched for JA biosynthetic genes (cluster U3_1). JA-related genes upregulated

at this time point include genes required for JA biosynthesis such as two lipoxygenases, two allene oxide cyclase genes, *AOC1* and *AOC4*, and 12-oxophytodienoate reductase. This increase correlates with a peak in levels of JA detected at 25 DAS (Figure 2). Also, genes implicated in controlling JA responses are up-regulated, including the TF *MYC2*, and jasmonate ZIM-Domain genes, *JAZ1*, *JAZ6*, and *JAZ8*. JAZ proteins are repressors of JA signaling, binding to *MYC2* and preventing its action (reviewed in Staswick, 2008). *MYC2* is also involved in expression of ABA response genes, and this may be the cause of the increased expression of ABA-related genes at this time point. ABA levels only show a large increase later in senescence (Figure 2), but several ABA-signaling genes (e.g., *ABI1* and *AFP1*) and dehydration response genes whose expression is induced by ABA (e.g., *RD20* and *RD26*; Fujita et al., 2004; Choudhury and Lahiri, 2011) are upregulated at 25 DAS. This suggests a potential coordination of JA and ABA responses at this early stage of senescence.

Many genes encoding TFs are first upregulated at 23 DAS, including four WRKY factors, eight NAC domain (for *Petunia hybrida* NAM and for *Arabidopsis* ATAF1, ATAF2, and CUC2) proteins, 10 zinc finger proteins, and the Nuclear Factor Y sub-unit NF-YA4, which has been implicated in regulating endoplasmic reticulum stress (Liu and Howell, 2010). Many of these genes show an increased expression followed by a fall in expression later in senescence (cluster U3_3), whereas others

show a continued increase from this time until later in senescence (see Supplemental Data Set 3 online). These are likely candidates for the control of later senescence-related processes.

At 25 DAS, the cluster of genes upregulated (cluster U4) is enriched for metal ion binding proteins, including many genes encoding DNA binding proteins, TFs, calcium-signaling genes, etc. The subgroup of this cluster that shows a pattern of increased followed by decreased expression (cluster U4_1) is enriched for genes with a protein binding function. There are five C3HC4-type RING finger protein binding genes in this group, which presumably have a role in regulating specific protein levels via the ubiquitination pathway.

At 27 DAS, there is an interesting overrepresentation of genes involved in carotene metabolism (cluster U5). The three genes involved are a β -carotene hydroxylase and two carotenoid cleavage dioxygenase genes, *CCD7* and *CCD8*. Carotenoids are precursors of signaling molecules that regulate shoot branching in *Arabidopsis*, and *CCD7* and *CCD8* mutants, *max3* and *max4*, respectively, show increased lateral branching (Ongaro and Leyser, 2008). These genes are involved in the production of a strigolactone-related signaling molecule (Gomez-Roldan et al., 2008). Interestingly, another gene that has a shoot-branching role, *MAX2*, was originally identified as *ORE9*, encoding an F box leucine-rich repeat protein required for normal leaf senescence (Woo et al., 2001). Mutants in *MAX2* show increased branching, indicating that this protein is a regulator of the strigolactone signal. The *ORE9/MAX2* gene also shows senescence-enhanced expression and is upregulated at 21 DAS. Enhanced expression of all three genes that regulate shoot branching in a senescing leaf and the fact that a mutant in *ORE9/MAX2* shows delayed senescence indicate that there may be a role for the novel strigolactone-like hormone in regulating an aspect of leaf senescence.

Caspase activity is also an enriched GO term in genes upregulated at 27 DAS due to increased expression of two of the nine *Arabidopsis* genes encoding potential caspase counterparts, metacaspases *MC6* and *MC9*. The other seven *Arabidopsis* metacaspase genes do not show differential expression during senescence. In other organisms, caspases play an essential role in controlling and executing programmed cell death (PCD), and two metacaspase genes, *MC1* and *MC2*, have been shown recently to control pathogen-induced PCD in *Arabidopsis* (Coll et al., 2010). Links between senescence and plant PCD are tenuous, but the coregulated expression of these two caspase-like genes at this time point may indicate that they have a role in the degradative processes and/or cell death that occur in leaf senescence. The autophagy gene *ATG7* also shows initial enhanced expression at this time point as described above, and this may be a significant link showing that, under our growth conditions, the first degradative processes of senescence are initiated at 27 DAS.

At the next time point (29 DAS, cluster U6), senescence-related degradation processes are shown by enrichment in genes for cell wall degradation. Six upregulated genes encode pectinesterase, involved in the degradation of plant cell wall pectin components. These enzymes and others such as xylosidase, glucosyl hydrolase, β -glucosidase, pectate lyase, and

pectin methylesterase inhibitor, may have a role in controlling the degradation of cell wall components and releasing sugars for respiration (Lee et al., 2007). Similarly, genes upregulated at 31 DAS (cluster U7) are highly enriched for catalytic activity, reflecting the considerable degradation that is underway. These include additional carbohydrate-degrading enzymes such as pectinesterases, glycosyl and glucosyl transferases, and polygalacturonase, and several proteases including the well-characterized senescence-enhanced Cys protease *SAG12*, which may have a role in chloroplast degradation (Martínez et al., 2008). Two upregulated genes, *LACS6*, encoding a long-chain acyl-CoA synthetase (Shockey et al., 2002), and *ACX1* (acyl-CoA oxidase), encoding the enzyme that catalyzes the first step in fatty acid β -oxidation in the peroxisome (Fulda et al., 2002), are involved in mobilizing membrane lipids via β -oxidation, likely to provide an energy source to fuel the senescence process. This could act as an initiator of precursors for jasmonate biosynthesis because levels of this hormone increase after this time point (Figure 2).

By the later time points in this experiment, after 31 DAS, the senescing leaf becomes more and more heterogeneous, with some cells within the leaf being at a more advanced stage of senescence than others and more variability between biological replicates. This means that there is less clarity in the functions of different groupings of genes that are differentially expressed at each time point after 31 DAS. GO term enrichment analysis of the 31, 33, and 35 DAS groups combined (clusters U7, U8, and U9) illustrates the degradation and mobilization of nutrients, with 44% of these genes involved in catalytic activity, with lipid catabolism highly represented, and 10% involved in transport. Response to chemical stimulus is also high, with two of the three *Arabidopsis* genes annotated detection of ethylene stimulus, i.e., ethylene *ETR1* and *ACC OXIDASE2* (*ACO2*) being upregulated late in senescence, indicating that the ethylene regulation of senescence may have a significant role at this time.

A surprising group of genes identified by this analysis is downregulated for most of the time course followed by a significant increase in expression at 35 or 37 DAS (clusters U8_1 and U9). This group is highly enriched for genes involved in the cytoskeleton (see Supplemental Figure 3B online), with members of the α -tubulin family (*TUA2*, 4, and 5), actin genes, *ACT3* and *ACT11*, and two aurora genes (*AURORA1* and *AURORA2*) encoding kinase proteins that have a role in histone phosphorylation and have been reported to be associated with microtubule spindles and abundantly transcribed only in dividing cells (Demidov et al., 2005, 2009). Downregulation of this group of genes after completion of the cell division and expansion stages of leaf development is to be expected, but the increase in expression at the end of senescence is unanticipated. *AURORA1* has been shown to phosphorylate histone H3 at Ser10 (Demidov et al., 2009), and, in mammalian cells, this modification has been suggested to have a crucial role in transcription and apoptosis as well as in cell division (Prigent and Dimitrov, 2003). These proteins may alter chromatin structure in late senescence to allow DNA fragmentation and eventual degradation. Histone modification and chromatin restructuring is a key regulator in *Arabidopsis* stress responses (Kim et al., 2010), and H3 phosphorylation increased in response to salinity, osmotic stress, and ABA treatment of cultured cells (Sokol et al., 2007).

The increased expression of actin and tubulin genes late in senescence could reflect an autophagy role (Monastyrska et al., 2009). Evidence from yeast and mammalian systems indicates that efficient autophagy requires microtubule action to facilitate autophagosome movement, and actin microfilaments have a role in selective types of autophagy in yeast. The *Arabidopsis* autophagy *ATG8* gene family shows significant homology to mammalian microtubule binding proteins and bind to microtubules in vitro (Ketelaar et al., 2004). Therefore, plant autophagy may involve the action of microtubules and microfilaments, explaining the increased expression of these genes late in senescence when the autophagic degradation of cellular compounds is active.

Chlorophyll degradation is a key step in the senescence process, and several of the genes involved are under transcriptional control (Hörtensteiner, 2009). The *STAYGREEN* gene, *SGR1* regulates the first step in the dismantling of chlorophyll from the chlorophyll binding proteins. Key genes involved in chlorophyll degradation, *SGR1*, *SGR2*, *NYC1* (chlorophyll *b* reductase), and *PaO* (pheophorbide *a* oxygenase), all show enhanced expression during senescence. All of these genes increase in expression during early time points, leveling out between 25 and 29 DAS, followed by a sudden increase in expression after 29 DAS (see Supplemental Figure 3C online). It is likely that it is the expression of *SGR* after 29 DAS that initiates the dismantling of the protein chlorophyll complexes, releasing chlorophyll for detoxification.

Additional information can be gained from the gradient analysis described above if the time of maximum gradient is also considered. For example, statistical analysis of gene clusters based on time of first differential expression indicated that photosynthetic genes were overrepresented in clusters showing downregulation at 29 to 33 DAS. However, if all the downregulated genes annotated as photosynthesis are examined, many of these show initial significant downregulation earlier in the time series (see Supplemental Figure 4A online), but the size of the clusters at these time points means this annotation does not show up as being significantly enriched. When the maximum absolute gradient for each of the photosynthesis genes was calculated (see Supplemental Figure 4B online), the vast majority of photosynthesis genes showed the most rapid drop in expression between 31 and 35 DAS, confirming the observation that photosynthesis-related gene activity is maintained until late in the leaf's development (Figure 6).

The maximum absolute gradient analysis was also applied to investigate genes responding to JA and ABA stimulus. In both cases, the majority of genes were first significantly upregulated at 21 and 23 DAS, early in senescence (see Supplemental Figures 4C and 4E online). However, although several of the genes had a maximum gradient early, at 23 DAS, there were also many showing maximum gradient much later in the time series, up to 35 and 37 DAS for the JA response genes and 33 and 35 DAS for the ABA response genes (see Supplemental Figures 4D and 4F online). This correlates with the data on the levels of JA and ABA shown in Figure 2 where a maximum level of both hormones is measured late in the process (increasing at 33 and 31 DAS, respectively). The timing of expression of specific hormone biosynthesis genes (see Supplemental Figure 5 online)

clearly illustrates the rapid increase in JA biosynthesis genes between 23 and 25 DAS, whereas ABA and SA biosynthesis genes show a later increase in expression with a maximum at the final stage of senescence. Although ethylene levels were not measured during the time course, the ethylene biosynthesis genes *ACS2* and *ACS7* also show increased expression from around 29 to 31 DAS, with a steady increase as senescence progresses. Thus, ABA, ethylene, and probably SA synthesis appear to be coordinately regulated in senescence, whereas JA synthesis shows a different pattern. Interestingly, some JA biosynthesis and signaling genes are only expressed at the early time point (e.g., *OPR3*), whereas others (e.g., *LOX3*) are also upregulated late, presumably enabling the accumulation of JA later during senescence.

Our detailed expression profiles and novel tools have enabled us to distinguish biological processes initiated at different stages of senescence and hence tease apart some of the components of this complex phenomenon. We now have a timeline that can be built upon to link these different processes and to identify the overarching regulatory mechanisms as well as candidate genes for specific senescence processes.

TF Binding Motifs Show Specific Enrichment in Differentially Expressed Gene Clusters

The SplineCluster analysis of differentially expressed genes (Figure 3) identified groups of genes that exhibit similar expression profiles and thus may be coregulated. Analysis of such coregulated gene sets should help pinpoint potential TF binding motifs important for gene expression during leaf development. To gain an initial understanding of the regulatory mechanisms of genes differentially expressed during senescence, promoters corresponding to 500 bp upstream of the predicted transcription start site of genes in each cluster were screened for overrepresentation of known TF binding motifs.

This analysis shows clearly that certain sequence motifs are selectively enriched in clusters that exhibit similar expression patterns (Figure 7, data shown in Supplemental Data Set 5 online), and there is an obvious difference in the range of motifs distributed over the different clusters. Consistent with the GO term analysis results, several of the downregulated clusters (clusters 1–24) are significantly enriched for sequence motifs associated with the regulation of photosynthesis and cell growth. For example, the G box variant motif is linked with the regulation of photosynthetic genes and responses to light (Martínez-García et al., 2000) and the TCP motif binds members of the TCP family of TFs, which have been implicated in the regulation of growth and cell division (Li et al., 2005). Binding sites for E2F TFs, key regulators of cell proliferation (Ramírez-Parra et al., 2003) are enriched in cluster 22, which is consistent with this cluster being enriched with genes annotated with the GO term cell cycle. Genes in this cluster are downregulated from the start of the measured time course, and this would be expected since cell division has ceased before the leaf is fully expanded.

Sequence regions upstream of genes in upregulated clusters (clusters 27–48) contain a number of sequence motifs that can bind TF families that are themselves upregulated during senescence. For example, NAC domain and WRKY TFs constitute a

large proportion of the senescence-regulated TFs and are known to play significant roles in regulating leaf senescence in *Arabidopsis* (Miao et al., 2004; Guo and Gan, 2006; Kim et al., 2009). Binding sites for NAC and WRKY TF families are overrepresented in several upregulated clusters sharing similar expression profiles. Sequence motifs associated with stress responses are also enriched. The heat shock element is overrepresented in a single cluster, and several heat shock factors are upregulated during senescence. The CGCG motif, which has been implicated as a calcium-signaling element in a range of stresses, is enriched in several upregulated clusters. This motif has been shown to bind CAMTA TFs (Yang and Poovaiah, 2002) involved in signaling responses to wounding, cold, and other stresses (Walley et al., 2007; Doherty et al., 2009). The ABA-responsive element (ABRE) is overrepresented in multiple upregulated clusters, and enrichment correlates with the observed increase in levels of ABA during senescence. ABRE-binding factors are known to activate target genes in an ABA-dependent manner (Nakashima et al., 2006). The ABRE contains an ACGT-core and, therefore, is a subset of the G box sequence (CACGTG). However, the pattern of overrepresentation of these two similar motifs across the senescence clusters is different, suggesting that divergent functional roles can be identified. G box-like motifs can bind many members of the bZIP and bHLH TF superfamilies (Toledo-Ortiz et al., 2003; Jakoby et al., 2002), and TFs from both these families are upregulated during senescence.

TF Families Are Active at Different Times during Senescence

To complement the analysis of TF binding motifs above, we investigated whether specific families of TFs were differentially expressed at particular times during senescence. The groups of genes identified as having the same initial timing of differential expression by the GP gradient tool analysis were further analyzed to identify time periods when families were overrepresented for genes with a positive or negative gradient (i.e., expression significantly increasing or decreasing). A heatmap, mapped to the significance of each family's activity, is shown in Figure 8, with the numerical data shown in Supplemental Data Set 6 online.

A number of TF families were significantly overrepresented for upregulated genes (adjusted $P < 0.01$), indicating a large amount of similar transcriptional activity and potential coregulation within these families. Specific members of the bZIP family have been shown to participate in defense against pathogens, development, stress treatments such as cold and drought, ABA signaling, and phenylpropanoid biosynthesis (Weisshaar and Jenkins, 1998; Jakoby et al., 2002). Another significantly overrepresented upregulated family is the large C3H superfamily, of which little is known of the function of many of its members. Several subfamilies related to the CCAAT box binding factor family were also significantly upregulated; factors in this family form the heterotrimeric NF-Y binding complex (consisting of NF-YA, NF-YB, and NF-YC subunits) that has been shown to influence flowering time and stress responses in plants (Wenkel et al., 2006; Liu and Howell, 2010). Interestingly, it is the NF-YA subunits specifically that are enriched in the senescence-enhanced gene lists, with nine of the 10 genes in the genome showing increased expres-

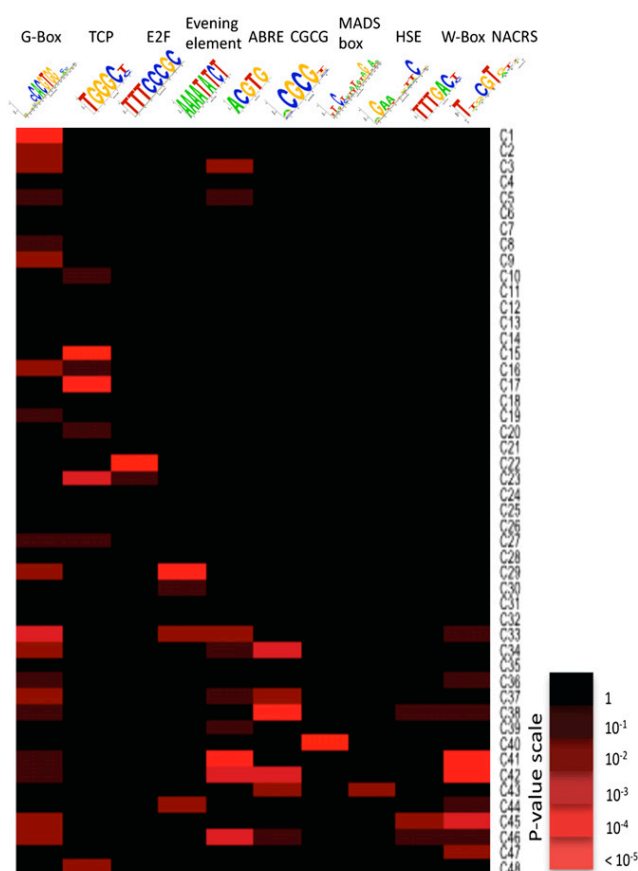


Figure 7. Over-Representation of Known TF Binding Motifs in Promoters of Coexpressed Genes.

Logo representations of known TF binding motifs are on the horizontal axis, and expression profile for each cluster (see Figure 3) is on the vertical axis. Colored boxes represent pairs of motif and expression cluster with a significant statistical link. Shown are a limited number of representative motifs and clusters (see Supplemental Data Set 5 online for full results).

sion during senescence. In comparison, few of the NF-YB and NF-YC genes are altered in expression, with only three NF-YB genes and one NF-YC gene being upregulated. This implies that it may be the regulated expression of the NF-YA subunit that controls the activity of the NF-Y complex during senescence.

The large NAC family also had a significant early overrepresentation, with over 30 of the members of this family being altered in expression at various times during senescence. Members of this family are known to have a large number of regulatory interactions with a diverse range of biological processes, including senescence, defense, and abiotic stress (Olsen et al., 2005, and references quoted above). Several other families show upregulated transcription later in the time course. The WRKY family shows an overrepresentation, with many members of this family being upregulated from around 25 DAS. WRKY TFs have been shown to be important for senescence (Robatzek and Somssich, 2001; Miao et al., 2004); others are induced by

infection by viruses or bacteria (Eulgem et al., 2000) and are downstream of defense-signaling mitogen-activated protein kinase pathways and involved in the regulation of SA- and JA-dependent defense signaling pathways (Ülker and Somssich, 2004; Eulgem and Somssich, 2007). The large AP2-EREBP family becomes significantly overrepresented around 27 DAS; members of this family are induced in several cases by hormones such as JA, SA, and ethylene, along with other signals related to pathogens, wounding, and abiotic stresses, and have influence on other stress and disease resistance pathways (Kizis et al., 2001; Gutterson and Reuber, 2004). Therefore, cascades of cellular information flow during the progress of leaf senescence can be predicted by this analysis, such as upregulation of NAC or WRKY genes influencing various hormone responses, followed by upregulation of AP2-EREBP TFs. This knowledge is key for future modeling of senescence transcriptional networks.

ANAC092 Target Genes Are Highly Enriched in Clusters Overrepresented for NAC Binding Motifs

The motif and TF analyses described above pinpoint NAC domain genes as being of key importance in regulation of senescence and we follow this observation up in more detail as an example of the increased understanding that this data set provides. A recent publication (Balazadeh et al., 2010) describes an elegant experiment using inducible expression of *ANAC092* to identify likely target genes. Of the 170 genes identified in that study as being upregulated after induced expression of *ANAC092*, 102 of these are senescence enhanced in our time course experiment; of these, 75%, including *ANAC092* itself, are to be found in the clusters enriched for NAC domain motifs (clusters 41, 42, 44, and 45; see Supplemental Data Set 7 online).

This provides clear evidence that the detection of enriched motifs within clusters is providing biologically relevant information and also indicates that *ANAC092*, probably together with other NAC domain proteins, has an influential role in regulating the expression of many genes at this stage of senescence. This information might be used in preliminary modeling experiments to predict interactions that regulate gene expression in these clusters.

DISCUSSION

In this article, we describe a high-resolution, highly replicated time-course analysis of gene expression during *Arabidopsis* leaf development from before complete expansion to full senescence. Over this time, the leaf develops from a sink that is importing nutrients for growth into an active source organ, performing maximum photosynthesis and exporting fixed carbon for further growth of the plant. This is followed, relatively rapidly in this short-lived plant, by the initiation of senescence whereby the leaf is converted from a source of photosynthetic carbon to a source of valuable macromolecules such as nitrogen, phosphorus, and minerals, as cellular components become degraded and mobilized from the leaf. Thus, in this short time period, the leaf undergoes enormous changes in metabolism and transport of metabolites.

To obtain insight into the timing and potential coregulation of the changes in genes and pathways in the complex process of senescence, it is essential to sample highly controlled replicate leaves and to measure at many time points. This is also essential if these data are to be used for network inference analysis. In the experiment reported here, we harvested the same leaf (leaf 7) from individual plants at different times over 3 weeks. Previous

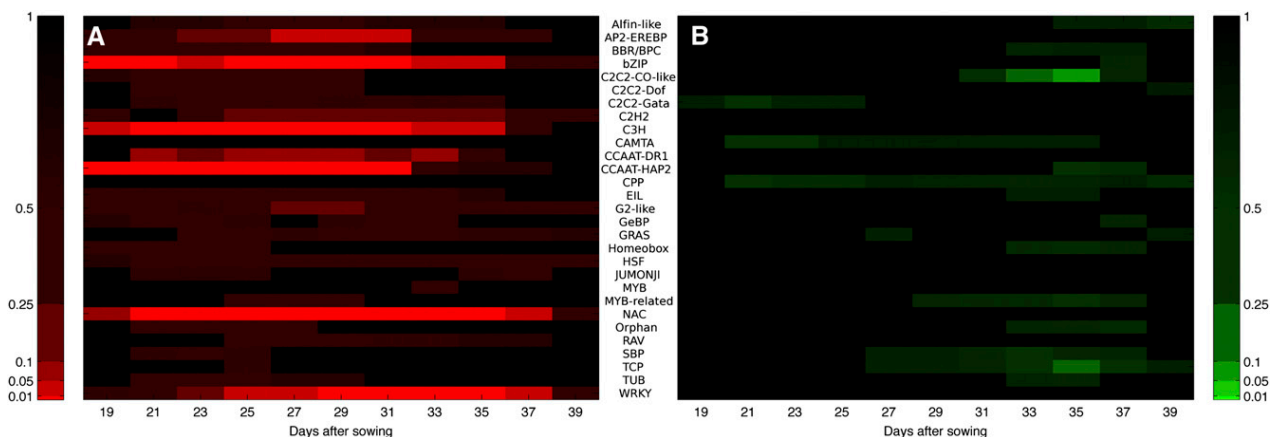


Figure 8. TF Families Significantly Over-Represented with Positive or Negative Gene Expression Gradients Highlighting Distinct Periods of Activity.

(A) A number of TF families significantly upregulated early, including the NAC and bZip families (19–21 DAS), remaining strongly upregulated throughout the experiment, and a small number of families only upregulated toward the middle and end of the experiment (27+ DAS), such as WRKY, AP2, and G2-like.

(B) A small number of weakly significant TF families downregulated early, and several families (C2C2-CO-like and TCP) significantly downregulated toward the end of the experiment (33+ DAS).

Color bars indicate P value (after FDR correction), with a range of significance thresholds (0.01, 0.05, 0.1, 0.25, and 0.5). Numerical data used to derive this figure are shown in Supplemental Data Set 6 online.

transcript profiling studies on developmental leaf senescence in *Arabidopsis* have analyzed a more limited number of time points (e.g., Buchanan-Wollaston et al., 2005; van der Graaff et al., 2006). Analysis of gene expression changes in field-grown *Populus* leaves has been performed over several time points during late summer and autumn (Andersson et al., 2004). In these examples, pooled leaves were used as biological samples; and replicates were limited; although these experiments gave a picture of the overall changes in gene expression that occur during senescence, they provide little information on the timing of the changes during the process. Senescing leaves are, by their nature, quite variable, particularly in mid or late senescence; therefore, to accurately determine coregulation of genes through senescence, it is essential to control and allow estimation of other sources of variability, such as may be caused by plant-to-plant differences and environmental factors by using a carefully designed sampling strategy.

Methods Developed to Enable Large-Scale, Two-Color Microarray Analysis and Identification of Differentially Expressed Genes

This large-scale microarray experiment with both biological and technical replication across multiple time points required the development of a novel and complex design approach to take full advantage of the two-color microarray system, providing efficient estimation of the differences in response between adjacent time points while still allowing effective comparison of all samples. The highly replicated experiment allowed the application of stringent statistical analysis to identify and characterize genes differentially expressed at different time points during senescence, which has not been possible with data from previous studies. The quantity of data generated necessitated the adaptation of existing analysis methods/algorithms, as well as the development of some new analysis tools.

The Bioconductor package MAANOVA was adapted to meet the specifications of the CATMA arrays, providing monitoring of slide and data quality and using information from the four technical replicates of each sample to remove the influence of the occasional outliers. The mixed model-fitting algorithm then enabled the estimation and testing of the differences caused by the treatment factors (day, time of day, the interaction [combined effect] of these factors, and the biological replicates), allowing for the complex design structure and the sources of variability (between slides and between dyes) imposed by using the two-color microarray system. There are considerable advantages to the experimental and analysis approaches used. The use of two-color arrays allows direct comparisons of samples between key time points and, through careful design of the pairs of samples compared on each array, across all time points by indirect association. By contrast, many applications of two-color arrays compare experimental samples by calculating the ratio of expression responses of each with a control sample hybridized on every slide, thus halving the amount of useful data obtained per slide (or doubling the cost of obtaining the same data).

Identification of genes showing interesting differential expression patterns was achieved by first assessing the significance of the variation due to different model terms (day, time of day, and

the interaction between them) relative to the between-biological replicate variation using the MAANOVA analysis, and then conducting a visual inspection of gene expression responses over time for genes giving a less significant test result ($0.0001 < P < 0.05$) for the effect of day. Of course, high levels of biological (between-plant) variability can lead to large changes in gene expression not being identified as statistically significant and, hence, genes not identified as being differentially expressed. Further exploration of approaches to control this biological variation, through both the statistical design of future experiments and the development of novel analysis methods, is important for future successful research.

The mixed model-fitting algorithm implemented in the MAANOVA package allows separation of the variation due to different sources within the treatment combinations (i.e., day, time of day, the interaction between these factors, and biological replicates), and hence the identification of genes showing different generic patterns of differential expression (see Supplemental Figure 1 online). An advantage of this approach, over a simple comparison of the responses across all 22 time points, is that genes showing only diurnal (time of day) variation can be easily identified and ignored in subsequent modeling of potential senescence-related gene networks, with those showing combined effects of day and time of day also easily identified and included. A disadvantage is that the analysis does not formally include any allowance for the ordering of the samples through time (e.g., the effect of day is essentially just an assessment of the average variability between the 11 mean values, and reordering the days would not change the test statistic and hence level of significance). A further development of this approach, allowing both separation of effects within a factorial treatment structure and estimation of the underlying shape of response over time, possibly following the approach proposed by Eastwood et al. (2008), should lead to a more reliable identification of genes showing important patterns of differential expression, although issues with high levels of biological variability would still result in some false-negative test results. Better estimation of the shape of expression profiles could also contribute to improved clustering of genes with similar shapes of expression profiles. In the absence of such a modeling approach, the approach used here, combining the highly significant results of the formal analysis with a visual inspection, is likely to result in the identification of most of the important genes showing differential expression related to senescence, while minimizing the number of false positives.

Analysis of Differentially Expressed Genes Revealed a Chronology of Processes and Signals

Analysis of individual clusters identified in the SplineCluster analysis, particularly those for downregulated genes, identified groups of genes involved in a common process such as photosynthesis, chlorophyll metabolism, etc. It seems likely that genes involved in the same process, with similar expression profiles, are coregulated rather than simply coexpressed during senescence, and this prediction is strengthened by the promoter motif analysis.

The GP gradient analysis, developed to enable more effective dissection of gene expression changes over time, identified

groups of genes that showed their first significant change in expression between the same pair of adjacent time points. The resulting clusters present a highly informative picture of the timeline of senescence, showing when individual pathways are upregulated or downregulated (Figure 6). Knowledge of such timing will prove a powerful tool for separation of pathways into groups to allow identification of upstream genes that control them.

Comparing the two approaches used to cluster the differentially expressed genes, it is clear that they will generate different sets of clusters. SplineCluster groups genes with overall profile shapes that are similar based on the fitted regression coefficients, which should therefore mean that genes in the same cluster will have similar changes in expression between every pair of adjacent time points. However, the approach does not take any account of the biological (between-plant) variability, so that only the initial filter will determine the significance associated with the overall differential expression, and so a gene with highly significant variation in expression could be clustered with one just breaking the significance threshold. By contrast, the GP gradient analysis groups genes that have the first significant changes in expression in the same direction at the same time. However, unless the gradient information is also taken into account, these may not always be showing the most dramatic change in expression at the same time point. Therefore, both approaches have value in identifying coregulated genes, but both have the potential to inappropriately group genes.

The clustering results have been discussed in detail above and have identified groupings of genes that had not been observed previously with more limited time series data. For example, it is clear that the extensive overall reduction in expression of genes involved in chloroplast activities occurs via a timed process. Chlorophyll biosynthesis genes are downregulated before carbon fixation genes, and these are downregulated well before the majority of key genes encoding proteins involved in photosynthesis, including chlorophyll binding proteins and components of PSI and II. Autophagy genes are enhanced from the start, but the level of the key gene *ATG7* starts to rise at 29 DAS. This is also the time at which chlorophyll degradation genes show a rapid induction of expression. Other metabolic pathways such as strigolactone synthesis, hormone biosynthesis, cell wall degradation, cytoskeleton, and microtubule activity, to name just a few, are implicated at different times during the senescence process.

Microarray Data Analysis Tools Used to Develop and Test Hypotheses for Transcriptional Control during Senescence

Analysis of the core promoters of coexpressed genes revealed potential regulatory sequence motifs that are likely to contribute to the coregulation of genes involved in the senescence process. Known sequence motifs are enriched in the promoters of genes that share similar expression profiles and correlate with the biological processes associated with such genes. The importance of the NAC and, to a lesser extent, the WRKY TF families in promoting senescence in *Arabidopsis* is illustrated through the specific and highly significant enrichment for potential binding sites for these regulators in the promoters of certain clusters of upregulated genes. In addition, most of the genes implicated as under the control of the senescence-enhanced ANAC092 TF

(Balazadeh et al., 2010) occur in these same clusters, showing the importance of this TF and other NAC family members in regulating gene expression during senescence. The identity of specific TFs that target these known motifs is unknown, and further bioinformatic analysis and modeling as well as laboratory experiments is required to characterize them fully. Regulation via known *cis*-regulatory elements is not sufficient to explain the expression patterns of all genes, and unknown sequence motifs likely contribute toward regulating specific groups of genes within the senescence process.

In comparison with previous senescence gene expression studies, the study reported here collected highly replicated gene expression responses at a high temporal resolution across the period during which the senescence response develops. Thus, these data are more suitable than previously collected data sets

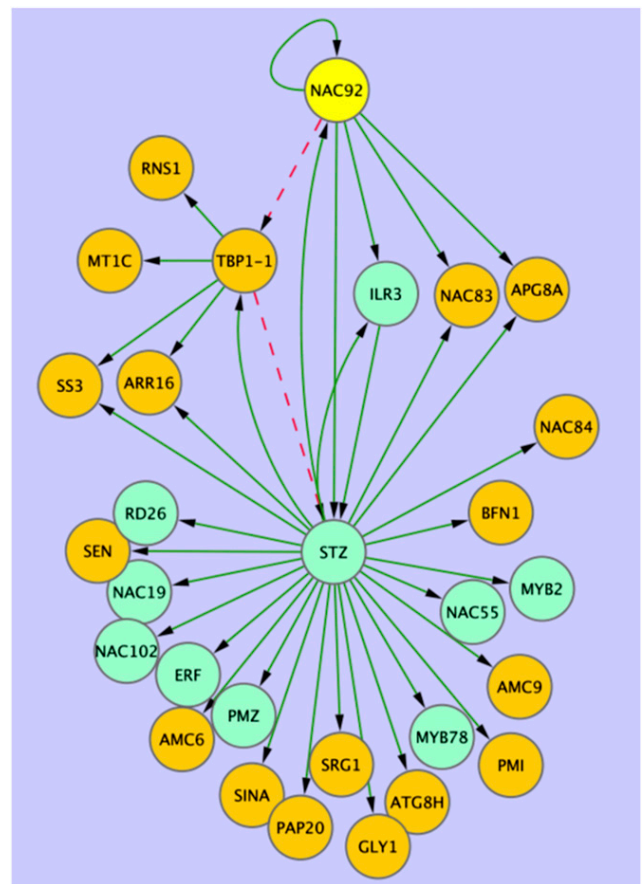


Figure 9. Network Model Inferred from Microarray Data.

Variational Bayesian state space modeling was used to generate a network model using senescence-enhanced genes selected from Spline-Clusters 41, 42, 44, or 45 in the microarray data. Genes showing induced expression in an *ANAC092* (yellow node) inducible overexpression experiment (see Supplemental Data Set 6 online; Balazadeh et al., 2010; orange nodes) were combined with selected TFs from the same clusters (green nodes). Green edges represent positive interactions, while red dashed edges predict negative effects. Genes are identified in Supplemental Data Set 7 online.

for applying statistical analyses aimed at predicting the gene regulatory networks operating during senescence. This is of immense value for the next step in the development of a model for the regulation of leaf senescence. However, even this substantial microarray experiment still imposes data limitations that make the application of network inference nontrivial; gene expression measurements of just 88 biological samples are insufficient to accurately model the regulation of thousands of genes. As the number of possible networks grows superexponentially with the number of genes involved, the correlations in the patterns of response across these 88 samples are certainly insufficient to accurately identify a unique network model describing the regulation of the 6326 genes identified as differentially expressed. This is known as the “curse of dimensionality” (Bellman, 1961). Therefore, the design of future studies needs to consider the balance between experimental cost and informative data for network inference, especially as a greater number of samples brings diminishing returns with respect to the number of additional genes it allows a researcher to model. It is also important to balance the need for good information about the biological (between-plant) variability, technical replicate variability, and temporal changes in gene expression during the senescence process.

The data presented here can be used to produce constrained network models for many small sets of genes. An example of the type of network model that can be inferred from these data is shown (Figure 9). A variational Bayesian state space modeling method (Beal et al., 2005) was applied to a selection of genes using the 11-time point data series (see Supplemental Table 2 online). Genes selected were present in clusters enriched for the NAC motif (i.e., clusters 41–45) and also present in the *ANAC092* overexpression data described above (Balazadeh et al., 2010; see Supplemental Data Set 7 online) or were annotated as TFs. The resulting model correctly predicts a positive influence of *ANAC092*, either direct or indirect, on multiple known downstream target genes (Figure 9, orange nodes). The model also makes several new hypotheses for experimental testing. For example, it predicts the influence of a zinc finger protein (STZ) in the expression of *ANAC092* and its downstream genes. Although no experimental evidence exists for this regulatory link, both STZ and *ANAC092* are induced in expression during salt stress, and knock out mutants in both genes have enhanced tolerance to stress in *Arabidopsis* (Mittler, et al., 2006; Balazadeh et al., 2010). Regulation by *ANAC092* of several other TFs known to be stress related is also predicted in this network model. *ANAC019*, *ANAC055*, and *RD26* (*ANAC072*) have all been shown to have a role in drought stress (Tran et al., 2004); MYB2 has a role in ABA signaling and salt stress (Abe et al., 2003; Yoo et al., 2005), and PMZ, a zinc finger protein, has a role in stress-induced senescence (Breeze et al., 2008). The gene regulatory network model also predicts feedback and feed-forward connections between *ANAC092*, STZ, and *TBP1-1*, which encodes a telomere binding protein. These types of interactions are crucial for the robustness of gene regulatory networks and would be almost impossible to predict from biological data alone. Thus, this model, generated with a small subset of the array data, correctly predicts known gene-gene interactions and generates complex novel predictions for experimental testing.

Many different models can be obtained with different collections of genes, and these transcriptional network models can be expanded using information on coregulated pathways and promoter motif analysis to identify likely downstream targets of key TFs. This model development and experimental testing is underway to generate validated gene regulatory network models underlying senescence.

METHODS

Plant Growth

Arabidopsis thaliana plants were grown as described in Breeze et al. (2008). Leaf 7 was tagged with thread 18 DAS. Sampling of leaf 7 started at 19 DAS and continued every other day until full senescence was reached (39 DAS). Leaves were harvested twice on each sampling day, 7 and 14 h into the light period. This resulted in samples being obtained at 22 distinct time points. At each time point, leaf 7 was sampled from 20 plants among the 720 being grown in the controlled-environment growth room, the plants being randomly selected to avoid any potential effects of position within the growth room. Leaves were rapidly weighed and photographed with a size scale before being frozen in individual tubes in liquid nitrogen. Leaf length was estimated against this scale from the photographs.

Protein and Chlorophyll Measurements

Total protein was extracted from five individual leaf samples by grinding the sample in liquid nitrogen before the addition of 500 μ L of extraction buffer (50 mM lithium phosphate [pH 7.2], 1 mM monoiodoacetic acid, 120 mM mercaptoethanol, 5% [v/v] glycerol, 1 mM PMSF, and 0.2% lithium dodecyl sulfate). At this stage, a 100- μ L aliquot of the extract was taken for chlorophyll analysis. The protein extract was boiled for 45 s and centrifuged for 20 min at 12,800g. Total protein was measured using the RC DC protein assay (Bio-Rad) according to the manufacturer's instructions. In addition, levels of the small and large subunits of Rubisco were assessed by diluting the protein extracts to normalize for leaf weight and then running an equal volume of each extract (equivalent to 0.5 mg of fresh tissue) on polyacrylamide gels (Invitrogen Novex 4-12% Bis-Tris gel), staining with Coomassie blue, and scanning the relevant protein band. Protein levels were assessed densitometrically using image analysis software (GeneTools; Syngene) against a calibration curve of bovine serum albumin (LSU) and lysozyme (SSU).

Chlorophyll was measured from five individual leaves using the total protein extracts. The chlorophyll was extracted using 80% acetone, vortexed, and then stored at -20°C for 1 h in the dark. The samples were then centrifuged for 3 min at 12,800g, and the absorbance of 1 mL was measured at 663 and 646 nm. Chlorophyll concentrations were calculated using the equations: total chlorophyll (mg/L) = $20.2A_{646} + 8.02A_{663}$, chlorophyll *a* (mg/L) = $13.19A_{663} - 2.57A_{646}$, and chlorophyll *b* (mg/L) = $22.1A_{646} - 5.26A_{663}$.

Hormone Measurements

The hormones ABA, SA, and JA were measured in five individual leaves. Each leaf was freeze dried and 10 mg of freeze dried tissue was used for hormone extraction as described in Forcat et al. (2008). Analysis was performed using an HPLC-ESI/MS-MS.

Statistical Treatment of Leaf Morphological and Biochemical Measurements

Leaf morphology (weight and length) and biochemical assay (total protein, chlorophyll *a+b*, Rubisco LSU and SSU, and hormone) data were

subjected to ANOVA to assess for differences in response over the time course using GenStat (VSN International). Data from hormone assays were subjected to a \log_{10} transformation (including the addition of a small constant to cope with zero observations) prior to analysis to satisfy the assumption of homogeneity of variance. LSDs were calculated at a 5% significance level to allow easy comparison of differences between adjacent time points. Significant effects noted in the results relate to either F tests for the overall variability over time or *t* tests for comparisons between adjacent time points.

Microarray Analysis

RNA Preparation and Labeling

Total RNA was isolated from four individual leaves from each sampled time point (arbitrarily labeled as biological replicates A, B, C, and D) using TRIzol reagent (Invitrogen), purified with RNeasy columns (Qiagen), and amplified using the MessageAmp II aRNA Amplification kit (Ambion) in accordance with the kit protocol with a single round of amplification. Cy3- and Cy5-labeled cDNA probes were prepared by reverse transcribing 5 μ g of aRNA with Cy3- or Cy5-dCTP (GE Healthcare) and a modified dNTP mix (10 mM each dATP, dGTP, and dTTP; 2 mM dCTP) using random primers (Invitrogen) and SuperScript II reverse transcriptase (Invitrogen), with the inclusion of RNase inhibitor (RNaseOUT; Invitrogen) and DTT. Labeled probes were purified using QiaQuick PCR Purification columns (Qiagen), freeze-dried, and resuspended in 50 μ L of hybridization buffer (25% formamide, 5 \times SSC, 0.1% SDS, and 0.5 μ g/ μ L yeast tRNA; Invitrogen).

Microarray Experiments

The microarray experiments were performed using the CATMA (version 3) microarray (Allemeersch et al., 2005; <http://www.catma.org>). CATMA probe annotations were updated using the TAIR9 release: oligo sequences of CATMA array probes were mapped to individual mRNA sequences of transcripts from the TAIR9 genome assembly using BLASTn (Altschul et al., 1997), with e-value cutoff of 0.01. Additionally, results were filtered to exclude alignments shorter than 30 bp or with less than 80% sequence identity. The best matching gene model (by e-value of hit to transcript) was identified for each probe. In addition, probe sequences were mapped to TAIR9 genomic DNA to clarify cases where a probe had been designed to a region of an earlier genome assembly now unannotated in TAIR9.

A novel experimental design strategy (A. Mead, unpublished data), based on the principle of the “loop design” (Kerr and Churchill, 2001), was developed to enable efficient extraction of information about key sample comparisons using a two-color hybridization experimental system. With 88 distinct samples (four biological replicates at each of 22 time points) to be compared, the experimental design included 176 two-color microarray slides, allowing four technical replicates of each sample to be observed. The detailed structure of the design, indicating how pairs of treatments were allocated to arrays, is described in Supplemental Methods 1 online, with an illustrative diagram shown in Supplemental Figure 6 online. According to a randomization of this experimental design, pairs of labeled samples were hybridized to slides overnight at 42°C. Following hybridization, slides were washed and scanned using an Affymetrix 428 array scanner at 532 nm (Cy3) and 635 nm (Cy5). Scanned data were quantified using Imagene 7.5.0 software (BioDiscovery, Inc.).

MAANOVA Analysis

A local adaptation of the MAANOVA package (Wu et al., 2003) was used to analyze the quantified microarray data, providing data quality assur-

ance, slide normalization through LOWESS data transformation, mixed model fitting, and identification of genes showing significant differential expression via F tests of fixed (treatment) terms included within the model. MAANOVA was selected to analyze the data because it is able to provide an accurate analysis of the effects on gene expression of multiple sources of variation (both fixed, treatment, terms, and random sources of background variation) in the experimental design, harnessing the power of direct comparisons between pairs of samples obtained using two-channel microarrays (Churchill, 2004). Full details of the data quality checking procedures, of the mixed model fitting approach to describe the observed gene expression data, and of the construction of F tests for fixed treatment terms are given in Supplemental Methods 1 online. Having fitted the mixed model to each gene, predicted means were calculated for each of the 88 samples, assuming the full treatment model (effects of day, time of day, the interaction between them, and the nested biological replicates) to produce a four-replicate 22-time point data set for each gene, or assuming a reduced treatment model (effects of day and the nested biological replicates) to produce an eight-replicate 11-time point data set for each gene. These data sets were then used in subsequent analyses.

Selection of Differentially Expressed Genes

The most significant differentially expressed genes were identified initially and this was followed by visual analysis of genes close to the borderline of significance. First, CATMA probes with no corresponding gene model in the TAIR9 annotation were ignored; also, replicate CATMA probes were removed (the most gene-specific probe being identified in each case). In total, 4989 genes had an adjusted day main effect F test P-value < 0.0001 (after multiple testing correction via a step-down FDR-controlling procedure (Westfall et al., 1998; Benjamini and Liu, 1999), equivalent to responses showing a significant test result at an FDR of $P < 0.0001$), and these were included in the initial list of differentially expressed genes. The patterns of expression of all genes with an adjusted day main effect F test P-value between 0.0001 and 0.05 (further responses showing a significant test result at an FDR of $P < 0.05$) were then screened visually to remove any showing either a small or a very variable change in expression over time. The final list of 6323 differentially expressed genes, with adjusted F test statistics, is shown in Supplemental Table 1 online.

Gene Expression Profile Clustering

Clustering of coregulated genes was performed by the application of SplineCluster (Heard et al., 2006), a Bayesian model-based hierarchical clustering algorithm for time series data, using the mean of the biological replicates for each gene. Recent functionality added to SplineCluster, (Heard, 2011) improves the gene allocation to clusters. Where a gene has become an outlier for its allocated cluster, it is reallocated to alternative clusters to maximize the log marginal likelihood once more. This option was used on all SplineCluster analyses presented in this article. The 22-time point data (averaged across the four biological replicates) was clustered using a prior precision of 5×10^{-4} , while the other data set composed of 11 time points was averaged across all eight morning and afternoon biological replicates before being clustered using a prior precision of 1×10^{-4} . These prior precisions were selected as they produce ~ 50 clusters for each of the two data sets.

GP Gradient Analysis

To identify an ordering of events, the rate of change of gene expression (gradient) was inferred using a GP regression approach (see Supplemental Methods 1 online), which has the notable advantage of incorporating all biological replicates. Furthermore, since the marginal distribution of a

GP is itself a Gaussian distribution, the probability that the gradient (at any particular time) lies sufficiently far from zero may be calculated analytically. When data are sufficiently time resolved, the GP model may therefore be used to identify times when the gradient of a gene expression profile is significantly positive (increased), negative (decreased), or not statistically different from zero (steady), whereas for less time-resolved data, it may identify times of significant change to the global trend.

GO Analysis

GO annotation analysis on gene clusters was performed using the BiNGO 2.3 plugin tool in Cytoscape version 2.6 with GO_full and GO_slim categories, as described by Maere et al. (2005). Over-represented GO_Full categories were identified using a hypergeometric test with a significance threshold of 0.05 after a Benjamini and Hochberg FDR correction (Benjamini and Hochberg, 1995).

Promoter Analysis

Plant position-specific scoring matrices (PSSMs) were collected from the TRANSFAC database, version 2010.3, (Matys et al., 2006) and the PLACE database (Higo et al., 1999). This set was supplemented with PSSMs for a heat shock element (TRANSFAC matrix record M00146) and two NAC TF binding sites (Olsen et al., 2005) since these important motifs were absent from the databases. PSSMs were clustered, and a representative of each cluster was chosen for screening. Promoter regions corresponding to 500 bp upstream of the transcription start site were retrieved from the Ensembl Plants sequence database (release 50).

For any given PSSM and promoter, we scanned the sequence and computed a matrix similarity score (Kel et al., 2003) at each position on both strands. P values for each score were computed from a score distribution obtained by applying the PSSM to a random sequence of 100 million bases in length generated by a 3rd order Markov model learned from the whole *Arabidopsis* genome. We took the top k nonoverlapping hits and performed the binomial test for the occurrence of k sites with observed n values within a sequence of length 500 bp. The parameter k is optimized within the range 1 to 5 for minimum binomial P-value. This allows detection of binding sites without a fixed threshold per binding site. Using a threshold ($P < 0.05$), the presence or absence of a PSSM was scored for each promoter based on the binomial probability.

For each PSSM, its frequency in promoters of each cluster was compared with its occurrence in all promoters in the entire genome. Motif enrichment was calculated using the hypergeometric distribution (phyper function in the R stats package). Hypergeometric P-values were corrected for the number of clusters tested using Bonferroni correction. Corrected P-values ≤ 0.05 were considered significant. Sequence logos were generated using code modified from Lenhard and Wasserman (2002). Sequence analysis was performed within the APPLES software framework (S. Ott, unpublished data).

TF Family Analysis

Gene expression activity was analyzed for 1733 TFs, grouped into 50 families defined in the *Arabidopsis thaliana* Transcription Factor Database, AtTFDB (Palaniswamy et al., 2006; 1843 TF, 50 families as of June 2010). Of these, 1733 were probes on the CATMA array using the GP gradient model. Families overrepresented for genes with significantly positive or negative gradients at each time point, using all genes within the experiment as a reference, were identified using the hypergeometric distribution (computed using the *hygeomdist* function in MS Excel 12.2.3) with Benjamini and Hochberg FDR correction. A heatmap of adjusted P-values, using five levels of significance (0.01, 0.05, 0.1, 0.25, and 0.5) was then generated, using only those values that correspond to

overrepresented counts (e.g., the proportion of positive/negative gradient TFs for a given family is larger than the proportion of positive/negative gradient genes in the entire data set for each time point).

Variational Bayesian State Space Modeling

Data for eight biological replicates from the 11-time point data series were used to generate a network model using the method published in Beal et al., 2005. ANAC092 was used as the gene on which to base the model and two groups of genes were selected to accompany it. First, several genes were selected that showed rapidly increased expression following induced expression of ANAC092 in green leaves (from Balazadeh et al., 2010), many of which were also in clusters enriched for NAC domains. Therefore, these are likely to be direct or indirect targets of ANAC092 activation. Second, a group of TFs that show coexpression with ANAC092 selected from the clusters 41 through 45 was included. Ten models were run from different random seeds and connections occurring in more than 50% of models at a confidence level of $>95\%$ were included in the network shown in Figure 9.

Data Repository

The microarray data used in this article have been deposited in NCBI's Gene Expression Omnibus (Edgar et al., 2002) and have been given a GEO Series accession number, GSE22982.

Accession Numbers

Arabidopsis gene names and identifiers referred to in this article are: ANAC092 (At5g39610), *WRKY53* (At4g23810), *ANAC029* (At1g69490), *SAG12* (At5g45890), *LHY* (At1g01060), *CCA1* (At2g46830), *PRR7* (At5g02810), *PHYA* (At1g09570), *CRY1* (At4g08920), *PIF4* (At2g43010), *ELF4* (At2g40080), *PCL1* (At3g46640), *LUT1* (At3g53130), *LUT2* (At5g57030), *LUT5* (At1g31800), *ARR4* (At1g10470), *ARR6* (At5g62920), *ARR7* (At1g19050), sedoheptulose biphosphatase (At3g55800), *HCF101* (At3g24430), *HCF109* (At5g36170), *HCF152* (At3g09650), *HCF173* (At1g16720), *HCF208* (At5g52110), *GLK2* (At5g44190), *GLK1* (At2g20570), *LHCB2.2* (At2g05070), *LHCB4.2* (At3g08940), *LHCB6* (At1g15820), *ATG7* (At5g45900), *ATG8H* (At3g06420), *ATG8A* (At4g21980), *ATG8B* (At4g04620), *DREB2A* (At5g05410), *LSD1* (At4g20380), *AtMPK7* (At2g18170), *HSP70* (At3g12580), *AHK3* (At1g27320), *ERD1* (At5g51070), *ERD14* (At1g76180), *RAB18* (At5g66400), *ABF2* (At1g45249), *AOC1* (At3g25760), *AOC4* (At1g13280), 12-oxophytodienoate reductase (At2g06050), *MYC2* (At1g32640), *JAZ1* (At1g19180), *JAZ6* (At1g72450), *JAZ8* (At1g30135), *COI1* (At2g39940), *LOX3* (At1g17420), *ABI1* (At4g26080), *AFP1* (At1g69260), *RD20* (At2g33380), *RD26* (At4g27410), *NF-YA4* (At2g34720), *CCD7* (At2g44990), *CCD8* (At4g32810), *ORE9* (At2g42620), *AtMC6* (At1g79320), *AtMC9* (At5g04200), *LACS6* (At3g05970), *ACX1* (At4g16760), *ETR1* (At1g66340), *ACO2* (At1g62380), *TUA2* (At1g50010), *TUA4* (At1g04820), *TUA5* (At5g19780), *ACT3* (At3g53750), *ACT11* (At3g12110), *AtAURORA1* (At4g32830), *AtAURORA2* (At2g25880), *SGR1/NYE1* (At4g22920), *SGR2* (At4g11910), *NYC1* (At4g13250), and *PaO* (At3g44880).

Author Contributions

All authors had a role in discussion of results. V.B.-W., J.B., K.D., B.T., S.J., D.R., S.O., and D.L.W. designed the research; E.B., E.H., C.Z., K.M., A.T., and C.J. did the experimental work; A.M. designed the array experiment; S.M., L.H., and A.M. developed microarray data extraction and primary analysis methods; S.M. designed Gene Viewer; S.M., D.L.W., and S.K. developed and applied clustering methods; Y.-s.K., C.H., V.B.-W., E.B., R.H., C.A.P., and D.J. performed data analysis and

biological interpretation; C.A.P. and D.L.W. developed the gradient analysis tool; R.H. and S.O. did the promoter motif analysis; D.J. developed the TF family analysis method; J.D.M. and R.L. organized data handling and web page; and V.B.-W. wrote the majority of the paper with input from E.B., E.H., S.M., A.M., C.H., Y.-s.K., R.H., S.O., D.J., C.A.P., S.K., K.D., J.D.M., and D.L.W.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure 1. Venn Diagram Summarizing Numbers of Genes Showing Significantly Different Expression for Different Combinations of Treatment Terms in the MAANOVA Fixed Model.

Supplemental Figure 2. Expression Patterns of Selected Genes Showing Time-of-Day Changes Only.

Supplemental Figure 3. Expression Patterns of Selected Genes during Leaf Senescence.

Supplemental Figure 4. Gradient Analysis on Selected Groups of Genes.

Supplemental Figure 5. Expression Patterns of Selected Hormone Biosynthesis Genes.

Supplemental Figure 6. Representation of the Experimental Design for the Microarray Experiment.

Supplemental Table 1. Enriched GO Terms in Genes Downregulated or Upregulated during Senescence.

Supplemental Table 2. Genes Used in the Variational Bayesian State Space Modeling Method Model Shown in Figure 9.

Supplemental Data Set 1. Genes Differentially Expressed during Senescence.

Supplemental Data Set 2. Enriched GO Terms in Each Cluster Shown in Figure 3.

Supplemental Data Set 3. Gradient Data for Differentially Expressed Genes.

Supplemental Data Set 4. Cluster Patterns Identified from the Gradient Analysis Showing Enriched GO terms.

Supplemental Data Set 5. Known DNA Sequence Motif Enrichment.

Supplemental Data Set 6. TF Family Analysis Data.

Supplemental Data Set 7. Senescence-Enhanced Genes Differentially Expressed in Inducible ANAC092 Line.

Supplemental Methods 1. Microarray Experiments and GP Regression and Gradient.

ACKNOWLEDGMENTS

We thank Mary Coates (University of Warwick) for helping with manuscript preparation and Miriam Gifford (University of Warwick) for critical reading. We thank Mark Bennett (Imperial College) and Prof. Murray Grant (Exeter University) for performing the hormone measurements. E.B., E.H., C.Z., C.J., and K.M. were funded for this work by a Biotechnology and Biological Sciences Research Council (BBSRC) core strategic grant to Warwick HRI; S.M., S.K., and R.H. are funded by the Engineering and Physical Sciences Research Council/BBSRC-funded Warwick Systems Biology Doctoral Training Centre; L.H. was funded by a BBSRC studentship; J.B., K.D., V.B.-W., D.R., D.L.W., S.O., C.H., Y.-s.K., C.A.P., D.J., J.D.M., R.L., and A.T. are part of the BBSRC-funded grant Plant Response to Environmental Stress Arabidopsis (BB/F005806/1).

Received January 21, 2011; revised January 21, 2011; accepted February 28, 2011; published March 29, 2011.

REFERENCES

- 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**: 63–78.
- Allemeersch, J., et al. (2005). Benchmarking the CATMA microarray. A novel tool for *Arabidopsis* transcriptome analysis. *Plant Physiol.* **137**: 588–601.
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**: 3389–3402.
- Andersson, A., et al. (2004). A transcriptional timetable of autumn senescence. *Genome Biol.* **5**: R24.
- Balazadeh, S., Siddiqui, H., Allu, A.D., Matallana-Ramirez, L.P., Caldana, C., Mehrnia, M., Zanor, M.L., 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**: 250–264.
- Beal, M.J., Falciani, F., Ghahramani, Z., Rangel, C., and Wild, D.L. (2005). A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics* **21**: 349–356.
- Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*. (Princeton, NJ: Princeton University Press).
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**: 289–300.
- Benjamini, Y., and Liu, W. (1999). A step-down multiple testing procedure that controls the false discovery rate under independence. *J. Stat. Planning Inference* **82**: 163–170.
- Breeze, E., Harrison, E., Page, T., Warner, N., Shen, C., Zhang, C., and Buchanan-Wollaston, V. (2008). Transcriptional regulation of plant senescence: From functional genomics to systems biology. *Plant Biol. (Stuttg.)* **10** (Suppl. 1), 99–109.
- 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., and Leaver, C.J. (2005). Comparative transcriptome analysis reveals significant differences in gene expression and signaling pathways between developmental and dark/starvation-induced senescence in *Arabidopsis*. *Plant J.* **42**: 567–585.
- Choudhury, A., and Lahiri, A. (2011). Comparative analysis of abscisic acid-regulated transcriptomes in *Arabidopsis*. *Plant Biol. (Stuttg.)* **13**: 28–35.
- Churchill, G.A. (2004). Using ANOVA to analyze microarray data. *Biotechniques* **37**: 173–175, 177.
- Coll, N.S., Vercammen, D., Smidler, A., Clover, C., Van Breusegem, F., Dangl, J.L., and Epple, P. (2010). *Arabidopsis* type I metacaspases control cell death. *Science* **330**: 1393–1397.
- Demidov, D., Hesse, S., Tewes, A., Rutten, T., Fuchs, J., Ashtiyani, R.K., Lein, S., Fischer, A., Reuter, G., and Houben, A. (2009). Aurora1 phosphorylation activity on histone H3 and its cross-talk with other post-translational histone modifications in *Arabidopsis*. *Plant J.* **59**: 221–230.
- Demidov, D., Van Damme, D., Geelen, D., Blattner, F.R., and Houben, A. (2005). Identification and dynamics of two classes of aurora-like kinases in *Arabidopsis* and other plants. *Plant Cell* **17**: 836–848.
- Dóczy, R., Brader, G., Pettkó-Szandtner, A., Rajh, I., Djamei, A.,

- Pitzschke, A., Teige, M., and Hirt, H.** (2007). The Arabidopsis mitogen-activated protein kinase kinase *MKK3* is upstream of group C mitogen-activated protein kinases and participates in pathogen signaling. *Plant Cell* **19**: 3266–3279.
- Doelling, J.H., Walker, J.M., Friedman, E.M., Thompson, A.R., and Vierstra, R.D.** (2002). The APG8/12-activating enzyme APG7 is required for proper nutrient recycling and senescence in Arabidopsis thaliana. *J. Biol. Chem.* **277**: 33105–33114.
- 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**: 972–984.
- 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.
- Edgar, R., Domrachev, M., and Lash, A.E.** (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**: 207–210.
- Epple, P., Mack, A.A., Morris, V.R., and Dangl, J.L.** (2003). Antagonistic control of oxidative stress-induced cell death in Arabidopsis by two related, plant-specific zinc finger proteins. *Proc. Natl. Acad. Sci. USA* **100**: 6831–6836.
- Eulgem, T., Rushton, P.J., Robatzek, S., and Somssich, I.E.** (2000). The WRKY superfamily of plant transcription factors. *Trends Plant Sci.* **5**: 199–206.
- Eulgem, T., and Somssich, I.E.** (2007). Networks of WRKY transcription factors in defense signaling. *Curr. Opin. Plant Biol.* **10**: 366–371.
- Forcat, S., Bennett, M.H., Mansfield, J.W., and Grant, M.R.** (2008). A rapid and robust method for simultaneously measuring changes in the phytohormones ABA, JA and SA in plants following biotic and abiotic stress. *Plant Methods* **4**: 16.
- Fujita, M., Fujita, Y., Maruyama, K., Seki, M., Hiratsu, K., Ohme-Takagi, M., Tran, L.S., 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**: 863–876.
- Fulda, M., Shockey, J., Werber, M., Wolter, F.P., and Heinz, E.** (2002). Two long-chain acyl-CoA synthetases from *Arabidopsis thaliana* involved in peroxisomal fatty acid β -oxidation. *Plant J.* **32**: 93–103.
- Gomez-Roldan, V., et al.** (2008). Strigolactone inhibition of shoot branching. *Nature* **455**: 189–194.
- Grbic, V., and Bleeker, A.B.** (1995). Ethylene regulates the timing of leaf senescence in Arabidopsis. *Plant J.* **8**: 595–602.
- Guo, Y., and Gan, S.** (2006). *AtNAP*, a NAC family transcription factor, has an important role in leaf senescence. *Plant J.* **46**: 601–612.
- Gutterson, N., and Reuber, T.L.** (2004). Regulation of disease resistance pathways by AP2/ERF transcription factors. *Curr. Opin. Plant Biol.* **7**: 465–471.
- Hanaoka, H., Noda, T., Shirano, Y., Kato, T., Hayashi, H., Shibata, D., Tabata, S., and Ohsumi, Y.** (2002). Leaf senescence and starvation-induced chlorosis are accelerated by the disruption of an Arabidopsis autophagy gene. *Plant Physiol.* **129**: 1181–1193.
- Harmer, S.L.** (2009). The circadian system in higher plants. *Annu. Rev. Plant Biol.* **60**: 357–377.
- Hazen, S.P., Schultz, T.F., Pruneda-Paz, J.L., Borevitz, J.O., Ecker, J.R., and Kay, S.A.** (2005). *LUX ARRHYTHMO* encodes a Myb domain protein essential for circadian rhythms. *Proc. Natl. Acad. Sci. USA* **102**: 10387–10392.
- He, Y., Fukushima, H., Hildebrand, D.F., and Gan, S.** (2002). Evidence supporting a role of jasmonic acid in Arabidopsis leaf senescence. *Plant Physiol.* **128**: 876–884.
- Heard, N.A.** (2011). Iterative reclassification in agglomerative clustering. *J. Comput. Graph. Stat.* <http://dx.doi.org/10.1198/jcgs.2011.09111>
- Heard, N.A., Holmes, C.C., and Stephens, D.A.** (2006). A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An application of Bayesian hierarchical clustering of curves. *JASA* **101**: 18–29.
- Higo, K., Ugawa, Y., Iwamoto, M., and Korenaga, T.** (1999). Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res.* **27**: 297–300.
- Hörtensteiner, S.** (2009). Stay-green regulates chlorophyll and chlorophyll-binding protein degradation during senescence. *Trends Plant Sci.* **14**: 155–162.
- Hörtensteiner, S., and Feller, U.** (2002). Nitrogen metabolism and remobilization during senescence. *J. Exp. Bot.* **53**: 927–937.
- Hwang, I., and Sheen, J.** (2001). Two-component circuitry in Arabidopsis cytokinin signal transduction. *Nature* **413**: 383–389.
- Ichimura, Y., Kirisako, T., Takao, T., Satomi, Y., Shimonishi, Y., Ishihara, N., Mizushima, N., Tanida, I., Kominami, E., Ohsumi, M., Noda, T., and Ohsumi, Y.** (2000). A ubiquitin-like system mediates protein lipidation. *Nature* **408**: 488–492.
- Jakoby, M., Weisshaar, B., Dröge-Laser, W., Vicente-Carbajosa, J., Tiedemann, J., Kroj, T., and Parcy, F.; bZIP Research Group.** (2002). bZIP transcription factors in *Arabidopsis*. *Trends Plant Sci.* **7**: 106–111.
- Kel, A.E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V., and Wingender, E.** (2003). MATCHM: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* **13**: 3576–3579.
- Kerr, M.K., and Churchill, G.A.** (2001). Experimental design for gene expression microarrays. *Biostatistics* **2**: 183–201.
- Ketelaar, T., Voss, C., Dimmock, S.A., Thumm, M., and Hussey, P.J.** (2004). Arabidopsis homologues of the autophagy protein Atg8 are a novel family of microtubule binding proteins. *FEBS Lett.* **567**: 302–306.
- Kizis, D., Lumberras, V., and Pagès, M.** (2001). Role of AP2/EREBP transcription factors in gene regulation during abiotic stress. *FEBS Lett.* **498**: 187–189.
- Kikis, E.A., Khanna, R., and Quail, P.H.** (2005). *ELF4* is a phytochrome-regulated component of a negative-feedback loop involving the central oscillator components *CCA1* and *LHY*. *Plant J.* **44**: 300–313.
- Kim, H.J., Ryu, H., Hong, S.H., Woo, H.R., Lim, P.O., Lee, I.C., Sheen, J., Nam, H.G., and Hwang, I.** (2006). Cytokinin-mediated control of leaf longevity by *AHK3* through phosphorylation of *ARR2* in Arabidopsis. *Proc. Natl. Acad. Sci. USA* **103**: 814–819.
- Kim, J.H., Woo, H.R., Kim, J., Lim, P.O., Lee, I.C., Choi, S.H., Hwang, D., and Nam, H.G.** (2009). Trifurcate feed-forward regulation of age-dependent cell death involving miR164 in Arabidopsis. *Science* **323**: 1053–1057.
- Kim, J.M., and DellaPenna, D.** (2006). Defining the primary route for lutein synthesis in plants: The role of Arabidopsis carotenoid β -ring hydroxylase CYP97A3. *Proc. Natl. Acad. Sci. USA* **103**: 3474–3479.
- Kim, J.M., To, T.K., Nishioka, T., and Seki, M.** (2010). Chromatin regulation functions in plant abiotic stress responses. *Plant Cell Environ.* **33**: 604–611.
- Lee, E.J., Matsumura, Y., Soga, K., Hoson, T., and Koizumi, N.** (2007). Glycosyl hydrolases of cell wall are induced by sugar starvation in Arabidopsis. *Plant Cell Physiol.* **48**: 405–413.
- Lenhard, B., and Wasserman, W.W.** (2002). TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics* **18**: 1135–1136.
- Li, C., Potuschak, T., Colón-Carmona, A., Gutiérrez, R.A., and**

- Doerner, P.** (2005). Arabidopsis TCP20 links regulation of growth and cell division control pathways. *Proc. Natl. Acad. Sci. USA* **102**: 12978–12983.
- Lim, P.O., Kim, H.J., and Nam, H.G.** (2007). Leaf senescence. *Annu. Rev. Plant Biol.* **58**: 115–136.
- Liu, J.X., and Howell, S.H.** (2010). *bZIP28* and NF-Y transcription factors are activated by ER stress and assemble into a transcriptional complex to regulate stress response genes in Arabidopsis. *Plant Cell* **22**: 782–796.
- Maere, S., Heymans, K., and Kuiper, M.** (2005). BiNGO: A Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21**: 3448–3449.
- Martínez, D.E., Costa, M.L., and Guiamet, J.J.** (2008). Senescence-associated degradation of chloroplast proteins inside and outside the organelle. *Plant Biol. (Stuttg.)* **10** (Suppl. 1), 15–22.
- Martínez-García, J.F., Huq, E., and Quail, P.H.** (2000). Direct targeting of light signals to a promoter element-bound transcription factor. *Science* **288**: 859–863.
- Matys, V., et al.** (2006). TRANSFAC and its module TRANSCOMP: Transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34** (Database issue): D108–D110.
- Meierhoff, K., Felder, S., Nakamura, T., Bechtold, N., and Schuster, G.** (2003). HCF152, an *Arabidopsis* RNA binding pentatricopeptide repeat protein involved in the processing of chloroplast *psbB-psbT-psbH-petB-petD* RNAs. *Plant Cell* **15**: 1480–1495.
- Meurer, J., Meierhoff, K., and Westhoff, P.** (1996). Isolation of high-chlorophyll-fluorescence mutants of *Arabidopsis thaliana* and their characterisation by spectroscopy, immunoblotting and northern hybridisation. *Planta* **198**: 385–396.
- 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**: 853–867.
- Mittler, R., Kim, Y., Song, L., Coutu, J., Coutu, A., Ciftci-Yilmaz, S., Lee, H., Stevenson, B., and Zhu, J.K.** (2006). Gain- and loss-of-function mutations in *Zat10* enhance the tolerance of plants to abiotic stress. *FEBS Lett.* **580**: 6537–6542.
- Monastyrska, I., Rieter, E., Klionsky, D.J., and Reggiori, F.** (2009). Multiple roles of the cytoskeleton in autophagy. *Biol. Rev. Camb. Philos. Soc.* **84**: 431–448.
- Morris, K.A.H., MacKerness, S.A., 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**: 677–685.
- Nakashima, K., Fujita, Y., Katsura, K., Maruyama, K., Narusaka, Y., Seki, M., Shinozaki, K., and Yamaguchi-Shinozaki, K.** (2006). Transcriptional regulation of ABI3- and ABA-responsive genes including *RD29B* and *RD29A* in seeds, germinating embryos, and seedlings of Arabidopsis. *Plant Mol. Biol.* **60**: 51–68.
- Noodén, L.D., Singh, S., and Letham, D.S.** (1990). Correlation of xylem sap cytokinin levels with monocarpic senescence in soybean. *Plant Physiol.* **93**: 33–39.
- Olsen, A.N., Ernst, H.A., Leggio, L.L., and Skriver, K.** (2005). NAC transcription factors: Structurally distinct, functionally diverse. *Trends Plant Sci.* **10**: 79–87.
- Ongaro, V., and Leyser, O.** (2008). Hormonal control of shoot branching. *J. Exp. Bot.* **59**: 67–74.
- Palaniswamy, S.K., James, S., Sun, H., Lamb, R.S., Davuluri, R.V., and Grotewold, E.** (2006). AGRIS and AtRegNet. A platform to link cis-regulatory elements and transcription factors into regulatory networks. *Plant Physiol.* **140**: 818–829.
- Prigent, C., and Dimitrov, S.** (2003). Phosphorylation of serine 10 in histone H3, what for? *J. Cell Sci.* **116**: 3677–3685.
- Ramírez-Parra, E., Fründt, C., and Gutierrez, C.** (2003). A genome-wide identification of E2F-regulated genes in Arabidopsis. *Plant J.* **33**: 801–811.
- Robatzek, S., and Somssich, I.E.** (2001). A new member of the Arabidopsis WRKY transcription factor family, AtWRKY6, is associated with both senescence- and defence-related processes. *Plant J.* **28**: 123–133.
- Sakuma, Y., Maruyama, K., Qin, F., Osakabe, Y., Shinozaki, K., and Yamaguchi-Shinozaki, K.** (2006). Dual function of an Arabidopsis transcription factor *DREB2A* in water-stress-responsive and heat-stress-responsive gene expression. *Proc. Natl. Acad. Sci. USA* **103**: 18822–18827.
- Shockey, J.M., Fulda, M.S., and Browse, J.A.** (2002). Arabidopsis contains nine long-chain acyl-coenzyme A synthetase genes that participate in fatty acid and glycerolipid metabolism. *Plant Physiol.* **129**: 1710–1722.
- Sokol, A., Kwiatkowska, A., Jerzmanowski, A., and Prymakowska-Bosak, M.** (2007). Up-regulation of stress-inducible genes in tobacco and Arabidopsis cells in response to abiotic stresses and ABA treatment correlates with dynamic changes in histone H3 and H4 modifications. *Planta* **227**: 245–254.
- Staswick, P.E.** (1994). Storage proteins of vegetative plant tissues. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **45**: 303–322.
- Staswick, P.E.** (2008). JAZing up jasmonate signaling. *Trends Plant Sci.* **13**: 66–71.
- Tanaka, R., and Tanaka, A.** (2007). Tetrapyrrole biosynthesis in higher plants. *Annu. Rev. Plant Biol.* **58**: 321–346.
- Toledo-Ortiz, G., Huq, E., and Quail, P.H.** (2003). The Arabidopsis basic/helix-loop-helix transcription factor family. *Plant Cell* **15**: 1749–1770.
- Tran, L.S., Nakashima, K., Sakuma, Y., Simpson, S.D., Fujita, Y., Maruyama, K., Fujita, M., Seki, M., Shinozaki, K., and Yamaguchi-Shinozaki, K.** (2004). Isolation and functional analysis of Arabidopsis stress-inducible NAC transcription factors that bind to a drought-responsive cis-element in the early responsive to dehydration stress 1 promoter. *Plant Cell* **16**: 2481–2498.
- Ulker, B., and Somssich, I.E.** (2004). WRKY transcription factors: From DNA binding towards biological function. *Curr. Opin. Plant Biol.* **7**: 491–498.
- 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**: 776–792.
- Walley, J.W., Coughlan, S., Hudson, M.E., Covington, M.F., Kaspi, R., Banu, G., Harmer, S.L., and Dehesh, K.** (2007). Mechanical stress induces biotic and abiotic stress responses via a novel cis-element. *PLoS Genet.* **3**: 1800–1812.
- Waters, M.T., Wang, P., Korkaric, M., Capper, R.G., Saunders, N.J., and Langdale, J.A.** (2009). GLK transcription factors coordinate expression of the photosynthetic apparatus in Arabidopsis. *Plant Cell* **21**: 1109–1128.
- 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**: 455–469.
- Weisshaar, B., and Jenkins, G.I.** (1998). Phenylpropanoid biosynthesis and its regulation. *Curr. Opin. Plant Biol.* **1**: 251–257.
- Wenkel, S., Turck, F., Singer, K., Gissot, L., Le Gourrierec, J., Samach, A., and Coupland, G.** (2006). *CONSTANS* and the CCAAT box binding complex share a functionally important domain and interact to regulate flowering of *Arabidopsis*. *Plant Cell* **18**: 2971–2984.
- Westfall, P.H., Krishen, A., and Young, S.S.** (1998). Using prior

- information to allocate significance levels for multiple endpoints. *Stat. Med.* **17**: 2107–2119.
- Woo, H.R., Chung, K.M., Park, J.-H., Oh, S.A., Ahn, T., Hong, S.H., Jang, S.K., and Nam, H.G.** (2001). ORE9, an F-box protein that regulates leaf senescence in Arabidopsis. *Plant Cell* **13**: 1779–1790.
- Wu, H., Kerr, K., Cui, X., and Churchill, G.** (2003). MAANOVA: A software package for the analysis of spotted cDNA microarray experiments. In *The Analysis of Gene Expression Data: Methods and Software*, G. Parmigiani, E. Garrett, R. Irizarry, and S. Zeger, eds. (New York: Springer), pp. 313–341.
- Yang, T., and Poovaiah, B.W.** (2002). A calmodulin-binding/CGCG box DNA-binding protein family involved in multiple signaling pathways in plants. *J. Biol. Chem.* **277**: 45049–45058.
- Yoo, J.H., et al.** (2005). Direct interaction of a divergent CaM isoform and the transcription factor, MYB2, enhances salt tolerance in Arabidopsis. *J. Biol. Chem.* **280**: 3697–3706.
- Yoshimoto, K., Hanaoka, H., Sato, S., Kato, T., Tabata, S., Noda, T., and Ohsumi, Y.** (2004). Processing of ATG8s, ubiquitin-like proteins, and their deconjugation by ATG4s are essential for plant autophagy. *Plant Cell* **16**: 2967–2983.