

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/60669>

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.

## Library Declaration and Deposit Agreement

### 1. STUDENT DETAILS

*Please complete the following:*

Full name: ..... **Aleksey Jironkin** .....  
University ID number: ..... **0507638** .....

### 2. THESIS DEPOSIT

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

2.2 The hard copy will be housed in the University Library. The digital version will be deposited in the University's Institutional Repository (WRAP). Unless otherwise indicated (see 2.3 below) this will be made openly accessible on the Internet and will be supplied to the British Library to be made available online via its Electronic Theses Online Service (EThOS) service.

[At present, theses submitted for a Master's degree by Research (MA, MSc, LLM, MS or MMedSci) are not being deposited in WRAP and not being made available via EThOS. This may change in future.]

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

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

(a) Hard Copy

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

I agree that my thesis may be photocopied.

YES / NO (*Please delete as appropriate*)

(b) Digital Copy

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

Please choose one of the following options:

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

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

OR My full thesis cannot be made publicly available online but I am submitting a separately identified additional, abridged version that can be made available online.

YES / NO (*Please delete as appropriate*)

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

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

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

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

### **4. DECLARATIONS**

(a) I DECLARE THAT:

- I am the author and owner of the copyright in the thesis and/or I have the authority of the authors and owners of the copyright in the thesis to make this agreement. Reproduction of any part of this thesis for teaching or in academic or other forms of publication is subject to the normal limitations on the use of copyrighted materials and to the proper and full acknowledgement of its source.
- The digital version of the thesis I am supplying is the same version as the final, hard-bound copy submitted in completion of my degree, once any minor corrections have been completed.
- I have exercised reasonable care to ensure that the thesis is original, and does not to the best of my knowledge break any UK law or other Intellectual Property Right, or contain any confidential material.
- I understand that, through the medium of the Internet, files will be available to automated agents, and may be searched and copied by, for example, text mining and plagiarism detection software.

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

- I grant the University of Warwick and the British Library a licence to make available on the Internet the thesis in digitised format through the Institutional Repository and through the British Library via the EThOS service.
- If my thesis does include any substantial subsidiary material owned by third-party copyright holders, I have sought and obtained permission to include it in any version of my thesis available in digital format and that this permission encompasses the rights that I have granted to the University of Warwick and to the British Library.

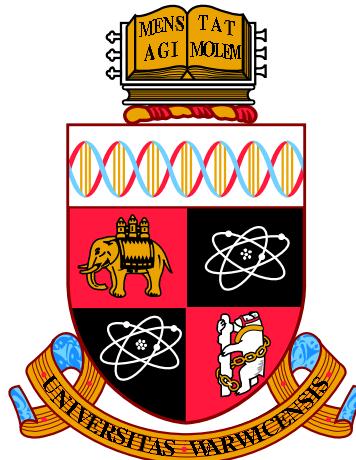
### **5. LEGAL INFRINGEMENTS**

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

---

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

Student's signature: .....  Date: 8/03/2014



**Computational And Experimental Analysis Of  
Plant Promoters: Identifying Functional Elements**

by

**Aleksey Jironkin**

**Thesis**

Submitted to the University of Warwick

for the degree of

**Doctor of Philosophy**

**Systems Biology**

September 2013

THE UNIVERSITY OF  
**WARWICK**

# Contents

<b>List of Tables</b>	<b>vi</b>
<b>List of Figures</b>	<b>viii</b>
<b>Acknowledgments</b>	<b>xii</b>
<b>Declarations</b>	<b>xiii</b>
<b>Abstract</b>	<b>xiv</b>
<b>Abbreviations</b>	<b>xvi</b>
<b>Chapter 1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	2
1.2 Regulation of gene expression . . . . .	3
1.2.1 TF families in Arabidopsis . . . . .	7
1.3 Gene regulatory networks . . . . .	10
1.3.1 Types of gene regulatory networks . . . . .	11
1.3.2 Strategies for uncovering gene regulatory networks . . . . .	12
1.4 Stress response in Arabidopsis . . . . .	14
1.4.1 PAMP Triggered Immunity . . . . .	14
1.4.2 Effector Triggered Immunity . . . . .	15
1.4.3 The role of hormones in stress response . . . . .	16
1.4.4 Hormone crosstalk fine-tunes the defence response in Arabidopsis . . . . .	20
1.5 Infection by Botrytis . . . . .	22
1.5.1 Botrytis Infection process . . . . .	22
1.5.2 Changes in Arabidopsis transcriptome in response to Botrytis	22
1.6 Aims and objectives . . . . .	23
1.7 Organisation of this thesis . . . . .	24

<b>Chapter 2 Conserved Noncoding Sequences Highlight Shared Components of Regulatory Networks in Dicotyledonous Plants</b>	<b>26</b>
2.1 Introduction . . . . .	26
2.1.1 Phylogenetic Footprinting . . . . .	27
2.1.2 Selection of Compatible Plant Species for Phylogenetic Footprinting . . . . .	27
2.1.3 Previous studies . . . . .	28
2.2 Methods . . . . .	31
2.2.1 Databases . . . . .	31
2.2.2 APPLES Framework . . . . .	31
2.2.3 Ortholog and Paralog Identification . . . . .	31
2.2.4 Sequence Alignments . . . . .	32
2.2.5 Converting Alignment Scores to Conservation Scores . . . . .	33
2.2.6 Filtering Out Potential Protein Coding Regions . . . . .	35
2.2.7 PSSM Clustering . . . . .	37
2.2.8 Motif Overpresentation . . . . .	37
2.2.9 GO Term Analysis . . . . .	38
2.2.10 Prediction of Nucleosome Positioning . . . . .	38
2.2.11 DNase-Seq Analysis . . . . .	39
2.3 Results . . . . .	40
2.3.1 Multispecies Analysis Yields Hundreds of CNSs . . . . .	40
2.3.2 CNSs Show Positional Bias toward TSSs . . . . .	44
2.3.3 CNSs Are Highly Enriched in TFBS Motifs . . . . .	46
2.3.4 Identification of Previously Experimentally Validated Promoter Binding Elements . . . . .	48
2.3.5 Prediction of Nucleosome Positioning . . . . .	48
2.3.6 GO Term Overrepresentation Unveils Key Biological and Molecular Functions of Genes Associated with CNSs . . . . .	50
2.3.7 Predicted CNSs Occur in Open Chromatin Areas . . . . .	51
2.4 Discussion . . . . .	53
2.4.1 Neo-/Non-/Sub-functionalisation of Conserved Genes . . . . .	53
2.4.2 Genome Availability and Annotation Quality Impact on CNS Predictions . . . . .	56
2.4.3 CNSs Are Likely To Be Functional Regions Of ncDNA . . . . .	57
2.4.4 Effectiveness of Alignment-Based Methods in CNS Discovery	61
2.5 Conclusions . . . . .	62

<b>Chapter 3 Elucidating Functional Elements and Gene Regulatory Network Using Yeast One-Hybrid Screens</b>	<b>63</b>
3.1 Introduction . . . . .	63
3.1.1 Available Experimental Techniques For Probing Protein-DNA interactions . . . . .	64
3.1.2 Biological Context . . . . .	68
3.2 Methods . . . . .	71
3.2.1 Gene Selection . . . . .	71
3.2.2 Yeast One-Hybrid . . . . .	71
3.2.3 Image Based Positive Result Inference . . . . .	79
3.2.4 pairwise mini-library Y1H screen . . . . .	80
3.3 Results . . . . .	83
3.3.1 Initial Gene Selection using WIGWAMS tool . . . . .	83
3.3.2 Clock Regulated Stress Response . . . . .	86
3.4 High-throughput identification of direct protein-DNA interactions using Y1H library screen . . . . .	88
3.4.1 Subscreen of TCP Transcription Factors using Y1H library screen . . . . .	89
3.4.2 Confirmation of Observed Y1H Interactions Using Pairwise Screens . . . . .	90
3.4.3 WRKY TF Subscreen . . . . .	94
3.4.4 Summary of the Y1H Screen Results . . . . .	97
3.5 Discussion . . . . .	106
3.5.1 Image Based Method For Identification of Positive Results In Y1H . . . . .	106
3.5.2 Y1H Screen Reliability and Reproducibility . . . . .	110
3.5.3 Importance of Correct 3AT Concentration . . . . .	111
3.5.4 Combined Y1H Screen Can Uncover Common Regulators . . . . .	118
3.5.5 Y1H Library Screen Can Uncover Subgroups of Co-Regulated Genes . . . . .	119
3.5.6 Circadian Clock Timing Stress Response . . . . .	121
3.6 Conclusions . . . . .	123
<b>Chapter 4 Computational Approaches To Identify Regulatory Elements In Arabidopsis</b>	<b>126</b>
4.1 Introduction . . . . .	126
4.1.1 TF Binding Site Availability . . . . .	127

4.1.2	Verification in plants . . . . .	128
4.1.3	DNase Assays . . . . .	129
4.2	Methods . . . . .	133
4.2.1	Promoter DNA analysis . . . . .	133
4.2.2	DNase Analysis . . . . .	133
4.2.3	Plant growth . . . . .	134
4.2.4	Fungal growth . . . . .	134
4.2.5	Phenotype Analysis . . . . .	134
4.2.6	Microarray analysis . . . . .	135
4.2.7	Microarray scanning . . . . .	137
4.2.8	Expression Analysis . . . . .	137
4.2.9	Gene Ontology (GO) analysis . . . . .	138
4.2.10	Pathway Analysis with MapMan . . . . .	138
4.3	Results . . . . .	139
4.3.1	New TF specific motifs conserved within the promoter fragments of gene tested in Y1H screen . . . . .	139
4.3.2	<i>De novo</i> motifs show DNase I footprint in genome-wide locations in <i>Arabidopsis</i> leaves and buds tissue . . . . .	142
4.3.3	Mutations of the new binding sites alter protein-DNA interactions of associated TFs . . . . .	146
4.3.4	Phenotype screen of <i>erf14</i> , <i>pif7</i> and <i>athb25</i> KO plants show increased susceptibility to infection with <i>Botrytis</i> . . . . .	150
4.3.5	Microarray analysis of <i>erf14</i> KO plants reveal new targets of the TF . . . . .	151
4.3.6	Reported Y1H interactions for <i>AtERF14</i> and <i>ESE1</i> are confirmed in protoplasts . . . . .	158
4.4	Discussion . . . . .	161
4.4.1	Characterisation of novel <i>cis</i> -acting elements in the promoters of gene screened in the Y1H experiments . . . . .	161
4.4.2	Sequence specificity of AP2 domain proteins . . . . .	163
4.4.3	DNase I analysis . . . . .	164
4.4.4	The <i>de novo</i> motifs interact with specific TFs . . . . .	165
4.4.5	Summary of new motif interaction patterns . . . . .	167
4.4.6	Role of <i>PIF7</i> and <i>AtHB25</i> in <i>Botrytis</i> infection . . . . .	176
4.4.7	Role of <i>AtERF14</i> in <i>Botrytis</i> infection . . . . .	176
4.4.8	Role of <i>ESE1</i> in <i>Botrytis</i> infection . . . . .	181

4.4.9	Context of regulation by <i>AtERF14</i> and <i>ESE1</i> TFs and their respected motifs . . . . .	185
4.5	Conclusions . . . . .	186
<b>Chapter 5 General Discussion</b>		<b>187</b>
<b>Appendix A Conserved Noncoding Sequences Highlight Shared Com-     ponents of Regulatory Networks in Dicotyledonous Plants</b>		<b>191</b>
<b>Appendix B Elucidating Functional Elements and Gene Regulatory     Network Using Yeast One-Hybrid Screens</b>		<b>196</b>
<b>Appendix C Computational Approaches To Identify Regulatory Ele-     ments In Arabidopsis</b>		<b>201</b>

# List of Tables

2.1	Summary of aligned regions and associated genes from orthologous promoters . . . . .	36
2.2	Summary of aligned regions and associated genes from orthologous promoters . . . . .	43
2.3	Expected true positive CNSs . . . . .	43
2.4	GO Term analysis of the genes in the CNS set . . . . .	51
3.1	Y1H amplification primers . . . . .	72
3.2	SOC media . . . . .	74
3.3	Cloning summary of fragments that were screened at Warwick laboratory. Green represents successful transformation into the specified product. . . . .	75
3.4	Cloning summary for the first cluster destined for the USA. Green represents successful transformation into the specified product. . . . .	76
3.5	Cloning for the second cluster destined for the USA. Green represents successful transformation into the specified product. . . . .	77
3.6	TRAFCO mix . . . . .	82
3.7	Genes in the “JAZ” cluster . . . . .	83
3.8	Genes in the “TCP” cluster . . . . .	85
3.9	Gene in the “WRKY” cluster . . . . .	87
3.10	TCP mini-library Y1H results . . . . .	91
3.11	Summary of the sequence verification of TFs for the mini-library . . . . .	92
3.12	Summary of the overall Y1H results . . . . .	93
3.13	Summary of the WRKY mini-screen . . . . .	94
3.14	FPR and FNR definitions . . . . .	113
3.15	Summary of FPR and FNR for Y1H library screen . . . . .	114
3.16	Summary of FNR changes with colony numbers in Y1H library screen	118
4.1	WRKY motifs derived using MEME and Y1H results . . . . .	140

4.2	AP2/ERF motifs derived using MEME and Y1H results . . . . .	141
4.3	Motifs present derived using MEME and Y1H results for PIF7, bZIP52, AtHB25 and NAC098 . . . . .	143
4.4	Library arrangement used in mutagenesis . . . . .	148
4.5	Differentially expressed genes in <i>erf14</i> 24 hpi with Botrytis . . . . .	154
4.6	Most overrepresented GO terms <i>erf14</i> differentially expressed genes. .	155
4.7	GO Term analysis of genes differentially expressed in <i>erf14</i> . . . . .	155
4.8	GO Term analysis of differentially expressed genes in protoplasts over-expressing <i>AtERF14</i> . . . . .	157
4.9	JAZ1 and MYB15, that were found to be interacting with <i>AtERF14</i> in Y1H, are also significantly differentially expressed in <i>erf14</i> KO lines compared to WT and in protoplasts overexpressing <i>AtERF14</i> . .	158
4.10	Direct targets of <i>AtERF14</i> in protoplasts and <i>erf14</i> plants . . . . .	160
A.1	Manually curated list of plant “Master Regulators” . . . . .	192
A.2	Numbers of aligned regions and associated genes from paralogous promoters . . . . .	193
B.1	Summary of Y1H library results. . . . .	197
C.1	All of the samples are derived from the normal distribution as determined by one-sample Kolmogorov-Smirnoff test. The values in the brackets show number of samples in each category. . . . .	202

# List of Figures

1.1	Overview of gene regulation and transcription mechanism . . . . .	5
1.2	Arabidopsis circadian clock model . . . . .	12
1.3	An overview of plant defence response mechanisms . . . . .	15
1.4	An overview of JA signalling mechanism . . . . .	19
1.5	Networking by phytohormones in the plant immune response . . . . .	21
2.1	Arabidopsis, papaya, poplar and grape phylogenetic tree . . . . .	29
2.2	The alignment score histograms between Arabidopsis and 3 other species . . . . .	34
2.3	Venn diagrams of distribution of Arabidopsis orthologs across comparator species . . . . .	40
2.4	Positional Bias of the CNSs towards the TSSs . . . . .	45
2.5	Arabidopsis CNSs enrichment for TF binding sites . . . . .	47
2.6	Predictions of nucleosome occupancy in the CNSs . . . . .	49
2.7	Histogram of DNase sequencing read from Arabidopsis leaf tissue in the CNSs . . . . .	52
2.8	Subfunctionalisation in regulatory regions of Arabidopsis paralogs .	55
3.1	Diagram of the steps in the Y1H library screen . . . . .	69
3.2	Expression of genes in “JAZ” cluster . . . . .	84
3.3	Expression of genes in “TCP” cluster . . . . .	85
3.4	mRNA expression levels of <i>WRKY11</i> and <i>WRKY40</i> during infection with <i>Botrytis</i> . . . . .	87
3.5	mRNA expression levels of the genes in the “WRKY” cluster . . .	88
3.6	Cumulative distributions of histograms of pixel intensities for Y1H-161	95
3.7	Cumulative histograms of intensities for spots predicted to be positive interactions by the automatic algorithm for Y1H-161 fragment. . . .	98
3.7	Cumulative histograms of intensities for spots predicted to be positive interactions by the automatic algorithm for Y1H-161 fragment (Cont.).	99

3.8	Final Y1H screen results after library and pairwise screens above each promoter fragment. Green - fragment location and identifier, blue - CNS location and score. . . . .	101
3.8	Final Y1H screen results after library and pairwise screens above each promoter fragment. Green - fragment location and identifier, blue - CNS location and score. . . . .	102
3.8	Final Y1H screen results after library and pairwise screens above each promoter fragment. Green - fragment location and identifier, blue - CNS location and score. . . . .	103
3.8	Final Y1H screen results after library and pairwise screens above each promoter fragment. Green - fragment location and identifier, blue - CNS location and score. . . . .	104
3.8	Final Y1H screen results after library and pairwise screens above each promoter fragment. Green - fragment location and identifier, blue - CNS location and score. . . . .	105
3.9	Typical image of Y1H pairwise plate . . . . .	107
3.10	Image of Y1H-177 SD-LTH plate . . . . .	109
3.11	mRNA expression levels in response to infection with <i>B. cinerea</i> of common Y1H regulators (blue) and corresponding target genes (red). .	120
3.12	mRNA expression levels during infection with Botrytis of two TFs (blue) predicted to be regulating a small subgroup of genes (red) in Y1H screens. . . . .	121
3.13	Gene regulatory network from Y1H interactions . . . . .	125
4.1	An overview of DNase-Seq . . . . .	131
4.2	DNase footprint for SEP3 binding site . . . . .	132
4.3	DNase footprint profiles of ORA59, AtERF14, ESE1 and AtERF7 binding motifs . . . . .	144
4.3	DNase footprint profiles for newly discovered motifs . . . . .	145
4.4	Mutagenesis setup for Y1H-174 and AtERF14, ESE1 and AtERF7 .	147
4.5	Images of WT and mutagenised promoters on SD-LTH agar plates vs mini-library arrangements (Table 4.4). Bright spots indicate growing yeast colonies. . . . .	149
4.6	Pairwise of serial dilutions of Y1H-174 WT and <b>m5</b> vs select TFs .	150
4.7	The growth of lesion sizes is slowed in the <i>pif7</i> and <i>athb25</i> 72 hours post infection. Botrytis Lesion size comparisons between Col0, <i>erf14</i> , <i>pif7</i> and <i>athb25</i> at 48, 57.5 and 72 hpi. . . . .	152

4.8	qPCR confirming AtERF14 KO in Arabidopsis . . . . .	153
4.9	mRNA expression levels of <i>AtERF14</i> (red) and associated targets (blue) during infection with Botrytis. . . . .	160
4.10	DNase I cutting profile of published and <i>de novo</i> motifs in Arabidopsis leafs for ORA59 TF. . . . .	165
4.11	DNase profile of new and unknown motif . . . . .	166
4.12	<i>At1g19180</i> intergenic region with predicted motifs and DNase profile	168
4.13	<i>At1g80840</i> intergenic region with predicted motifs and DNase profile	169
4.14	<i>At2g35930</i> intergenic region with predicted motifs and DNase profile	170
4.15	<i>At2g44840</i> intergenic region with predicted motifs and DNase profile	171
4.16	<i>At3g23250</i> intergenic region with predicted motifs and DNase profile	172
4.17	<i>At3g25760</i> intergenic region with predicted motifs and DNase profile	173
4.18	<i>At3g25780</i> intergenic region with predicted motifs and DNase profile	174
4.19	<i>At4g31550</i> intergenic region with predicted motifs and DNase profile	175
4.20	mRNA expression profiles of <i>AtERF14</i> (red) and associated targets (blue), found to be differentially expressed in <i>erf14</i> KO line, and have correlated expression during the infection with Botrytis supporting regulatory link. . . . .	178
4.20	mRNA expression profiles of <i>AtERF14</i> (red) and associated targets (blue), found to be differentially expressed in <i>erf14</i> KO line, and have correlated expression during the infection with Botrytis supporting regulatory link. . . . .	179
4.20	mRNA expression profiles of <i>AtERF14</i> (red) and associated targets (blue), found to be differentially expressed in <i>erf14</i> KO line, and have correlated expression during the infection with Botrytis supporting regulatory link. . . . .	180
4.21	mRNA expression profiles of <i>ESE1</i> (red) and associated targets (blue), found to be differentially expressed in protoplasts overexpressing <i>ESE1</i> TF, and have correlated expression during the infection with Botrytis supporting regulatory link. . . . .	182
4.21	mRNA expression profiles of <i>ESE1</i> (red) and associated targets (blue), found to be differentially expressed in protoplasts overexpressing <i>ESE1</i> TF, and have correlated expression during the infection with Botrytis supporting regulatory link. . . . .	183

4.21 mRNA expression profiles of <i>ESE1</i> (red) and associated targets (blue), found to be differentially expressed in protoplasts overexpressing <i>ESE1</i> TF, and have correlated expression during the infection with Botrytis supporting regulatory link. . . . .	184
A.1 Normalised distribution of CNS distances from the TSS . . . . .	194
A.2 Distribution of CNS lengths . . . . .	195

# Acknowledgments

First of all I would like to thank my supervisors, especially Katherine for providing support and guidance throughout my PhD. At the same time I would like to thank Systems Biology DTC and Vicky Buchanan-Wollaston for the support provided by the department.

I would also like to thank all the people I had pleasure of meeting and working with, including but not limited to: Johanna Rhodes, Miguel, Peter Krusche, Adam Talbot, Sarah Harvey, Nigel Dyer, Steve Kiddle, Richard Hickman, Ana Mendes, Ben Wareham, Damon Daniels, Ben Hamilton, Leslie, all the members of the Systems Biology DTC, PRESTA group. Special thanks go out to Peijun Zhang, Claire Hills, Alison Jackson, Justyna, Emily Breeze, who have provided immense amount of help and support in the lab and writing. Without Alison Jackson's help in the lab many of the results presented here would not have happened. I would like to extend a separate thanks to all my close friends: Jo, Phil, Ali, John, Chris, Max and Mato: we had great time off work with "Game Of Thrones". Thanks to all of my housemates over the years for putting up with me! Last but not least I would like to thank my main man Cai Wingfield for his helpful comments and discussions.

Finally, I would like to thank my mum, grandmother, aunt, uncle Nastia and Zlata for all the years of support they have given me, including the offer to write this thesis for me. Zoe for being the biggest joy of my life.

This thesis was typeset with L<sup>A</sup>T<sub>E</sub>X 2 <sub>$\varepsilon$</sub> <sup>1</sup> by the author.

---

<sup>1</sup>L<sup>A</sup>T<sub>E</sub>X 2 <sub>$\varepsilon$</sub>  is an extension of L<sup>A</sup>T<sub>E</sub>X. L<sup>A</sup>T<sub>E</sub>X is a collection of macros for T<sub>E</sub>X. T<sub>E</sub>X is a trademark of the American Mathematical Society. The style package *warwickthesis* was used.

# **Declarations**

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

# Abstract

Understanding the regulatory DNA sequences are becoming increasingly important in understanding the way plants integrate signalling cues mediated through the actions of the transcription factors (TFs). This thesis presents an interdisciplinary investigations into regulatory elements found in the promoter regions of a model organism *Arabidopsis thaliana*.

The intergenic DNA sequences are studied between sets of orthologous genes in *A. thaliana* and 3 other related species to uncover hundreds of evolutionary conserved noncoding sequences (CNSs). The CNSs are found to be more skewed towards the annotated transcription start sites (TSSs) and enriched in previously identified transcription factors binding motifs. Furthermore, the nucleosomes are predicted to have strong presence in the uncovered CNS than random intergenic sequences alone. Altogether the evidence presented in the thesis points to the functional nature of the CNSs.

Then, the promoters of genes thought to be co-regulated together and transcriptionally active during infection with fungal pathogen *Botrytis cinerea* are experimentally tested for direct protein-DNA interaction using high-throughput Yeast One-Hybrid (Y1H) library screens against the TFs found in *A. thaliana*. The resulting predictions were further validated using pairwise Y1H screen to suggest potential common regulation by *ORA59*, *PIF7*, *ESE1*, *At4g38900* and *ERF14*, and uncovering a complex gene regulatory network (GRN) associated with the tested genes.

The promoter fragments together with the predictions from the Y1H screens were used in the computational analysis to establish transcription factor specific binding motifs. Some of the newly predicted motifs were mutated and tested again for altered binding of the associated TFs. Furthermore, *in planta* mutations of the TFs predicted to be interacting with the promoters of the genes in the Y1H screens were found to have significant impact on the susceptibility of *A. thaliana* to infection with *B. cinerea*, further informing gene regulatory network active in response to biotic stress.

## Resulting publications and those in preparation

L. Baxter<sup>1</sup>, **A. Jironkin**<sup>1</sup>, R. Hickman<sup>1</sup>, J. Moore, C. Barrington, P. Krusche, N.P. Dyer, V. Buchanan-Wollaston, A. Tiskin, J. Beynon, K. Denby, S. Ott, “Conserved noncoding sequences highlight shared components of regulatory networks in dicotyledonous plants”, *Plant Cell*, vol. 24, no. 10, pp. 3949–65, Oct. 2012.

K. Polanski, J. Rhodes, C. Hill, P. Zhang, D. Jenkins, S. Kiddle, **A. Jironkin**, J. Beynon, V. Buchanan-Wollaston, S. Ott, K. Denby, “Wigwams: identifying gene modules co-regulated across multiple biological conditions”. Submitted to *Bioinformatics*.

## APPLES software

The APPLES framework can be downloaded, with the installation instructions, from:

[http://www2.warwick.ac.uk/fac/sci/systemsbiology/staff/ott/tools\\_and\\_software/apples/](http://www2.warwick.ac.uk/fac/sci/systemsbiology/staff/ott/tools_and_software/apples/)

# Abbreviations

- ABA Abscic Acid  
AD Activation domain  
BRE IIB recognition element  
bHLH basin Helix Loop Helix  
bZIP bsic Zipper  
CBE Cold binding element  
CNS Conserved noncoding sequence  
ChIP Chromatin Immunoprecipitation  
DBD DNA binding domain  
DB DNA binding  
DPE Downstream promoter element  
DREB Dehydration responsive element binding  
ETI Effector triggered immunity  
EMSA Electrophoresis mobility shift assay  
ERF Ethylene response factor  
ET Ethylene  
eY1H enhanced Yeast One-Hybrid  
FGF Fibroblast response factor  
FNR False negative rate  
FPR False positive rate  
JA Jasmonic acid

hpi hours post infection  
GM Genetic modification  
GRN Gene regulatory network  
IFPRI International food policy research institute  
LRR Leucine rich region  
mRNA messenger RNA  
miRNA micro RNA  
ncDNA noncoding DNA  
PAMP Pathogen associated molecular patterns  
PSSM Position specific scoring matrix  
PTI PAMP triggered immunity  
SA Salicilic acid  
siRNA short interfering RNA  
TAIR The Arabidopsis information resource  
TBP TATA binding protein  
TF Transcription factor  
TSS Transcription start site  
T-DNA Transfer-DNA  
T3SS Type 3 secretion system  
UTR untranslated region  
WGD Whole genome duplication  
WM Weight matrix  
WIGWAM WIGWAMS identifies genes working across multiple stresses  
Y1H Yeast One-Hybrid



# Chapter 1

## Introduction

Direct alteration of genomic DNA offers a powerful mechanism for improving a species' fitness, survivability and adaptation potential in the wake of an ever changing environment. For plants this means an increased resilience to pathogens and improved crop yields. Until recently, cross breeding of different plant varieties allowed for the transfer of desirable traits from one crop or wild variety to another. However, this is a slow and often unsuccessful process. In the early twentieth century researchers determined the importance of certain amino acids for nutrition. Tryptophan (Trp) was the first amino acid to be recognized as essential for normal growth of young animals when Willcock and Hopkins (1906) and later Osborne and Mendel (1914) observed its ability to stimulate weight gain in mice and rats when added to low Trp rations. Subsequent studies in a variety of species confirmed that Trp was essential for normal growth and furthermore, was required for maintenance of nitrogen equilibrium in mature animals. Some years after those early animal studies, Rose (1957) demonstrated that Trp was an essential amino acid for humans. Soon after that, scientists began to search for a variety of maize that would yield higher levels of Trp to supplement our daily diet. In the late 1960s, researchers found a variety of maize with a mutation in *opaque-2* gene that had higher amounts of Trp (Mertz et al., 1964). However, this highly desirable trait also had some undesirable side effects. The dull, chalky, soft *opaque-2* maize kernels yielded 15% to 20% less grain weight than wild-type grain. However, scientists from the International Maize and Wheat Improvement Center (Mexico City) working with *opaque-2* maize observed small islands of translucent starch in some *opaque-2* endosperms. Using conventional breeding methodologies supported by rapid chemical analysis of large numbers of samples, scientists were able to slowly accumulate modifier genes to convert the original soft *opaque-2* endosperm into vitreous, hard endosperm types.

This conversion took nearly 20 years. If genetic engineering techniques had been available then, the genes that controlled high Trp levels could have been inserted into high-yielding hard-endosperm phenotypes, saving decades of labour intensive plant breeding.

## 1.1 Motivation

Food security remains one of the biggest challenges facing mankind. The problem of food security is two-fold- firstly, sustaining food production for a growing population and secondly, distributing food high in nutrients to the places where it is needed the most. The world population is predicted to grow by 2.5 billion to reach a staggering figure of 9.5 billion by the year 2050, even by conservative estimates (United Nations, 2012). According to the data from the IFPRI, hunger is at an “alarming” to “extremely alarming” state already in large parts of central and southern Africa (International Food Policy Research Institute, 2011). Therefore, the “Millennium Development Goals” set out as the number one target to halve world poverty and hunger by 2015 (United Nations, 2013). Among the key challenges in securing sustainable crops is improving their tolerance to a large variety of biotic and abiotic stresses including, but not limited to infection with *Botrytis cinerea*. *B. cinerea* is a necrotrophic fungus that affects over 200 plant varieties worldwide (van Kan, 2006; Williamson et al., 2007). Infection with the fungus leads to devastating pre- and post-harvest losses and severe financial losses to farmers (Williamson et al., 2007). Hence, improving our understanding of the mechanisms of infection and the defence response will provide possible future solutions in improving crops’ resilience to biotic stress. As with the example of *opaque-2* mutant in maize, improving plant responses to biotic and abiotic stresses often coincides with decreased biomass (Herms and Mattson, 1992). Each stress elicits a complex cellular and molecular response system implemented by the plant in order to prevent damage and ensure survival, but often at the detriment of growth and yield (Herms and Mattson, 1992). Therefore, manipulating plant responses to internal and external changes has the potential to significantly impact on the survivability and longevity of the plant.

Thale cress (*Arabidopsis thaliana*) has become a *de facto* model organism used to understand plant biological processes. The genome of *A. thaliana* (*Arabidopsis*) was sequenced and published in 2000 (Arabidopsis Genome Initiative, 2000) and has since been extensively annotated. *Arabidopsis* has a relatively short

life cycle of approximately 6 weeks from germination to seed maturation and can be easily cultivated in restricted spaces, such as growth rooms and chambers. Developments in the “floral dip” technique by Clough and Bent (1998) means that Arabidopsis is subject to efficient transformations utilising *Agrobacterium tumefaciens* leading to a large number of mutant lines and genomic resources becoming available from “The Arabidopsis Information Resource” (TAIR). Arabidopsis is a member of the *Brassicaceae* family, which also includes cabbage and radish and is therefore closely related to other crop species, suggesting that Arabidopsis may serve as a good proxy for uncovering biological and molecular workings of the plant. The knowledge gained in the studies of Arabidopsis can potentially be directly applied to other related crops, provided that functionally equivalent genes exist and can be identified in other plant species.

## 1.2 Regulation of gene expression

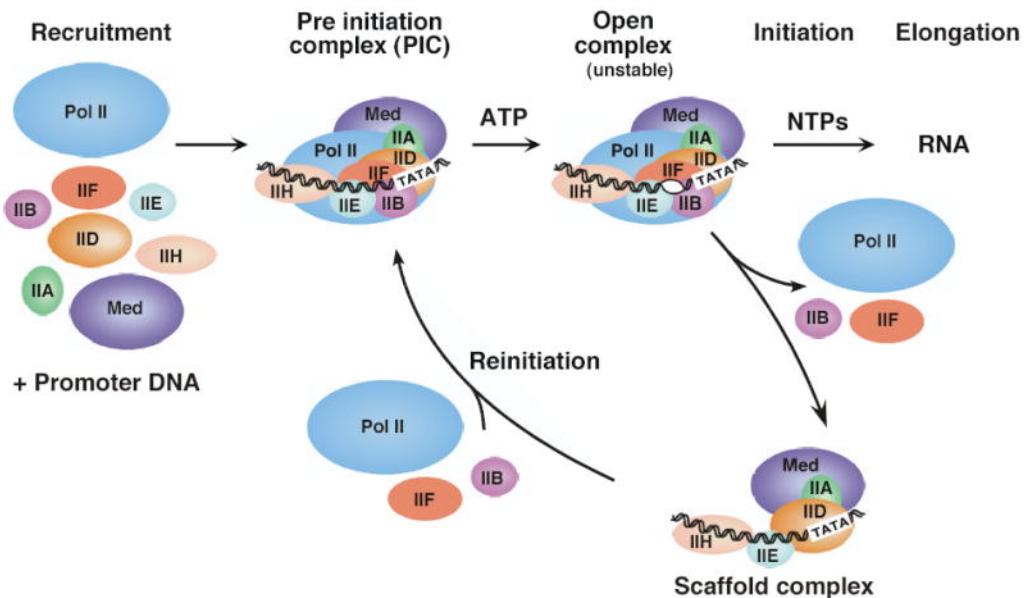
The gene is one of the key units in molecular biology. A gene is defined as a sequence of nucleotides which when converted into mRNA (messenger RNA) serves as a template for building the associated protein. Introns and exons are also contained within the coding region of a gene, splicing of the introns out of the protein coding sequence from the mRNA molecules gives rise to multiplicity of different proteins from the same template coding sequence. The process of generating mRNA from DNA is performed by RNA polymerase (Lehman et al., 1958; Bessman et al., 1958) and is called transcription. Gene expression/transcription can be enhanced or silenced through transcription factors (TFs). TFs contain DNA binding domains (DBD) which can bind directly to the DNA sequence and can also interact with other proteins and TFs.

One of the earliest studied DNA sequences found to be important for transcription of genes and bound by a TF is the 5'-TATAAA-3' regulatory sequence, also known as the TATA-box (Breathnach and Chambon, 1981). The TATA sequence is well conserved among the eukaryotes and appears 30 bp upstream from the transcription start site (TSS), forming part of the core promoter sequence, Figure 1.1b. Xu et al. (1991) have shown that the asymmetry of the TATA-box is the major determinant of the direction of transcription. TFs binding to the TATA-box are known as TATA binding proteins (TBP) and act as general activators of transcription. TBPs were first isolated from yeast (Cavallini et al., 1989; Eisenmann et al., 1989; Hahn et al., 1989; Horikoshi et al., 1989; Schmidt et al., 1989) and then

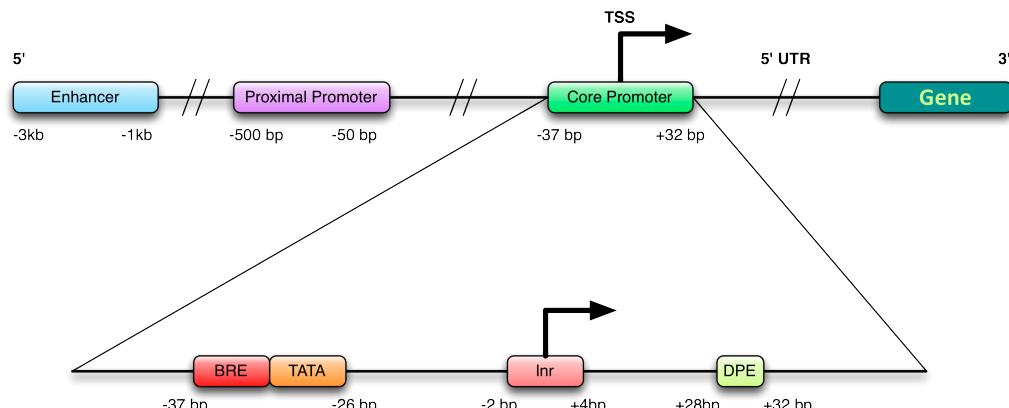
discovered in humans (Hoffman et al., 1990; Kao et al., 1990; Peterson et al., 1990). Arabidopsis TBPs have been shown to be functionally equivalent to those found in animals (Mukumoto et al., 1993), and function as a subunit of Transcription Factor II D (TFIID) along with other TBP associated factors (TAFs) (Albright and Tjian, 2000). The TATA-box acts as a starting point for the assembly of the transcription machinery, which TFIID itself is part of, whereby RNA Polymerase II, Mediator, general TFs, transcription factor II A (TFIIA), transcription factor II B (TFIIB), transcription factor II E (TFIIE), transcription factor II F (TFIIF) and transcription factor II H (TFIIC) assemble in a specific order (reviewed in Hahn (2004)).

Other polymerases include Polymerase I, which transcribes genes encoding the 45S precursor into large ribosomal RNA (reviewed in ) and Polymerase III, which transcribes many small RNA genes including those for tRNA and 5S ribosomal RNA (reviewed in ). Plants also contain additional RNA polymerases: Polymerase IV is involved in the transcription of microRNAs (miRNAs) (Herr et al., 2005; Onodera et al., 2005) whereas some organelles, such as chloroplasts, use plastid encoded polymerase (PEP) (Lysenko and Kuznetsov, 2005). In turn, mRNA is translocated outside of the nucleus into the cytoplasm and turned into proteins by the process of translation (reviewed in (Malys and McCarthy, 2011)). Together, the process of conversion from DNA into proteins through transcription then translation forms the central tenet of the “Central Dogma” (Crick, 1970).

Regulation of gene expression can be carried out by TFs binding either in the immediate vicinity of the TSS and interacting directly with the transcriptional machinery subunits, or in the upstream elements such as proximal promoters or enhancers, Figure 1.1b, however, not all promoters contain all identified elements. For example, some genes have very short intergenic regions leaving room only for the core promoter; alternatively, only 29% of genes in Arabidopsis contain a TATA-box within the core promoter region (Molina and Grotewold, 2005) and instead have unmethylated CG-rich regions (CpG islands) which have been found to bind TFIID proteins, in turn initiating the transcription process (Kim et al., 2005c). It has been proposed that transcription of the genes that are always required to be switched on (the so-called ”housekeeping” genes), is carried out through CpG islands, whereas TATA-box elements are required for expression of genes in response to signals and cues. Regulation of gene expression also occurs during the translation process, for example miRNAs and short interfering RNAs (siRNAs), which are usually 21-22 nucleotides long, function as regulators of translation by affecting the stability of the mRNA molecules in the cytoplasm as well as the rate of mRNA degradation



(a) Pathway of transcription initiation and re-initiation for RNA Polymerase II. Pol II, Mediator and TFIID, B, D, E, F and H subunits assemble around the TATA-box sequence in the core promoter region, immediately upstream from the TSS, to form the Pre initiation complex (PIC). With help from ATP, the complex opens and becomes unstable, initiating transcription. After the transcription process has been initiated, Pol II, TFIIB and TFIIF are released from the DNA, whilst Med TFIID, E and H remain bound to the promoter sequence allowing for re-recruitment of the missing factors. Once all subunits have assembled again into PIC transcription can be initiated once more. With permission from Macmillan Publishers Ltd, from Hahn (2004).



(b) General promoter structure consists of upstream enhancers and proximal promoter where TFs can bind followed by the core promoter region surrounding the transcription start site (TSS) with 5' UTR and protein coding region immediately downstream of the TSS. The core promoter contains IIB recognition element (BRE) adjacent to the TATA-box, initiator (Inr) site and downstream promoter element (DPE) (Adapted with permission from Smale and Kadonaga (2003)).

Figure 1.1: Overview of gene regulation and transcription mechanism.

(Valencia-Sanchez et al., 2006; Pillai et al., 2007; Standart and Jackson, 2007; Jackson and Standart, 2007; Nilsen, 2007). The combination of mRNA production and degradation rates defines gene expression.

TFs preferentially bind to a specific pattern of nucleotides known as binding motifs, or simply motifs. The length of binding sites varies greatly ranging from 5 to 30 nucleotides, with an average of 9.9 base pairs for eukaryotic genomes (Stewart et al., 2012). Specificity of the binding motifs comes from non-covalent interactions between the DBD found in TFs and side-chains in the major and minor grooves of the B-DNA double helix, which is the form predominantly found in living cells (Leslie et al., 1980). Chemical properties exhibited by different nucleotides contribute to sequence specific protein-DNA interactions, for example the pattern of hydrogen bond formation from GC nucleotides are central for recognition by the SP1 TF in humans (Letovsky and Dynan, 1989). However, some nucleotides share chemical properties and this can therefore give rise to several nucleotides being recognised by the same DBD, leading to degeneracy in the binding motifs, for example the human tumor suppressing p53 TF contains non specific nucleotides in the binding motif 5'-PuPuPuC(A/T)(T/A)GPyPyPy-3' (el Deiry et al., 1992).

Motifs describe TF binding site preferences by summarising different instances where such a binding site occurs. This is also called the consensus sequence and is frequently written in the form of a regular expression pattern and often only uses significantly conserved bases. However, a more detailed summary can be constructed when different instances are combined in a matrix form. The matrix describes how often each nucleotide occurs at each position summed across all instances of the binding site, forming an  $N \times 4$  matrix where  $N$  is the length of the binding site. Such a matrix is often referred to as position specific scoring matrix (PSSM) or weight matrix (WM). Furthermore, the same information can be described visually as a sequence logo, where the specificity of each nucleotide at each position is described by the height in terms of entropy, also known as information content (Schneider and Stephens, 1990). Additional benefits of describing binding sites using PSSMs is that statistical analysis techniques can be applied for discovering new motifs (e.g. using MeMe tool (Bailey and Elkan, 1994)) or for comparing the similarity between existing DNA sequences and PSSMs using Kullback-Leibler or Hellinger distance metrics. Alternatively, Hidden Markov Models (HMMs) can be constructed to incorporate possible dependency among the neighbouring basis, which subsequently improves estimating TF binding locations (Salama and Stekel,

2010).

### 1.2.1 TF families in *Arabidopsis*

As mentioned above, TFs play a crucial role in activating or silencing gene expression in response to the time of day or to a specific signalling molecule. It has been estimated that *Arabidopsis* contains approximately 1533 TFs (Riechmann et al., 2000), however more recent studies of TFs put the estimate between 1510 and 1922 (Xiong et al., 2005; Guo et al., 2005), the variation being largely due to the comparative methods used and definitions of unclassified TFs. These figures mainly break down into MYB (150) and MYB-related (49), AP2/EREBP (146), bHLH (127), C2H2 (134), NAC (107), MADS (104), bZIP (72) and WRKY (72) superfamilies of TFs (Guo et al., 2005). Each family has a unique role associated with it, for example TFs containing the MADS domain frequently function in plant and flower development (Rounsley et al., 1995) and functional redundancy exists within the family (Pelaz et al., 2000b). Alternatively, some superfamilies are involved in regulating the plant's response to a variety of different conditions and are often not functionally redundant. For example, WER and GL1 MYB TFs are functionally interchangeable but not functionally redundant, since they are expressed in different tissues (Lee and Schiefelbein, 2001).

#### **AP2/EREBP**

The first members of the family were identified from the homeotic gene APETALA2 (AP2) (Jofuku et al., 1994) and ethylene-responsive element binding proteins (EREBPs) in tobacco (Ohme-Takagi and Shinshi, 1995). AP2/EREBP TFs are characterised by the presence of one or more conserved AP2 DBDs within the protein coding sequence. Ethylene response factors (ERFs) form a clade within the AP2/EREBP superfamily and are expressed in response to the gaseous hormone ethylene (ET). ERF genes show a variety of stress related expression patterns and are regulated by disease-related stimuli such as ET, salicylic acid (SA), jasmonic acid (JA) and infection by a virulent pathogen as has been shown for several genes, e.g. *ERF1*, *Pti4* and *AtERF1* (Chen et al., 2002; Brown et al., 2003; Gu et al., 2000; Oñate-Sánchez and Singh, 2002). ERFs have also been shown to be activated by wounding of the plant (Cheong et al., 2002). Other family members, such as dehydration-responsive element binding (DREB) and C-repeat binding factor (CBF) are expressed in drought and cold conditions respectively (Sakuma et al.,

2002). Although many ERFs are thought to be transcriptional activators, *ERF3* and *ERF4* have been shown to repress the expression of their target genes (Fujimoto et al., 2000). Interestingly, those two genes were also found to be activated by ET, JA and an incompatible pathogen infection (Brown et al., 2003).

### bHLH

The basic helix-loop-helix (bHLH) TF family is one of the largest in Arabidopsis. The bHLH family is defined by its conserved structure domain, which consists of ~ 60 amino acids (aa) with two functionally distinct regions. The basic region, located at the N-terminal end of the domain, is involved in DNA binding and consists of ~ 15 aa with a high number of basic residues. The HLH region at the C-terminal end, functions as a dimerisation domain (Murre et al., 1989; Ferré-D'Amaré et al., 1994) and is composed mainly of hydrophobic residues that form two amphipathic  $\alpha$ -helices separated by a loop region of variable sequence and length (Nair and Burley, 2000). Crystal structure analysis of bHLH TFs has revealed a unique binding technique, consisting of an interaction of the homo- or heterodimers, formed by two bHLH proteins, leading to the formation of the DNA binding domain, where the basic region of each TF binds to half of the binding motif (Ma et al., 1994; Shimizu et al., 1997). Subsequently, bHLH TFs preferentially bind to an E-box motif (5'-CANNTG-3'), although a more common variation of the same motif is the palindromic G-box (5'-CACGTG-3') (Menkens et al., 1995). Certain residues within the basic region serve to recognise the core binding motif, whereas other residues dictate the specificity for the type of E-box recognised (Robinson et al., 2000). Members of the bHLH family include phytochrome interacting factors 3 (PIF3), which have been found to regulate circadian rhythms by interacting with PhyB and binding to G-box motifs found in the promoters of *RBCS-1A*, CCA1, LHY, and SPA1 (Martínez-García et al., 2000). PIF4 is closely related to PIF3 (Ni et al., 1998) and can also bind PhyB and G-box sequences, but not simultaneously (Huq and Quail, 2002). bHLH TFs also play a role in stress response, for example mutants of the *ATR2* gene have been found to have the stress markers *PDF1.2* and *LOX1* up-regulated suggesting a role in the JA signalling pathway (Smolen et al., 2002).

### bZIP

Basic leucine zipper (bZIP) TFs are characterised by the presence of a basic region followed by the leucine zipper domain. The domain structure is somewhat

similar to that of bHLH TFs in principle and like bHLH TFs, bZIPs also bind the C-box (5'-GACGTC-3'), G-box (5'-CACGTG-3') and A-box (5'-TACGTA-3') sequences (Martínez-García et al., 1998). Moreover, members of the two families of TFs has been found to act antagonistically. PIF1/PIF3 and HY5/HYH function antagonistically during the seedling greening process and in the production of reactive oxygen species (ROS), highlighting their role in adapting to changing light conditions (Chen et al., 2013). Factors interacting specifically with or insensitive to abscisic acid (ABA) form a subgroup within the bZIP superfamily of TFs (reviewed in (Jakoby et al., 2002)). AREB1, AREB2 and AREB3 have been shown to regulate *RD29B* in response to drought and high salinity (Uno et al., 2000), whilst TGA2 and TGA3 regulate the expression of *PR1* (Johnson et al., 2003). TGA2, TGA5 and TG6 TFs are differentially expressed as regulators of the detoxification process in plants (Mueller et al., 2008).

## WRKY

For a long time WRKY TFs had been thought to be present only in plants, however recent studies have isolated several WRKY proteins in non-plant eukaryotes, including *Giardialamblia* and the slime mould *Dictyostelium discoideum* (reviewed in (Zhang and Wang, 2005)). The name of the WRKY family is derived from the most prominent feature of their domain, a conserved WRKYGQK aa sequence within the total 60 aa domain, which is often followed by a C2H2 or C2HC zinc binding motif (Eulgem et al., 2000). WRKY TFs exhibit preferential binding to the W-box motif (5'-(T)TGAC(C/T)-3') (Rushton et al., 1996). Although, all WRKY TFs recognise the core TGAC sequence, some WRKYs, for example WRKY11, show high sensitivity to subtle changes in the W-box and the nucleotides immediately outside of it (Rushton et al., 1996). Previous reports suggest that WRKY TFs are involved in the regulation of genes containing a W-box and are often associated with the defence response (reviewed in (Rushton and Somssich, 1998; Yang et al., 1999)). For example, triple knock-outs of *wrky18wrky40wrky60* resulted in a reduction of bacterial growth when infected with *Pseudomonas syringae*, however the mutants were much more susceptible to infection with Botrytis (Xu et al., 2006). The SA induced expression levels of *PR1* gene were negatively correlated with the extent of resistance to the fungal pathogen among the wild type and WRKY triple mutants. Conversely, expression levels of *PDF1.2*, induced by JA signalling, were positively correlated with resistance to Botrytis. WRKY18, 40 and 60 have been found to interact in yeast-two hybrid experiments suggesting a differential and complex behaviour in response to different types of biotic stress. WRKY TFs have also been

shown to be targeted directly by the mitogen-activated protein kinase (MAPK), which activates camalexin biosynthetic genes following infection with a pathogen (Ren et al., 2008). *WRKY33* is targeted by MAPK3 and MAPK6 (Mao et al., 2011), which in turn activates expression of *PAD3* in *Arabidopsis* (Qiu et al., 2008).

## MYB

The largest superfamily of TFs is MYB, which can be classified into three subfamilies according to the number of adjacent repeats in the MYB domain (one, two or three) (Jin and Martin, 1999; Rosinski and Atchley, 1998). MYB-like proteins with one repeat are referred to as “MYB1R” factors, with two repeats as “R2R3-type MYB” factors and with three repeats as “MYB3R” factors. The MYB-like proteins with a single repeat (or occasionally just a partial repeat) are fairly divergent and include factors that bind the consensus sequence of plant telomeric DNA (5'-TTTAGGG-3') (Yu et al., 2000). It has also been shown that MYB1R factors (e.g. MYBST1 or StMYB1R1) can act as transcriptional activators (Baranowskij et al., 1994) and some are associated closely with the activity of the circadian clock (e.g. CIRCADIAN CLOCK ASSOCIATED 1 (CCA1) and LATE ELONGATED HYPOCOTYL (LHY)) (Schaffer et al., 2001). CCA1 and LHY1 bind to DNA, indicating that they might act by modulating transcription (Wang et al., 1997; Schaffer et al., 1998). Additionally, some MYB TFs play a critical role in the response to pathogen infection. For example, MYB108 (also known as BOS1) knock-out plants are significantly more susceptible to infection with *Botrytis* and *Alternaria brassicicola* (Mengiste et al., 2003). BOS2, BOS3 and BOS4 have similar effects on susceptibility to biotic stress and all BOS loci genes are thought to function as part of the ethylene and jasmonate signalling pathways (Mengiste et al., 2003; Veronese et al., 2004).

### 1.3 Gene regulatory networks

Regulation of gene expression is a fundamental process by which an organism controls appropriate spatial and temporal expression of genes. Typically, the nodes of regulatory networks are functional genes such as structural proteins and enzymes whereas TFs are largely responsible for the overall topology of the network. Transcriptional gene regulatory networks (GRNs) have been found to largely follow power-law, whereby a few nodes are associated with a large number of connections and visa versa (Guelzim et al., 2002; Teichmann and Babu, 2004). In other words, the degree of the node is the number of edges associated with it. Such networks

resemble “scale-free” networks which are characterised by an abundance of nodes with small-degree but the frequency of high-degree nodes decreases relatively slowly. Thus, nodes that have degrees much higher than average, also called ‘hubs’, exist. Because of the heterogeneity of “scale-free” networks, random node disruptions do not lead to a major loss of connectivity, but the loss of the hubs causes the breakdown of the network into isolated clusters (Albert and Barabási, 2002). Various experimental data also suggests that TFs are the hubs in GRNs (Blais and Dynlacht, 2005). Consistent with the notion of hub importance in the GRNs, changes to parts of the network have been shown to have dramatic implications on the overall developmental of embryos (Davidson and Erwin, 2006). Moreover, the development of different cell types derived from the same plant stem cell is heavily dependent on the overall architecture of the GRN within an organism (Espinosa-Soto et al., 2004). Temporal changes in gene expression patterns have been found to control stress-related (Breeze et al., 2011), seasonal (Aikawa et al., 2010) and circadian (Locke et al., 2006) changes in *Arabidopsis*. It has also been suggested that changes in GRN are largely responsible for significant speciation events (Chen and Rajewsky, 2007) as well as loss or gain of certain traits (Crombach and Hogeweg, 2008), where previously point mutations were thought to be solely responsible for the altered protein function of a single gene. Although there are examples where point mutation in a gene produced significant phenotypes for example a single mutation in the equine *DMTR3* gene is a prerequisite for lateral gait (Andersson et al., 2012). However, changes to the GRN of an organism appear to be more beneficial for increasing robustness from an evolutionary prospective (Crombach and Hogeweg, 2008).

### 1.3.1 Types of gene regulatory networks

GRNs are often presented as graphs or qualitative models summarising experimental findings. One of the most studied systems in plants is the function of the circadian clock which is presented qualitatively in Figure 1.2. However, one of the shortcomings of qualitative models such as this is a lack of key details, for example the model presented in Figure 1.2 does not describe the timings associated with the expression of the genes involved in this system, nor that the whole cycle repeats every 24 h. These dynamic details arise from the quantitative models and the parameters associated with them. By modelling the predicted expression levels of the genes in the model, complex hypotheses and new models can be formulated to include previously available data and subsequently verified using new experimental data.

However, using mRNA expression levels to build GRNs remains a challeng-

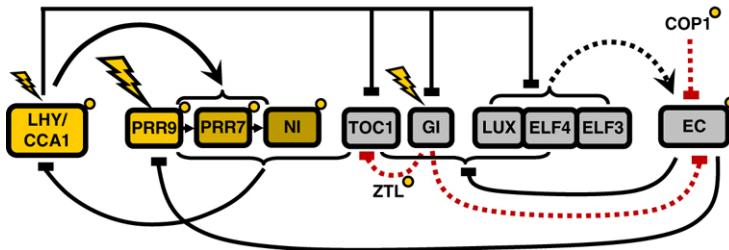


Figure 1.2: The revised outline of the *Arabidopsis* circadian clock. Elements of the morning and evening loops are shown in yellow and grey, respectively. Proteins are shown only for EC, ZTL and COP1 for simplicity. Transcriptional regulation is shown by solid lines. EC protein complex formation is denoted by a dashed black line. Post-translational regulation of TOC1 and the EC by GI, ZTL and COP1 are shown by red dashed lines. Acute light responses in gene transcription are shown by flashes. Post-translational regulation by light is shown by small yellow circles. Reproduced with permission from the Nature Publishing Group from Pokhilko et al. (2012).

ing and non-trivial process. As discussed above, the levels of mRNA are a function of transcription and degradation rates. Only the former is dependent on the TF activity and little is known about degradation rates, which are specific for the gene of interest. In turn, protein production from mRNAs is subject to post-transcriptional control. Protein themselves may require activation prior to being functionally active, adding further layers of dependency and regulation for the transcription of target genes in the case of TFs. Genes are often controlled by more than one TF, which leads to a combinatorial explosion when gene expression levels are modelled quantitatively. Furthermore, increased complexity resulting from the combinatorial variation of gene expression influences the number of possible genes to model simultaneously. The more detailed the model, the more computing power is required, leading to fewer genes that could be modelled with high precision. Altogether, modelling approaches rely on the availability of high quality genome-wide data, preferably with high-resolution temporal profiles, in order to establish causal relations between TFs and their target genes.

### 1.3.2 Strategies for uncovering gene regulatory networks

Alternative to modelling gene expression and GRNs arising from these *in silico*, are experimental approaches which can be classified as “TF-centred” and “gene-centred”. A TF-centred approach focuses on identifying all possible downstream genome-wide targets of the TF. Alternatively, a gene-centred approach focuses on

identifying all TFs that are able to interact with the ncDNA regions of the gene. In practise, both approaches are used to construct comprehensive GRNs.

Chromatin immunoprecipitation (ChIP) assays are a widely used and powerful tool to identify hundreds of genome-wide TF binding locations (Morohashi et al., 2009). ChIP works by chemically cross-linking a TF to the genomic DNA with formaldehyde, followed by immunoprecipitation of the TF together with the associated DNA fragments. The location of the precipitated fragments is either determined by massively parallel sequencing and subsequent alignment to the reference genome (ChIP-Seq) or by using microarrays (ChIP-ChIP) (Kaufmann et al., 2010). Although ChIP techniques allow the identification of hundreds of targets across a genome, one of the major drawbacks is the reliance on antibodies to immunoprecipitate the TF with bound fragments. A high degree of sequence homology amongst plant TFs makes it difficult to identify an antibody specific for an individual TF. Furthermore, the TF of interest may be expressed at low levels in the host cells and therefore it may be difficult to isolate a sufficient amount of chromatin.

An alternative approach to identify GRNs from a “TF-centred” prospective, is to first establish the binding motif of a TF, or to select existing motifs available from online databases such as PLACE, JASPAR or TRANSFAC (Higo et al., 1999; Bryne et al., 2008; Wingender et al., 2000) and then scan the genome for occurrences of this motif. If the motif is present in the promoter sequence of a gene, it forms a link between the TF and the associated gene in the GRN (Walhout, 2006). A number of techniques are available for determining the DNA sequence bound by a TF including EMSA, DNase assays and protein-binding microarrays (Fried and Crothers, 1981; Galas and Schmitz, 1978; Godoy et al., 2011), although many TFs do not have a binding motif associated with them. To improve the predictions made using this method, the regions scanned can be restricted to core promoters and/or known open chromatin areas previously determined by DNase hypersensitivity assays (Boyle et al., 2008). However, predictions on the regulation of a particular gene by a specific TF are coarse and require further validation using conventional *in planta* or *in vivo* techniques.

In the gene-centred view, the yeast one-hybrid (Y1H) system is one of the most popular for identification of direct protein-DNA interactions (Meijer et al., 1998). This approach has been used successfully to isolate many plant TFs that directly interact with regulatory DNA sequences (Tran et al., 2007; Chen et al.,

2010; Zhu et al., 2010). Furthermore, the simplicity of the method allows not only members of the same TF family to be tested for interaction with the DNA sequence, but also multiple TFs can be tested simultaneously, providing a comprehensive view of gene-centred regulatory networks.

## 1.4 Stress response in Arabidopsis

Unlike animals, plants lack a mobile immune system to defend themselves against hostile stimuli. Faced with a threat from bacteria, pathogen or fungi, plants rely on a number of defence response mechanisms, which can be broadly characterised as PAMP Triggered Immunity (PTI) and Effector Triggered Immunity (ETI).

### 1.4.1 PAMP Triggered Immunity

Extracellular receptor like kinases (RLKs) typically contain a signalling sequence, a transmembrane region and a C-terminal domain with eukaryotic protein kinase signatures, similar to the animal receptor tyrosine kinase (van der Geer et al., 1994). As such, RLKs perceive pathogen-associated molecular patterns (PAMPs) through their extracellular domains and propagate the signal into the cytoplasm, where they are carried through by MAP kinases finally activating transcription of certain pathogen-response genes, leading to PAMP triggered immunity (PTI). One such RLK found in Arabidopsis is FLS2, containing a leucine repeat rich (LRR) region, which recognises the short, 22 aa, polypeptide (flg22) corresponding to the highly conserved bacterial flagellin amino terminus (Gómez-Gómez and Boller, 2000). flg22 has been found to activate host receptors (Felix et al., 1999) and lead to the defence response (Gómez-Gómez and Boller, 2000). *fls2* mutant plants have been found to be flagellin insensitive (Gómez-Gómez and Boller, 2000). A further link was established between flagellin perception and restriction of pathogen growth (Zipfel and Felix, 2005). flg22-induced PTI to *Botrytis* requires *BOTRYTIS INDUCED KINASE1 (BIK1)*, an RLK found in Arabidopsis (Laluk et al., 2011). It is likely that there are more LRR-RLKs with recognition of specific pathogens yet to be discovered.

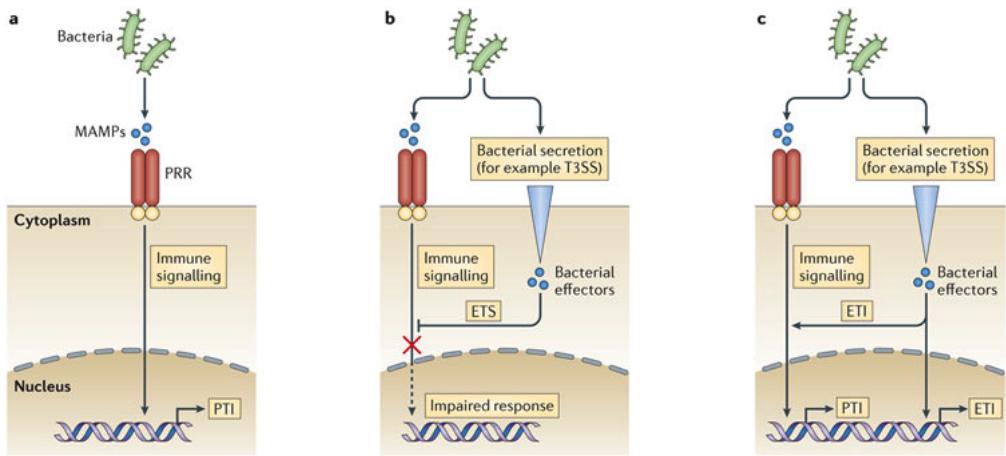


Figure 1.3: Overview of plant defence response mechanisms: a) extracellular signals are sensed by pathogen recognition receptors (PRRs) and signal is carried inside cytoplasm to activate PAMP Triggered Immunity (PTI), b) effector proteins released by type III secretion system (T3SS) block signal transduction inside cytoplasm, inhibiting defence response, c) effector proteins are sensed by the NB-LRR containing proteins inside the cell leading to transcription of defence genes, Effector Triggered Immunity (ETI). Adapted from Stuart et al. (2013) with permission of Nature Publishing Group.

### 1.4.2 Effector Triggered Immunity

However, ever changing adaptation of viruses, bacteria and pathogens through natural selection has lead to the successful development of means to suppress the PTI response in plants. Effector proteins are released by the invader in order to circumvent PTI response. An archetypal example of such a pathway is the *Arabidopsis RIN4* protein. *RIN4* is a known regulator of defence responsive genes in *Arabidopsis* against such pathogens as *Pseudomonas syringae* (Mackey et al., 2002). *P. syringae*, on the other hand, specifically targets *RIN4* in at least two different ways for inactivation by phosphorylation using AvrRpm1 and AvrB (Mackey et al., 2002) and cleavage by AvrRpt2 (Kim et al., 2005a). Both effectors are part of the type III secretion system (T3SS). However, this inactivation by phosphorylation and cleavage is detected by RPS2 and RPM1, which in turn activate the defence response in *Arabidopsis* (Kim et al., 2005b). Both RPS2 and RPM1 are nucleotide binding-leucine rich repeat (NB-LRR) containing proteins. In general, many of the *Arabidopsis* R genes that are part of the plant's sensory network include a NB-LRR region (Dangl and Jones, 2001).

### 1.4.3 The role of hormones in stress response

Phytohormones are small molecules that are essential for the regulation of plant growth, development, reproduction and survival. They act as signalling molecules mediating signals in PTI and ETI responses and occur in low concentrations. The plant is subjected to many abiotic stress conditions such as drought, cold and high light as well as being a target for numerous biotic invaders such as insects, pathogens and fungus. Biotic stress comes in two different types, biotrophic and necrotrophic. The former creates a symbiotic relationship with the plant, feeding on the available plant nutrients. Common examples come from the *Agrobacterium* genus (recently split into multiple genera including *Rhizobium* genus (Young et al., 2001)) and include *A. tumefaciens* which causes crown-gall disease in plants (Chilton et al., 1977), as well as *Rhizobium leguminosarum*) that forms a positive symbiotic relation with the roots of legumes and helps to fixate nitrogen. In contrast, necrotrophs kill cells and feed on the dead tissue in order to grow, for example *Botrytis cinerea* (Botrytis) that causes grey mould. Such a diverse range of potential threats means that plants have to correctly identify each type of threat. Arabidopsis, for example, uses programmed cell death (PCD) in order to stop infection from biotrophic pathogens (reviewed in (Greenberg, 1997)), such as *P. syringae*, which in turn uses the AvrP-toB effector to inhibit the PCD response (Abramovitch et al., 2003). However, PCD increases the susceptibility of Arabidopsis to infection with Botrytis, which exploits PCD for increased pathogenicity (Govrin and Levine, 2000). Therefore, it is vital that the plant responds in the correct manner in order to fight infection. One way that the plant perceives the type of stress is through hormones such as abscisic acid (ABA), ethylene (ET), jasmonic acid (JA), salicylic acid (SA), auxin, gibberellic acid (GA), cytokinin (CK) and brassinosteroids (BR) (reviewed in (Bari and Jones, 2009; Pieterse et al., 2009)).

#### Abscisic Acid

ABA is one of the key phytohormones involved in the signalling pathways in response to both biotic and abiotic stress, as well as integrating developmental queues. ABAinsensitive *abi4* mutants, for example, have pleiotropic defects in seed development, including decreased sensitivity to ABA, inhibition of germination and altered seed-specific gene expression (Finkelstein et al., 1998) and ABA-insensitive *abi1* displays a stunted phenotype (Barrero et al., 2005). Originally ABA was linked with water deficit, since after a 30 minute lag, ABA levels were seen to increase approx-

imately 100-fold in dehydrated plants (Guerrero and Mullet, 1986). Furthermore, ABA acts in stomatal closure by re-organising actin from a radial pattern into a random and short-oriented pattern (Eun and Lee, 1997). Closure of stomata helps to prevent transpirational water loss through the stomatal pores, making plants more drought resistant (Schroeder et al., 2001). ABA levels accumulate in senescing leaves suggesting that ABA plays a role in the induction of senescence in *Arabidopsis* (Breeze et al., 2011). Consistent with this hypothesis, exogenous application of ABA also induces premature senescence (Gepstein and Thimann, 1980). Expression profiling studies have found that many ABA inducible genes are also upregulated in senescence, which suggests further induction of abiotic stress signalling pathways (Buchanan-Wollaston et al., 2005; van der Graaff et al., 2006). Although ABA has been known to be primarily an abiotic hormone, there has been mounting evidence to suggest that it is also involved in the biotic stress response (reviewed in (Ton et al., 2009)) where it may play both a positive and negative role (reviewed in (Asselbergh et al., 2008)). Initially, ABA induced stomatal closure helps to protect the leaf against the spread of pathogen infection (Melotto et al., 2006), whilst accumulation of ABA during the infection process disrupts the defence response modulated by other hormones, such as ET or JA, resulting in increased susceptibility to *P. syringae* (de Torres-Zabala et al., 2007).

In addition to functional proteins, ABA also induces the expression of regulatory proteins. Genes induced by the ABA signalling cascades have been found to contain the ABA-responsive element (ABRE), which has been shown to be necessary and sufficient for transcriptional activation of those genes in the presence of elevated levels of ABA (Choi et al., 2000). The ABRE consensus motif (C/T)ACGTGGC is able to interact with bZIP TFs known as ABRE-binding factors (ABFs). bZIP is a large family of TFs in *Arabidopsis* and all ABFs are part of the same phylogenetic clade (group A) within the bZIP family (Jakoby et al., 2002). ABF2, ABF3 and ABF4 have been shown to be master regulators of the ABA response through the ABRE motif and activate gene expression under abiotic stress (Yoshida et al., 2010). As well as ABRE motifs, many ABA-regulated genes also contain binding sites for other TFs. For example, binding sites for MYC2 and MYB2 TFs have been reported to have functional roles in ABA mediated induction of stress related genes such as *RD22* and *ADH1* (Abe et al., 2003). The dehydration response element (DRE) has been found to act in concert with the ABRE motif to positively regulate ABA-mediated response to abiotic stress (Narusaka et al., 2003). Additionally, coupling element 1 (CE1) has been shown to act together with ABRE in the regulation

of *HVA22* in Arabidopsis, however TFs binding to CE1 were not known when this research was carried out (Shen and Ho, 1995). Lee et al. (2010) have recently shown that ET responsive TFs, such as *ERF13*, are able to bind the CE1 element, suggesting crosstalk between ABA and ET through ABRE and CE1 motifs respectively.

### Ethylene

ET is a gaseous hormone that regulates a variety of plant growth stages and development, for example ET positively regulates root hair development (Tanimoto et al., 1995). However, ET also plays one of the key roles as a signalling molecule in response to wounding, dehydration, cold and salt stress (Morgan and Drew, 1997). Ethylene receptors such as ETR1, ETR2, ERS1, ERS2 and EIN4 located on the endoplasmic reticulum (ER) membrane, maintain the constitutive triple response 1 (CTR1) protein in an active form, inhibiting any further downstream components such as ethylene insensitive 3 (EIN3) which is constantly degraded by EIN3 binding F-box (EBF1) and EBF2 via the proteasome-mediated degradation pathway. In the presence of ET, CTR1 is inactivated and EIN2, an ER localised protein, is not phosphorylated by CTR1 allowing it to interact with EBF1 and EBF2 which in turn prevents them from degrading EIN3 TF. EIN3 activates transcription of a variety of ethylene response factor (ERF) genes, for example *ERF1* (Solano et al., 1998). These TFs regulate the expression of genes that encode stress-related proteins. Generally ERF proteins have been found to contain the AP2 DBD, which targets the GCC-box (5’-(A/G)CCGCC-3’) motif present in many stress-related genes ref(Okumaro). Studies of several ERF genes have shown loss-of-function plants to be more susceptible to biotic stress (Oñate-Sánchez and Singh, 2002; Lorenzo et al., 2003), suggesting an important role for ET in Arabidopsis stress response. ET also has a major role during leaf senescence as *ein2* mutants show delayed senescence (Oh et al., 1997). ET biosynthesis genes are up-regulated during the senescence process and the hormone is a major positive regulator of leaf senescence, levels of which rise during senescence (van der Graaff et al., 2006).

### Jasmonic Acid

JA mediates signalling associated with the wound response and regulates downstream elements in response to infection with Botrytis, amongst other biotic invaders (Reymond et al., 2000), through the action of various jasmonate Zim domain

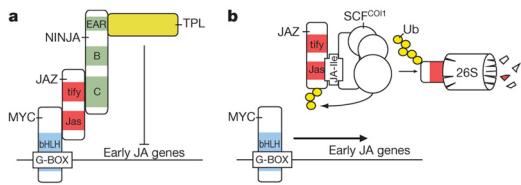


Figure 1.4: An overview of JA signalling mechanisms. a) In the absence of JA, basic helixloophelix (bHLH) MYC factors interact with the Jas domain of JAZ proteins that interact through their TIFY motif with domain C of NINJA. The EAR motif of NINJA is essential for interaction with the TPL co-repressors. b) In the presence of Jasmonoyl-isoleucine, JAZ proteins interact with the ubiquitin ligase SCF<sup>COI1</sup>, leading to proteosomal JAZ degradation and subsequent release of the NINJATPL complex from the MYC factor and activation of JA-responsive gene expression (with permission from Nature Publishing Groups, (Pauwels and Goossens, 2008)).

(JAZ) proteins. JAZ proteins generally act as transcriptional repressors by interacting with a TF such as MYC2 through the Jas domain, additionally interacting with NINJA and TPLS proteins to repress genes normally regulated by MYC2. In the presence of JA, JAZ proteins interact with the ubiquitin ligase SCF<sup>COI1</sup>, leading to proteosomal JAZ degradation and subsequent release of the NINJATPL complex from the MYC factor and activation of jasmonate-responsive gene expression, Figure 1.4. JAZ proteins are able to interact with a variety of different TFs and therefore have multiple sequence motifs associated with this process. For example, MYC proteins usually consist of a bHLH domain architecture that binds a G-box motif, 5'-CACGTG-3'. JAZ proteins also interact with AP2 domain containing proteins such as ERF1, ERF2, ORA47 and ORA59, which act through the GCC-box motif. The JA induced TFs regulate multiple classes of genes which encode proteins that function in the plant defence response. Whole genome expression profiling studies have revealed that JA induces the expression of genes involved in the production of stress-associated metabolites including glucosinolates, phenylpropanoids and anthocyanins (Sasaki-Sekimoto et al., 2005; Pauwels and Goossens, 2008). JA also plays an active role in controlling cell growth and proliferation through the repression of cell cycle genes (Pauwels and Goossens, 2008). In addition to regulating biotic stress response pathways, JA has also been implicated in regulating responses to abiotic stresses such as high salinity and osmotic stress (Xu et al., 1994; Lehmann et al., 1995). Microarray studies have identified JA as being functionally important in senescence since levels of JA accumulate in senescing leaves (Breeze et al., 2011).

## **Salicilic Acid**

Although SA has been reported to have a functional role in senescence, plant development and photosynthesis (Morris et al., 2000; Rivas-San Vicente and Plasencia, 2011), it is more often identified as a key player in response to biotic stress. Specifically, defence against biotrophic pathogens, in contrast to JA which regulates genes in response to necrotrophic invasions. SA functions through the *NPR1* gene. Once activated by SA, *NPR1* is translocated into the nucleus where it acts as a co-activator of SA responsive genes. For example, *NPR1* interacts with TGA2 and TGA3 enhancing their effect on the transcription of *pathogenesis related 1 (PR1)* (Johnson et al., 2003; Dong, 2004; Spoel et al., 2009).

### **1.4.4 Hormone crosstalk fine-tunes the defence response in *Arabidopsis***

However, response to biotic stress is not subject to the action of a single hormone or signalling pathway, instead the response is fine-tuned by the balanced action of all of the hormones. The roles of SA, JA and ET as dominant local and systemic induced defence signalling hormones has been well documented (Loake and Grant, 2007; Pozo et al., 2004; van Loon et al., 2006). JA-ET are often found to be signalling in synergy together. For example, activation of the *Arabidopsis* defence gene PDF1.2 requires both ET and JA signalling components (Penninckx et al., 1996). So far, two members of the plant specific AP2/ERF family of TFs, ERF1 and ORA59, has been found to be principal integrators of the JA and ET signalling network (Lorenzo et al., 2003; Pre et al., 2008). However, JA alone negatively regulates activity of PDF1.2 in a MYC2 dependent manner (Lorenzo et al., 2004). This disparity allows separate branches of the plant defence response to be activated in the presence of JA, ET or both. Alternatively, SA represses the JA and ET induced expression of PDF1.2 through SA-dependent expression of GRX480 (Ndamukong et al., 2007) and WRKY70 TF (Li et al., 2004), highlighting the difference in defence response mechanisms between biotrophs and necrotrophs, Figure 1.5. Taken together, this suggests a highly interconnected defence signalling and response network exists in plants (Katagiri, 2004).

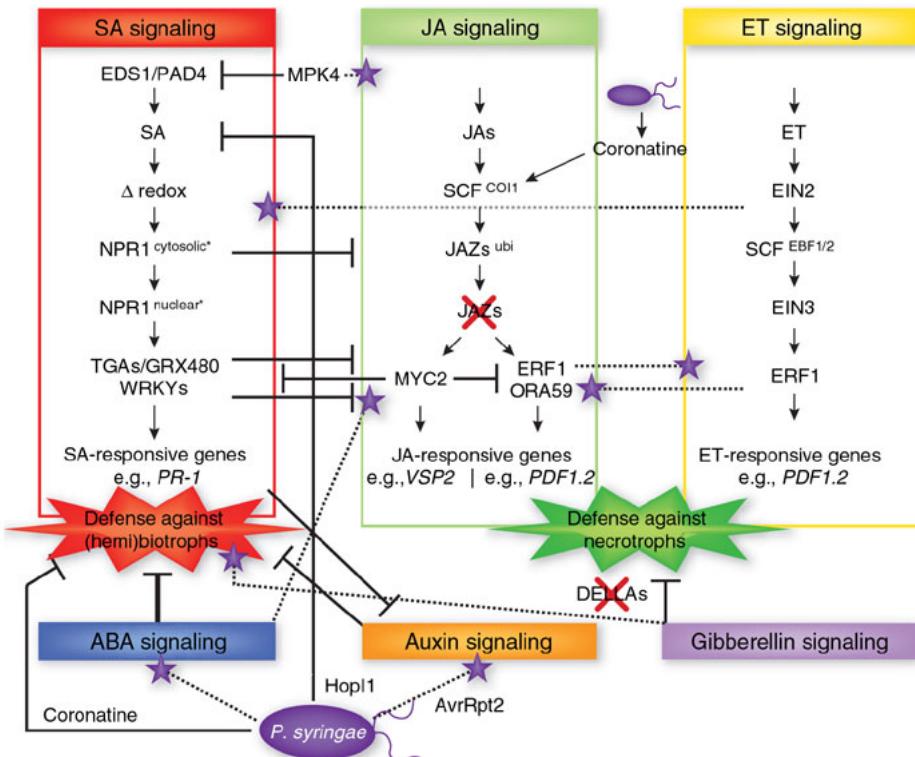


Figure 1.5: Networking by phytohormones in the plant immune response. Cross-communication between hormone signalling pathways provides the plant with a large regulatory capacity that may tailor its defence response to different types of attackers. On the other hand, pathogens such as *P. syringae* produce effector proteins (for example, coronatine, HopI1 and AvrRpt2) that manipulate the signalling network to suppress host immune responses and promote virulence. The SA, JA and ET signalling pathways represent the backbone of the defence signalling network, with other hormonal signalling pathways feeding into it. Only those signal transduction components that are relevant to this review are shown. , negative effect; purple stars, positive effect. With permission from the Nature Publishing group, Pieterse et al. (2009).

## 1.5 Infection by Botrytis

Botrytis is a necrotrophic plant fungus that infects a variety of economically important crops including grapes, vegetables, berries and stone fruit (Williamson et al., 2007; Jarvis, 1977). Botrytis causes grey mould formation and growth on the infected plants, substantially reducing crop yields (Elad et al., 2007b; Williamson et al., 2007). However, Botrytis also infects the model plant *Arabidopsis* (Koch and Slusarenko, 1990). *Arabidopsis* is related to a number of other plant species that also serve as a host for Botrytis, suggesting that findings in *Arabidopsis* can be applied directly to other species with the help of genetic modifications (GMs) and breeding.

### 1.5.1 Botrytis Infection process

Botrytis infects host species using spores/conidia, which germinate on the exterior of leaves or flowers and form appressoria (van Kan, 2006), an infection structure that differentiates on the surface and forms a penetration peg that breaches the cuticle. The timing of the germination and penetration of Botrytis spores/conidia is not known, however conidia has been shown to germinate within six hours in water (Hawker and Hendy, 1963). Botrytis experiences a fast growth phase during spore/conidia germination and hyphae formation, followed by a lag phase 20-28 hours post infection (hpi). During the lag phase, dark lesions on the surface of the tissue are formed, corresponding to penetration of the plant cells by appressoria. The lesions grow in size and biomass by consuming nutrients available from the host plant (Hancock and Lorbeer, 1963).

### 1.5.2 Changes in *Arabidopsis* transcriptome in response to Botrytis

As a result of invasion by a bacteria or fungus, *Arabidopsis* undergoes drastic changes to its transcriptome (Tao et al., 2003; Windram et al., 2012). A variety of different defence responsive genes have previously been identified. For example, PATHOGENESIS RELATED 1 (PR1) is induced by SA and is a marker of systemic acquired resistance (SAR) in *Arabidopsis* (Cao et al., 1994). SAR activated genes often encode proteins with antimicrobial activity, such as PR1 (Van Loon and Van Strien, 1999). Activation of these genes has been shown to be dependent on NPR1 and TGA TFs (Kesarwani et al., 2007). Similar to PR1 genes, phytoalexins are low molecular weight antimicrobial metabolites produced by plants in response to pathogen

attack (Paxton, 1981). *PHYTOALEXIN DEFICIENT 3* (PAD3) gene encodes a cytochrome P450 monooxygenase, which is required for the biosynthesis of camalexin, a type of phytoalexin (Zhou et al., 1999). Plant defencin proteins are believed to contribute to the defence arsenal of plants directed against microbial phytopathogens. As such, detailed expression analysis of the *PLANT DEFENCIN 1.2* (PDF1.2) gene has demonstrated that it is expressed in *Arabidopsis* leaves upon fungal attack, not only locally in inoculated leaves, but also systemically in uninoculated leaves of pathogen infected plants. In contrast to most systemically pathogen-induced genes described so far, *PDF1.2* is not activated by exogenous salicylic acid (Penninckx et al., 1996), instead it responds to methyl jasmonate (Manners et al., 1998). In addition, many more TFs have also been identified as playing an important role in the regulation of the plants susceptibility to Botrytis: *MYC2*, *ANAC019*, *ANAC055*, *ANAC092*, *TGA3*, *EIN3*, *ERF1*, *MYB46*, *MYB108*, *ZFAR1*, *WRKY70*, *WRKY33*, *ORA59*, *CAMTA3* and *ATAF1*. Furthermore, *ANAC019*, *ATAF1*, *ERF1*, *MYB108*, *MYC2*, *WRKY70* and *ZFAR1* have been shown to be differentially expressed in other microarray experiments (Lorenzo et al., 2004; Bu et al., 2008; Windram et al., 2012; Zhu et al., 2011; Berrocal-Lobo et al., 2002; Ramírez et al., 2011; Mengiste et al., 2003; AbuQamar et al., 2006; Zheng et al., 2006; Pre et al., 2008; Galon et al., 2008; Wang et al., 2009). These TFs combine together to form a complex regulatory network controlling the plant's response to a variety of stresses and infection with Botrytis in particular. TFs often regulate more than one gene at a time and therefore identifying further regulatory targets would add to the comprehensive picture of the defence regulatory network, additionally aiding the modelling approaches of stress GRNs.

## 1.6 Aims and objectives

One of the key paradigm shifts in recent years came from controversial conclusions in the ENCODE project about the nature of “junk DNA” (ENCODE Project Consortium et al., 2012), and that regulation of gene expression plays a bigger role in the large observed variation in phenotypes and responses than previously thought. Conventional gene knock-outs and subsequent phenotypic analysis has been partially successful in identifying major players in plants’ defence response. However, regulatory reprogramming which *Arabidopsis* undergoes in response to Botrytis infection is not well understood and therefore this work aims to provide new information about gene regulatory networks and their role in Botrytis infection. Firstly, transcriptional

regulation of genes often takes place in the non-coding DNA (ncDNA) regions of a genome. Therefore, the aim is to identify potentially functional regions occurring in ncDNA sequences that are also conserved among closely related dicotyledonous plants. Using this information together with the expression of genes in response to Botrytis, a gene regulatory network of a select subset of genes will be constructed using Y1H library screens. This high-throughput technique allows a comprehensive GRN to be built from the ground up. The promoter fragments of genes tested in the Y1H experiments will serve as inputs to a bioinformatics pipeline to identify specific TF binding motifs, which will be further validated in the context of Y1H screens. Finally, plants with knockouts in TFs found to be interacting in the Y1H screens, will be tested for susceptibility to infection with Botrytis, establishing their role in the response to biotic stress.

## 1.7 Organisation of this thesis

Understanding the regulatory mechanisms underlying stress response in *Arabidopsis* helps inform a better understanding of gene regulatory networks. In addition, identification of key regulatory TFs in response to infection with Botrytis would lead to novel ways of modifying plant genomes in order to decrease susceptibility to infection with this necrotrophic fungus. This thesis presents an interdisciplinary investigation into the regulatory code responsible for controlling transcription among related plant species and in response to infection with Botrytis. Following from the introductory section, chapter 2 focuses on identifying conserved non-coding sequences (CNSs) and presents evidence that newly founded CNSs correspond to functional areas in promoters of the corresponding genes and has potential for multiple TF binding sites. In chapter 3, a select promoter set of genes differentially expressed in response to Botrytis, some containing newly found CNSs, are tested to uncover common TFs binding using a high throughput Y1H library screen. Furthermore, newly found protein-DNA interactions are validated using pairwise Y1H screens. Interactions discovered in chapter 3 are used in conjunction with the corresponding DNA sequences in order to identify sequence specific binding sites for individual TFs. *De novo* predicted binding motifs are tested using mutated promoter fragments and pairwise Y1H screens against the reported TF to assess any changes in binding capacities of previously found interactions. Furthermore, *Arabidopsis* plants harbouring T-DNA insertions in some TFs binding to promoters of multiple genes used in chapter 3 are tested for altered susceptibility to infection with Botrytis.

Finally, chapter 5 summaries the content of the findings presented in this thesis and draws final conclusions of the investigations carried out.

## Chapter 2

# Conserved Noncoding Sequences Highlight Shared Components of Regulatory Networks in Dicotyledonous Plants

### 2.1 Introduction

Genome sequencing has greatly expanded our understanding of organisms on a macro and micro scale, and how the two interact together. For example, in the 1950's eminent mathematician and code-breaker Alan Turing proposed a mechanism by which a system of chemical substances (which he termed morphogens) reacting and diffusing together could be driven unstable by unpredictable (random) influences resulting in spatially varying patterns of chemical concentrations (Turing, 1952). Turing was referring to patterns seen on a wide variety of animals, like zebras, tigers and snails. However, only recently scientists were able to pinpoint some of the key genes that are responsible for such patterning, e.g. Fibroblast Growth Factor (FGF) and *Shh* act as an activator-inhibitor pair to achieve stripe pattern in zebrafish (Economou et al., 2012). The prevailing opinion at the time was that the key to understanding biology in human and other cells is in understanding the coding sequences of genes. Knowing the sequence of all genes in human cells would inform us of all complex behaviours associated with the proteins of these genes. However, recent findings suggest that the key may lie outside of protein coding

sequences. Noncoding DNA (ncDNA) sequences upstream of the coding region of genes, so-called “Junk DNA”, play an integral part in regulating the expression of genes (ENCODE Project Consortium et al., 2012).

The current state of experimental biology does not allow us to rigorously test all ncDNA for potential interactions with all known TFs. Therefore, narrowing down potential locations where functional interactions may occur would provide a footing for further experimental validation of proposed functional regions. This chapter will focus on the computational identification of potentially conserved noncoding sequences (CNSs) in the promoters of *Arabidopsis* and three selected plant species using a comparative genomics approach. Four genomes were analysed using novel comparison methods utilising the power of computational clusters in order to make analysis time feasible. Furthermore, the genes identified as containing CNSs will be tested for potential functional importance. CNSs themselves were assessed for the presence of potential binding sites for known TFs.

### **2.1.1 Phylogenetic Footprinting**

Central to the analysis of the conserved regions between species is the concept of “Phylogenetic Footprinting”. It hypothesises that functional regions of the ncDNA are under higher selective pressure, and therefore evolve at a slower rate, than non-functional ncDNA. Thus, detecting a sequence that has remained conserved across evolutionarily divergent clades implies that the sequence has functional significance, whereas non-functional areas will eventually disappear as a result of genetic drift, where random mutations accumulate more often (Tagle et al., 1988). Phylogenetic footprinting simplifies the task of finding regulatory elements by identifying CNSs initially using orthologous sequences and then refining the search space to informative regions (Frazer et al., 2003). For CNSs upstream of a gene’s transcription start site (TSS), this conserved function is likely to be regulatory.

### **2.1.2 Selection of Compatible Plant Species for Phylogenetic Footprinting**

A major point of debate is selection of the appropriate species/genomes for comparison. On one hand, if two genomes have diverged only “recently”, there would not have been enough time for the sufficient number of mutations to accumulate, and therefore a large number of regions could be considered as conserved. On the other hand, if two genomes have diverged a “long” time ago, very little would be

conserved between them and comparisons may not be as informative (Duret and Bucher, 1997). Additionally, species that have diverged a long time ago may have developed a mechanism that performs a similar task, but functions through a different set of *cis*-regulatory elements. A careful balance needs to be achieved in order for comparison to be both meaningful and informative in terms of the functional CNSs. Additional consideration needs to be given to the whole genome duplication (WGD) events. Such events can give rise to many-to-many relationships between CNSs across multiple genomes.

*Arabidopsis thaliana* is a member of the mustard family whose genome sequencing was completed in 2000 (Initiative, 2000). Since the publication of *A. thaliana* sequence it has become one of the most comprehensively studied and annotated genomes in *Viridiplantae*, such that it became a model organism. In this study, the genome sequence of *A. thaliana* was compared to the sequences of three other dicot plant species: papaya (*Carica papaya* (Ming et al., 2008)), poplar (*Populus trichocarpa* (Tuskan et al., 2006)) and grape (*Vitis vinifera* (Velasco et al., 2007)), that diverged from a common ancestor with *Arabidopsis* 72 million years ago (Mya), 109 Mya, and 117 Mya, respectively (Hedges et al., 2006). A previous study has estimated that  $\approx$  100 Mya is an appropriate divergence limit for reliable CNS discovery using phylogenomic comparisons of plant upstream regions for species within this clade (Reineke et al., 2011). Additionally, poplar has undergone a WGD event and *A. thaliana* had two WGD events since its divergence from its most recent ancestor, papaya Figure 2.1.

### 2.1.3 Previous studies

Some previous studies have focused on extracting information potentially conserved intragenomically, finding paralogs within *Arabidopsis* arising from WGD events (Freeling et al., 2007; Thomas et al., 2007; Haberer et al., 2004). However, as a result of recent tetraploidy of *Arabidopsis* these studies have focused on identifying CNSs potentially present in paralogs. At the same time, large stretches of the ncDNA sequences are conserved between paralogs which leads to a strict definition of a CNS ( $\geq 70\%$  identity;  $\geq 100$  bp in length (Loots et al., 2000)) and coarse identification of long CNSs. Using orthologs offers an opportunity to explore previously uncharacterised CNSs and the evolution of ncDNA sequences in general. Limited scope studies uncovering CNSs using orthologous sequences have focused on specific gene families across various plant species: rice-*Arabidopsis* (Liu et al., 2001), cauliflower-*Arabidopsis* (Colinas et al., 2002), within cereals (Guo and Moose, 2003) and within

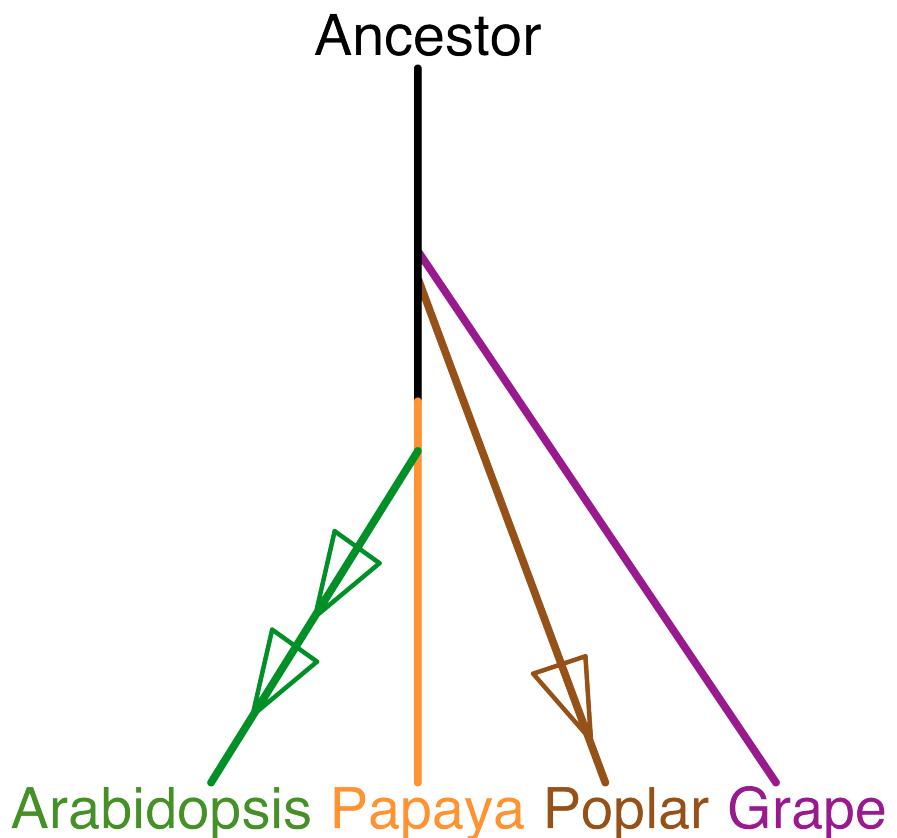


Figure 2.1: Phylogenetic tree of papaya, poplar and grape that diverged from last common ancestor with *Arabidopsis* 72, 109 and 117 Mya. Triangles represent whole genome duplication events. With permission from American Society of Plant Biologists from Lyons et al. (2008).

grasses (Inada et al., 2003). However, phylogenomic comparisons are heavily influenced by the algorithms used for sequence alignment, such as BLAST (Altschul et al., 1990), which is not sensitive to weakly aligning short regions, and instead focuses on global alignments, where sequences are aligned as a whole (Bray et al., 2003). In order to find similarity between orthologous promoters, a fast implementation of the alignment plot method (Krusche and Tiskin, 2010) is used, based on the seaweed algorithm of Tiskin (2008). The alignment plot method has been previously used to accurately predict evolutionarily conserved promoter regions in *LHY*, *TOC1*, *LUX*, *CAB2* and *ABI3*, all of which matched the experimentally validated regulatory regions for those genes (Picot et al., 2010; Spensley et al., 2009).

## 2.2 Methods

### 2.2.1 Databases

Genome databases for *Arabidopsis thaliana*, grape (*Vitis vinifera*), and poplar (*Populus trichocarpa*) were downloaded in MySQL format from Ensembl<sup>1</sup>, release 62, and installed locally. The papaya (*Carica papaya*) draft genome sequences (super-contigs.filtered\_012808.fasta, contigs.filtered\_012808.fasta) and annotation (super-contigs.evm\_27950.gff3, contigs.evm\_27950.gff3) were downloaded from the internet<sup>2</sup> and were used to create a local Ensembl-format database using the Ensembl pipeline and customized Perl scripts. All four databases were prepared by Laura Baxter for the analysis. The Arabidopsis promoter binding element database (AtProbe) was accessed from the internet<sup>3</sup> and TRANSFAC (2009.4 release) were used for motif analysis (Wingender et al., 2000).

### 2.2.2 APPLES Framework

Analysis of Plant Promoter-Linked ElementS (APPLES) framework was written in collaboration with Richard Hickman, Laura Baxter, Christopher Barrington, Sascha Ott, Jonathan Moore, Nigel Dyer, Peter Krusche and myself. It was designed to study intergenic sequences in easily accessible manner. The framework is written in Perl to make it more accessible for a wider biological community, with computationally intensive tasks written in high performance languages such as C and C++. The framework takes advantage of object-oriented design principles and consists of a variety of modules and submodules, allowing it to be easily extendible and reusable.

### 2.2.3 Ortholog and Paralog Identification

A pan-rosid syntenic gene set created by Haibao Tang was obtained from CoGePedia<sup>4</sup>. This uses the QUOTA-ALIGN algorithm (Tang et al., 2011) to identify inferred syntenic regions when no homologous gene is present and enforce a set syntenic relationship based on the whole-genome duplication history of each genome (1:1:2:4 in grape, papaya, poplar, and Arabidopsis). This data set was combined with a set of orthologous genes identified using an implementation of the reciprocal best hit method (Moreno-Hagelsieb and Latimer, 2008). In summary, a FASTA file of all proteins in each genome is made and formatted into a BLAST database. BLASTP

---

<sup>1</sup>Ensembl - <http://plants.ensembl.org/index.html>

<sup>2</sup>Papaya genome - <http://www.life.illinois.edu/plantbio/People/Faculty/Ming.htm>

<sup>3</sup>AtProbe website- <http://exon.cshl.org/cgi-bin/atprobe/atprobe.pl>

<sup>4</sup>CoGePedia website - [http://genomevolution.org/wiki/index.php/Syntenic\\_gene\\_sets](http://genomevolution.org/wiki/index.php/Syntenic_gene_sets)

(Altschul et al., 1990) is performed between each set of proteins, selecting the best hit for each protein. The BLAST results are compared, and where the best match of protein A in genome 1 is protein B in genome 2 and vice versa (i.e., reciprocal best BLAST hit), an ortholog assignment is made. Using this method produced three lists of *Arabidopsis* genes with a corresponding orthologous gene from each of the target genomes, which were merged to produce a single list of 15,386 *Arabidopsis* genes that had at least one orthologous gene in one target species.

An accurate list of manually curated paralogous pairs was obtained from Thomas et al. (2007). For each member of a paralog pair, and where the synteny-based map indicated multiple *Arabidopsis* genes are orthologs, each *Arabidopsis* gene is also assigned the orthologs of its gene-duplicate partner(s). In the combined ortholog map for *Arabidopsis* against poplar, grape, and papaya, 21,034 *Arabidopsis* genes have at least one ortholog assigned in at least one species.

#### 2.2.4 Sequence Alignments

Perl script has been written to automatically access APPLES framework, core of which has been previously written by Laura Baxter and Richard Hickman (Baxter et al., 2012) to retrieve the TSSs and upstream sequences for each *Arabidopsis* gene and its ortholog(s) from sequence databases (see above). For *Arabidopsis*, poplar, and grape, TSSs correspond to Ensembl annotations, and for papaya, these correspond to the 5'-most feature (mRNA or CDS) in the gff3 file. A maximum of 2000 and minimum of 200 nucleotides were taken, but truncating the sequence to the neighbouring gene if within 2 kb. The sequence alignment scores were calculated using an implementation of the seaweed algorithm (Krusche and Tiskin, 2010) in C, with a sliding window length of 60 nucleotides. The first 60 bp window in specie A is aligned with the first 60 bp window in species B and score is calculated using the following rules: +1 for a match, 0 for a mismatch, and 0.5 for a gap. Thus, for a 60-bp window, the highest score possible is 60. If the score is higher than any previous score calculated then it is stored. The window in specie B is shift by 1 bp and the alignment score is calculated once again and checked if it is the new maximum. This process is repeated until available sequence in specie B has all been traversed and scored giving a maximum alignment score for the first window in specie A. The window is shifted by 1 bp, the window in specie B is reset back to the start of the sequence and the process is repeated again. Therefore, score for each window in A is the maximum scored alignment between it and all possible 60 bp windows in specie B.

### 2.2.5 Converting Alignment Scores to Conservation Scores

The alignment score is converted into a conservation score using a sigmoidal function with upper and lower thresholds. The upper threshold indicates that any alignment score found above this threshold is assigned a conservation score of 1. Conversely, the lower threshold indicates that any alignment score found below this threshold is assigned a conservation score of 0. The upper and lower thresholds were calculated for each species pairing using the distribution of alignment scores from randomly assigned gene pairs (random orthologs).

The upper thresholds (48, 47 and 47 for papaya, poplar and grape respectively) were established by manual inspection of the alignment score histograms (Figure 2.2), taking the score above which no random gene pair produced a significant alignment. The lower bounds (38, 38 and 38 for papaya, poplar and grape respectively) were chosen as the point where the control set and the ortholog set begin to show significantly different numbers of alignments. These thresholds were used for the real orthologs to find the conservation score of each CNS. Repetitive sequences are penalized in the conversion procedure. A region is called repetitive if it is annotated as a repeat in the Ensembl sequence database (identified by RepeatMasker, based on species-specific libraries of repeats). Repetitive sequence in a window shifts the sigmoidal curve proportionally to the right, so a region containing repeats requires a higher alignment score than a window of non-repetitive sequence to obtain the same conservation score. During the conversion procedure, where significantly high-scoring window pairs positionally overlap, they are merged into a single contiguous region. In the multi-species analysis, the conservation scores between each of the three target species and *Arabidopsis* (where available) are combined into a single conservation score using Equation 2.1.

$$1 - \prod_i (1 - P_i), \quad (2.1)$$

where  $P$  is the maximum conservation score for a region in one species pair, and  $i$  is each species pair. For example, in the case of three species with conservation score 0.2 ( $P_1 = P_2 = P_3 = 0.2$ ), the overall conservation score is 0.488, whereas a conservation score of 0.5 in just one species ( $P_1 = 0.5, P_2 = P_3 = 0$ ) yields an overall conservation score of 0.5. Implementation of this conversion mechanism was done by Christopher Barrington and forms part of APPLES framework.

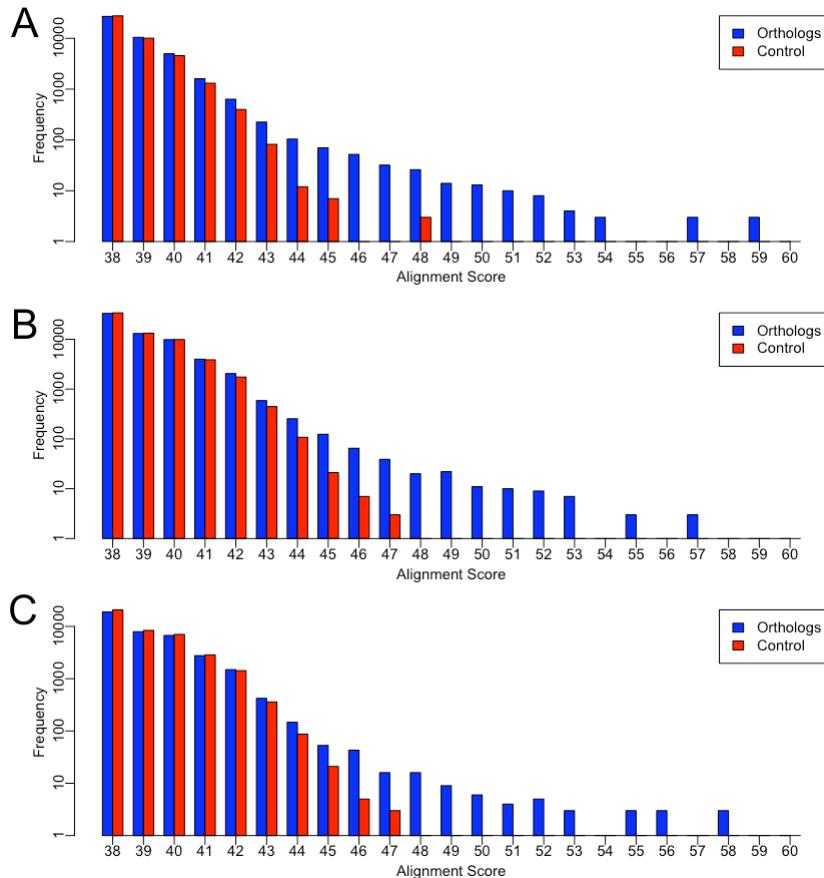


Figure 2.2: A large number of *Arabidopsis* promoters contain evolutionarily conserved noncoding sequences. (A) to (C) Histograms of the alignment score distributions between *Arabidopsis* promoters and orthologous promoters in other species. (A) *Arabidopsis* vs Papaya, (B) *Arabidopsis* vs Poplar, (C) *Arabidopsis* vs Grape. For each *Arabidopsis* promoter the alignment score of the highest-scoring 60 bp-window is recorded. Scores are based on match/mismatch/gap values of +1/0/-0.5 respectively. The maximum score for a 60 bp-window is 60. Alignment scores of orthologous promoters and promoters of randomly assigned gene pairs (control) are shown.

### **2.2.6 Filtering Out Potential Protein Coding Regions**

The threshold of 0.3 was chosen as it has resulted in a low predicted false positive rate between CNSs found in real and random orthologs. To exclude potential protein coding regions from the set of conserved regions, Laura Baxter has removed any gene and all of its associated conserved regions from the 0.3 threshold set and above if any of its regions had a significant BLASTX hit to any Viridiplantae sequence in the National Center for Biotechnology Information database. To establish an appropriate e-value cutoff for a significant hit, Laura Baxter has also randomly permuted each sequence in the 0.3 threshold set and performed the BLASTX search using this set of sequences to obtain the distribution of e-values for random sequences. Laura Baxter then performed the same BLASTX search on the real sequences, using the minimum e-value from the random set ( $4.00 \times 10^{-8}$ ) as the cutoff for a significant hit, Table 2.1.

Conservation core Threshold	No. Genes before CDS Filtering	No. Aligned Regions before CDS Filtering	No. Genes after CDS Filtering	No. Aligned Regions after CDS Filtering
1	148	157	136	143
0.9	384	414	365	392
0.8	481	517	460	492
0.7	578	630	554	602
0.6	782	850	758	822
0.5	1230	1372	1202	1340
0.4	1319	1481	1291	1448
0.3	1672	1902	1643	1865

Table 2.1: Numbers of aligned regions and associated genes from orthologous promoters before and after filtering for putative coding sequences.

### 2.2.7 PSSM Clustering

A total of 2595 known position specific weight matrices were retrieved, including all matrices from TRANSFAC v2009.4 (Wingender et al., 2000), JASPAR (Bryne et al., 2008), and PLACE (Higo et al., 1999) databases. To reduce redundancy of this set of weight matrices and the computational load of performing the overrepresentation test using binomial statistics for all motifs, the matrices were clustered into 728 clusters using the sum of discrete Hellinger distance metric for each nucleotide (Equation 2.2, (Hellinger, 1909)) with a threshold of 2.3 for each cluster, and a single representative motif was selected from each. Under the threshold of 2.3 similar weight matrices clustered together, while more distant matrices formed parts of different clusters. The weight matrices were sorted in ascending order of information content calculated for each matrix in the cluster and median matrix was chosen as a representative member of the cluster.

$$\begin{aligned} H^2(P, Q) &= \sum_{n \in \{A, C, G, T\}} H_n^2(P_n, Q_n) \\ H_{n \in \{A, C, G, T\}}^2(P_n, Q_n) &= \frac{1}{2} \sum_{i=1}^L (\sqrt{p_{n,i}} - \sqrt{q_{n,i}})^2 \end{aligned} \quad (2.2)$$

, where  $P$  and  $Q$  are PSSM matrices of the same size  $L$  times 4 being compared  $P_n$  and  $Q_n$  are  $1 \times L$  vectors and  $L$  is length of the binding motif.

### 2.2.8 Motif Overpresentation

The set of 728 motifs, representative of 2595 available motifs (see above), was used in the binomial overrepresentation tests, where each motif is tested against the set of CNSs with 0.7 threshold (or against a set of randomly selected regions). The best 100 matches of the motif in the sequence set were identified and sorted in descending order of significance. The binomial distribution (Equation 2.3) was used to compute an overall overrepresentation score, taking into account the number of instances of the motifs, the overall length of sequences in the set and probability of this motif occurring by chance alone within the sequence of the same length generated using second order Markov Model. The score was computed for the top n motif matches where  $1 \leq n \leq 100$  was chosen to optimize the overrepresentation score.

$$P(N = n) \sim Bi(n, p) \quad (2.3)$$

, where  $n$  is the number of occurrences of a motif within the CNS and  $p$  is maximum probability from across all  $n$  sites.

For each of the 602 CNSs from the Arabidopsis 0.7 threshold set (chosen for the highest acceptable false positive rate), genes were randomly assigned from the same genome to make background sets for comparison. The locations of the control CNSs in the randomly assigned genes were chosen to be the same as the locations of the corresponding CNSs in the real set to make the comparison more stringent. The overrepresentation test was run 100 times to assess the distribution of motif overrepresentation scores in the random sets, and it was run once on the set of CNSs. The probability threshold determined for each motif in the real set of CNSs was applied individually for each motif in the randomly located CNSs to remove sites with lower significance than found in the real set. The individual P values for each motif site were calculated using the *pnorm* function in R (Gentleman et al., 2004). Known repeats in all sequences were masked using the repeat annotations in the sequence databases.

### 2.2.9 GO Term Analysis

GO term analysis was performed using the BiNGO plugin (version 2.3) (Maere et al., 2005) for Cytoscape (version 2.6) (Shannon et al., 2003). The set of 554 Arabidopsis genes (0.7 threshold) was compared for overrepresentation using a hypergeometric test statistic using the set of Arabidopsis genes with an identifiable ortholog as the reference set. Benjamini and Hochberg false discovery rate correction for multiple testing was applied, with significance level of 0.05 (5%). The tests were performed using three ontology files that come as part of BiNGO (updated August 2010): “GO\_Biological\_Process”, “GO\_Molecular\_Function” and “GO\_Cellular\_Component.”

### 2.2.10 Prediction of Nucleosome Positioning

Sequences of 10 kb, with the CNS positioned centrally in each sequence, were used as input. Where a gene was associated with more than one CNS, one was randomly selected. Nucleosome occupancy probabilities were calculated at each nucleotide position and the results averaged across the CNS set (554 sequences, 0.7 threshold). Ten sets of control sequences were created, whereby for each conserved sequence, regions of 10 kb were selected upstream of 10 randomly picked genes in the Arabidopsis genome, such that the centre of the selected region is at the same position relative

to the TSS as the centre of its comparable conserved sequence and not allowing the sequence in the centre to contain repetitive sequences. Average nucleosome occupancy probabilities were calculated for each of the 10 control sets of 554 sequences, and the mean and sd of these averaged values were plotted. As the prediction software does not tolerate input sequences containing non-ACGT characters, up to two sequences were omitted from each set prior to analysis (< 0.4% of sequences). 10 kb sequences were prepared by Alex Jironkin and Richard Hickman computed the nucleosome occupancy probabilities, using nucleosome prediction software (Kaplan et al., 2009) with default parameters <sup>5</sup>.

### 2.2.11 DNase-Seq Analysis

Genomic coordinates associated with the CNS set (0.7 threshold), excluding mitochondrial and chloroplast genes that do not have publicly available DNase-Seq data, were retrieved and numbers of sequencing reads associated with published DNase-Seq data (Zhang et al., 2012) (retrieved from the GEO database (Edgar et al., 2002) accession: GSM847326) were averaged for every CNS to obtain an average number of reads per region.

---

<sup>5</sup>Nucleosome prediction software download page - [http://genie.weizmann.ac.il/software/nucleo\\_prediction.html](http://genie.weizmann.ac.il/software/nucleo_prediction.html)

## 2.3 Results

### 2.3.1 Multispecies Analysis Yields Hundreds of CNSs

The first phase of establishing CNSs between selected species is to calculate orthologous genes. Orthologous genes determined by Reciprocal Best Hit (RBH) and pan-rosid syntetic orthologs, estimated using QUOTA-ALIGN (Tang et al., 2011), were merged together for a complete list of orthologous genes across four species of interest (see Methods). A total of 21,034 genes were found to have one or more orthologous genes in papaya, poplar and/or grape. 92% of all compared genes from *Arabidopsis* were found to map to at least one ortholog in another species (Figure 2.3, left).

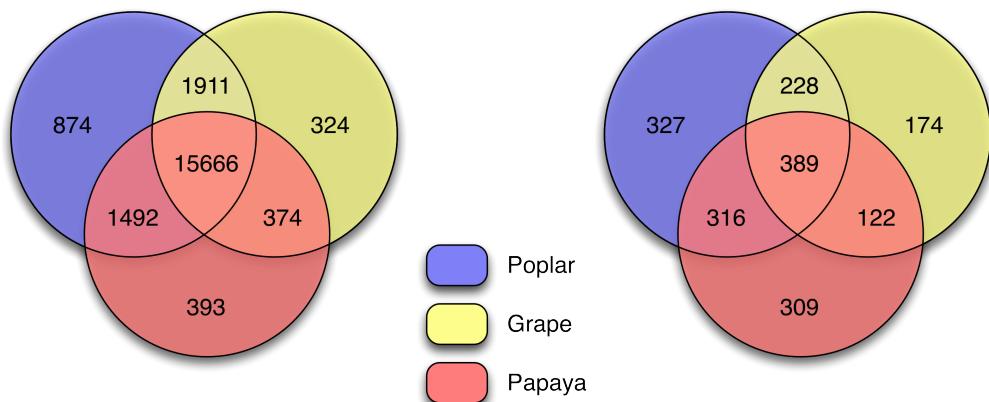


Figure 2.3: Left - distribution of *Arabidopsis* orthologs across comparator species, showing how many *Arabidopsis* genes have an ortholog in one, two or all three comparator species (21034 *Arabidopsis* genes). Right - distribution of species contributing to *Arabidopsis* CNS at the 0.3 threshold (1865 CNSs).

Phase two of the analysis was to determine alignments of the promoter sequences from *Arabidopsis* with the promoter sequences from orthologous genes in other genomes. Promoter sequences are defined as sequences upstream from the annotated TSSs up to 2kb in length (average intergenic sequence in *Arabidopsis*), or shorter if another gene lies closer than 2kb. Sequence alignments were produced using an implementation of the seaweed algorithm (Krusche and Tiskin, 2010). The algorithm computes the optimal sequence alignments for all pairs of 60bp sequence windows for a typical promoter. Using the seaweed algorithm allows for highly sensitive detection of the conserved regions regardless of the position along the promoter sequence. On the other hand, traditional algorithms, like Needleman-Wunsch (Needleman and Wunsch, 1970), would miss pockets of conserved sequences be-

cause the algorithm focuses on the global alignment pattern. The short promoter sequences, 100-200 bp in length, may contain conserved binding site locations, even if the promoter as a whole, typically 2kb in length, has undergone a large number of mutations and therefore no longer results in significant alignments with the promoters in orthologous sequences. The CNSs may also be shuffled in the intergenic sequence itself following the WGD events and chromosome rearrangements (Murat et al., 2010). In order to determine the significance of the newly found CNSs, an equivalent control set was generated, whereby for each of the 21,034 *Arabidopsis* genes that map to an orthologous gene, a pseudo-ortholog was assigned at random from each of the three other species on a gene-by-gene basis, as in the real set, and the alignments were computed as before (see Methods). Distribution of alignment scores from real and pseudo-orthologs has revealed significant differences, with real orthologs producing a greater number of alignments and higher overall scores (Figure with the 3 distributions). The higher level of alignments in the real orthologs suggests the presence of evolutionary conserved sequences within the promoter fragments, otherwise not present in the pseudo-ortholog assignments. High scoring alignments only occur in the real orthologous sequences, and are missing from the pseudo-orthologs, where alignment scores drop-off abruptly at the high end of scoring, 48, 47 and 47 for papaya, poplar and grape respectively (Fig with the 3 distributions). In contrast, real orthologous sequences produce alignments with top scores of up to 59 out of 60 (based on the window length used). This suggests that a large number of promoter sequences found in *Arabidopsis* are sequence conserved in other species and may potentially indicate conserved function.

The significance of the alignment scores may not be immediately visible and depends on the evolutionary distance from one species to another. Thus, a direct comparison of raw alignment scores is not possible and may be misleading. To overcome this obstacle, a concept of conservation score was developed (see Methods) that allows the integration of the evolutionary distance as well as taking into account alignment scores across other species used in the analysis. Additionally, conservation scores also allows the comparison of weakly aligned sequences in multiple species together with the strongly aligned regions between two species. Conservation score has a range [0, 1] and signifies how strongly a sequence alignment is expected to reflect conservation, i.e. sequence similarity as a result of evolutionary constraints. The higher the conservation score, the greater is the expectation that the alignment score observed represents true sequence conservation; conversely, alignment scores commonly found by chance determine the lower end of the conservation spectrum.

Also, as conservation scores are computed, any overlapping regions are merged together into a single region to avoid false significance which may arise from counting regions that differ only by few base pairs multiple times. By comparing the distributions of conservation score from real and pseudo-orthologs we are able to determine threshold levels with the desired false positive rate (FPR) (Table 2.2). A threshold of 0.3 was used to prevent sequences with weak alignment scores (from 0 to 0.2) “piggy-backing” together with the strongly aligned sequences into the CNS regions, therefore no data for thresholds of 0.1 and 0.2 was computed at all as these weakly aligning regions are thought to contain no meaningful information within them.

Conservation Score Threshold	Orthologs		Random Gene Pairs		False Positive Rate
	No. of Genes after CDS Filtering	No. Aligned Regions after CDS Filtering	No. Genes	No. Aligned Regions	
1	136	143	0	0	0
0.9	365	392	7	7	$3 \times 10^{-4}$
0.8	460	492	23	23	$1.1 \times 10^{-3}$
0.7	554	602	36	36	$1.7 \times 10^{-3}$
0.6	758	822	117	119	$5.6 \times 10^{-3}$
0.5	1202	1340	412	431	$1.96 \times 10^{-2}$
0.4	1291	1448	467	492	$2.22 \times 10^{-2}$
0.3	1643	1865	657	700	$3.12 \times 10^{-2}$

Table 2.2: Numbers of aligned regions and associated genes from orthologous promoters (after filtering for putative coding sequences) and from promoters of random gene pairs at different thresholds of conservation score.

Threshold	FPR	Expected True Positive Genes	Expected True Positive Regions
1	0	100%	100%
0.9	$3 \times 10^{-4}$	98.1%	98.2%
0.8	$1.1 \times 10^{-3}$	95%	95.3%
0.7	$1.7 \times 10^{-3}$	93.5%	94%
0.6	$5.6 \times 10^{-3}$	84.6%	85.5%
0.5	$1.96 \times 10^{-2}$	65.7%	67.8%
0.4	$2.22 \times 10^{-2}$	63.8%	66%
0.3	$3.12 \times 10^{-2}$	60%	62.5%

Table 2.3: Expected true percentage of positive genes and associated regions after taking into account number of negatives after random controls for each threshold and FPR.

Taking the least stringent threshold level into account, we find 1865 CNSs present upstream of annotated TSSs for 1643 *Arabidopsis* genes. Figure 2.3 (right) shows the relative contribution of each of the other species used to compare orthologous genes, in the 1865 CNSs. Alignments from promoters of two or more species contributed the CNSs in 57% of the cases at the 0.3 threshold.

Distribution of FPRs for randomly assigned orthologs (Table 2.2) suggests the threshold to be used in future analyses. In particular, the FPR for correctly identifying conserved regions is very low for stringent thresholds. Taking a 0.9 threshold as an example, 392 CNSs were identified as being present in 365 genes in real orthologs, in contrast only 7 CNSs covering 7 genes were identified using randomly assigned orthologs ( $FPR 3 \times 10^{-4}$ ). Taking even lower thresholds, e.g. 0.6, there are 822 regions spanning 758 genes in the real orthologs and 119 regions in 117 genes in the random set ( $FPR 5.6 \times 10^{-3}$ ), showing a significant difference between the control and real ortholog assignments. The threshold of 0.7 was deemed to be suitably significant ( $FPR 1.7 \times 10^{-3}$  applied to the real-ortholog set yields 6.5% false positives) and therefore 602 CNSs found in 554 *Arabidopsis* genes were selected as a robust candidate CNS set for subsequent studies.

### 2.3.2 CNSs Show Positional Bias toward TSSs

The distances between the start of the conserved regions and each *Arabidopsis* gene's TSS associated with the corresponding CNS were recorded in both real- and pseudo-ortholog sets. A threshold of 0.3 was applied to the pseudo-ortholog set in order to achieve a comparable number of regions to the CNS set, where a previously established threshold of 0.7 was applied.

A restriction was applied to only include genes with at least 500bp in the intergenic region so as to limit a potential bias that may be caused by the genes with generally short intergenic distances. A clear positional bias is observed towards the first 100bp to 200bp upstream from the TSS in the CNS set derived from the real orthologs (Figure 2.4A). On the other hand, no such positional bias is observed in the CNS set derived from the pseudo-orthologs (0.3 threshold) and the distribution of distances from the TSSs is approximately uniform (Figure 2.4B). The length of the intergenic region has no significant effect on the positional bias, or lack thereof (Figure A.1). Such positional bias towards the TSS is indicative of a potential *cis*-acting regulatory function of the CNSs. In particular, some CNSs (14%) are within 50bp of the TSS and have a TATA-box motif present, which is important for polymerase assembly. The potential functional importance is consistent with

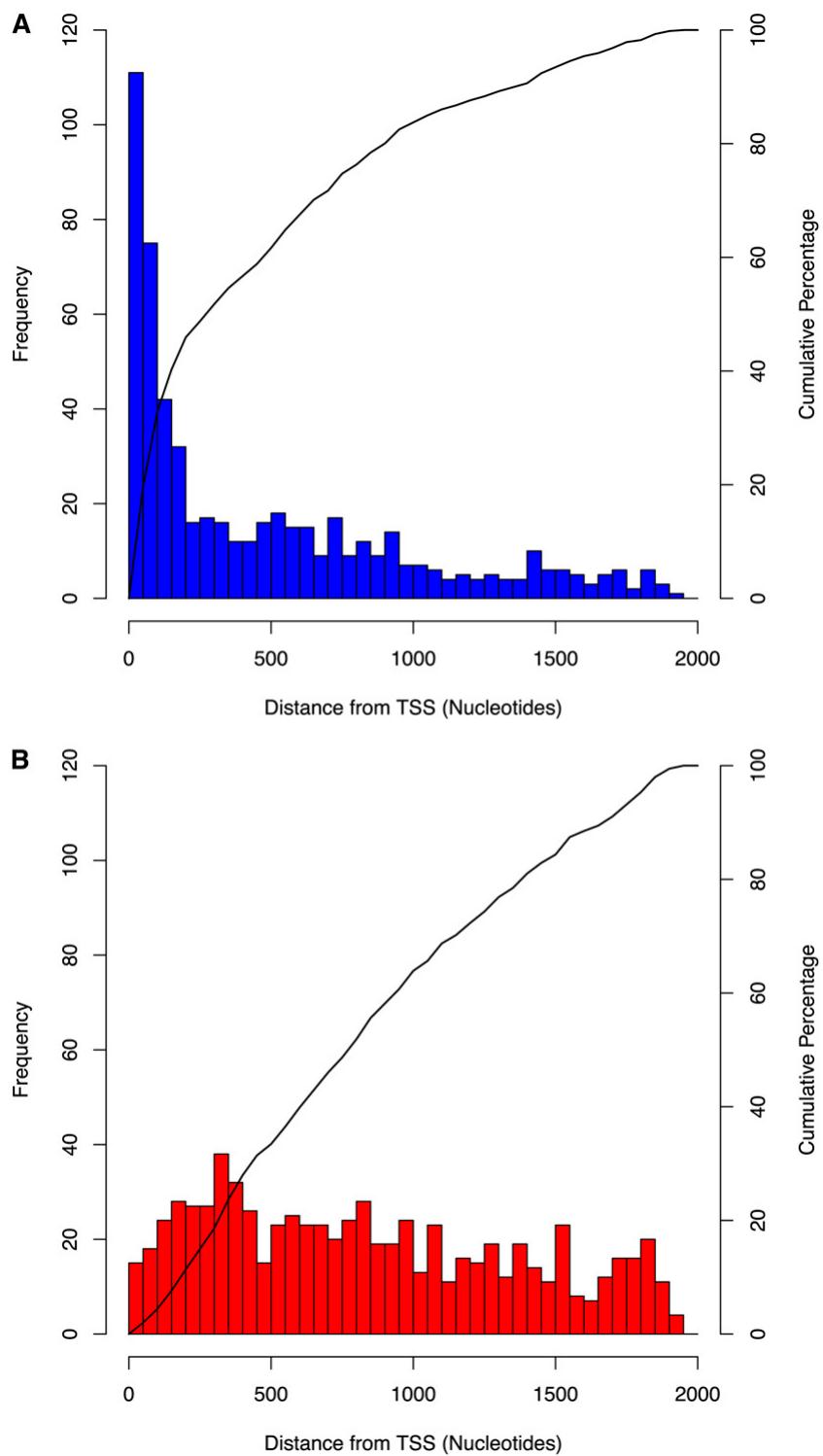


Figure 2.4: Alignments produced in orthologous promoters reveal a positional bias toward the TSS. Distribution of distances between conserved regions and the TSS in *Arabidopsis* promoters. Only distances where the intergenic length is at least 500 nucleotides are plotted. A - Distances observed in orthologous promoters, 0.7 threshold (566/602 distances plotted). B - randomly assigned gene pair promoters, 0.3 threshold (684/700 distances plotted)<sup>45</sup>

the previous findings that the TATA-box occurs within 50bp of the TSS (Gannon et al., 1979) and is likely to form part of the core promoter region. However, the majority of the CNSs (76%) lay further than 50bp away from the TSS and therefore fall outside of the core promoter region. Additionally, 36% of the CNSs are beyond 500bp upstream from the TSS.

### 2.3.3 CNSs Are Highly Enriched in TFBS Motifs

Functional promoter regions are expected to contain a larger number of TF binding sites than regions derived from the random ortholog assignments, which may still contain TF binding sites as the regions were derived from real upstream sequences. The presence of known motifs would also provide additional support to the premise that CNSs are potentially functionally active. The set of 602 CNSs (0.7 threshold) was tested for an enrichment of known TF binding sites and compared to enrichments found in the CNSs from a control set. Prior to testing for enrichment, known motifs were clustered based on the Hellinger metric and a representative from each cluster with a mean information content was selected to be tested further (see Methods). A total of 728 eukaryotic TF binding sites, represented by the PSSMs, were tested for presence in the CNS set. A negative control set was designed to mimic the CNS set in as many ways as possible. The control sequences were chosen in the promoters of randomly selected *Arabidopsis* genes to be the same length and the same distance upstream from the TSS as the regions from the CNS set. The control sequences were also chosen to include only non-repetitive regions. Therefore, the control regions are identical to the 602 CNSs in every respect except for their conservation. 100 control sets were tested in order to obtain robust statistics (see Methods).

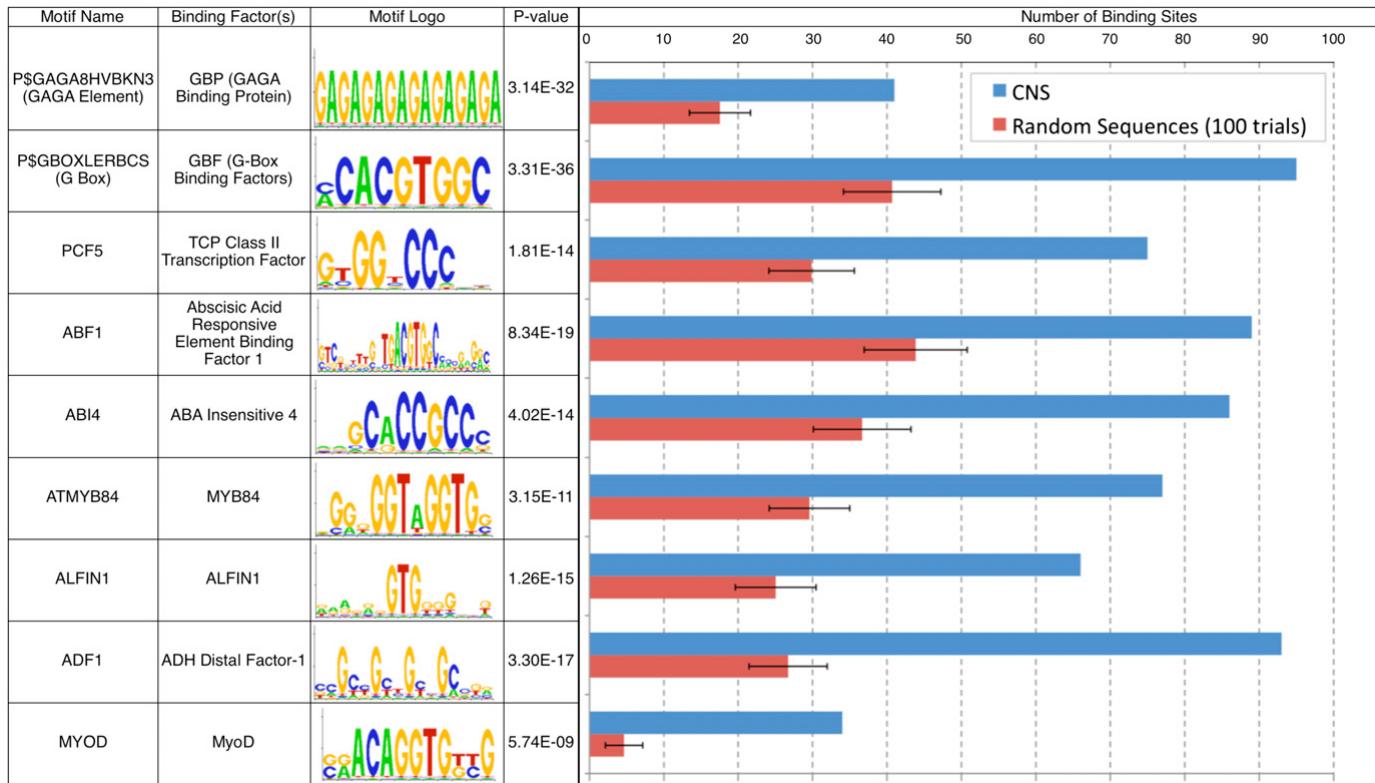


Figure 2.5: Arabidopsis CNSs are strongly enriched for specific TF binding sites. Data for selected TFBS motifs are shown. Number of binding site occurrences in CNSs from orthologous promoters (blue) compared with the number in random control sequences (red; mean of 100 trials and sd shown). Sets of control sequences were picked from Arabidopsis promoter regions and match the CNS in length, number, relative position to TSS, and underrepresentation of known repetitive elements.

### **2.3.4 Identification of Previously Experimentally Validated Promoter Binding Elements**

Experimentally verifying promoter binding elements *in planta* and *in vivo* is a difficult and time consuming task. As a result, limited data is available to directly confirm the functionality of a CNS. Previously published and publicly available online databases were used in order to establish if the CNS set (0.7 threshold) contains any previously experimentally proven binding regions. AtProbe is a small database hosted by the Zhang laboratory that focuses on experimentally validated binding elements. The database contains information for 76 Arabidopsis genes and is manually curated from the primary literature sources. One of the genes, *AP1* (APETALA 1), is present in the CNS set identified earlier. Moreover, the CNS identified 433bp upstream from the TSS, entirely covers the experimentally verified binding element LFY, that is recognised by *LEAFY* and controls flower development in Arabidopsis (Parcy et al., 1998). An analogous method was used in the past to identify potential conserved sequences in four Arabidopsis genes, with well characterised promoters (Spensley et al., 2009; Picot et al., 2010); and our results match the previously published regions for *TOC1*, *LUX* and *ABI3* (all of which are present in the 0.7 threshold CNS set). Orthologs were not identified for *CAB2*, used in the same study.

### **2.3.5 Prediction of Nucleosome Positioning**

The positioning of nucleosomes plays a vital part in gene regulation, dividing promoter regions and influencing the activation of transcription (Jiang and Pugh, 2009). The nucleosome positions are thought to be determined by certain DNA sequence preferences. A model developed previously by Kaplan et al. (2009) which determines nucleosome-DNA interactions was used to compute occupancy probabilities for the set of CNSs (0.7 threshold), given the functional importance of nucleosomes and the potential functional significance of the CNSs. The model reflects sequence features preferential for nucleosome positioning and is independent of the alignment and conservation scores used in this study. The model outputs the probability that each base pair is part of a region bound by a nucleosome at each position along the input sequence. A total of 10kb sequence with the CNS positioned in the middle of the sequence was constructed for each CNS (554 sequences) and the mean score calculated for each position in the CNS set. Ten comparable control sets were generated to reflect the CNS set (see Methods) and scored using the same procedure. Average nucleosome occupancy probability in the CNS regions were compared to the average of the control sequence sets (mean across 10 sets), Figure 2.6.

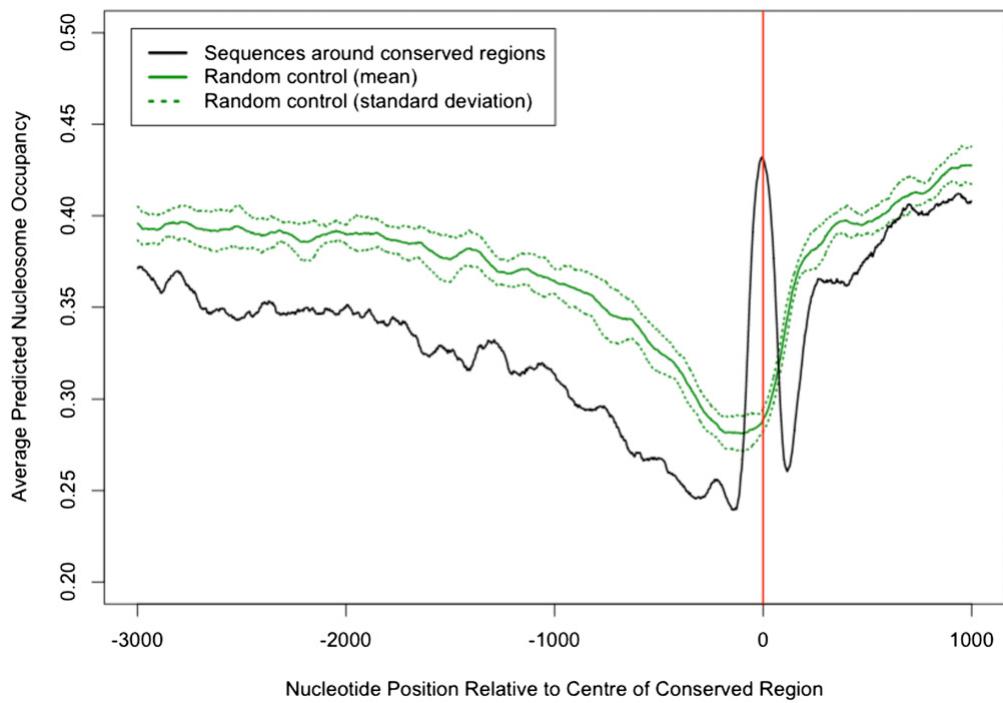


Figure 2.6: Predictions of nucleosome occupancy confirm the significance of the identified CNSs. Average predicted nucleosome occupancy for 554 10 kb sequences surrounding CNSs (black line) and 10 equivalent sets of control sequences (solid green line represents mean of the 10 control sequence sets; dashed line shows sd) was calculated. The 3 kb to +1 kb regions are plotted for clarity, as values plateau either side of this for the remainder of the 5 kb to +5 kb range examined. Red line at nucleotide position 0 indicates the centre position of the CNS or control sequences, with TSS therefore being positioned to the right of this.

Previous studies in yeast (Albert et al., 2007; Kaplan et al., 2009) and humans (Ozsolak et al., 2007) have identified that “nucleosome-free” regions immediately upstream from the TSS are depleted of nucleosome occupancy. This abrupt decline in occupancy probability can be seen in both the control and sequences around the CNS (Figure 2.6). However, in the CNS set there is a clear peak in predicted nucleosome occupancy that directly coincides with the CNS regions, which indicates that the CNSs tend to have higher propensity to be occupied by nucleosome than expected by chance alone. As nucleosomes are known to play functionally important roles in transcriptional regulation (Jiang and Pugh, 2009), their strong presence in the CNSs further strengthens the evidence that CNSs are potentially functional regions and involved in the regulation of their respective genes.

### 2.3.6 GO Term Overrepresentation Unveils Key Biological and Molecular Functions of Genes Associated with CNSs

Overrepresentation of Gene Ontology (GO) terms is commonly used to find particular biological or molecular processes overrepresented within a set of genes. GO terms associated with 554 genes from the CNS set (0.7 threshold) were tested for overrepresentation in three categories: Biological Processes, Molecular Function and Cellular Component, to identify common roles for these genes (see Methods).

Looking at overrepresented terms for Biological Processes reveals two main areas where genes from the CNS set may function. Firstly, the ten most strongly overrepresented terms (adjusted P-value  $7.59 \times 10^{-36}$ , 134 unique genes) are associated with a variety of regulatory processes. In particular, “regulation of transcription” is highly overrepresented with 83 genes in the CNS set (adjusted P-value  $1.09 \times 10^{-25}$ ). Other overrepresented terms in the regulatory category include various biological, cellular, biosynthetic and metabolic processes. The second major area of overrepresentation is closely related to the developmental processes, e.g. organ development (adjusted P-value  $7.22 \times 10^{-26}$ ), system development (adjusted P-value  $7.22 \times 10^{-26}$ ), shoot development (adjusted P-value  $2.86 \times 10^{-16}$ ), flower development (adjusted P-value  $1.33 \times 10^{-11}$ ), leaf development (adjusted P-value  $1.47 \times 10^{-10}$ ) and meristem development (adjusted P-value  $4.05 \times 10^{-7}$ ).

Molecular function GO terms show transcription-related activities overrepresented the most (adjusted P-value  $1.23 \times 10^{-57}$ ) closely followed by DNA and nucleic acid binding (adjusted P-value  $7.79 \times 10^{-48}$  and  $8.66 \times 10^{-25}$  respectively).

Finally, “Nucleus” was revealed to be the most overrepresented term in the Cellular Components category of the GO annotations (adjusted P-value  $3.53 \times$

P-Value	GO Term	P-Value	GO Term
$7.59 \times 10^{-36}$	regulation of biological process	$1.23 \times 10^{-57}$	transcription regulator activity
$3.44 \times 10^{-35}$	biological regulation	$7.43 \times 10^{-51}$	transcription factor activity
$9.31 \times 10^{-32}$	regulation of cellular process	$7.79 \times 10^{-48}$	DNA binding
$8.49 \times 10^{-27}$	regulation of cellular metabolic process	$8.66 \times 10^{-25}$	nucleic acid binding
$2.50 \times 10^{-26}$	regulation of metabolic process	$1.29 \times 10^{-17}$	binding
$7.22 \times 10^{-26}$	system development		
$7.22 \times 10^{-26}$	organ development		
$1.08 \times 10^{-25}$	regulation of biosynthetic process		
$1.09 \times 10^{-25}$	regulation of transcription		
$2.16 \times 10^{-25}$	regulation of macromolecule biosynthetic process		
$6.06 \times 10^{-25}$	regulation of macromolecule metabolic process		
$3.70 \times 10^{-24}$	regulation of gene expression		
$6.46 \times 10^{-23}$	multicellular organismal process		
$8.84 \times 10^{-22}$	multicellular organismal development		
$4.94 \times 10^{-20}$	developmental process		

(a) Biological Processes

(b) Molecular Function

(c) Cellular Components

Table 2.4: Go Term analysis of the genes in the CNS set (0.7) threshold reveals a variety of regulatory biological processes and molecular functions. As well as being primarily localised to the nucleus.

$10^{-11}$ ).

### 2.3.7 Predicted CNSs Occur in Open Chromatin Areas

One of the indicators of transcriptional activity along the ncDNA is accessibility to DNase I restriction enzyme. In particular, open areas of chromatin, upstream from the annotated TSS that are accessible to the restriction enzyme are markers of potential TF binding sites. Unlike ChIP-Seq or ChIP-PCR methods where an antibody is used to pull down specific DNA binding proteins, the genome wide DNase I cleavage sites cannot report the identity of the individual TFs that are potentially active within the area of cleavage. However, DNase I data is able to mark open/closed chromatin areas, signifying transcriptional activity inside (Boyle et al., 2008). Sequencing reads, obtained from previously published DNase-Seq data (Zhang et al., 2012), from the high confidence CNS set show a markedly different, and statistically significant (two-sample Kolmogorov-Smirnov P-value  $6.9 \times 10^{-3}$ ) distribution of reads as compared to the control set (Figure 2.7). CNSs from the control set have a low number of reads associated with them, as expected (Figure 2.7, green), on the other hand the CNSs from the 0.7 threshold set derived from real orthologs have a higher mean read length and also contain some regions that are associated with much higher read numbers than average (Figure 2.7, blue), and which are therefore more likely to be in transcriptionally active promoter areas.

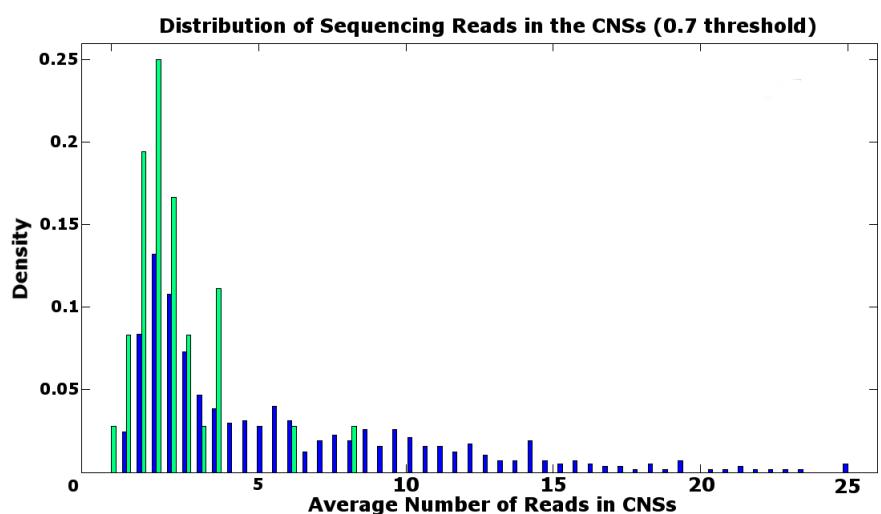


Figure 2.7: Distribution of sequencing reads from promoters of *Arabidopsis* genes associated with control (green) and real CNS (blue) sets in the leaf tissue.

## 2.4 Discussion

This chapter presents a comprehensive study of ncDNA sequences across multiple dicotyledonous plant genomes: Arabidopsis, papaya, poplar and grape. An inclusive ortholog map between Arabidopsis and the three other species was constructed using combinations of reciprocal best BLAST hit and pan-rosid synteny based assignments. The resulting ortholog map covers 21,034 Arabidopsis genes of the total 27,416 protein coding genes (77% of TAIR 10 genes) that have at least one identifiable ortholog in papaya, poplar and/or grape. The number of orthologous genes may potentially increase as the sensitivity of the ortholog assignment methods and the annotation of related genomes improves. However, genes without identifiable orthologs in any species are unlikely to be informative in discovering new CNSs by means of phylogenetic footprinting. Therefore, the majority of the informative gene sets have been captured in the present study, given current genome data and annotations (TAIR 10).

### 2.4.1 Neo-/Non-/Sub-functionalisation of Conserved Genes

An important aspect of WGD events is the potential for paralogs to undergo subfunctionalisation (division of functions), neofunctionalisation (gaining of new function) and/or nonfunctionalisation (loss of function). Changes in protein function could be mediated by the changes in the *cis*-acting elements, e.g. changes in a TF binding site could mean that a different TF expressed at a different time or under different conditions will bind to the ncDNA of the gene, leading to it being expressed at a different time or in a different tissue changing the protein's mode of action. It is difficult to distinguish between nonfunctionalisation and neofunctionalisation and determination is not possible in the absence of detailed expression data for all orthologous genes across all comparator species, which is beyond the scope of the study presented here. However, subfunctionalisation may occur in paralogous genes that are derived from an ortholog in another species. An insight can be gained into potential subfunctionalisation events by comparing CNSs found in paralogs with CNSs found in orthologs.

Thomas et al. (2007) used bl2seq in order to identify paralogous CNSs in Arabidopsis for a set of 3179 gene pairs retained from the  $\alpha$  tetraploidy event. Our study was able to identify all of the regions previously identified in the 2 kb upstream gene sequence, as well as additional CNSs. Using the list of paralogous

pairs produced by Thomas et al. (2007), alignment scores were computed for the 2 kb promoter regions upstream from the annotated TSS. A control set was produced by randomly permuting the pairs, and alignments were computed as before. Both sets of alignment scores were converted into conservation scores and thresholded as before. Using this framework, the FPRs were extremely low at all thresholds ( $< 0.0012$ ; Table A.2). At the 0.3 threshold, 1573 regions were found upstream of 1149 genes (Table A.2) and had an average length of 98 bp. The paralogous CNSs were then compared against orthologous CNSs. Of 3019 genes with both a paralog and an ortholog (as defined in this study), 565 have paralogous CNSs and 291 have orthologous CNSs above the 0.3 threshold. The overlap of these sets is highly significant, with 133 genes having both types of CNSs ( $P < 3.79 \times 10^{-29}$ , hypergeometric test; only one paralog of each pair was included in the set of 3019). Among this set of 133 genes, paralogous CNSs and orthologous CNSs are seen to be overlapping in the promoters of 85 genes. From GO analysis, these 85 genes are enriched for terms including regulation of biological process, regulation of transcription, and system development (data not shown). This is consistent with the idea that some types of genes, such as TFs and genes controlling developmental processes, are generally under greater regulatory constraint on the transcriptional level than other genes and that some of this constraint is often maintained after gene duplication.

### Evidence of Potential Subfunctionalisation of Genes in Arabidopsis

By manually inspecting the positioning and distribution of alignments in the set of 85 genes (overlap between paralogous and orthologous CNSs), a potential example of subfunctionalisation driven by changes in *cis*-acting elements has been uncovered (Figure 2.8). Figure 2.8A shows a case of nonoverlapping orthologous conservation in a paralogous pair of genes: *LUX ARRHYTHMO* and *BOA* (Brother of *LUX ARRHYTHMO*). Whilst the orthologous gene in poplar has two CNSs, only one conserved sequence is found in each of the paralogs in Arabidopsis. This observation leads to the hypothesis that each conserved sequence contributes a part of the expression pattern. If these sequences function in a largely independent (additive) manner, then the joint expression pattern of the two paralogs in Arabidopsis may resemble the expression pattern of the single gene in poplar.

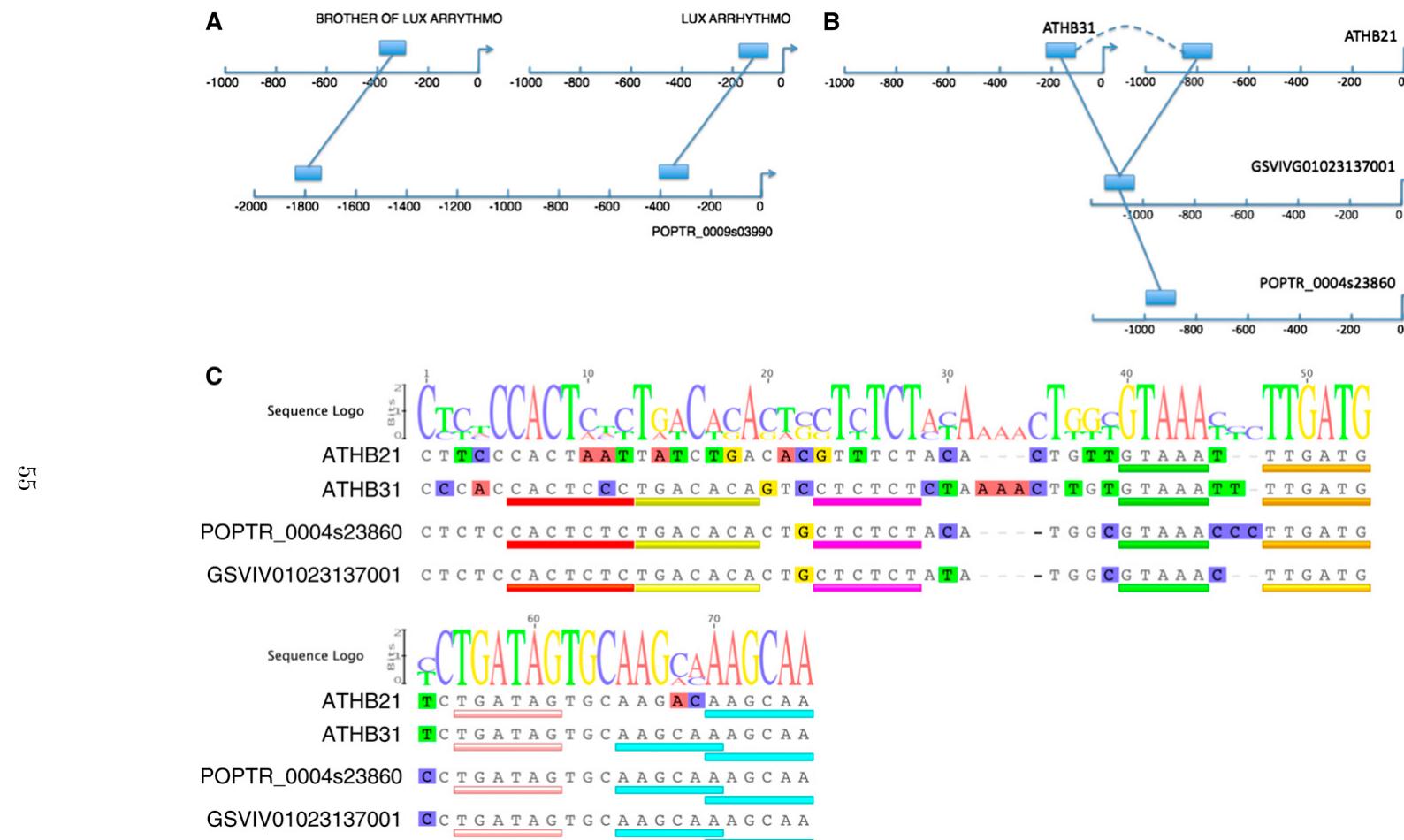


Figure 2.8: Analysis of the CNSs reveals potential subfunctionalisation in regulatory regions of *Arabidopsis* paralogs. (A) and (B) Positions of CNSs upstream of paralogous *Arabidopsis* genes and their orthologs. Arrows indicate TSS positions. Solid lines joining blocks indicate CNSs between orthologs, and dashed curved line in (B) indicates conservation between paralogs. (C) Alignment of CNS depicted in (B). Size of letters in the sequence logo indicates conservation of individual nucleotides. Colored bars indicate positions of potential binding sites based on alignment conservation (yellow, purple, green, and orange bars) and matches with known motifs (P300 in red, GATA in pink, and CBNAC in turquoise).

Additionally, two homeobox genes *ATHB21* and *ATHB31* were identified to be paralogous within Arabidopsis and to share common orthologous CNSs in two other species: grape (*GSVIVG01023137001*) and poplar (*POPTR\_0004s23860*) (Figure 2.8B). However, the pattern of conservation at the nucleotide level in the CNS shows the potential subfunctionalisation of binding motifs (Fig 2.8C). Within the most conserved 75 nucleotides, the poplar sequence has diverged from the grape sequence at only three positions, while the sequences upstream of *ATHB21* and *ATHB31* have diverged at 21 and 15 positions, respectively (Figure 2.8C). This is consistent with the hypothesis of reduced selective pressure at the loci of the Arabidopsis paralogs after gene duplication. The sequence upstream of *ATHB21* is particularly diverged, and conserved regions at four sites have been lost in this paralog, suggesting that the set of TFs binding this region has changed.

#### 2.4.2 Genome Availability and Annotation Quality Impact on CNS Predictions

Due to large-scale duplication events in plant genomes and subsequent neo-/ sub-functionalisation of paralogs, similarity between sequences does not implicitly determine orthology in all cases. Therefore, care must be taken when assigning orthologs, and improvements to orthology assignment methodology would enhance the detection of CNSs. Duplicated promoters may also acquire or lose individual *cis*-elements, meaning that even if orthology between genes is assigned correctly, their individual promoters may have undergone many evolutionary changes since they last shared a common ancestral sequence, leading to an absence of CNSs. The inclusion of multiple species for comparison in the present study improves the chances of finding CNSs in at least one species but may not be sufficient in all cases. As more fully sequenced genomes become available, the addition of more comparator species genomes (within an appropriate evolutionary distance from Arabidopsis) would further improve CNS detection using the method presented in this study. The accuracy of genome annotations is another factor that may affect the ability to detect CNSs, particularly with regard to correctly defining TSS positions. All genome annotations in this study have some degree of EST support, and for Arabidopsis, ~ 66% of genes have a defined 5' untranslated region (UTR) boundary (Chung et al., 2006). In cases where predicted gene models are not supported by full-length ESTs, our TSS position will correspond to the ATG. In this study, if the TSS is correctly annotated in Arabidopsis or correctly annotated in the comparator species, then this is sufficient to exclude discovery of CNSs within 5' UTR regions of any of the

orthologous genes in question. In some cases, however, a CNS may fall within a 5' UTR. CNSs and motifs embedded in the 5' UTR may still play a role in transcript regulation, for example, in modulating transcript abundance (Liu et al., 2010; Wang and Xu, 2010).

#### **2.4.3 CNSs Are Likely To Be Functional Regions Of ncDNA**

The study presented in this chapter is largely driven by the hypothesis that functional areas of the ncDNA sequence are under more selective pressure than their nonfunctional counterparts and therefore, functional regions evolve at a slower rate (Tagle et al., 1988). Working within the APPLES framework, 1865 CNSs present upstream of 1643 *Arabidopsis* genes (Table 2.2) were identified. At a high confidence 0.7 threshold (FPR:  $1.7 \times 10^{-3}$ ) a subset of 554 genes with 602 CNSs was identified and is believed to be part of the regulatory machinery shared among dicot plants. A strict control mechanism was developed such that 94% of the discovered CNSs at 0.7 threshold are expected to be true conserved sequences and not due to chance. Furthermore, a relatively new and high-throughput alignment plot method was used in this study to evaluate millions of alignment scores for all pairs of short sequence fragments, thereby providing a comprehensive and sensitive detection mechanism of alignment conservation. The evolutionary distance between *Arabidopsis* and the three other species used for comparison is reflected in the conversion from the raw alignment scores into conservation scores, providing additional enhancement to discover weakly conserved regions between multiple species.

#### **Positional Bias and Length of CNSs Are Indicators of Their Functional Importance**

Sequence conservation implies potential functional conservation, and several results presented in this study provide evidence for functional involvement of CNSs in transcriptional regulation. Firstly, CNSs show a clear and significant bias towards the first 100bp upstream from the annotated TSS, whereas the strict control set produced alignments with a uniform distribution in the promoter regions. The positioning of a subset of CNSs corresponds to the core promoter region (14% are within 50bp of the annotated TSS and contain a TATA-box motif), although the majority of the CNSs lie beyond the core promoter region ( $76\% > 50$  bp distant from the annotated TSSs). In general, the existence of CNSs upstream of the annotated TSS is consistent with the hypothesis that they contain embedded *cis*-regulatory elements to which TFs can bind. Comparing the alignment length distributions,

it was also found that alignments from orthologs (Figure A.2) are on average significantly longer than alignments from random gene pair sequences (Figure A.2), with mean lengths of 93 ( $\pm 25$  bp) and 66 ( $\pm 13$  bp) respectively. The difference in length distributions is consistent with the view that the CNS set, derived from real ortholog assignments, are meaningful sequences whose lengths are determined by the nature of their biological function (i.e., they contain multiple TF binding sites), while the latter set comprises randomly occurring alignments that are expected to be short in length. Transcriptional complexes are assembled from multiple proteins, and long stretches of conserved sequence will allow enough space for a number of these proteins to bind to the DNA in the regulatory region. Therefore, a significant length of CNS suggests complex function necessitating a large stretch of nucleotides to facilitate binding.

### TF Binding Motif Overrepresentation in CNSs

The CNSs from a high confidence threshold were tested for overrepresentation of known eukaryotic motifs compiled from TRANSFAC, JASPAR and PLACE databases. In order to reduce redundancy in the existing set of TF binding motifs, weight matrices representing them were clustered together and a representative member was selected from each one. A total of 728 eukaryotic motifs were tested for overrepresentation in the CNS set and 182 motifs were found to be significantly overrepresented. The CNSs contained 106% more matches for these motifs than an equivalent control set. A diverse range of motifs were found to be present and overrepresented in the CNS set, e.g. *ABF1*, *ABI4*, and GAGA elements (P-values  $< 1 \times 10^{-4}$ ). This result points to the functional nature of the CNSs being part of the transcriptional regulatory machinery, regulating the genes with which they are associated. In this study, only single motifs were tested for overrepresentation. However, it is common for a region of 100 bp to contain more than one TF binding site. Therefore, further insight could be gained by examining motif multiplicity and combinatorics. For example, two TF binding sites may not be overrepresented individually, but a combination of the two factors binding within a certain distance of each other might be. This would also be consistent with the “DNA-templated protein assembly” hypothesis put forward by Kaplinsky et al. (2002), whereby long CNSs can act as templates for the assembly of regulatory protein complexes that may not associate on their own.

Our findings are in slight contrast to the results obtained by Thomas et al. (2007), where CNSs found in paralogs are shorter, median length 25bp. Our study

uses a higher sensitivity method for alignment, which therefore provides an explanation for the discrepancy in the results. At the 0.3 threshold, the average length of CNSs obtained for paralogs by our method is 96bp. The short regions obtained by Thomas *et al.* may correspond to an individual TF binding, whereas CNSs found using our method contain multiple binding elements. For example, *TOC1* contains evening elements (experimentally proven to be necessary for circadian activity (Alabadí *et al.*, 2001)), as well as G/C-box elements and *DOF* binding sites (Picot *et al.*, 2010).

### Nucleosome Presence in CNSs

Further evidence for the functional nature of the CNSs presented in this study came from the statistical link found between a peak in the predicted nucleosome occupancy correlating with the location of CNSs in *Arabidopsis*. The reasons for such associations are not known, but it may be that since nucleosomes occlude underlying DNA sequences, and therefore restrict access to TF binding sites present in the CNSs, they are effectively turning off the associated genes. The CNSs presented here, with lower predicted nucleosome occupancy, are  $\sim$  93bp long and contain multiple binding sites within them providing support for the hypothesis based on the model of nucleosome-mediated cooperativity between TFs, whereby TFs can bind to nucleosomal/closed DNA subsequently displacing nucleosomes and making it further available for other TFs to bind, as described by Mirny (2010).

These findings suggest that multiple TF binding sites within a region no longer than the 147 bases occupied by one nucleosome may be required (and in terms of regulatory logic, essential) to displace nucleosomes, rendering their associated genes transcriptionally active. The prediction tool ((Kaplan *et al.*, 2009) was claimed to work well for the subset of the genome featuring well-positioned nucleosomes, though not in the majority of genomic sequence where nucleosome positioning is thought not to be determined by DNA motifs (Stein *et al.*, 2010). Hence the predictions are likely to be accurate if the CNS set overall contains features that reflect the intrinsic DNA sequence preferences of the nucleosome.

### GO Terms

The genes containing CNSs were subjected to GO term overrepresentation analysis in order to discover any potential biological and molecular functions associated with

the genes whose regulation is potential conserved across species. Strongly overrepresented GO terms for the different categories of regulatory processes, such as “DNA binding”, “promoter binding” and “regulation of transcription”, suggest that the genes found to contain CNSs are involved in the regulation of gene expression. Furthermore, by cross-referencing the list of 554 genes associated with high confidence CNSs reveals 208 (37.5%) genes that match the manually curated set of 2468 genes representing known TFs in Arabidopsis (unpublished). Additionally, genes containing high confidence CNSs contain a significantly high proportion of genes described as “Master Regulators” (Table 4.20n). Therefore, the findings presented here strongly suggest that the CNSs are part of the set of key transcriptional sequences, but also that the CNSs are associated with the genes more likely to be placed at the top of the transcriptional hierarchy in Arabidopsis. The finding that TFs tend to be CNS rich is consistent with a previous study of paralogs in Arabidopsis (Thomas et al., 2007), which was also noted to be true in grasses (Inada et al., 2003).

Other highly significant GO terms found to be overrepresented among the genes involved in development and morphological processes, related to organs, flowers and reproductive structures (Table 2.4). Development is a tightly regulated process, and as such there may be a strong selective pressure on the gene regulatory sequences. This hypothesis is supported by findings in the human genome that the most highly conserved noncoding sequences are associated with the developmental regulators, suggesting a key role in orchestrating early embryo development (Elgar and Vavouri, 2008). Our findings provide further support for this hypothesis, as developmental regulatory genes are in high abundance among the genes containing high confidence CNSs (72 genes).

### **The CNSs correspond to open chromatin areas in Arabidopsis**

DNase I sensitivity methods were originally developed to uncover short DNA binding sequences associated with particular protein-DNA interactions (Gross and Garrard, 1988; Urnov, 2003). DNase I hypersensitivity assays have recently been developed to outline regulatory regions along the promoter DNA sequences (Song and Crawford, 2010). From hypersensitive sites (HS) alone one cannot distinguish the particular proteins that can bind to DNA sequences, but one can predict active regions of various length thought to be involved in transcriptional regulation of the associated genes. This method has been successfully used to map protein-DNA interactions in yeast (Hesselberth et al., 2009) and human cells (Boyle et al., 2011) some of which

were validated using a ChIP-Seq approach. Using DNase-Seq data for Arabidopsis leaves (Zhang et al., 2012)), the average distribution of sequencing reads for high confidence CNSs showed that the conserved regions are more likely to be in transcriptionally active areas in the leaves of the plant than the random control set (Figure 2.7). Additionally, it is possible to estimate that 108 regions (29%) contain a much higher number of sequencing reads associated with them than the maximum average number of reads found in the corresponding control set. Although DNase-Seq is a snapshot of transcriptional activity in a particular cell type at a particular time point, nevertheless this result points to the functional nature of CNSs found in Arabidopsis.

#### 2.4.4 Effectiveness of Alignment-Based Methods in CNS Discovery

All functional genes are subject to transcriptional regulation, however not all genes are annotated with an identifiable promoter region upstream of the TSS. Additionally, previous studies have shown that regulatory regions may lie within the introns (Schauer et al., 2009) or the 3' UTR (Cawley et al., 2004). The findings presented here have been derived from the upstream promoter sequences. The method used can equally be applied to any orthologous or paralogous sequences that are thought to contain functionally conserved regulatory regions, and doing so may uncover more functional CNSs.

The high rate of “binding site turnover” means that regulatory elements mutate rapidly over evolutionary time but maintain their functional role, albeit having little sequence conservation, as demonstrated in *Drosophila* (Moses et al., 2006). Therefore, the approach presented here could not account for the full complement of conserved regulatory elements within plant genomes. However, the loss-free nature of the method means that it is able to find all alignment conserved sequences, when provided with the appropriate set of orthologous sequences for comparison. Alternative methods, such as the alignment-free model developed by Koohy et al. (2010) can be used in order to find elements that are functionally, but not alignment, conserved. Methods that combine comparative genomics with other resources, such as gene expression data (as in (Vandepoele et al., 2006; Heyndrickx and Vandepoele, 2012; Spangler et al., 2012)) are also useful in aiding the discovery and analysis of regulatory modules.

## 2.5 Conclusions

In this chapter, highly conserved noncoding sequences (CNSs) identified using a comparative genomic approach, were predicted to be involved in the transcriptional regulation of their associated genes. In addition, it was found that the CNS-associated genes themselves commonly had a role in transcriptional regulation. The finding that regulatory genes are themselves highly regulated makes biological sense; as plants rely on their regulatory machinery to integrate signals from internal and external stimuli to formulate a complex response, it is intuitive to put those genes under strict control. Taking into account the CNS length and binding site content, the prediction can be made that each gene is likely to have a number of regulators that can interact directly with the promoter DNA, and others that potentially operate indirectly through protein protein interactions with DNA-bound regulators. Furthermore, several thousand binding sites have been predicted to be mediating TF-gene links in the gene regulatory network of *A. thaliana*. The implication of this finding is that the strongly maintained CNSs and the genes they are associated with play an intrinsic role in the regulatory network that is shared among dicot plants.

## Chapter 3

# Elucidating Functional Elements and Gene Regulatory Network Using Yeast One-Hybrid Screens

### 3.1 Introduction

Over the next few decades the world population is predicted to grow by another 2.5 billion, to reach 9.5 billion. Sustaining a growing population means increasing crop yields in a sustainable fashion. One of the challenges associated with increased crop yields is increasing resistance of plants to a variety of biotic and abiotic stresses. Past research links increased stress resilience with decreased growth and yield (Herms and Mattson, 1992). Therefore understanding plant responses to stresses means that a better balance can be attained in the future crop with the help of genetic engineering. The necrotrophic fungus *Botrytis cinerea* accounts for an estimated 15-40% of harvest losses in grape varieties and 20-25% in strawberry crops and has a broad host range, infecting more than 200 plant species also including tomato (Elad et al., 2007a). Studies on the model organism Arabidopsis have identified a number of genes that play a major role in resistance to Botrytis, including PDF1.2 (Penninckx et al., 1996; Zimmerli et al., 2001), BOS1 (Mengiste et al., 2003) and PAD3 (Ferrari et al., 2007). However, many more genes and mechanisms that are involved in increased resistance to the fungus are yet to be characterised. Additionally, mechanisms of gene regulation are not well understood and simply increasing expression of defence resistant genes leads to adverse effects on the overall growth phenotype of

the plant (Clarke et al., 2001; Hua et al., 2001; Jambunathan et al., 2001). Similarly, gain-of-function double mutants restore growth but increase the susceptibility of the plant to a variety of infections (Shirano et al., 2002; Zhang et al., 2003). Therefore a better understanding of the regulation of gene expression would allow for fine tuning of the expression of disease resistant genes whilst minimising the negative effects such as reduced growth.

In the previous chapter putative regulatory regions along the promoter DNA in Arabidopsis were identified using the APPLES software package. The promoter DNA sequences of Arabidopsis were compared to the promoter sequences of a variety of other, closely and distantly related, plant species. Hundreds of CNSs were identified with varying degrees of conservation across multiple species and multiple lines of evidence point at the functional nature of the CNSs. The aim of this chapter is, firstly, to identify a small set of genes that are regulated by the same TF(s) using time-series mRNA expression profiles associated with the response to the infection with Botrytis. The promoters of the identified genes will be interrogated further for information determining their regulation using Yeast One-Hybrid (Y1H) library screens. The library screen allows a picture of the gene regulatory network (GRN) to be built from the bottom up, focusing on all possible protein-DNA interactions associated with particular promoters. TF(s) regulating selected genes may also serve as master regulators in the stress related network against Botrytis as a whole and therefore would be good targets for further experimentation. This knowledge would contribute to our understanding of the regulation of Botrytis defence responses by identifying direct protein-DNA interactions for differentially expressed genes regulated in the infection process.

### 3.1.1 Available Experimental Techniques For Probing Protein-DNA interactions

Different experimental approaches can be broadly divided into two separate groups; first, techniques identifying functional elements within promoter region of interest, by way of serial deletions. Shortened promoter sequences fused to reporter genes, e.g. *Luc* or  $\beta$ -glucuronidases (**GUS**), are tested *in planta*, where direct protein-DNA interactions are not known, or *in vivo* together with suspect TFs that are thought to be interacting with promoter of interest, thus validating the hypothesis. Second type of techniques focuses on known direct protein-DNA interactions, e.g. EMSA, or *de novo* discovery, e.g. Yeast One-Hybrid or ChIP assays.

## **Identification Of Functional Elements In A Promoter Of Interest**

In order to assess the activity of a gene promoter as a whole or a smaller part of it, the promoter DNA sequence associated with the gene of interest is fused to a reporter gene and used as an “indirect measure of gene activity. Reporter genes are derived from a variety of organisms and their enzymatic activity or fluorescent emissions are not usually found in most eukaryotes. Therefore, they serve as a good proxy to measure promoter activity and this activity is approximately proportional to transcriptional initiation frequency. One such technique involves gradually deleting/truncating promoter sequence in a 5' to 3' direction until reporter activity increases or decreases or is completely abolished. Serial promoter deletions have been used to uncover functional elements within a promoter of interest. For example, serial promoter deletions allowed the discovery of both necessary and sufficient regions in the promoter of *GC1* gene in Arabidopsis (Yang et al., 2008). However, if a functional core or enhancer element exists in the middle of the promoter fragment and functions under different conditions to the experimental set up, then it will be wrongly discarded as non-functional, using the serial deletions method. A more sophisticated version of the serial deletion technique uses the deletion or mutation of a certain feature (sequence) within the promoters of interest, for example, two *cis*-acting elements in the promoter of *rd29A* in Arabidopsis were found by promoter deletion and base substitutions using **GUS** promoter fusions as a reporter (Narusaka et al., 2003). A similar technique of promoter deletion and mutation was used to functionally dissect the G-box and novel coupling element (CE1) in the abscisic acid (ABA)-inducible gene *HVA22* in barley (Shen and Ho, 1995). In addition to gene activation, it is also possible to study gene repression using promoter fusion constructs. For example, promoters containing GAL4+GCC boxes were fused to the *LUC* reporter gene and had reduced expression in the presence of TF(s) containing the EAR repression domain, as compared to known activators of expression through a GCC box (Ohta et al., 2001).

The results obtained from full/partial/mutated promoter-reporter fusions provide information about the promoter activities when interacting TFs are known and controlled *in vivo* or not known *in planta*. However, a comprehensive GRN of all potential protein-DNA interactions using the promoter associated with the gene of interest and all Arabidopsis proteins is not possible in a high-throughput manner. The outcome of promoter-reporter fusions are the validation of known or predicted links in GRN, not *de novo* discovery of new links in the network. Additionally, only a small number of promoters can be screened at any one time, greatly limiting the

number of interactions that can be tested.

### **Electrophoretic Mobility-Shift Assay**

One of the earliest methods to test for direct protein-DNA interactions was Electrophoretic Mobility-Shift Assay (EMSA) (Garner and Revzin, 1981; Fried and Crothers, 1981). EMSA takes advantage of the different diffusion rates through the polyacrylamide gel for the protein bound and unbound to a DNA sequence. Larger, protein-bound DNA move at a slower speed in the polyacrylamide gel as compared to the corresponding protein-free DNA fragments. If left for 1.5h-2h period, bound and unbound samples will produce bands at different heights, indicating the presence and rough nature of the protein(s) bound to the DNA. For example, Arabidopsis genes *ABF1* and *ABF3* have exhibited sequence-specific binding activity to the G/ABRE motif *in vitro* (Choi et al., 2000). EMSA provides a robust and sensitive way to study direct protein-DNA interactions, including studying protein complexes interacting with DNA sequences using improved EMSA (Deckmann et al., 2012). However, in order for a TF to be tested using EMSA, it has to first be isolated from nuclear extracts, which is a cumbersome and time consuming process. Therefore, limiting the number of different TFs that can be tested simultaneously. EMSA is an excellent technique for confirming predicted protein-DNA interactions, however, it is not suitable for high-throughput *de novo* interaction discovery.

### **Chromatin Immunoprecipitation**

Another powerful technique in assessing the binding potential of a TF to a DNA sequence are Chromatin Immunoprecipitation (ChIP) assays (Gilmour and Lis, 1984). Proteins present are first cross-linked to the chromosomal DNA by formaldehyde. Cross-linked protein-DNA complexes are then extracted using nuclear extraction methods and chromosomal DNA is sheered by sonication into small fragments. Sonoicated fragments are immunoprecipitated using an antibody specific for the protein of interest. After reverse cross-linking, short DNA fragments are quantified using qPCR (ChIP-qPCR). Direct genome-wide targets of *DELLA* in Arabidopsis were found using the ChIP-qPCR technique (Zentella et al., 2007). Instead of qPCR, sonicated DNA can be sequenced on massively parallel scale, also providing genome-wide targets for the protein of interest (ChIP-Seq (Kaufmann et al., 2010)). The ChIP-Seq technique has been recently been utilised to map interactions of the core

Arabidopsis circadian clock genes (Huang et al., 2012). The ChIP-Seq technique can also identify specific regulatory sequences which the TF is able to bind. Unlike other methods described thus far, ChIP techniques provide a genome-wide snapshot of direct targets of a protein of interest and are therefore able to identify downstream targets of a TF in a GRN. However, the major limitation of the ChIP approach is the requirement for existence of protein specific antibodies in order to be able to pull out protein cross-linked with DNA. ChIP assays can also be used in studying protein complexes comprising of two or more proteins (Re-ChIP-Seq, (Ross-Innes et al., 2010)), where antibodies are available for all members of the complex. Additionally, ChIP assays cannot be performed in a high-throughput manner, in terms of the number of different TFs tested, even if antibodies do exist, since it is not possible to identify which TF binds in a given location with a large number of reads. Therefore, ChIP is not appropriate when all potential interactions are to be tested for a DNA sequence of interest.

### **Yeast 'n'-Hybrid Screens**

Various Yeast 'n'-Hybrid technologies have been available for over two decades. This method of screening was first formally described by Fields and Song Fields and Song (1989) to detect protein-protein interactions (Yeast Two-Hybrid). In order to discover if protein X and Y interact, each one is independently fused with two other proteins, one containing a DNA-binding (DB) domain which allows for the creation of a DB-X fusion to the promoter of the reporter gene; and a second protein containing the activation domain (AD) that allows for the polymerase machinery to assemble and transcribe the reporter gene, typically *lacZ*. The reporter gene is only transcribed when DB-X and Y-AD fusions interact with each other, specifically when X interacts with Y since DB and AD do not interact on their own. When X-Y interaction takes place, the AD comes close enough to the promoter DNA to allow for the transcriptional machinery to assemble. Modern variations use *HIS3* and *LEU2* gene products as powerful growth selection markers. The Arabidopsis protein-protein interactome map has been developed using this Y2H approach (Arabidopsis Interactome Mapping Consortium, 2011). Further systems were developed for RNA-protein interactions (Yeast Three-Hybrid (Sengupta et al., 1999)). Moreover, counter-selection was designed to study particular residues that were important for the DNA binding (Reverse Yeast Hybrid Systems (Vidal, 1997)). Instead of a reporter gene, a toxic gene is introduced and yeast only grows when protein-, DNA- or RNA-protein interactions do not take place.

An extension by simplification of the original Two-Hybrid screen, was to altogether remove the DB-X fusion and use TF-AD instead of the Y-AD complex (Yeast One-Hybrid, Y1H) (Meijer et al., 1998). Instead of protein-protein interactions, the Y1H system is designed to probe for direct protein-DNA interactions, as TFs already contain DNA-binding domains and putative DNA sequences can be fused to the reporter gene *HIS3* or *lacZ*, serving as bait for TFs. Thus, yeast only grows on the selective media when the TF of interest interacts with the promoter DNA. The original Y1H screen was modified for high-throughput screening by PRESTA, where instead of a single TF, a number of different TFs are tested in the same well and growing colonies are sequenced to identify interacting TFs(Ou et al., 2011). Pruneda-Paz *et al.* observed that the TCP TF could interact with a fragment of the *CCA1* promoter in a library Y1H screen. This binding was validated *in vitro* by EMSA and subsequently *in planta* by ChIP-PCR (Pruneda-Paz et al., 2009). High-throughput Y1H screens have also been used to start mapping human protein-DNA interactions (Reece-Hoyes et al., 2011a), although the study in human TFs uses “enhanced” Y1H (eY1H) where both *lacZ* and *HIS3* are used as reporter constructs simultaneously (Reece-Hoyes et al., 2011b).

### 3.1.2 Biological Context

Evidence for transcriptional regulation is always associated with certain external conditions, chemical treatments or specific cell types. However, it is not clear how an interaction would behave under context that differs from the original study. For example, *ARF2* gene in Arabidopsis has been shown to target genes in seedlings (Vert et al., 2008) and transgenic *ARF2* knock-out plants have been shown to be more susceptible to infection with Botrytis. The biological context of seedlings, or infection with Botrytis, may affect the genes *ARF2* is able to transcriptionally regulate. Therefore, targets identified in seedlings are “out of context”, when considered during the infection with Botrytis. Alternatively, evidence obtained from *in vitro* experiments, such as Y1H assays, is “context-free”, i.e. it is not known under what conditions this interaction will take place, if any at all.

In summary, there are a number of experimental techniques available to test protein-DNA interactions. Promoter-reporter fusions provide information about the functional nature of the promoter as a whole and are geared towards testing known or predicted novel protein-DNA interactions (Ohta et al., 2001). However,

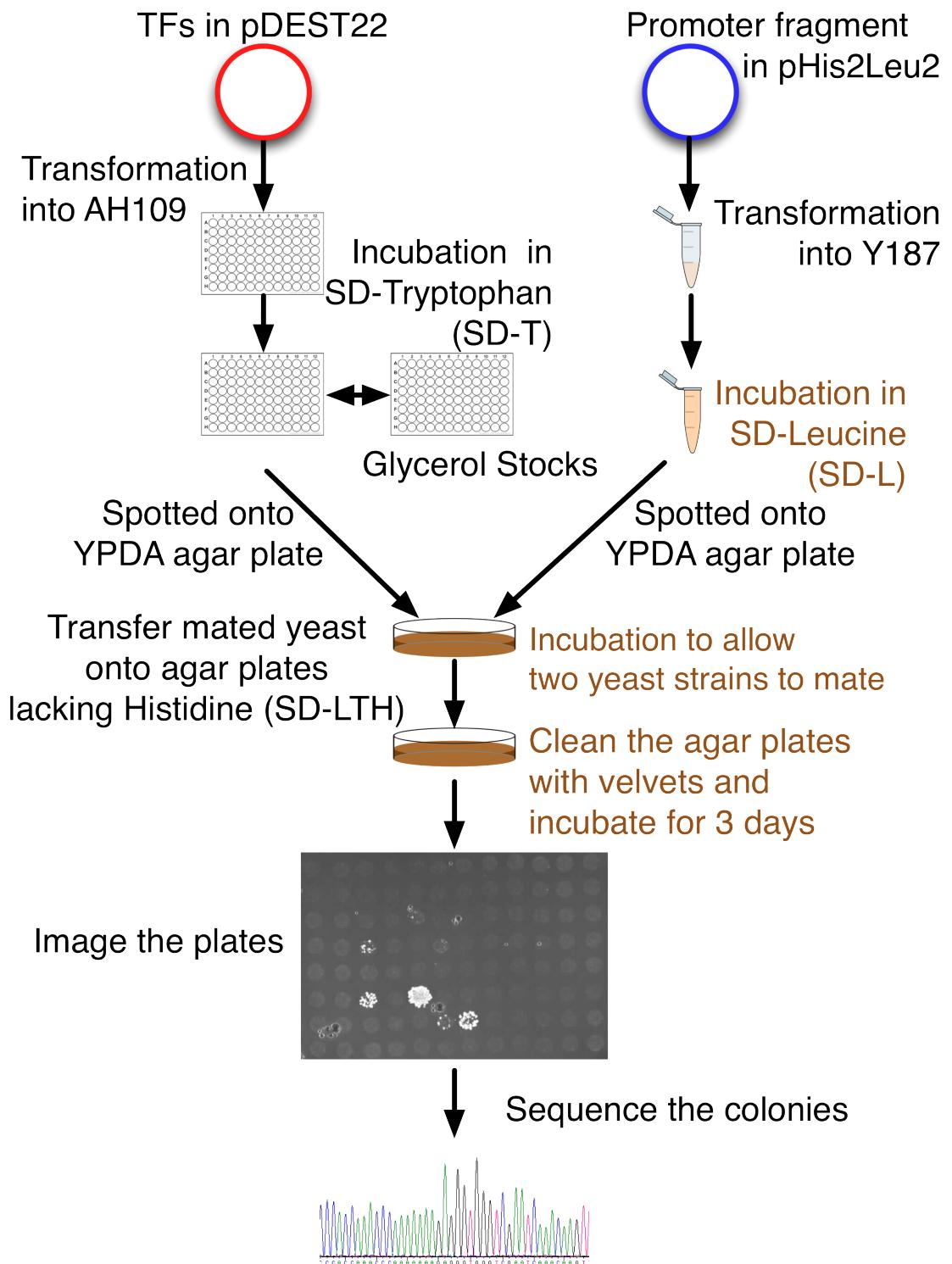


Figure 3.1: Diagram of the steps in the Y1H library screen, adapted from Ou et al. (2011).

the technique does not allow high-throughput *de novo* interaction discovery. Similarly, EMSA has high specificity of direct protein-DNA interactions (Choi et al., 2000) but is limited to short binding site sequences and only allows a limited number of TFs to be tested simultaneously. Alternative techniques such as ChIP allow the determination of genome-wide binding events for a single TF, provided that an antibody exists (Zentella et al., 2007), however it cannot be scaled up in terms of the number of different TFs being tested without losing specificity. ChIP methods are appropriate when all potential targets of a single TF need to be established. On the other hand, the pooled Y1H technique permits multiple promoters to be screened against over a thousand TFs simultaneously and in a high-throughput manner, in order to discover *de novo* interactions. This technique has been successfully used and validated in the past (Pruneda-Paz et al., 2009) and additionally has been adapted to map human protein-DNA interactions (Reece-Hoyes et al., 2011a).

The aim of this body of work is to screen a large number of promoter (30) associated with genes that are hypothesised to be regulated by the same TF(s). Genes will be selected based on their time-series expression profiles in response to infection with Botrytis. A collection of genes that have very similar expression pattern are more likely to be co-regulated through the same mechanism and the same TFs, than simply co-expressed together (Allococo et al., 2004). Therefore, the promoters of the selected genes will be screened using a pooled library screen of over 1300 Arabidopsis TFs previously cloned and transformed into yeast. Results from the library screen will be validated in a pairwise Y1H screen to increase the confidence in new GRNs.

## 3.2 Methods

### 3.2.1 Gene Selection

#### Gene Selection using WIGWAMS

Wigwam clustering method was originally developed by Jo Rhodes (unpublished) and is used to cluster genes across multiple time-series, for a potential multi-stress response. A single gene is used as a seed and remainder of the genome is compared for similarity based on the mRNA expression levels during each stress. Pearson correlation coefficient is calculated and genes are ordered in descending order of similarity using the correlation coefficient. Ordered gene clusters are generated for each individual stress. Increasing number of top genes from each cluster are statistically compared for significant overlap between clusters, if any, using hypergeometric test statistics. The analysis was performed by Jo Rhodes for all genes in *Arabidopsis* genome as a seed. Top 100 clusters were used to select for further analysis and potential Yeast One-Hybrid experiments. All of 100 clusters were analysed manually to determine suitable candidate for experiment taking into account the complexity of the expression patterns in stresses as well as amount of overlap with other stresses and nature of the overlapping genes, as determined by GO annotations, in order to maximise the potential for multi-stress functionality.

### 3.2.2 Yeast One-Hybrid

#### Cloning promoter fragments with Gateway homologous recombination

1000bp upstream from TSS of each gene were obtained using Biomart portal (Smedley et al., 2009), and manually inspected for a likely arrangements of 400bp fragments using SeqMan software, part of DNASTAR package.(Burland, 2000). Primers around the designated regions were designed to incorporate Gateway *attB* sites on either end, forward Primer: 5-GGGG-ACA-AGT-TTG-TAC-AAA-AAA-GCA-GGC-TNN-(template specific sequence)-3; reverse primer: 5-GGGG-AC-CAC-TTT-GTA-CAA-GAA-AGC-TGG-GTN-(template specific sequence)-3 and synthesised by IDT.

Promoter regions of the selected gene sets were amplified from genomic DNA (Col4, prepared by Alex Tebrett) using oligonucleotides in Table 3.1 and KOD HotStart polymerase (Roche Diagnostics, Welwyn) according to the manufacturer's instructions. After 2 step PCR process 5 µl of resulting product were combined with 5 µl of loading buffer and run on 2% Agarose Gel at 110V for 35 mins to check for presence of the band at 400bp using HyperLadder I(Invitrogen,

Gene ID	Primer Name	Direction	Sequence	Fragment ID
Gene ID	Primer Name	Direction	Sequence	Fragment ID
At5G50570	SABR-3072	S	5'-AAAAAAAGCAGGCCCTCCCTAATTGACGGTCATAAAGAGCAG-3'	Y1H_139
At5G50570	SABR-3073	AS	5'-CAAGAAAGCTGGTCGGGAGAACGCTATAAAAGTCTTCATG-3'	Y1H_139
At5G50570	SABR-3074	S	5'-AAAAAAAGCAGGCCCTCGTTAGGACAATTATGGG-3'	Y1H_140
At5G50570	SABR-3075	AS	5'-CAAGAAAGCTGGTCCTTGATCTTTACTGACCCATTATCC-3'	Y1H_140
At5G50570	SABR-3076	S	5'-AAAAAAAGCAGGCCCTCCCAATTATTGTTCATTTCATC-3'	Y1H_141
At5G50570	SABR-3077	AS	5'-CAAGAAAGCTGGGTCTGATGATAATTAGCTTACTAAGTAAG-3'	Y1H_141
At5G0590	SABR-3078	S	5'-AAAAAAAGCAGGCCCTCCATTATAGTATTAAATCATATAATAG-3'	Y1H_142
At5G0590	SABR-3079	AS	5'-CAAGAAAGCTGGGTCTTAGTCTCATTTGTAAGATAAATCTTC-3'	Y1H_142
At5G0590	SABR-3080	S	5'-AAAAAAAGCAGGCCCTCGTTAGGATTAGGTGGAAGAAAAG-3'	Y1H_143
At5G0590	SABR-3081	AS	5'-CAAGAAAGCTGGCTCTTCTTAAGTTCTGATTGATTC-3'	Y1H_143
At5G0590	SABR-3082	S	5'-AAAAAAAGCAGGCCCTCGTGAAGTCATTAGAAATAAGCAAATC-3'	Y1H_144
At5G0590	SABR-3083	AS	5'-CAAGAAAGCTGGGTCTGAAAGAGACTGACAGATTC-3'	Y1H_144
At4G31550	SABR-3084	S	5'-AAAAAAAGCAGGCCCTCCAAAGAAATAATCGAAATTTCG-3'	Y1H_145
At4G31550	SABR-3085	AS	5'-CAAGAAAGCTGGGTCCAACTAGATTACGATTAACATTC-3'	Y1H_145
At4G31550	SABR-3086	S	5'-AAAAAAAGCAGGCCCTCGTAAAGTAAGCTCAAATTATTCATC-3'	Y1H_146
At4G31550	SABR-3087	AS	5'-CAAGAAAGCTGGGTCTCTTAACAAAATATTCAACTTAG-3'	Y1H_146
At4G31550	SABR-3088	S	5'-AAAAAAAGCAGGCCCTCCAAATTCCAGCTGGCCCTTCTC-3'	Y1H_147
At4G31550	SABR-3089	AS	5'-CAAGAAAGCTGGTCGGAGAAAGAGAAGAGAAGAGATGCG-3'	Y1H_147
At3G25780	SABR-3090	S	5'-AAAAAAAGCAGGCCCTCGAAATAAGGACAAATGATGCGTAC-3'	Y1H_148
At3G25780	SABR-3091	AS	5'-CAAGAAAGCTGGTCGATCCACATCATGGTAATCATG-3'	Y1H_148
At3G25780	SABR-3092	S	5'-AAAAAAAGCAGGCCCTCGAGCTGCTGATAAAAAAGAGTGG-3'	Y1H_149
At3G25780	SABR-3093	AS	5'-CAAGAAAGCTGGGTCTCTGGTCGGTCTGTCGTCATTG-3'	Y1H_149
At3G25780	SABR-3094	S	5'-AAAAAAAGCAGGCCCTCCGAATAGAATTGTTGATACTAGTGG-3'	Y1H_150
At3G25780	SABR-3095	AS	5'-CAAGAAAGCTGGGTCTTGTGAGTTAGTAAATGAGTCTATT-3'	Y1H_150
At3G25760	SABR-3096	S	5'-AAAAAAAGCAGGCCCTCGAAGATTAGATTGCAACCTATTG-3'	Y1H_151
At3G25760	SABR-3097	AS	5'-CAAGAAAGCTGGTCAGAAAGAACATAAAACTCCAAAAC-3'	Y1H_151
At3G25760	SABR-3098	S	5'-AAAAAAAGCAGGCCCTCGGTTTCAGCCAATAACGGCGTCG-3'	Y1H_152
At3G25760	SABR-3099	AS	5'-CAAGAAAGCTGGTCCTCCCACATTATTTAATTAGATAGACATC-3'	Y1H_152
At3G25760	SABR-3100	S	5'-AAAAAAAGCAGGCCCTCGTCACTAAACAAACTATTATC-3'	Y1H_153
At3G25760	SABR-3101	AS	5'-CAAGAAAGCTGGTCGAGTTTACGAAATGTCATGTG-3'	Y1H_153
At3G23250	SABR-3102	S	5'-AAAAAAAGCAGGCCCTCGAAATAAAATGTCGAGGAAATTTAG-3'	Y1H_154
At3G23250	SABR-3103	AS	5'-CAAGAAAGCTGGTCGAGATAAAATTAGAGATTGTATG-3'	Y1H_154
At3G23250	SABR-3104	S	5'-AAAAAAAGCAGGCCCTCTCCTTAAAGACTAAAGATGCGTC-3'	Y1H_155
At3G23250	SABR-3105	AS	5'-CAAGAAAGCTGGTCCTACTATTATCATGTCGAAATTG-3'	Y1H_155
At3G23250	SABR-3106	S	5'-AAAAAAAGCAGGCCCTCGAAATAGAAATACAAAACGTAC-3'	Y1H_156
At3G23250	SABR-3107	AS	5'-CAAGAAAGCTGGTCATAATTATCATGTCGATGAGATG-3'	Y1H_156
At2G44840	SABR-3108	S	5'-AAAAAAAGCAGGCCCTCCGATTAGTTTATTGTTAAATGG-3'	Y1H_157
At2G44840	SABR-3109	AS	5'-CAAGAAAGCTGGTCGATCATCTTGTGCGATTGTTAG-3'	Y1H_157
At2G44840	SABR-3110	S	5'-AAAAAAAGCAGGCCCTCCCTTATATTGCTCTCTCTC-3'	Y1H_158
At2G44840	SABR-3111	AS	5'-CAAGAAAGCTGGTCGCTGTTCTTGATATTGTAACCC-3'	Y1H_158
At2G44840	SABR-3112	S	5'-AAAAAAAGCAGGCCCTCGATTTTGGTAGTACAGATAGGCCAC-3'	Y1H_159
At2G44840	SABR-3113	AS	5'-CAAGAAAGCTGGTCGAAGAGATAAGTAGTTGTTAG-3'	Y1H_159
At2G35930	SABR-3114	S	5'-AAAAAAAGCAGGCCCTCGAAGACCATAACAAAATTATCCTC-3'	Y1H_160
At2G35930	SABR-3115	AS	5'-CAAGAAAGCTGGTCGTAAGTATTATTAATCAACAAATG-3'	Y1H_160
At2G35930	SABR-3116	S	5'-AAAAAAAGCAGGCCCTCGACCCATGTCGTTATATGGTTAG-3'	Y1H_161
At2G35930	SABR-3117	AS	5'-CAAGAAAGCTGGTCGGTTGACTTTCAAGAGAGATTG-3'	Y1H_161
At2G35930	SABR-3118	S	5'-AAAAAAAGCAGGCCCTCGACACAAAGCAGACAGTACACTC-3'	Y1H_162
At2G35930	SABR-3119	AS	5'-CAAGAAAGCTGGTCGAGGAAGAGAGAAGAGGAGTTGGG-3'	Y1H_162
At1G19180		S	5'-AAAAAAAGCAGGCCCTCTTCTTTAGGGGACCACTCACTAAC-3'	Y1H_172
At1G19180		AS	5'-CAAGAAAGCTGGTCCTATTATAAGTATTAACAGCGTG-3'	Y1H_172
At1G19180		S	5'-AAAAAAAGCAGGCCCTCGTGGGTTGACTTTGATGTATGAC-3'	Y1H_173
At1G19180		AS	5'-CAAGAAAGCTGGTCACGTAACGTTAGGCATAATTCTCG-3'	Y1H_173
At1G19180		S	5'-AAAAAAAGCAGGCCCTCGCTTATTATAACAAAAG-3'	Y1H_174
At1G19180		AS	5'-CAAGAAAGCTGGTCACACAAAGCTATATTATAATAG-3'	Y1H_174
At1G80840		S	5'-AAAAAAAGCAGGCCCTCCGTCACGATGGTATCGTCATAATGTC-3'	Y1H_175
At1G80840		AS	5'-CAAGAAAGCTGGTCCTAGATTTTTCAGACAATAATTATG-3'	Y1H_175
At1G80840		S	5'-AAAAAAAGCAGGCCCTCATGTAATCAACCAATTCTTATC-3'	Y1H_176
At1G80840		AS	5'-CAAGAAAGCTGGTCGTTAACAAACATTGGTGTG-3'	Y1H_176
At1G80840		S	5'-AAAAAAAGCAGGCCCTCGCAACTAACCCGACAGAAATGTC-3'	Y1H_177
At1G80840		AS	5'-CAAGAAAGCTGGTCGAAGAGAGAAAAGATTGGTTC-3'	Y1H_177

Table 3.1: Primers used to amplify corresponding fragments from genomic DNA Col4 previously prepared by Alex Tebrret.

Paisley). Samples with the band present were purified using QIAquick PCR purification kit according to the manufacturer's instructions. Purified PCR products and donor vector containing the *pHist2Leu2* plasmid (provided by Claire Hill) were combined using homologous recombination together with BP®Clonase (Invitrogen, Paisley) and incubated for 2hrs at 25°C C. 1 µl of the reaction were added to 10 µl defrosted α-Select gold efficiency competent cells (Bioline, London), mixed gently and incubated on ice for 30 mins. Cells were heat shocked at 42°C C for 30 sec then left to rest on ice for 2 mins. 250 µl of SOC media Table 3.2 was added to the cell and incubated with vigorous shaking at 37°C C for 2 hrs. 150 µl of incubated cells were transferred onto LB agar (Sigma-Aldritch, Gillingham) plates containing Zeocine®(25 µg/µl). Plates with cells were incubated overnight at 37°C C. Colony PCR was performed on up to 8 colonies for each transformation with Taq polymerase according to the manufacturer's instructions (with oligonucleotides: forward - 5'-CTATCAGGGCGATGGCCCCTA-3', reverse - 5'-AATGCACTAACGATTAGCG-3') to check for presence of the insert. Insert positive colonies were grown overnight in LB containing Zeocine®(25 µg/µl) at 37°C on a vigorous shaker. The plasmids were extracted with QIAprep miniprep kit (Qiagen, West Sussex) and sequenced using the primers above according to the manufacturer's instructions. 1 µl of correctly sequenced plasmids was used for transformations into destination vector, *pHist2Leu2* (provided by Claire Hill) and remaining plasmids were stored at -20°C C. 1 µl of sequence verified plasmids was combined with 1 µl of vector containing *pHist2Leu2* plasmid and 11 µl of LR®Clonase (Invitrogen, Paisley) and incubated for 2hrs at 25°C C. 1 µl of the reaction were added to 10 µl defrosted α-Select gold efficiency competent cells (Bioline, London), mixed gently and incubated on ice for 30 mins. Cells were heat shocked at 42°C C for 30 sec then left to rest on ice for 2 mins. 250 µl of SOC media Table 3.2 was added to the cell and incubated with vigorous shaking at 37°C C for 2 hrs. 150 µl of incubated cells were transferred onto LB agar (Sigma-Aldritch, Gillingham) plates containing Kanamycin (50 µg/µl). Plates with cells were incubated overnight at 37°C C. Two colonies were grown overnight in LB containing Kanamycin (50 µg/µl) at 37°C on a vigorous shaker. The plasmids were extracted with QIAprep miniprep kit (Qiagen, West Sussex) and sequenced using oligonucleotides according to the manufacturer's instructions. Sequence verified plasmids were stored at -20°C. Table 3.3 is a summary of the cloning experiments for "WRKY" cluster genes.

Table 3.2: SOC media

Reagents (Sigma-Aldrich, Gillingham)
2% (w/v) bacto-tryptone (20 g)
0.5% (w/v) bacto-yeast extract (5 g)
8.56 mM NaCl (0,5 g)
2.5 mM KCl (0.186 g)
10 mM MgCl <sub>2</sub> (0.952 g)
20 mM glucose (3.603 g)
ddH <sub>2</sub> O to 1000 mL

### USA plasmid preparation

Additionally, “Jaz” and “TCP” cluster genes were screened in to USA, courtesy of S.Kay, in order to maximise number of promoter fragments screened. Therefore primers were designed around the appropriate fragments and promoters were amplified and cloned using the same protocol into entry vector (*pDonorZeo*, see above). The genes from these two clusters were destined for yeast one-hybrid screen in the USA and therefore were transformed into USA destinations vectors (*pGlacZI* and *pPGA59NglucGW*). At the time, it was not known which of the two destination vector will be used for screen, therefore both were used and decision could be made at later stage on the correct plasmids to use. Table 3.4 and table 3.5 provides a summary of the transformation into *pGlacZI* and *pPGA59NglucGW* vectors.

ATG	Y1H ID	PCR Product	pDonorZeo	pHistLeu
<i>At5G50570</i> <b>SPL13</b>	Y1H-139			
	Y1H-140			
	Y1H-141			
<i>At5G50590</i> <b>HSD4</b>	Y1H-142			
	Y1H-143			
	Y1H-144			
<i>At4G31550</i> <b>WRKY11</b>	Y1H-145			
	Y1H-146			
	Y1H-147			
<i>At3G25780</i> <b>AOC3</b>	Y1H-148			
	Y1H-149			
	Y1H-150			
<i>At3G25760</i> <b>AOC1</b>	Y1H-151			
	Y1H-152			
	Y1H-153			
<i>At3G23250</i> <b>MYB15</b>	Y1H-154			
	Y1H-155			
	Y1H-156			
<i>At2G44840</i> <b>ERF13</b>	Y1H-157			
	Y1H-158			
	Y1H-159			
<i>At2G35930</i> <b>PUB23</b>	Y1H-160			
	Y1H-161			
	Y1H-162			
<i>At1G19180</i> <b>JAZ1</b>	Y1H-172			
	Y1H-173			
	Y1H-174			
<i>At1G80840</i> <b>WRKY40</b>	Y1H-175			
	Y1H-176			
	Y1H-177			

Table 3.3: Cloning summary of fragments that were screened at Warwick laboratory. Green represents successful transformation into the specified product.

ATG	Y1H ID	PCR	pDonorZeo	<i>pGlacZI</i>	<i>pPGA59NglucGW</i>	NOTES
<i>AT5G08470</i> <b>PEX1</b>	Y1H-118					16bp insert
	Y1H-119					
	Y1H-120					
<i>AT4G35450</i> <b>AFT</b>	Y1H-121					
	Y1H-122					
	Y1H-123					
<i>AT4G17600</i> <b>LIL3:1</b>	Y1H-124					
	Y1H-125					
	Y1H-126					
<i>AT3G16570</i> <b>RALF23</b>	Y1H-127					
	Y1H-128					Extra 'A'
	Y1H-129					
<i>AT2G41940</i> <b>ZFP8</b>	Y1H-130					
	Y1H-131					
	Y1H-132					
<i>AT1G19000</i>	Y1H-133					
	Y1H-134					
	Y1H-135					
<i>AT5G23280</i>	Y1H-163					
	Y1H-164					
	Y1H-165					
<i>AT1G09030</i> <b>NF-YB4</b>	Y1H-166					Done by Peijun Zhang
	Y1H-167					
	Y1H-168					

Table 3.4: Cloning summary for the first cluster destined for the USA. Green represents successful transformation into the specified product.

#### Small scale transformation of yeast strain *Y187* with *pHist2Leu2* plasmids

An  $\alpha$  strain of *s. cerevisiae*, *Y187* was grown overnight in 10 ml of YPDA (Clontech, Saint-Germain-en-Laye) at 30°C C on a vigorous shaker. 1 ml of the culture was centrifuged at 400 g for 5 mins for each ten transformations. Cells were resuspended in 1ml of 0.1 M LiAc, centrifuged and resuspended in 1ml of 0.1 M LiAc. Cells were incubated at 30°C C in water bath for 1 hour.

1  $\mu$ g of promoter fragments in *pHis2Leu2* was combined with 40  $\mu$ g of denatured salmon sperm carrier DNA (Clontech, Saint-Germain-en-Laye) and mixed with 290  $\mu$ l 50% (v/v) polyethylene glycol (PEG) 3350. The DNA/PEG mix was heated to 30°C C.

100  $\mu$ l of cell suspension was added to the DNA/PEG mix and mixed gently. Cell/DNA/PEG

ATG	Y1H ID	PCR	pDonorZeo	<i>pGlacZI</i>	<i>pPGA59NglucGW</i>	NOTES
<i>AT5G13220</i> <b>JAZ10</b>	Y1H-169					
	Y1H-170					
	Y1H-170					
<i>AT5G55120</i> <b>VTC5</b>	Y1H-354					
	Y1H-355					Missing 'A'
	Y1H-356					
<i>AT5G13550</i> <b>SULTR4:1</b>	Y1H-357					
	Y1H-358					Error in attB
	Y1H-359					
<i>AT4G30530</i>	Y1H-360					
	Y1H-361					
	Y1H-362					
<i>AT4G01850</i> <b>SAM2</b>	Y1H-363					
	Y1H-364					
	Y1H-365					
<i>AT2G22330</i> <b>CYP79B3</b>	Y1H-366					
	Y1H-367					
	Y1H-368					
<i>AT2G04400</i>	Y1H-369					
	Y1H-370					
	Y1H-371					
<i>AT1G72450</i> <b>JAZ6</b>	Y1H-372					
	Y1H-373					
	Y1H-374					
<i>AT1G51760</i> <b>IAR3</b>	Y1H-428					
	Y1H-429					
	Y1H-430					
<i>AT1G44350</i> <b>ILL6</b>	Y1H-431					
	Y1H-432					
	Y1H-433					

Table 3.5: Cloning for the second cluster destined for the USA. Green represents successful transformation into the specified product.

mix was incubated at 30°C C for 50 mins in water bath. Cells were heat shocked by incubation in 42°C C for 10 mins and then centrifuged at 1000 g for 5 mins. Supernatant was removed, and cells resuspended in sterile water were spread on SD minus Leucine (SD-L; minimal SD and amino acid dropout supplements from Clontech, Saint-Germain-en-Laye) agar plates and incubated at 30°C C until colonies appeared, typically 2 days.

### **Transcription factor library subculture**

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

### **Pooled library yeast one-hybrid by mating**

*S. cerevisiae* cultures, of Y187 strain that had been transformed with the promoter fragments containing *pHist2Leu2* plasmids, were made in 10 ml of SD-L (minimal SD amino acid dropout supplements from Clontech, Saint-Germain-en-Laye) media and incubated overnight on a shaker platform at 30°C. 3ul of the overnight culture was spotted onto each gridspot of a 96-well arrangement on a YPDA (Clontech, Saint-Germain-en-Laye) agar plate. 3ul of each well of the transcription factor library subcultured were spotted on top of the Y197 spots, at the corresponding library grid positions. Yeast were allowed to mate overnight by incubation at 30°C. YPDA (Clontech, Saint-Germain-en-Laye) agar plates were replicated using velvets onto agar plates containing the following growth media (minimal SD and amino acid dropout supplements from Clontech, Saint-Germain-en-Laye):

- SD minus Leucine and Tryptophan (SD-LT) - To check that both yeast strains have successfully mated
- SD minus Leucine, Tryptophan and Histidine (SD-LTH) - Selection plate
- SD-LTH with various concentration of 3-Amino-1,2,4-triazole (3AT). - Inhibit auto-activation levels.

Plates containing replicated mated spots were incubated at 30°C overnight. Then the plates were cleaned with up to 3 velvets before being incubated at 30°C for 3-4 days. Finally, the plates were imaged with the upper white light in a G:BOX (SynGene, Cambridge). Up to 5 growing colonies on SD-LTH and SD-LTH+3AT agar plates were picked and re-streaked onto the same selective media and grown in 30°C incubator until colonies appeared. A small amount, picked with 10 µl pipette tip was picked into 10 µl of 20mM NaOH on a 96-well PCR plates. Then the plate was shaken, sealed and incubated at 99°C for 10 mins. Then colony PCR was performed on 1.2 µl of the boiled yeast extracts using Taq polymerase (Invitrogen) (oligonucleotides: forward - 5'-CTAACGTTCATGATAACTTCATG-3; reverse - 5'-GAAGTGTCAACAAACGTATCTACC-3') according to the manufacturer's instructions. PCR products were cleaned using MultiScreen HTS PCR 96-well plate (Millipore, Watfod) according to the manufacturer's instructions. Cleaned PCR products were sequenced to identify interacting TFs, using the forwards oligonucleotides above and a BigDye® Terminator v3.1 cycle sequencing kit (Applied Biosystems, Warrington) according to the manufacturer's instructions. Colony PCR, purification and sequencing of the PCR products was done by Alison Jackson. Sequencing results were analysed using a custom written BLAST script (provided by Laura Baxter) against TAIR10 annotations.

### 3.2.3 Image Based Positive Result Inference

Images of the plates containing yeast were taken using upper white light setting in G:BOX (SynGene, Cambridge) in grey scale at the end of the Y1H screen (see above). Growing yeast colonies are lighter, whiter, in colour as compared to the background of the plate. Therefore, if the area where the yeast was spotted could be correctly identified, each one could have been statistically compared to the positive and negative controls to establish if there is a significant result for each spot. Each one is compared to negative control, this establishing bound of prediction. Comparing against negative control also allows to take into account general background growth associated with the auto-activation on a plate. A template was prepared and used to overlay and align on top of the cropped images using Microarray Profile plugin in ImageJ (Schneider et al., 2012) and grey scale histograms, in range [0 – 255, were retrieved using the same plugin from all spots in the template grid. The same size area is used to avoid normalisation due to variable spot size. From this intensity histogram, cumulative distribution of the intensities could be constructed by summing values from original histogram in increasing order from zero. Each cumulative histogram has the same domain but not the same range, as

sum of intensities is different between images, which makes comparison difficult and not very effective at this stage. Therefore, each cumulative histogram is normalised to range [0 – 1] by dividing through by the total intensity for each spot. As a result, all histograms have the same domain ([0 – 255]) and range, moreover, the range is [0 – 1] similar to the cumulative probability density function from probability theory. Two cumulative probability density functions can be compared in a hypothesis testing, where *null* hypothesis is that two cumulative distributions are the same and alternative hypothesis that they are not the same. Therefore, each spot is compared to the cumulative distribution of negative control to find if growth at that spot differs enough from background to be considered a positive result. A custom MATLAB (MATLAB, 2012) script was written to retrieve histograms for each spot in turn across all repeats, combine them by calculating an average across all repeats, then comparing cumulative distribution of the histogram for TF and negative control, present on the plate, using two-sample Kolmogorov-Smirnov test (*kstest2* function). The method allows to combine multiple trials into a single hypothesis test. Images from separate experiments can be “merged” together and analysed all at the same time. There are many different ways of combining images, after the spots have been isolated and histograms of intensities are computed, they are averaged across all trials to create an “average” histogram. When comparing against the negative control spot, number of trials is taken into account and “average” histogram for negative controls is also computed from the spots corresponding to the separate experiments. This way, only appropriate control and test spots are compared against each other for significance. The test outputs P-value associated with the probability of two cumulative distributions being the same. Additionally, maximum vertical distance between two distributions,  $\Delta d$ , is returned as one of the intermediate results of the test. P-value, or  $\delta d$ , can be used to convert to our internal scale of one to ten as it is monotonically increasing.

### 3.2.4 pairwise mini-library Y1H screen

#### Sequence verification of library TFs

Positive interactions from library screen were extracted from bacterial glycerol stocks by dipping a pipette tip and streaking onto LB agar (Sigma-Aldrich, Gillingham) plates containing Kanamycin (50 µg/µl), then incubated overnight at 37°C C, until colonies appeared. One colony from each plate was grown in LB (Sigma-Aldrich, Gillingham) containing Kanamycin (50 µg/µl) overnight at 37°C C with vigorous shaking. The plasmids were extracted with QIAprep miniprep kit (Qi-

agen, West Sussex) and sequenced using the oligonucleotides primers according to the manufacturer's instructions. Sequencing data was aligned against CDS sequence of expected TFs for corresponding well using BioMart (Smedley et al., 2009) and SeqMan (Burland, 2000), Table 3.11.

### **Large scaletransformation of yeast strain Y187 with *pDest22* plasmids**

An  $\alpha$  strain of *s. cerevisiae*, AH109 was grown overnight in 10 ml of YPDA (Clontech, Saint-Germain-en-Laye) at 30°C C on a vigorous shaker. 50  $\mu$ l of this culture was used to inoculate 200 ml of YPDA (Clontech, Saint-Germain-en-Laye) sufficient for two 96-well transformations. 200 ml of the cultures were grown in 21 flasks overnight at 30°C C with vigorous shaking. Newly grown cultures were centrifuged in 50 ml Falcon tube at 2500 rpm, YPDA discarded. Cells were resuspended in 5 ml of sterile water and combined into single tube. Then combined cells were centrifuged at 2500 rpm for 5 mins at room temperature. the cells were washed with 0.1 M LiAc, peletted as above and resuspended in 2 ml of 0.1 M LiAc. 1  $\mu$ l of lasmids containing TFs in *pDest22* were aliquoted into 96-well PCR plate using multichannel pipette. AH109 cell suspension was added to the TRAFCO mix, Table 3.6, and mixed gently by dispersing. TRAFCO/cell mix was aliquoted into each well in 96-well plate containing TF plasmids using multichannel pipette and shaken vigorously for 5 mins to mix. 100  $\mu$ l of 50% polyethylene glycol (PEG) 3350 was added to each well, sealed with adhesive foil and incubated at 42°C C for 1 hour with vigorous shaking. Then, cells were pelleted by centrifuging at 2500 rpm for 5 mins at room temperature. TRAFCO mix was gently removed using multichannel pipette, such that pelleted cell did not move. 14  $\mu$ l of sterile water were added to each well of to 96-well plate to resuspend the cells. 5  $\mu$ l of the cell suspension mix was spotted on SD minus Tryptophan ((SD-T; minimal SD and amino acid dropout supplements from Clontech, Saint-Germain-en-Laye) agar plates and incubated at 30°C C until colonies appeared, typically 2 days. Once colonies are visible corresponding well in 96-deep well plate containing 500  $\mu$ l of SD minus Tryptophan (SD-T; minimal SD and amino acid dropout supplements from Clontech, Saint-Germain-en-Laye) media were innoculated using 96-deep well replicator (V and P Scientific Inc, San Diefo) and grown at 30°C C for 2 days prior to screen. 100  $\mu$ l of grown cells were mixed together with 100  $\mu$ l of sterile 50% glycerol, covered with adhesive foil and stored at -80°C C.

Table 3.6: TRAFCO mix

Reagents (Sigma-Aldrich, Gillingham)
3ml 1M LiAc
1ml sterile water
4ml 2 µg/ml denatured salmon sperm carrier DNA (ssDNA)

### Pairwise yeast one-hybrid by mating

*S. cerevisiae* cultures, of Y187 strain that had been transformed with the promoter fragments containing *pHist2Leu2* plasmids, were made in 10 ml of SD-L (minimal SD amino acid dropout supplements from Clontech, Saint-Germain-en-Laye) media and incubated overnight on a shaker platform at 30°C. 3 µl of the overnight culture was spotted onto each gridspot of a 96-well arrangement on a YPDA (Clontech, Saint-Germain-en-Laye) agar plate. 3µl of each well of the transcription factor mini-library subculture were spotted on top of the Y197 spots, at the corresponding library grid positions. Yeast were allowed to mate overnight by incubation at 30°C.

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

- SD minus Leucine and Tryptophan (SD-LT) - To check that both yeast strains have successfully mated
- SD minus Leucine, Tryptophan and Histidine (SD-LTH) - Selection plate
- SD-LTH with various concentration of 3-Amino-1,2,4-triazole (3AT). - Inhibit auto-activation levels.

Plates containing replicated mated spots were incubated at 30°C overnight. Then the plates were cleaned with up to 3 velvets before being incubated at 30°C for 3-4 days. Finally, the plates were imaged with the upper white light in a G:BOX (SynGene, Cambridge). Up to 5 growing colonies on SD-LTH and SD-LTH+3AT agar plates were picked and re-streaked onto the same selective media and grown in 30°C incubator until colonies appeared. Positive spots were referenced against the mini-library template to establish ATG of the positive interactor.

### 3.3 Results

#### 3.3.1 Initial Gene Selection using WIGWAMS tool

The aim of the gene selection process was to select candidate subsets of genes that are likely to be co-regulated together and form tight clusters across multiple stresses, for a possible multi-stress response. WIGWAMS (Rhodes et al, unpublished) was developed to select gene clusters based on the similarity of mRNA expression patterns across multiple stresses (see Methods). The top 100 clusters were identified to contain significantly overlapping genes across two or more stresses (data not shown). The clusters were manually pruned for complex patterns of expression and for those genes containing stress-related Gene Ontology (GO) annotation terms. A final set of two clusters was chosen to determine the likely Protein-DNA interactions from the promoter fragments of the selected genes Table 3.7 and Table 3.8.

ATG	Name	ReMo
<i>At1g72450</i>	<b>JAZ6</b>	
<i>At1g44350</i>	<b>ILL6</b>	
<i>At1g51760</i>	<b>IAA</b>	
<i>At2g04400</i>	<b>IGPS</b>	
<i>At2g22330</i>	<b>CYP79B3</b>	
<i>At4g01850</i>	<b>SAM2</b>	
<i>At4g30530</i>		
<i>At5g13220</i>	<b>JAZ10</b>	
<i>At5g13550</i>	<b>SULTRA4</b>	
<i>At5g55120</i>		

Table 3.7: “JAZ” cluster selected using WIGWAM for multi stress response. Green boxes indicate presence of a CNS within the first 1000bp of the gene promoter.

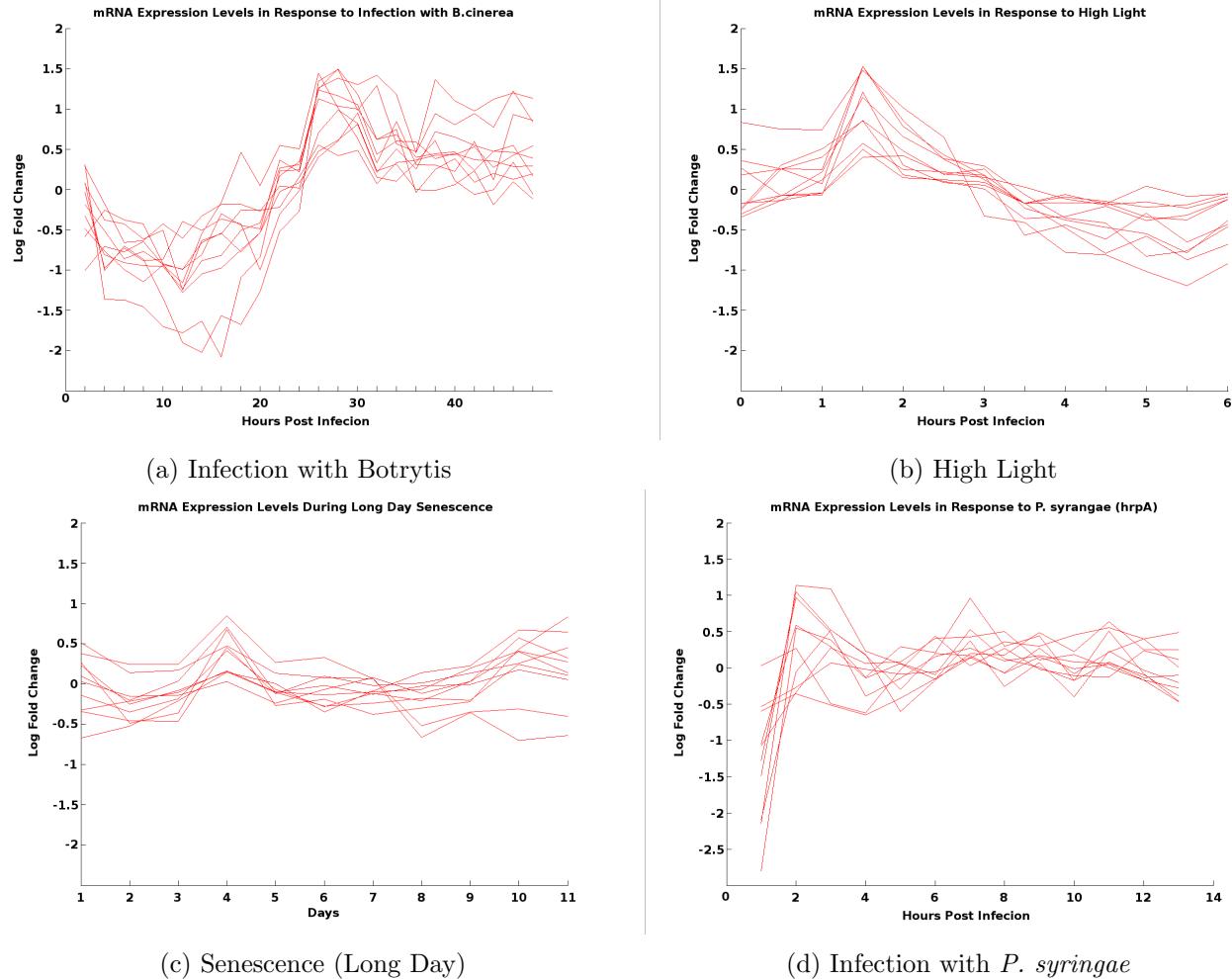


Figure 3.2: mRNA expression levels of genes selected using WIGWAM in response to multiple stresses in “JAZ” cluster.

ATG	Name	ReMo
<i>At5g23280</i>	<b>TCP</b>	
<i>At4g17600</i>	<b>LIL3</b>	
<i>At1g19000</i>	<b>MYB</b>	
<i>At2g41940</i>	<b>ZFP8</b>	
<i>At4g35450</i>	<b>AKR2</b>	
<i>At1g09030</i>	<b>NF-YB4</b>	
<i>At5g08470</i>	<b>PEX1</b>	
<i>At3g16570</i>	<b>RAFL23</b>	

Table 3.8: “TCP” cluster selected using WIGWAM for multi stress response. Green boxes indicate presence of a CNS within the first 1000bp of the gene promoter.

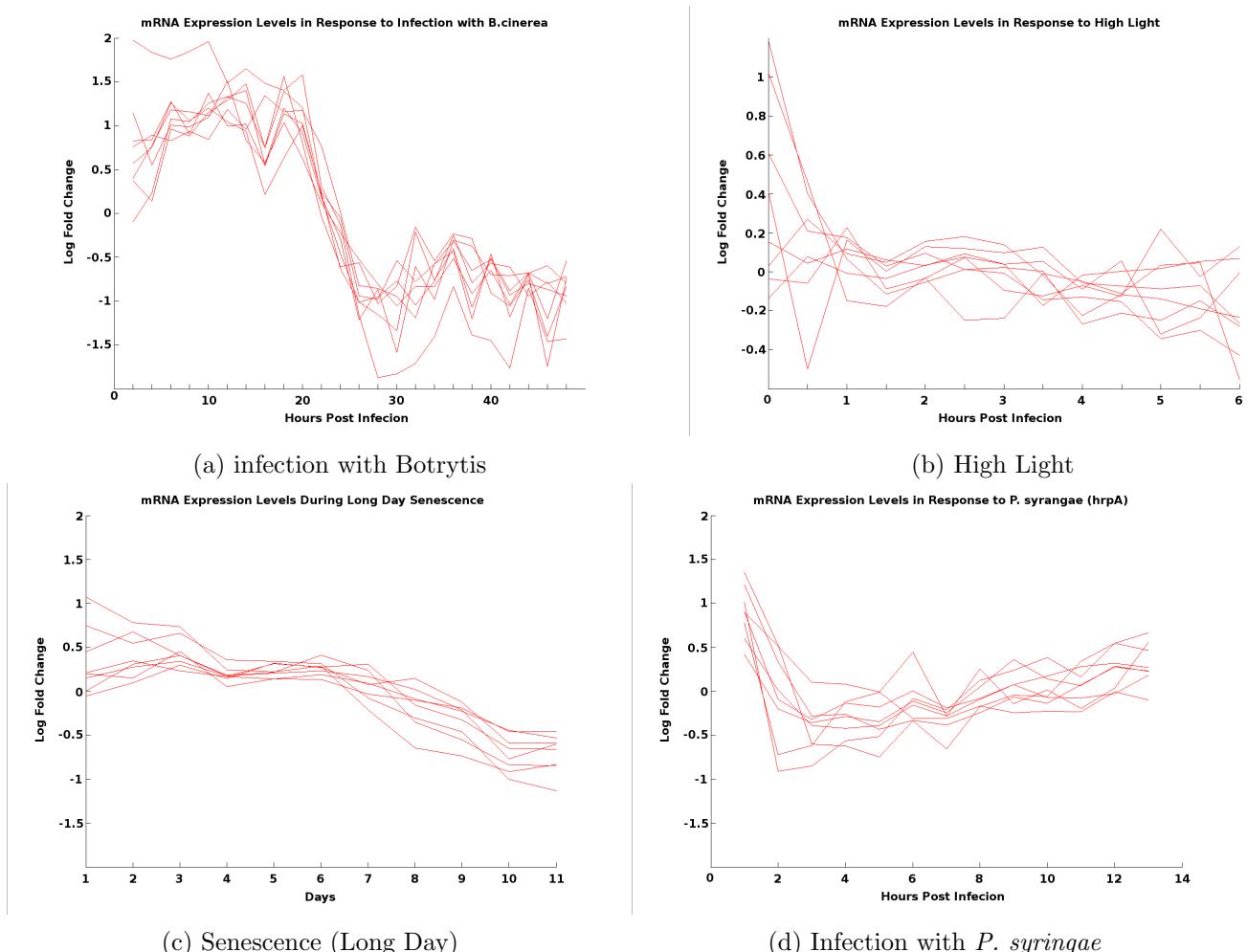


Figure 3.3: mRNA expression levels of genes selected using WIGWAM in response to multiple stresses in “TCP” cluster.

The “JAZ” cluster was selected based on the broad correlation of expression in four stresses: Botrytis infection, *P. syringae* infection, high light and long day senescence. The second cluster, “TCP”, was selected on the similarity of expression in the same four stresses as the “JAZ” cluster, however it differs in the pattern of expression of these genes. For example, during the infection with the necrotroph Botrytis, genes in the “JAZ” cluster are up-regulated after approximately 24 hours post infection, whereas genes in the “TCP” cluster are sharply down-regulated 24 hours post infection.

### 3.3.2 Clock Regulated Stress Response

Involvement of the plant’s circadian clock has been implicated as being coupled with the response to stress (Wang et al., 2011). Promoter fragments in *A thaliana* containing recently identified ReMos(Baxter et al., 2012) were analysed for the presence of the *CCA1* binding site 5'-AA[A/C]AATC[T/A]-3' (Andronis et al., 2008). If the binding site was found to be present in the potentially evolutionary conserved modules, it has a higher chance of still being functional. Two genes, *WRKY40* and *WRKY11*, were identified as having this site present (ReMo belief score > 0.9). mRNA expression profiles of these two genes were analysed across all time-course datasets available from PRESTA project. Two *WRKY* genes had a unique temporal expression profile in response to infection with *textitB.cinerea*. Moreover, both genes were differently expressed as compared to mock infection. Expression in other stresses was not differentially expressed for these two genes.

Expression profile of the two *WRKY* genes was used as a baseline and statistically compared against the remainder of the genome using Pearson’s correlation coefficient. Eight more genes were identified as being significantly correlated with the *WRKY* profiles.

Based on the presence of the *CCA1* binding sites and high correlation within the set of ten genes, these genes are predicted to be regulated by *CCA1* directly or indirectly by the circadian clock. Promoters of the ten genes will be screened to identify the underlying GRN using a Y1H library screen.

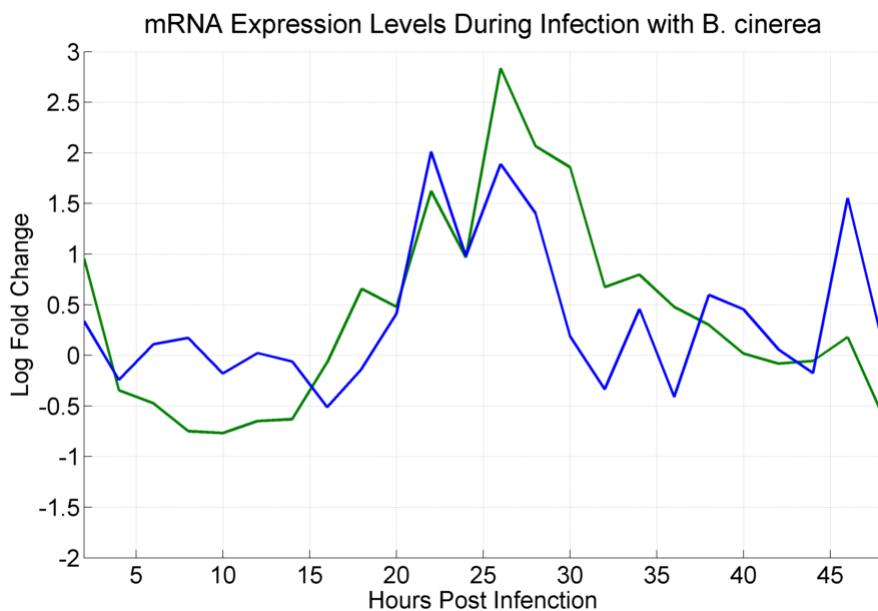


Figure 3.4: Distinct temporal expression mRNA expression profiles of *WRKY40* (green) and *WRKY11* (blue) during infection with *B. cinerea*

ATG	Name	ReMo
<i>At5g50570</i>	<b>SPL13</b>	
<i>At5g05090</i>		
<i>At4g31550</i>	<b>WRKY11</b>	
<i>At3g25780</i>	<b>AOC3</b>	
<i>At3g25760</i>	<b>AOC1</b>	
<i>At3g23250</i>	<b>MYB15</b>	
<i>At2g48440</i>	<b>ERF13</b>	
<i>At2g35930</i>	<b>PUB23</b>	
<i>At1g19180</i>	<b>JAZ1</b>	
<i>At1g80840</i>	<b>WRKY40</b>	

Table 3.9: Cluster of genes selected using WRKY40 and WRKY11 expression profiles in response to infection with Botrytis. Green boxes indicate presence of a CNS within the first 1000bp of the gene promoter.

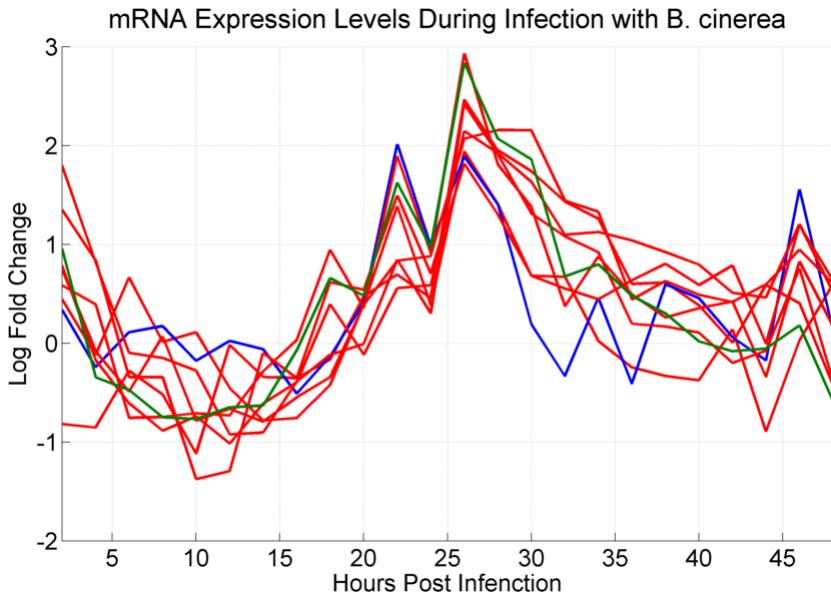


Figure 3.5: mRNA expression levels of the genes in the “WRKY” cluster in response to infection with *B. cinerea*. Green - WRKY40, blue - WRKY11

### 3.4 High-throughput identification of direct protein-DNA interactions using Y1H library screen

Promoter constructs of ten genes predicted to be regulated by the circadian clock were divided into three overlapping, 400bp fragments, spanning 1000bp upstream from the TSS of each gene. The fragments were designed to take into account existing ReMos, if any, as potential functional promoter regions, such that the conserved sequence was included entirely as part of at least one fragment. Primers for 30 fragments were designed to incorporate *attB* recombination sites on the ends of the promoter fragments (see Methods). The fragments were amplified from genomic DNA of Col-4 (extracted by Alex Tebrett previously), transformed into a Gateway entry vector (pDonorZeo), sequence verified to contain the correct fragment and transformed into a Gateway destination vector (pHisLeu). One fragment was amplified but could not be successfully transformed into the Gateway entry vector after several attempts. The remaining 29 fragments were screened against a library of Arabidopsis TFs in yeast (see Methods). Up to five colonies from each positive interaction were re-streaked onto selective media; two colonies from each positive set were sequenced to establish the identity of the interacting TF. The results of the library screen are summarised in Table B.1 for each gene.

The library screen has not revealed a master regulator(s) that could potentially interact and therefore regulate, all or the majority of the genes thought to be co-regulated. *ANAC092* and *ANAC038/309* had the largest coverage of the tested genes, binding to the promoter fragments of three different genes. Similarly, *ORA47*, *WRKY53*, *ZFP2*, *ANAC102* and *ATHB25* were found to directly interact with the promoters of two genes. The remaining interactions were associated with a single gene.

### 3.4.1 Subscreen of TCP Transcription Factors using Y1H library screen

Screens conducted prior to these experiments have shown that TEOSINTE BRANCHED1, CYCLOIDEA and PCF (TCP) have consistently given positive results for interactions in almost all screens performed to date. The reasons for this are unclear, but TCPs have been implicated in plant development and leaf differentiation processes (Koyama et al., 2010, 2007) which may explain the large number of positive results since they are involved in basic developmental processes. Concerns were also raised that the positive TCP interactions may obscure other interactions in the same wells as the TCPs. As a result TCPs were taken out from the main library and put into a separate library. This additionally reduces the sequencing costs, since the TCPs were no longer repeatedly sequenced since the TCP mini-library is screened pairwise and therefore interactions can be determined immediately. Each fragment is screened against the TCP mini-library alongside the common libraries. Table 3.10 shows the summary of all the fragments screened against the TCP TFs. As the results of the whole library screens were accumulating, it soon became apparent that not all TCPs were successfully removed, since some well locations consistently gave a positive result.

#### Sequence Verification of TFs in the Y1H library

Mini-libraries used in the pairwise screen only contain a single TF in each well, unlike library screen where up to 24 TFs are pooled in each well. Therefore, sequencing colonies is not required at the end of the mini-library screen since the identity of detected interactors are known for each well, unlike in the library screen in which the presence of multiple TFs means that at least two colonies are sequenced to establish the specific interacting TF/s. This presents an opportunity to additionally verify CDS sequences of the TFs prior to testing for positive interaction in the pairwise

screen. All unique positive results were extracted from glycerol stocks in bacteria, sequenced and compared to the published gene models. Table 3.11 shows a summary of all TFs used in the pairwise screen. As the results suggest, not all proteins have the correct sequences associated with them.

### 3.4.2 Confirmation of Observed Y1H Interactions Using Pairwise Screens

In order to verify interactions from the library screen, all unique TFs have been pooled to create a mini library. Plasmids (*pDEST22*) of previously verified TFs were transformed into a different yeast strain (AH109). Each promoter was tested individually against the mini-library arranged in a 96-well plate, with a single TF in each well. Additionally, the pairwise screen enables any missing interactions to be established that have not been previously detected in the library screen. The pairwise screen was performed a total of three times to increase confidence in the results. After the first pairwise screen, the library was reduced to contain only interactors giving positive results from the pairwise screen. The second and third screens would be tested against the reduced mini-library.

ATG	Fragment ID	TCP																									
		1	2	3.1	3.2	4.1	4.2	5	6	7	8	9	10.1	10.2	11	12	13	14	15	16	17	18	19	20	21	22	23
AT5G50570	Y1H-139	Green	Green	Red	Green	Green	Green	Green	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT5G50570	Y1H-140	Green	Green	Red	Green	Green	Green	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT5G50570	Y1H-141	NOT SCREENED																									
AT5G05090	Y1H-142	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT5G05090	Y1H-143	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT5G05090	Y1H-144	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT4G31550	Y1H-145	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT4G31550	Y1H-146	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT4G31550	Y1H-147	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT3G25780	Y1H-148	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT3G25780	Y1H-149	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT3G25780	Y1H-150	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT3G25760	Y1H-151	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT3G25760	Y1H-152	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT3G25760	Y1H-153	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT3G23250	Y1H-154	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT3G23250	Y1H-155	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT3G23250	Y1H-156	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT2G44840	Y1H-157	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT2G44840	Y1H-158	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT2G44840	Y1H-159	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT2G35930	Y1H-160	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT2G35930	Y1H-161	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT2G35930	Y1H-162	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT5G13220	Y1H-163	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT5G13220	Y1H-164	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT5G13220	Y1H-165	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT1G19180	Y1H-166	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT1G19180	Y1H-167	NOT SCREENED																									
AT1G19180	Y1H-168	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT1G19180	Y1H-169	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT1G80840	Y1H-170	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT1G80840	Y1H-171	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT1G80840	Y1H-172	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT1G80840	Y1H-173	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT1G80840	Y1H-174	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT1G80840	Y1H-175	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT1G80840	Y1H-176	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
AT1G80840	Y1H-177	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red

Table 3.10: All fragments were screened separately against a mini-library of TCP, with one TCP in each well. The TCPs were taken out from the main library due to their constant strong interactions. TCP 3.2, 4.1, 4.2, 14, 15, 16, 20 and 23 interact with almost every fragment and are hypothesised to be interacting in sequence independent manner as general TFs. Green represents interaction seen during the screen, red represents lack of any colonies during the screen.

ATG	Name	Location	CORRECT	Notes	ATG	Name	Location	CORRECT	Notes
AT3G12890	AML2	001-E03	██████	STOP	AT3G18960		013-E10	██████	
AT2G46160		001-G04	██████	5 mutations, 2 insertions	AT3G06760		013-F02	██████	internal primers
AT1G21960		001-H03	██████	9 mutations	AT4G30180		013-H09	██████	
AT4G37730	AtbZIP7	002-A04	██████		AT4G11140	CRF1	014-A03	██████	
AT1G50640	ERF3	002-B10	██████		AT4G39250	ATRL1	014-D02	██████	last 26nt are wrong
AT5G61270	PIF7	002-D11	██████		AT5G64340	SAC51	014-H02	██████	
AT3G12910		003-G07	██████	6 mutations	AT1G74930	ORA47	015-H01	██████	
AT4G32040	KNAT5	003-H02	██████		AT4G39100	SHL1	016-B03	██████	
AT1G66350	RGL1	004-B03	██████	STOP, 1 mutation	AT1G73360		016-G04	██████	AT5G41400 is located in this position
AT5G66770		004-B10	██████	STOP, need internal primers	AT1G12860	ICE2	017-D07	██████	
AT5G17490	RGL3	004-B11	██████	STOP, 1 mutation, need internal primers	AT1G02680	TAF13	017-E02	██████	
AT4G17920		004-G03	██████	1 mutation	AT5G24930	COL4	019-B12	██████	>10 mutations
AT1G68520		004-H07	██████		AT3G47600	MYB94	019-D2	██████	
AT5G21120	EIL2	005-G03	██████	STOP, internal primers	AT4G14410		019-E2	██████	
AT4G42060		006-B01	██████	STOP, 1 mutation, internal primers	AT4G13010		019-G7	██████	starts from 2nd 'ATG'
AT2G46830	CCA1	006-E09	██████	2 mutations, internal primers	AT2G22750		001-F01	██████	
AT5G11260	HY5	006-H10	██████	STOP, 4 mutations	AT2G41710		003-H06	██████	internal primers
AT2G17600		007-A01	██████	>10 mutations	AT5G60200	TMO6	004-F04	██████	
AT4G37790	HAT22	007-D02	██████	STOP, 1 deletion	AT2G34000		006-F11	██████	
AT4G29080	IAA27	007-E02	██████	STOP	AT4G05100	MYB74	008-D04	██████	
AT5G67190		007-F12	██████	STOP, 1 deletion	AT1G57560	MYB50	009-D05	██████	
AT2G03710	AGL3	008-B06	██████		AT3G01140	MYB106	009-E06	██████	wrong annotation
AT3G02310	AGL4	008-D06	██████		AT1G59640	ZCW32	009-H02	██████	1 mutation
AT1G06850	AtbZIP51	008-D06	██████	1 mutation	AT4G32890	GATA9	011-A11	██████	1 mutation
AT3G27785	MYB118	009-A09	██████		AT5G49300	GATA16	011-B12	██████	1 mutation
AT5G62320	MYB99	009-C11	██████		AT3G53600		011-D08	██████	
AT3G49690	MYB84	009-D08	██████		AT3G15540	IAA19	011-G11	██████	STOP
AT3G13540	MYB5	009-F06	██████		AT3G20310	ERF7	013-B08	██████	
AT3G58120	AtbZIP61	010-E07	██████		AT1G06160	ORA59	013-C04	██████	
AT5G06500	AGL96	010-F08	██████	1 mutation	AT4G36780	BEH2	013-G10	██████	4nt insert, internal primers
AT2G28340	GATA13	011-B03	██████	STOP, 2 mutations	AT2G18300		014-B11	██████	>10 mutations
AT1G66140	ZFP4	011-C04	██████	1 mutation	AT4G38900		015-H12	██████	internal primers
AT1G24625	ZFP7	011-C06	██████	STOP	AT1G76110		017-B07	██████	internal primers
AT4G25470	DREB1C	011-E08	██████	STOP, 1 mutation	AT4G31660		007-B03	██████	not sequenced
AT2G41070	AtbZIP12	011-E12	██████	STOP, 1 mutation	AT2G46270	GBP3	011-F10	██████	not sequenced
AT2G33310	IAA13	011-G06	██████	STOP, internal primers	AT5G42630	KAN4	013-F06	██████	not sequenced
AT3G04730	IAA16	011-G08	██████	STOP	AT1G75390	AtbZIP44	015-F05	██████	not sequenced
AT3G61830	ARF18	012-D02	██████	3 mutations, internal primers	AT5G17810	WOX12	017-D06	██████	not sequenced
AT2G33860	ARF3	012-H02	██████	STOP, internal primers	AT5G63890	AtHDH		██████	NOT IN THE LIBRARY
AT4G13620		013-C11	██████	1 mutation	AT2G28240			██████	UNKNOWN

Table 3.11: Majority of the TFs had correct sequence within them or incorrect "STOP" codon at most. Prior to assembling TFs into a mini-library, we took this opportunity to verify sequences of the TFs. Mutations within coding sequence of a gene may lead to misfolding, rendering DNA-binding domain inactive or, conversely, resulting in false positive results. "STOP" - 'STOP' codon was incorrect at the end of the coding sequence, "need internal primers" - internal primers are needed to verify sequence inside the coding region as current sequencing primers did not cover the entire sequence.

	<b>ORA59</b>	bZIP	PIF7	ESE	ATERF14	ATHB25	WRKY29	WRKY21	WRKY15	ANAC098	ATERF7
	AT1G06160	AT4G38900	AT5G61270	AT3G23220	AT1G04370	AT5G65410	AT4G23550	AT2G30590	AT2G23320	AT5G53950	AT3G20310
AT5G50570	SPL13										
AT5G05090	MYB										
AT4G31550	WRKY11										
AT3G25780	AOC3										
AT3G25760	AOC1										
AT3G23250	MYB15										
AT2G44840	ERF13										
AT2G35930	PUB23										
AT1G19180	JAZ1										
AT1G80840	WRKY40										

	AtbZIP52	ATHB22	HMG	AtHB23	WRKY75	WRKY41	ANAC038	WRKY8	WRKY28	WRKY57	bHLH
	AT1G06850	AT2G36610	AT1G76110	AT5G39760	AT5G13080	AT4G11070	AT2G24430	AT5G46350	AT4G18170	AT1G69310	AT2G22750
AT5G50570	SPL13										
AT5G05090	MYB										
AT4G31550	WRKY11										
AT3G25780	AOC3										
AT3G25760	AOC1										
AT3G23250	MYB15										
AT2G44840	ERF13										
AT2G35930	PUB23										
AT1G19180	JAZ1										
AT1G80840	WRKY40										

	WRKY74	WRKY53	ATHB12	MYB77	ATHB52	WRKY45	ATERF15	ORA47	ZCW32	WRKY51
	AT5G28650	AT4G23810	AT3G61890	AT3G50060	AT5G53980	AT3G01970	AT2G31230	AT1G74930	AT1G59640	AT5G64810
AT5G50570	SPL13									
AT5G05090	MYB									
AT4G31550	WRKY11									
AT3G25780	AOC3									
AT3G25760	AOC1									
AT3G23250	MYB15									
AT2G44840	ERF13									
AT2G35930	PUB23									
AT1G19180	JAZ1									
AT1G80840	WRKY40									

	HMG	ANAC015	NAC	WRKY65	WRKY17	WRKY69	WRKY68	WRKY22	WRKY30	MYB118
	AT1G76110	AT1G33280	AT3G12910	AT1G29280	AT2G24570	AT3G58710	AT3G62340	AT4G01250	AT5G24110	AT3G27785
AT5G50570	SPL13									
AT5G05090	MYB									
AT4G31550	WRKY11									
AT3G25780	AOC3									
AT3G25760	AOC1									
AT3G23250	MYB15									
AT2G44840	ERF13									
AT2G35930	PUB23									
AT1G19180	JAZ1									
AT1G80840	WRKY40									

Table 3.12: Summary of the Y1H experimental results presented in this chapter after library screens and two rounds of pairwise screening. *ORA59*, *ESE1* and *PIF7* regulate 90% of the genes thought to be co-regulated together. *At4g38900* and *AtTERF14* regulate 80% and 70% of target genes respectively. Further subgroups of genes can be identified regulated by 1 or more TFs. Green - positive result (growing colonies) were observed in at least one fragment for each gene, Red - no positive results (absence of colonies) were observed in any fragment for each gene.

### 3.4.3 WRKY TF Subscreen

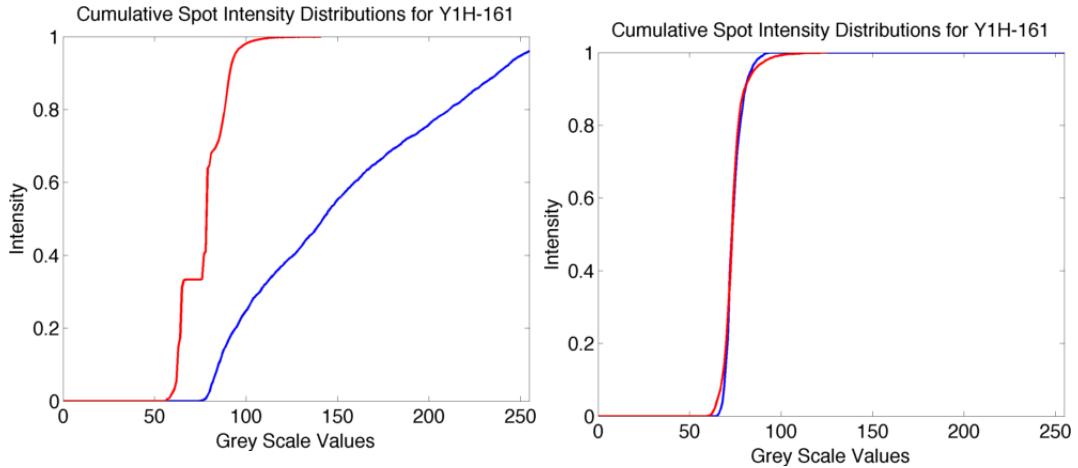
To identify whether all *WRKY* TFs bound to the same motif, TTGAC(Y(Eulgem et al., 2000), or if there existed some clade specificity, two promoter constructs were chosen that were shown to bind different set of *WRKY* genes in the library and pairwise Y1H screens.

Clade	Y1H_161	Y1H_175	Clade
I	<b>WRKY3</b>	<b>WRKY2</b>	I
I	<b>WRKY4</b>	<b>WRKY3</b>	I
IIc	<b>WRKY8</b>	<b>WRKY4</b>	IIc
IID	<b>WRKY11</b>	<b>WRKY8</b>	
		<b>WRKY12</b>	IIc
IId	<b>WRKY15</b>	<b>WRKY15</b>	IId
		<b>WRKY18</b>	IIa
IId	<b>WRKY21</b>	<b>WRKY21</b>	IId
IIc	<b>WRKY23</b>	<b>WRKY23</b>	IIc
I	<b>WRKY26</b>	<b>WRKY26</b>	I
IIe	<b>WRKY27</b>		
IIc	<b>WRKY28</b>	<b>WRKY28</b>	IIc
I	<b>WRKY33</b>	<b>WRKY33</b>	I
I	<b>WRKY45</b>	<b>WRKY41</b>	III

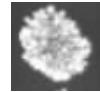
Table 3.13: A subset of WRKY TFs was screened against 2 Y1H fragments (Y1H-161 and Y1H175) for potential common regulators. The library screen revealed these fragments to be interacting with the same TFs leading us to hypothesis that these fragments contain WRKY binding motif suitable for all WRKYS. However, this subscreen revealed some specificity to individual WRKY TFs.

### New Method For Identification Of Positive Pairwise Y1H Interactions

A new method was developed to identify positive interactions from pairwise Y1H screens using images of the plates with the yeast growing on selective media. The method uses a combination of manual grid alignments for spots on an image of the selective plate. The histograms of pixel intensities are extracted and compared to a negative control on the same plate using a statistically robust comparison method (see Methods). The new method can successfully distinguish between a positive result and a lack of interactions, Figure 3.6. When a positive interaction is compared with the negative control containing an empty vector and therefore no



(a) Cumulative distributions of negative control(red) shows markedly different profile of pixel intensities as positive result *WRKY15*(blue). (b) Cumulative distributions of negative control(red) shows the same distribution of pixel intensities as negative result *TAF13*(blue).



(c) Image of the positive result *WRKY15* used for above analysis.



(d) Image of the negative control used for the above analysis.



(e) Image of the negative result *TAF13* used for the above analysis.

Figure 3.6: Cumulative distributions of histograms of pixel intensities for positive and negative results identified by the new method for Y1H-161 fragment together with the associated spot images.

interaction, the difference in the cumulative histogram distributions is immediately obvious. The negative control quickly rises because of the large number of dark grey pixels, at around 60 (background). Whereas, distribution of pixels from the positive result increases more gradually and over a larger range of pixel intensities, capturing brighter pixels, Figure 3.6a. On the other hand, when the control is compared to a spot with no yeast growth, the resulting cumulative histogram distributions are almost identical, Figure 3.6b. Additionally, the new method provides the scores and P-values associated with a confidence associated with the positive results.

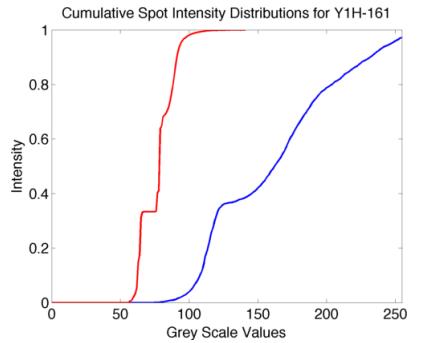
After confirming that the new method can successfully determine positive from control interactions, it was applied to a randomly selected image to automatically compare all TFs in the grid against the control spot (grid location: H12), Figure 3.7. 96 (95 TFs + negative control) TFs were screened pairwise against a promoter fragment (Y1H\_161) and only 9 were found to be bound to it. Manual

inspection of the image revealed 9 positive interactions. The new method identified 10 positive interactions from 95 potential interactions. 9 interactions were the same as determined by manual inspection, 1 interaction was due to bubbles trapped in the agar plate.

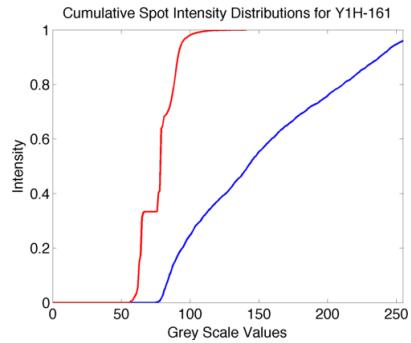
The new method was used to identify positive interactions combined from all repeats. In order to cross validate the new approach, positive interactions were manually cross referenced with the images and scoring sheets used in the lab.

### 3.4.4 Summary of the Y1H Screen Results

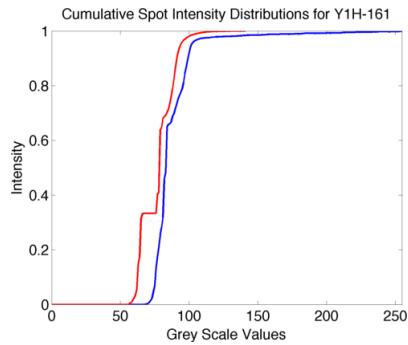
Ten genes have been identified as likely to be co-regulated together based on their high correlation of mRNA expression during infection with Botrytis. Moreover, these same genes contained the *CCA1* binding site in the CNS locations. Promoters of these ten gene have been split across thirty fragments, twenty nine of which have been screened using a library of over 1300 Arabidopsis TFs. A library Y1H screen did not reveal common TFs across all selected genes' promoter regions, however, the screen allowed the narrowing down of the selection of TFs that can potentially interact with the promoter fragments. All unique TFs arising from positive interactions identified by the library screen have been arranged into a mini-library, sequence verified and screened again pairwise against the twenty nine promoter constructs. Pairwise screened was designed to verify interactions detected in the library screen as well as to identify any missing interactions. To increase confidence of the final pairwise results, the screens were repeated three times. Results from the pairwise screens have identified a set of four TFs that were found to be binding to the promoters of up to 9 of the 10 selected genes, supporting the original hypothesis that the selected genes are potentially co-regulated. Additionally, the screen was able to identify subgroups of TFs within a TF family, that are able to bind to a select number of promoters, whereas others in the family do not. E.g. *WRKY15*, *WRKY21* and *WRKY29* only bind the promoters of four genes, whereas no other WRKY TF screened bind the same group of promoters, suggesting that *WRKY15*, *WRKY21* and *WRKY29* bind to the same promoter sequence. *WRKY15* and *WRKY21* belong to the same clade (IId), *WRKY29* belongs to a similar (IIe) clade, both supporting the idea that these three WRKY TFs can bind to a specific sequence found in the four promoter regions that they bind. Similarly, there are some TFs that exclusively bind to a small number of promoters within the set. For example, *ORA47* and *ATERF15* bind to promoters of two unique sets of genes, potentially indicating their unique ability to regulate these genes under certain conditions. Furthermore, *ORA47* binds exclusively to the promoters of *AOC1* and *AOC3* both known to be involved in the production of secondary metabolites involved in jasmonic acid biosynthesis and a previous study have shown that *ORA47* also binds to the promoter of *AOC2* (Zarei et al., 2011), indicating that combination



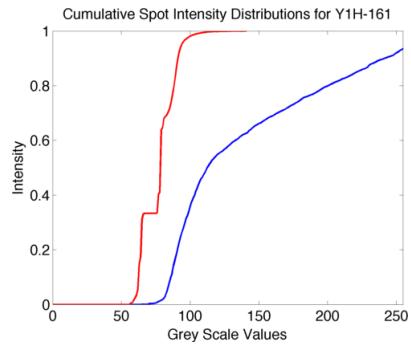
(a) *At1G69310* (red, grid location: B01) and (b) *At2g23320* (red, grid location: C01) and negative control (blue).



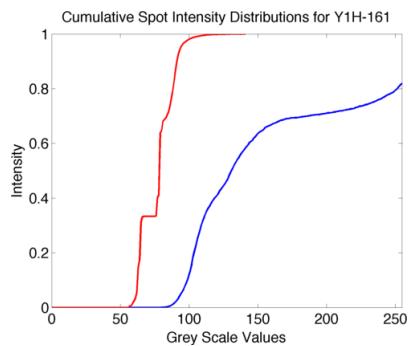
(a) *At1G69310* (red, grid location: B01) and (b) *At2g23320* (red, grid location: C01) and negative control (blue).



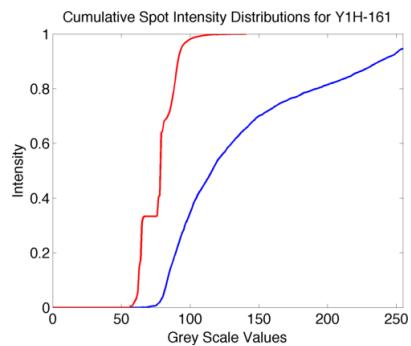
(c) *At2g24570* (false positive) (red, grid location: D01) and negative control (blue).



(d) *At2g30590* (red, grid location: E01) and negative control (blue).

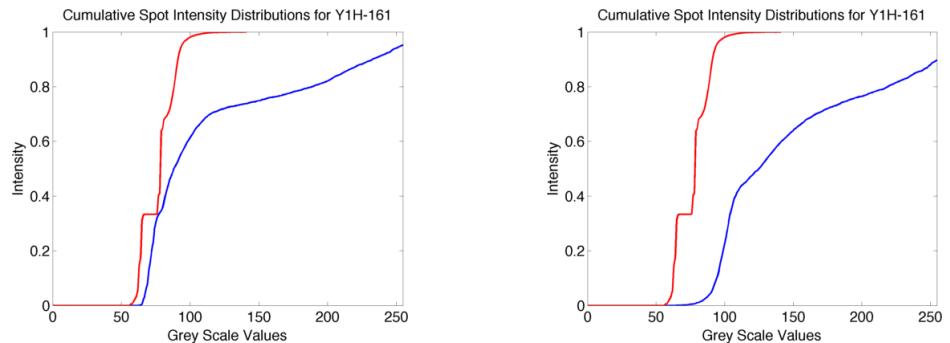


(e) *At3g01970* (red, grid location: G01) and negative control (blue).

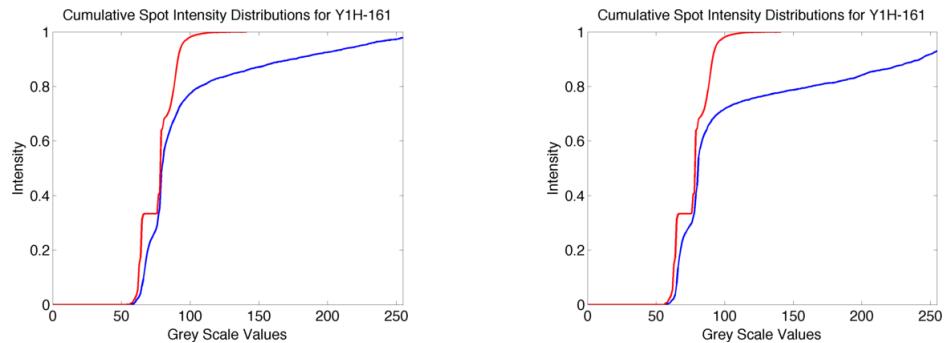


(f) *At4g18170* (red, grid location: D02) and negative control (blue).

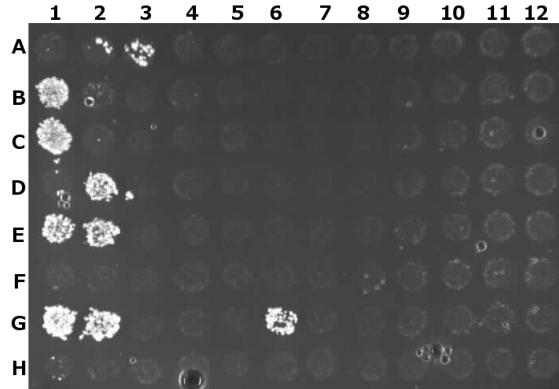
Figure 3.7: Distribution of pixel intensities from positive results (red) selected by the algorithms for Y1H-161 and corresponding distribution of pixel intensities from negative control.



(g) *At4g23550* (red, grid location: E02) and (h) *At5g13080* (red, grid location: G02) and negative control (blue).



(i) *At5g46350* (red, grid location: A03) and (j) *At5g61270* (red, grid location: G06) and negative control (blue).

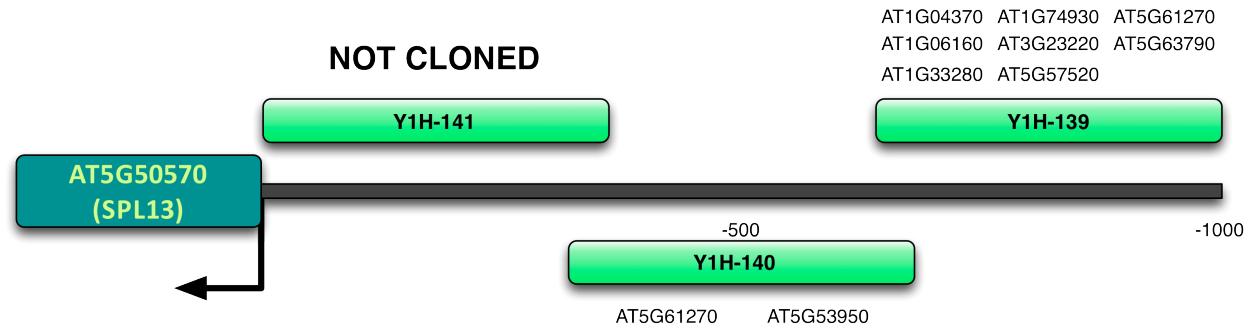


(k) Image of the Y1H-161 SD-LTH plate against 95 TFs and negative control (H12) used for automatic processing showing nine positive results and one false positive.

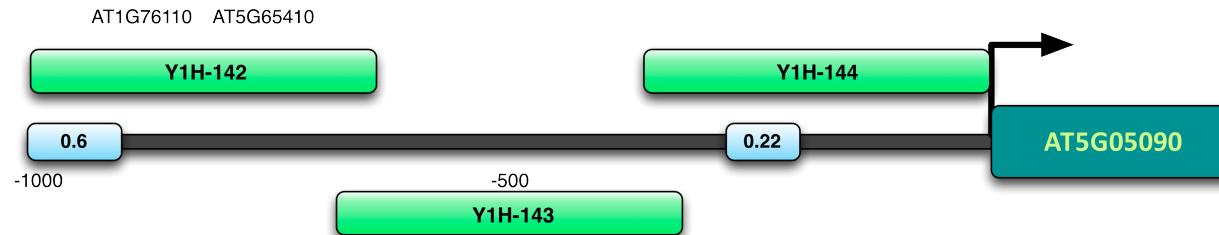
Figure 3.7: Distribution of pixel intensities from positive results (red) selected by the algorithms for Y1H-161 and corresponding distribution of pixel intensities from negative control.

of library and pairwise screen can identify potential common regulators of multiple genes as well as specific interactions.

The final results, combining library and all replicates of pairwise screens are shown in Figure 3.8 and summarised in Figure 3.12.

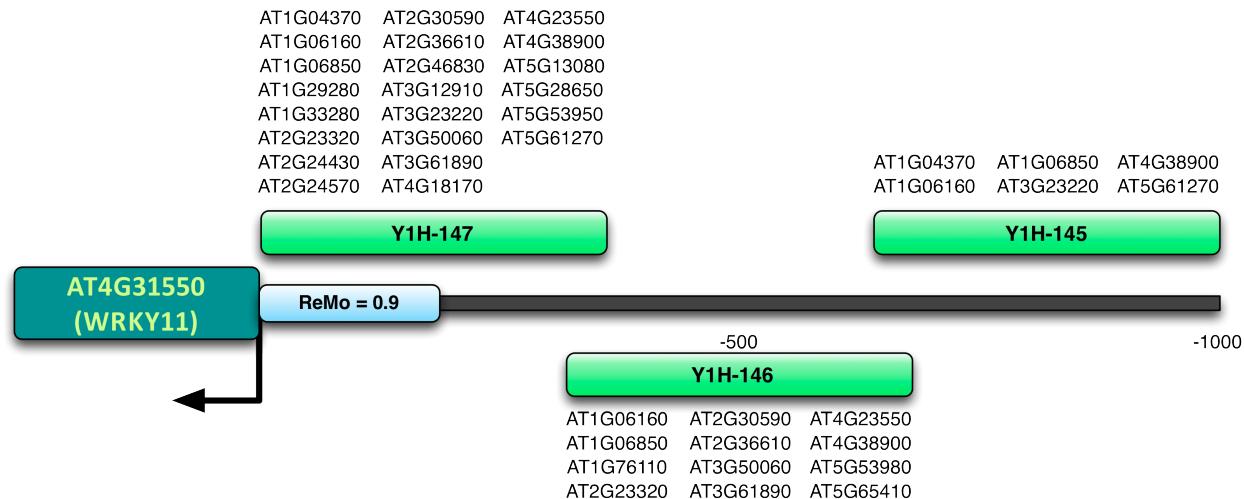


(a) Final results of Y1H screens for *At5g50570* revealing Y1H-139 to interact with large number of TFs. Y1H-141 was not amplified form the genomic DNA and was not screened.

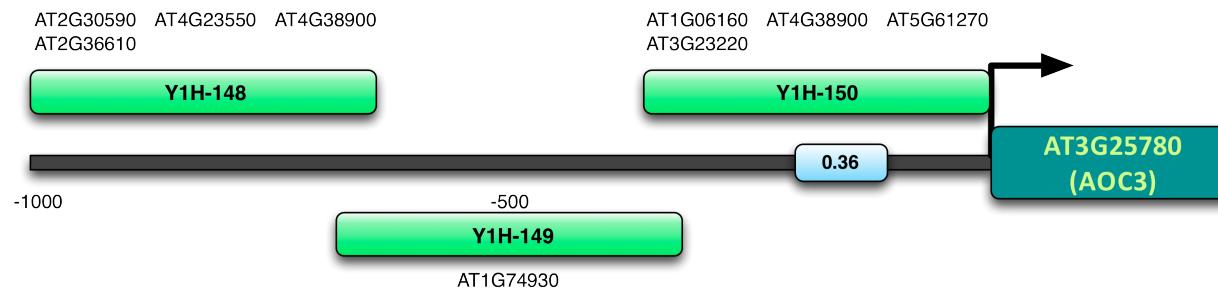


(b) Final results of the Y1H screens for *At5g05090* revealing little interaction with the 1000bp of the promoter sequence.

Figure 3.8: Final Y1H screen results after library and pairwise screens above each promoter fragment. Green - fragment location and identifier, blue - CNS location and score.



(c) Final results of the Y1H screens for *At4g31550* show that promoter fragment nearest to the TSS interacted with large number of TFs.



(d) Final results of the Y1H screens for *At3g25780* revealing a small number of interaction of 1000bp promoter sequence and library TFs.

Figure 3.8: Final Y1H screen results after library and pairwise screens above each promoter fragment. Green - fragment location and identifier, blue - CNS location and score.

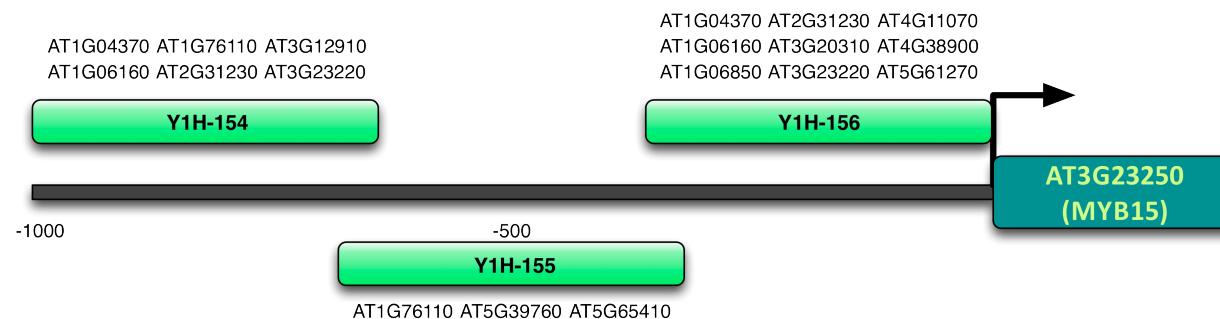
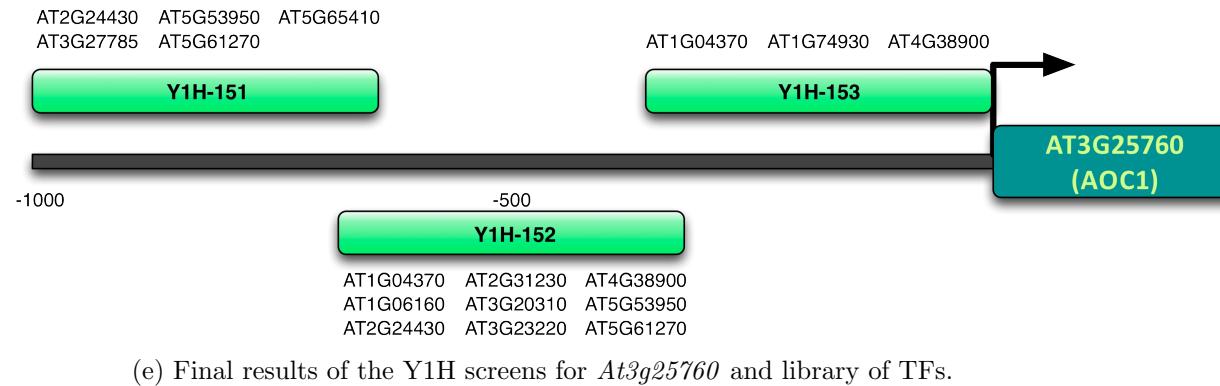
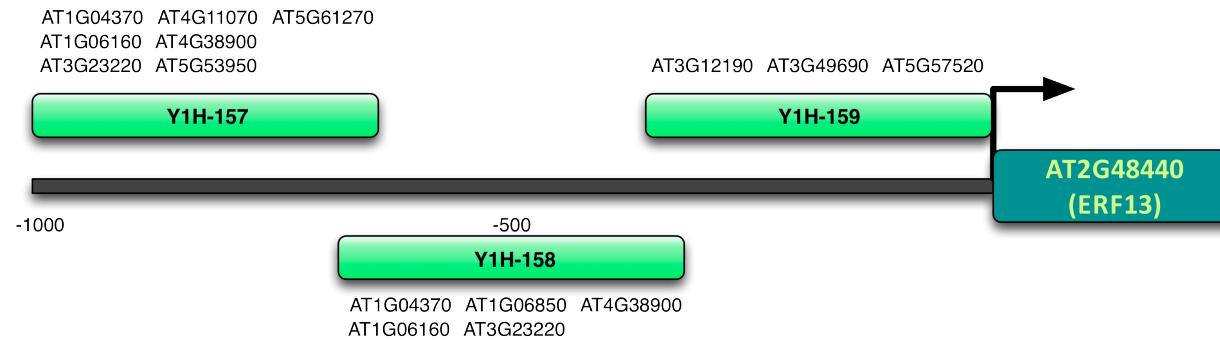
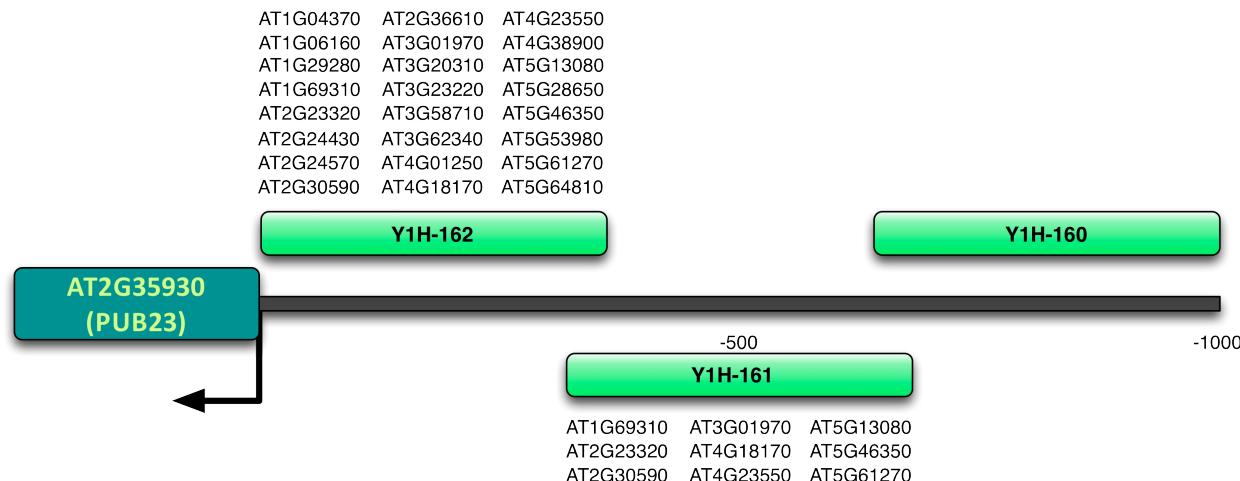


Figure 3.8: Final Y1H screen results after library and pairwise screens above each promoter fragment. Green - fragment location and identifier, blue - CNS location and score.



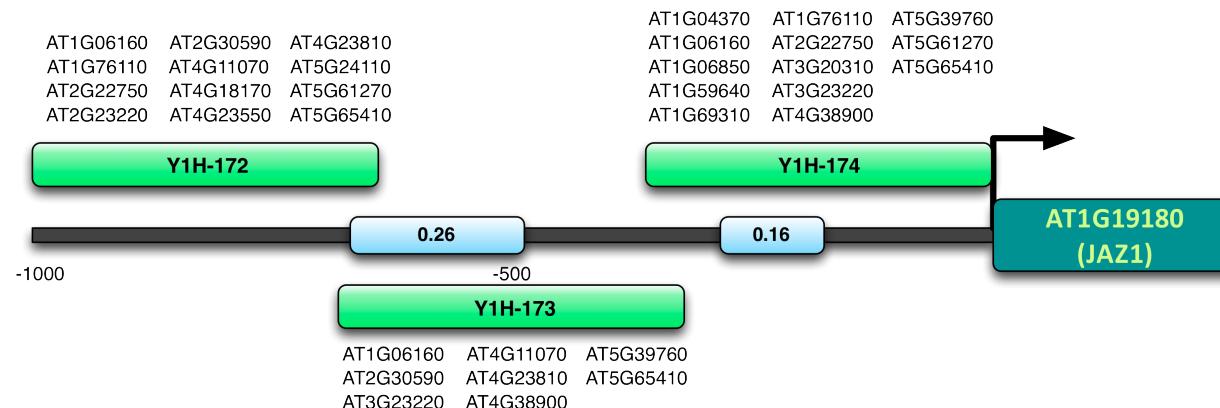
(g) The fragment furthest from the TSS of *At2g48440* gene shows the largest number of interactions with library TFs.

104



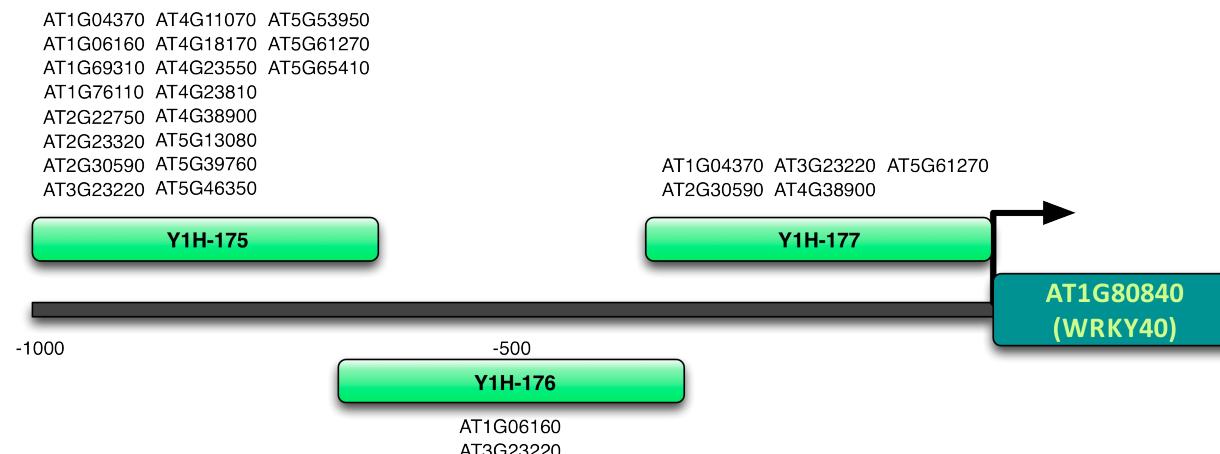
(h) The fragment closest to the *At2g35930* TSS shows the largest number of interactions with library TFs suggesting regulatory activity.

Figure 3.8: Final Y1H screen results after library and pairwise screens above each promoter fragment. Green - fragment location and identifier, blue - CNS location and score.



(i) Final results of the Y1H screens for *At1g19180* promoter and library TFs.

105



(j) The fragment furthest from the annotated TSS of *At1g80840* gene revealed to have the largest number of interactions with library of TFs.

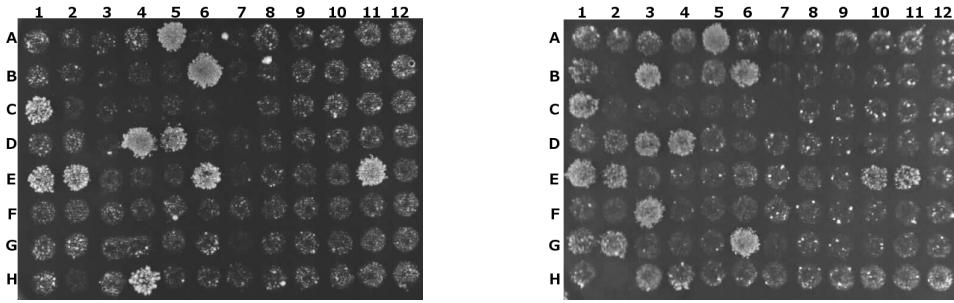
Figure 3.8: Final Y1H screen results after library and pairwise screens above each promoter fragment. Green - fragment location and identifier, blue - CNS location and score.

## 3.5 Discussion

One of the criteria for selecting genes in the “WRKY” cluster, Table 3.9, was the high degree of similarity in mRNA expression levels during infection with the necrotrophic fungus *B. cinerea*. A previous study has shown a high correlation of co-regulation instead of co-expression for the genes with very similar expression patterns. Using mRNA similarity and additional heuristics, a set of ten genes, thought to be involved in the stress response to the invading fungus, were selected to identify the underlying gene regulatory network using a Y1H Library screen (Ou et al., 2011). Over 1300 Arabidopsis TFs have been cloned and transformed into yeast cells to generate a high-throughput library to probe for direct Protein-DNA interactions. The 29 promoter constructs, covering ten genes have been amplified from genomic Col-4 DNA, cloned and transformed into yeast to be tested against a library of TFs. Furthermore, after the library screen, TFs that resulted in positive interactions have been extracted from the bacterial glycerol stocks, sequence verified and assembled into a separate mini-library to be tested again in triplicate with the 29 promoter constructs. The second Y1H screen was carried out in a pairwise manner, such that only a single TF was tested against the promoter DNA in each well, in contrast to the library screen in which up to 24 TFs are pooled into each well.

### 3.5.1 Image Based Method For Identification of Positive Results In Y1H

One of the big challenges after performing the experiment is to correctly identify positive and negative results. Previously, the positive interactions were found by counting the number of colonies on the plate where mated yeast has been growing for a number of days after the cleaning process. When the plates have been adequately cleaned and there is no auto-activation present, it is relatively easy to pick out strong positive results, since they appear as densely growing colonies. However, in a practical setting this is often not the case. Figure 3.9 shows a more representative example where positive results have to be identified. Firstly, scoring was introduced to differentiate weak and strong interactions on a scale from one to ten respectively. However, such scoring was not always consistent, between operators and over time. Variability can result from growth conditions in the incubator and on the plate (e.g. 3AT concentrations), and also depend on the individual scoring the results. Therefore, consistency is a major factor in analysing colony growth and the subsequent categorisation as a ‘positive’ or ‘negative’ result. In order to eliminate this variability an alternative method for the determination of positive interactions



(a) Y1H-146 fragment.

(b) Y1H-147 fragment.

Figure 3.9: A typical pairwise plate of one promoter fragment vs 95 TFs ( control at H12) on agar SD-LTH plate. It is difficult to identify all positive interactions visually due to auto-activation.

has been devised. Photographs of the final plates are always taken and kept for reference purposes. However, these pictures combined with some image analysis and statistical techniques could potentially be used to consistently identify positive interactions, which is especially important when considering weak interactions.

Use of the automated scoring and analysis technique, SpotOn, has recently been used in mapping human (Reece-Hoyes et al., 2011a) and worm (Reece-Hoyes et al., 2011b) TFs to bait DNA sequences in a high-throughput manner. The method used in these two studies and the new method presented in this chapter differ in a number of important aspects. Firstly, SpotOn uses colour information obtained from the blue colonies produced by the *lacZ* reporter gene, whereas only grey scale information is available from the study presented here. Presence of colour information means that SpotOn is able to hone in on individual colonies, as opposed to combining all colonies within the same spot/well, as used in the method here. Although both methods use colour and grey scale intensities to determine positive interactions, both apply different statistical techniques to the intensity information. SpotOn uses negative control spots with empty reporter constructs as a normalisation factor and a Z-score is derived from normalised samples, taking into account growth variation in row/column and negative controls. Higher Z-scores correspond to high confidence positive results. On the other hand, the new method compares negative controls against test spots using Kolmogorov-Smirnov Two-Sample test statistics.

## **Spot Detection**

A number of different approaches have been tried to automatically identify circular spots where the yeast was spotted, for example, using Sobel filters to outline regular circular structures. However, no adequate settings were found. An alternative approach was to use raw image recognition library OpenCV which provides libraries for feature detection, for example using Haar wavelets. In order to use feature detection, positive and negative sets containing true positives and true negative examples need to be prepared and feature detection can be trained using these sets. The larger the training sets are, then the better feature detection becomes. However, constructing a large enough set from the existing data would have been as time consuming as picking out wells by hand. Moreover, as most of the data used for training couldn't have been used for detection meant that success of the feature detection would have been difficult to access. Therefore, an ImageJ coupled with microarray plugin that allows for easy drawing of the round spots on a large scale allowed for picking a relatively high throughput way of outlining mated spots. Once the spots were outlined, the same plugin outputs statistics about the spot, including a histogram of the intensity values and intermediate statistics such as mean, variance and standard deviation of intensities. Histograms are used for cumulative distribution computations and in further downstream analysis.

## **Positive Interaction Identification Among Auto-Activation**

An ability to identify positive interactions amongst the noise present in the fragments that have high levels of auto-activation is one of the major advantages of using automatic identification and classification. For example, the Y1H\_177 fragment consistently auto-activated during the screen, but appropriate levels of 3AT that maintained a high fraction of positive results were difficult to determine, Figure 3.10. The method developed here allows even noisy spots to be compared with negative controls since both should contain roughly the same amount of growth due to auto-activation. In addition, positive results should also contain more growing yeast due to the positive interaction of the TF with the promoter fragment. This can be difficult using the traditional, by eye, method.

The new method has been applied to the images for this fragment and found to contain a few significant interactions that were previously discarded in the traditional technique.

## Limitations Of Automatic Approach

The new methods have been shown to be able to distinguish between positive and negative results, Figure 3.6, as well as allowing the automatic determination of all potential positive results on a single plate, Figure 3.7. However, one of the major challenges that hindered successful prediction of positive interactions is the presence of bubbles in the agar plates. When bubbles are formed during plate pouring and are not dissipated, they leave a circular shaped indentation in the plate. When the picture of the plate is taken after the incubation period with upper white light (see Methods), the edges of the circular indentations produce brighter outlines as light is refracted internally more around these edges than inside of the indentations, giving a “halo” effect. Furthermore, the analysis relies on the premise that brighter pixels represent growing yeast, and therefore this “halo” effect introduces artificially brighter pixels. The spots with the “halo” effect produce lower P-values that are still statistically significant. This effect can be corrected by using appropriate spot recognition, as mentioned above. Automatic classifiers can be trained to recognise bubble indentations and flag them appropriately, given that an appropriate training set exists. During the development of the new method, no such training set existed, therefore it was not possible to train automatic classifier. However, following the pairwise Y1H results presented here, there are over 6500 examples of spots containing positive, negative and auto-activating results, which can be used to successfully train an automatic classifier and design a pipeline that would take images of plates containing yeast and output unbiased estimates for positive and negative interactions on the plate. This method can also be extended to be used in any pairwise Yeast ’n’ Hybrid screen. Additionally, the classifier could be trained to be used with library screens using images from the library screens

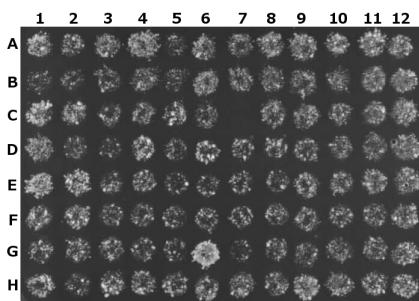


Figure 3.10: Pairwise screen of Y1H\_177 fragment against a mini-library of TFs. All spots, with exception of C07, have some level of growth due to auto-activation in yeast. Negative control is located at H12.

presented here, although spots are not as well defined in the library images as in the pairwise screen.

### 3.5.2 Y1H Screen Reliability and Reproducibility

In the original design, a Y1H screen using 1300 TFs served as a large scale, high throughput way of testing which proteins have the potential to bind to the promoter fragments. Conventional wisdom associated with high throughput Y1H screens was that although some interactions may not be picked up, overall many more interactions will be tested and positive results will constitute to the majority of the true regulators. However, once the results from the whole library screen were obtained, the validity of all interactions came into question due to the results from smaller scale experiments that were carried out in parallel by other members of the lab and subsequently tested in a pairwise fashion, in which not all positive results were reproducible. Additionally, the *in vitro* nature of the screen may also play an important part in establishing the validity of observed interactions. For example, the host system (yeast) may fold transcribed protein in a different way to the conformations found in Arabidopsis, which may lead to a different DNA-binding domain structure and result in protein binding to a sequence not usually associated with that domain. The Y1H system is appropriate when a TF is thought to function independently of other TFs, otherwise the Y1H system does not have the capacity to assemble transcriptional complexes from a variety of exogenous proteins. Moreover, the endogenous TF may interact in unpredictable ways, forming chimera complexes with the TF of interest, resulting in false positive results. Both types of faults could occur in library and pairwise screens, therefore, pairwise screens are not true indicators of false positive and negative interactions.

The second, pairwise, screen was devised in order to confirm true interactors as well as to potentially identify new interactors that were previously not detected. This allowed for additional confirmation of the existing results as well as checking the correctness of the TF in the library. Although TFs were individually transformed, the TFs are assumed to be represented by the correct coding sequence. Table 3.11 shows a summary of the sequence verification of the TFs that were used in the pairwise screen. Approximately one third of the TFs were correct from start to end and correspond to at least one known annotated transcript. Another third were correct but missing a "STOP" codon. Finally, the remaining third had various numbers of mutations, often changing a correct amino acid to an incorrect one at that position. Here we assumed that clones were correct even with the "STOP" codon

missing, because it should have a minimal effect on the final protein. Moreover, the lab is continually improving the number of correct, sequence verified, TF families, which meant that these clones can be included in pairwise screens instead of being taken from the library, which may or may not be correct. Of over 1300 potential regulators, only 123 unique TFs that were positive interactions across all 29 library screens had to be tested on the same 29 fragments. This resulted in a much smaller scale and thus higher precision screen, as well as eliminating the need and the cost of sequencing, since pairwise screens yield instant results as the location of each individual TF is known. The results of the pairwise screen show that only 43 unique interactions were reproduced from a total of 123. Moreover, many new interactors that were previously missed in the whole library screen, were detected in the pairwise screen. Subsequently these 41 interactors were tested again in a second and a third pairwise screen to confirm the reproducibility of the pairwise screen. Different replicates of pairwise screens have shown that some interactions are not reproducible across multiple screens. A total of 204 interactions were found across 29 promoter fragments in 3 screens, 50 (25%) were only identified as positive interactions in one of the three screens, 28 of these interactions had very faint yeast growth associated with them in one or two repeats and therefore were not considered as positive results. 39 (20%) were identified in two out of the three replicate screens. Using all results taken together, including potentially weak interactions seen in at least two screens (171) gives 86% reproducibility of the pairwise screen. A previous study in worm found similar reproducibility of 90% using an enhanced Y1H screen (Reece-Hoyes et al., 2011b).

### 3.5.3 Importance of Correct 3AT Concentration

One of the fundamental challenges in a Y1H experiment is correctly identifying positive results amongst background auto-activation levels. Since *HIS3* has a leaky expression<sup>1</sup> in many yeast strains, it is recommended to inhibit the basal expression of this gene with 3-amino-triazole (3AT), a known competitive inhibitor of the *HIS3* gene product. For a Y2H screen prior to starting a large-scale transformation procedure it can be very informative to perform a pilot transformation that enables titration of the optimal amount of 3AT needed. Using too much 3AT will result in a loss of weak interactions, whereas the use of no or too little 3AT results in high numbers of false positives. The optimal concentration is largely dependent on the

---

<sup>1</sup>Endogenous proteins in yeast cell may recognise bait constructs and express *HIS3* gene without true Protein-DNA interaction taking place. This phenomena is often termed - leaky expression or auto-activation in the context of Yeast 'n'-Hybrid.

yeast strain and promoter DNA used as the target for screening. Fundamentally this is very difficult to achieve because *a priori* likely positive interactions are not known, to be used as various controls for pilot transformations. Moreover, screen optimisation takes a considerable amount of time for a single bait, optimisation for 29 fragments is not feasible in a reasonable time frame. Therefore, appropriate 3AT concentrations were adjusted as positive results from the screen became apparent. For example, mated fragment TF constructs are replicated onto SD-LT, -LTH (selective plate), -LTH+25mM 3AT, -LTH+50mM 3AT and -LTH+100mM 3AT. If a promoter fragment is observed to be auto-activating on a selective plate, it is replica plated again (from an existing YPDA plate) onto a plate containing half 3AT concentration from the plate that inhibited all interactions. The process is repeated until 3AT inhibition is deemed to be significant enough to suppress auto-activate while maintaining a large fraction of true positive interactions. An alternative strategy would be to gradually increase the levels of 3AT until there is no activation across all of the wells, or conversely, start with a high 3AT concentration and gradually decrease it until yeast colonies appear. Ideally, some combination of these two methods should be used to narrow down the true interactors, however, in practise the former technique was used and the corresponding effect on the false positive frequency is not clear. At least conceptually, weaker interactions may be eliminated in such a process, as they would be comparable to auto-activation levels and therefore would be eliminated when yeast is grown on 3AT concentrations higher than that required to eliminate auto-activation.

### **Assessment of Y1H False Positive/Negative Rates**

Success and failure rates have only been characterised in two studies involving the Y1H system with TFs from *C. elegans* (Deplancke et al., 2004; Reece-Hoyes et al., 2011b). Positive results from the library screen and subsequent confirmation with pairwise libraries provides clues to the true false positive and negatives rates of the screen. Accurately identifying the False Negative Rate (FNR) of the library screen using pairwise results will help to determine the fraction of expected true positive interactions in the results obtained here and inform the outcome of future library screens. False positive and negative terms are often used in hypothesis testing, as such, the *null* hypothesis,  $H_0$ , must be defined. In the context of a Y1H screen, the *null* hypothesis is defined as the promoter DNA fragment does not interact with a TF. As the screen is performed and enough evidence is amassed to reject this *null* hypothesis in favour of an alternative,  $H_1$ , that a given TF does interact with the DNA fragment.

	$H_0$ is TRUE	$H_0$ is FALSE
Fail to Reject $H_0$	True Negative	Type II (False Negative)
Reject $H_0$	Type I error (False Positive)	True Positive

Table 3.14: Definitions of True Positive, True Negative, False Positive and False Negative results.

- $H_0$  - TF does not interact with DNA
- $H_1$  - TF interacts with DNA

Following from the above definitions of the *null* and alternative hypothesis, false positive and negatives can be defined following schema in 3.14. It is worth noting that Type I and Type II errors are often described in terms of rejecting *null* hypothesis or failing to reject it.

Furthermore, False Positive and False Negative definitions can be refined further in the context of the Y1H screen.

### False Positive (Type I Error)

Rejecting  $H_0$ , when it is true. In other words, positive results from Y1H screen (a number of colonies), when there is no significant interaction.

### False Negative (Type II Error)

Failing to reject  $H_0$ , when it is false. There was no interaction in the corresponding well, but in the future screen it has been found that protein-DNA interaction does occur.

Using the definitions from 3.14 and the results from the library, the number of positive and negative results can be similarly represented. Additionally, pairwise screening is used to check the results from the library screen. In the absence of a gold standard positive and negative control sets, we assume that the pairwise screen is a true indicator of positive and negative results in the context of a Y1H screen. Additionally, only positive results that pass three replicates of the pairwise screen are used, similar to the methodology used in the Y2H screen (Arabidopsis Interactome Mapping Consortium, 2011).

True positives are results which have been identified as positive, rejecting  $H_0$ , in the library screen and also found as a positive result in the pairwise screen. False negative - is a number of new positive results found in the pairwise screen, that

	$H_0$ is TRUE	$H_0$ is FALSE
Fail to Reject $H_0$	TN = <b>n/a</b>	FN = <b>156</b>
Reject $H_0$	FP = <b>83</b>	TP = <b>48</b>

Table 3.15: Numbers of positive and negative results from combined library and pairwise screens.

were originally discarded as negative results in the library screen. False positive - is a number of positive results identified from the library screen but could not be confirmed using a pairwise selection process. Finally, the number of true negatives is not known.

### False Negative Rate

$$\text{Type II Error} = \text{FNR}(\beta) = \frac{\text{FN}}{\text{TP} + \text{FN}}, \quad (3.1)$$

where FN - number of false negatives, TP - number of true positives.

Following from 3.1,  $\text{FNR} = 0.76$ . FNR can also be thought of in terms of sensitivity:

$$\text{sensitivity} = 1 - \text{FNR} = 1 - 0.76 = 0.24 \quad (3.2)$$

On the other hand, experimental technique has a much greater impact on the overall results. For example, in the first step of the Y1H protocol, inoculation of the TF library from the glycerol stocks, the real number of TF picked up from the glycerol is not known and cannot be determined afterwards. Whereas, in the pairwise screen, if the well was not inoculated, this can be seen in the SD-LT selective plate because the mated yeast will not grow due to the lack of tryptophan produced by the TF construct. In contrast, the whole library screen has up to 24 TFs and inoculating with at least one out of 24 will result in growth on the selective, SD-LT, media. So although we assume that all TFs are present and transferred from the glycerol, that is impossible to verify. Assuming for a moment that all TFs were transferred from the glycerol and grew approximately in a uniform proportion to each other in  $500\mu\text{l}$  of liquid media, only  $3\mu\text{l}$  are spotted on top of the promoter construct on the plate with the selective media. Whether or not all TFs are picked up in those  $3\mu\text{l}$  is once again impossible to verify. Steps are taken to minimise these effects, for example, the liquid TFs in liquid media are shaken in order to disperse them uniformly in the well. However, experimental technique plays a major part in

the potential number of false negatives observed.

Another aspect of the whole library screen is the relatively high number of false negatives. After conducting pairwise screens, the number of new interactions was almost double that of the already existing ones. Overall, a Y1H screen using 1300 TFs is useful as a first stage in narrowing down potential true interactors. Specifically, if many different genes are tested which are thought to be *a priori* co-regulated, the number of false positives and false negatives play a much smaller part in the overall quality outcome of the experiment due to the fact that the same test is conducted many times, 29 in this case, and interactors that should appear in common have a higher probability of occurring at least once. Even with high false positive and false negative rates, after the first round of testing all interactors that appeared at least once are tested again with all promoter DNA fragments. So even though the first experiment produced inconclusive results about the degree of commonality across co-regulated genes and missed potential regulators, pairwise re-testing improves the predictions and results in new interactions.

An additional factor that may influence false negatives is the location of the binding site away from the TSS. There is some evidence to suggest that the more distal the binding site is, the less likely transcription is to be initiated (R. Hickman, S. Kiddle unpublished). The TF may still be binding, but being a long distance away from the TSS means that the activation domain does not come into close proximity and therefore transcription is not initiated. It is not clear how far away a binding site needs to be in order for transcription not to be initiated at all. The strength of the protein-DNA interaction may also play a role in false negative results. Weaker interacting TFs may be more easily displaced by other proteins in the cell which may therefore prevent transcription of the reporter gene. Moreover, the negative impact of both weak interactions and binding site location is probably additive and adds to the false negatives that would not be detected in the Y1H screen.

### **False Positive Rate**

Although it is not possible to comment on the False Positive Rate (FPR) of the screen as a whole, the number of False Positive results may provide a clue to the overall specificity of the test. Given a positive result it is not immediately obvious how to show that given interaction does not occur under any circumstances, conditions or biological contexts. It is simply impossible to test every alternative and

exhaustively show an interaction to be not true. A limited study has been done by B. Deplancke and colleagues Deplancke et al. (2004), where promoters of *C. elegans* were tested for interactions with a library of TFs. 2 out of 6 interactors tested had a significant effect on the expression of the target gene. This suggests a FPR of 67% in their experiments, meaning that some interactions are prevented from acting upon the promoter DNA in certain, possibly all, experimental conditions. This observation strengthens the idea that biological context needs to be considered when ascertaining the evidence for transcriptional regulation.

Out of 131 unique TFs identified as positive interactions in the library screen, only 46 were confirmed using pairwise screen of the same TFs, approximately  $\frac{2}{3}$  of the original data was rejected as false positives. There are many factors that may account for such a high FPR. For example, experimental technique, in particular the quality of the replicated plate cleaning with velvets (see Methods). Once the mated yeast is replica plated onto the selective media and incubated for 24 h, it is cleaned by applying the surface of the plate to a clean velvet. A layer of yeast is removed, so that it cannot be used as a source of nutrients for the layers below, and so if a positive result is observed it is due to the TF interaction and subsequent production of *HIS3* gene, not to the availability of histidine from dead yeast in the layer below. The cleanliness of the plate is usually visually judged and therefore subject to external conditions, for example, the thickness of the media and general ambient light conditions. When a plate has the extra yeast, single colonies may form that are due to the nutrients from the dead yeast and not a true interaction. Another possible avenue for a false positive may be a relatively high mutation rate of the host. A recent estimate puts spontaneous mutation rate in yeast to be  $U = 9 \times 10^{-5}$  (Wloch et al., 2001). Over a large number of colonies, a non-deleterious mutation may arise that allows the host system to activate the promoter DNA, or more likely, circumvent the knockout of the histidine biosynthesis to allow for production independent of the TF-promoter interaction, especially as the yeast is under very severe selective pressure to survive on selective growing conditions. One way to assess the possible impact of this is to look at the number of colonies that were seen in the whole library screen and whether the results were reproduced in the pairwise screen.

FPR can be calculated using Equation 3.3.

$$\text{Type I Error} = \text{FPR}(\alpha) = \frac{\text{FP}}{\text{TN} + \text{FP}}, \quad (3.3)$$

where FP - number of False Positives, TN - number of True Negatives.

In this case the FPR cannot be calculated, because the number of true negatives is not known. It has previously been assumed that all TFs, up to 24, stored in the glycerol stock in the libraries are successfully transferred into liquid media to grow for screening. However, a large number of positive results identified in the pairwise screen, but not originally seen in the library screen (false negatives) indicates that this assumption is no longer true. If there are more positive results picked out from the pairwise screen than originally identified from the library selection means that in all likelihood, not all TFs have been used to test for binding in the first instance. It is also not possible to reasonably estimate the number of true negatives in the screen, because not all TFs are tested pairwise, only the positive results from the library screen are validated using a pairwise approach. It may have been possible to give a rough estimate of the different number of unique TFs that are successfully used in the library screen by using a number of wells, for example 30, and inoculate liquid culture from the glycerol stocks for each well. After the yeast containing TFs have grown for 3 days, the same time as for the preparation of the library screen, they could be spread thinly on the selective media and as many colonies as possible picked out to estimate the fraction of unique TFs selected for inoculation. Over 30 independent wells would provide a good overall estimate for a distribution of TFs tested during the screen. This estimate would also permit the FPR to be estimated with some confidence.

### **Impact Of Colony Numbers On Positive Library Screen Interactions**

Factors not associated directly with protein-DNA interactions, for example a mutation in a selected yeast colony, may lead to false positive results in the library screen. Such one off events would only be characterised by a few colonies growing, instead of over four for true positive results, therefore in general, positive results derived from a small number of colonies are more likely to be associated with false positive results in the library screen. To test this hypothesis, positive results with different number of colonies associated with them were considered together with the corresponding pairwise results.

When removing results with only one colony, the FNR increases from 0.76 to 0.84, as was expected since some verified positive interactions which were only identified by a single colony in the library screen are now been classed as a negative result. Furthermore, when results with two or fewer colonies are removed, the FNR increases only slightly to 0.86. Similarly, for three or fewer colonies the FNR increases even further to 0.92. Inversely proportional to FNR is the sensitivity of

Colonies associated with positive results	FNR
> 0 (all)	0.76
> 1	0.84
> 2	0.89
> 3	0.92

Table 3.16: Summary statistics of colony number impact in the Y1H library screen on the false positive results identified from using pairwise screen.

the assay, which decreases as FNR increases, meaning that the test is less and less sensitive. However, even though the number of true negatives can not be reliably identified, it will remain constant when only considering results with a larger number of colonies in the library screen. Using Equation 3.3 for FPR, as the number of false positives decreases and number of true positive remains constant, the overall FPR falls with an increase in the associated colonies from the library screen. In summary, the sensitivity of the library screen decreases and specificity increases when using positive results associated with a high number of colonies, over three.

Different rates of FNR and FPR associated with different number of colonies suggest that there are two separate cases to consider when using the Y1H Screen. Firstly, if a single promoter is screened, it is important to increase the specificity of the screen as much as possible, whilst sensitivity is not so important, this would ensure higher confidence in the overall results. Whereas when screening a large number of promoters, specificity can be reduced on a per screen basis, whilst increasing sensitivity. Since a large number of promoters is screened, higher sensitivity allows the detection of a larger number of positive results. In addition, pairwise validation in both cases should be performed to confirm any library identified interactions and discover new interactions that have been missed during the library screen, as suggested by the high number of false negatives.

### 3.5.4 Combined Y1H Screen Can Uncover Common Regulators

The rationale for conducting a Y1H screen on a cluster of predicted co-regulated genes is to discover the elements that regulate genes in the cluster under a given condition, in this case, regulation in response to infection with *B.cinerea*. Given this, the expectation was potentially for a small number of TFs to be in common across all genes of interest and a larger number that were not as commonly occurring. Proteins that don't overlap across many genes may have a function in a different stress,

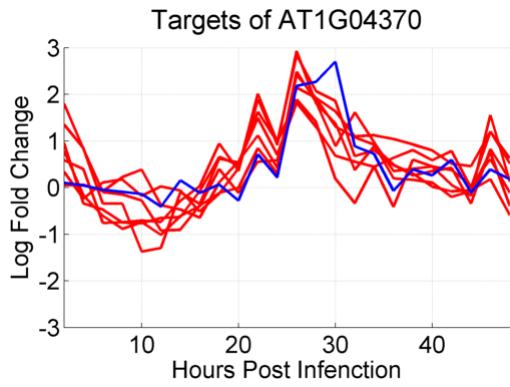
different tissue-type or may be regulating a non stress-related process, for example development, thus forming small subgroups. The final results after library and triplicate pairwise screens reveal a small number (five) of TFs regulating over 70% of the genes. Strikingly, some TFs also have very similar expression pattern in response to *B. cinerea* to the *in vitro* predicted targets. For example, *ORA59*, known to be involved in the stress response to the fungal infection (Pre et al., 2008), was observed to be interacting with nine out of ten genes, Figure 3.11b. *ESE1* has been previously described as being involved in salt stress response in Arabidopsis(Zhang et al., 2011), but differential expression during *B. cinerea* infection together with direct Protein-DNA interactions detected in the Y1H screen would suggest that *ESE1* also plays a role in response to biotic stress, Figure 3.11d. *ERF14* has been implicated in previous studies as an important stress regulator in infection response to *F. oxysporum*(Moffat et al., 2012). *ERF14* is shown to be interacting with seven out of ten genes thought to play an important part in response to *B. cinerea*, including *JAZ1* a well known repressor of stress responsive genes (Thines et al., 2007; Chini et al., 2007).

Apart from the known stress responsive interactors, there are additional TFs that have not been previously characterised as being involved in stress. For example, *AT4G38900* is a basic leucine zipper (bZip) protein which has not been well studied and has only been described based on its sequence homology with existing reference sequences (TAIR10 annotation). *AT4G38900* appears to be differently expressed between mock and *B. cinerea* infection and therefore may play a functional role in regulating stress responsive genes, Figure 3.11e.

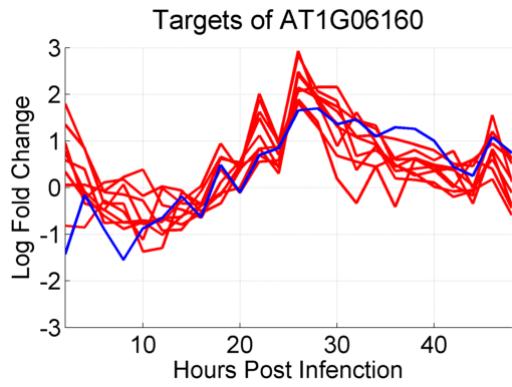
### 3.5.5 Y1H Library Screen Can Uncover Subgroups of Co-Regulated Genes

Y1H is a well established experimental technique for discovering Protein-DNA interactions *in vitro*. The library screen (Ou et al., 2011) has been shown to successfully predict correct interactions validated using other *in vitro* and *in planta* procedures (Pruneda-Paz et al., 2009). However, using Y1H on a set of promoters that are thought to be co-regulated has the potential for building a gene regulatory network from the ground-up. Simultaneously testing over 1300 TFs in a single screen offers a unique advantage in building a comprehensive “context-free” GRN from the observed interactions. Whereas other experimental techniques, for example ChIP-Seq, only considers potential genome-wide regulation by a single TF, for which an antibody is available.

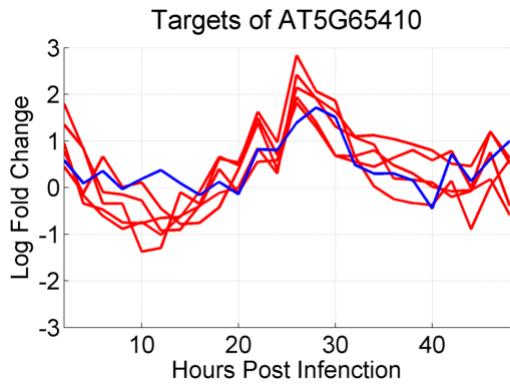
Additionally, the Y1H screen allows both common regulators, thought to be



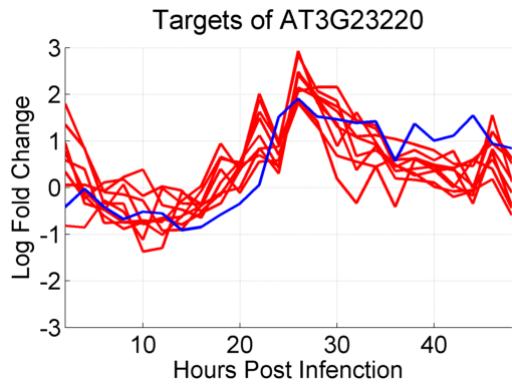
(a) AtERF14 (blue), target genes (red)



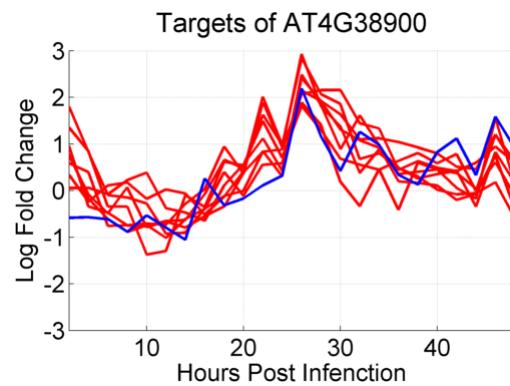
(b) ORA59 (blue), target genes (red)



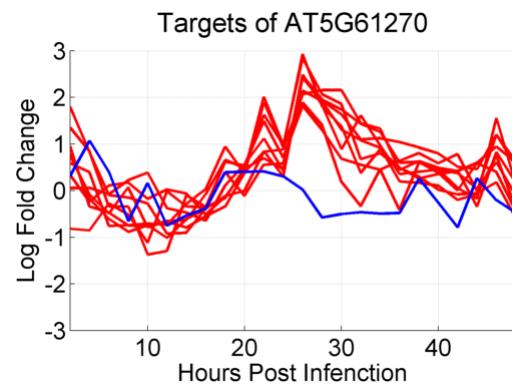
(c) AtHB25 (blue), target genes (red)



(d) ESE1 (blue), target genes (red)



(e) bZIP (blue), target genes (red)



(f) PIF7 (blue), target genes (red)

Figure 3.11: mRNA expression levels in response to infection with *B. cinerea* of common Y1H regulators (blue) and corresponding target genes (red).

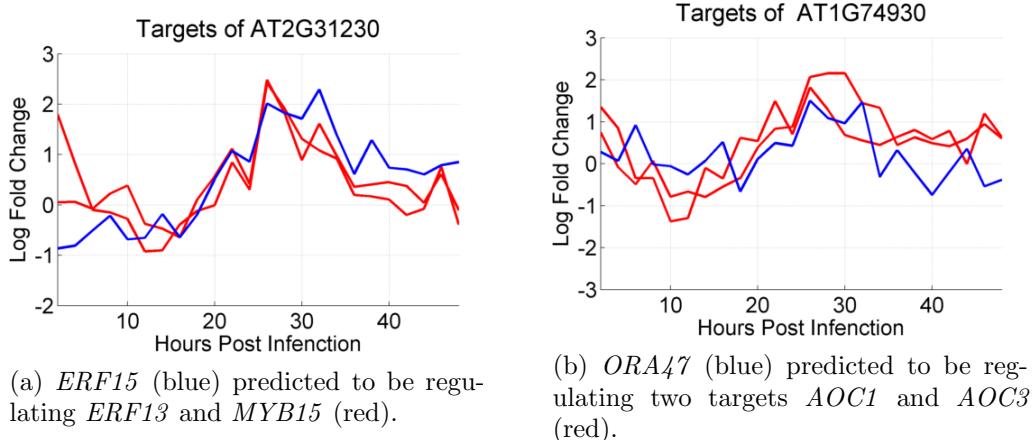


Figure 3.12: mRNA expression levels during infection with Botrytis of two TFs (blue) predicted to be regulating a small subgroup of genes (red) in Y1H screens.

involved in the transcriptional regulation of a co-regulated set of genes; and potential subgroups of genes that may be regulated by the same TFs in the same or different contexts, to be identified 3.12. Notably, *AOC1* and *AOC3* are known to be important elements in the JA biosynthetic pathway (Vick and Zimmerman, 1984; Schaller, 2001) and form a subgroup together with the TF *ORA47* that may additionally regulate their expression during the fungal infection process or elsewhere. Similarly, *ERF13* and *MYB15* are potentially co-regulated by *ERF15*. Although, to date, not published as having a role in regulating defence signalling, *ERF15* is the closest homologue of *ORA59* based on its nucleotide sequence and forms a separate clade with *ERF1* and *ORA59* based on their similarity at the conserved AP2 domain (McGrath et al., 2005; Nakano et al., 2006), both of which have been previously identified as playing an important role in stress response (Berrocal-Lobo et al., 2002; Pre et al., 2008)

### 3.5.6 Circadian Clock Timing Stress Response

Previous studies have reported the role of the circadian clock in precise timing of expression of stress responsive genes for maximum effect on the potential threat (Goodspeed et al., 2012; Pruneda-Paz et al., 2009; Wang et al., 2011). The presence of the *CCA1* binding site in the CNSs of *WRKY11* and *WRKY40* provided a hypothesis to probe deeper into the role of *CCA1* during fungal infection. One positive interaction has been observed between *CCA1* and *WRKY11* (Y1H-147 promoter fragment), however, this interaction was not observed in three subsequent pairwise screens, Table 3.12. Sequencing *CCA1* CDS from the glycerol stocks revealed the

presence of at least two mutations, Table 3.11, and further internal primers were required to complete the sequencing in the middle of the coding sequence. Observed mutations present in the CDS may prevent the protein from interacting correctly and may abolish binding and/or transcriptional activation altogether. On the other hand, library and pairwise screens revealed *PIF7* to be involved in direct Protein-DNA interactions with nine out of ten co-regulated proteins. *PIF7* is known to interact with phytochromes (Leivar et al., 2008) and function as a transcriptional repressor together with *TOC1* (Kidokoro et al., 2009). mRNA levels during the infection with *B. cinerea* do not change significantly, however, given that *PIF7* has been reported in the past as being a transcriptional repressor, existing *PIF7* protein and mRNA may becomes inactive, changing and stopping the repression of defence responsive genes, even though transcription levels of *PIF7* do not alter.

### 3.6 Conclusions

The importance of gaining an in-depth understanding of the regulatory mechanisms of gene transcription is becoming increasingly important. Ubiquitous gene overexpression or loss-of-function have been shown to have a major negative impact on growth and survivability of the plant (Herms and Mattson, 1992; Clarke et al., 2001; Hua et al., 2001; Jambunathan et al., 2001) especially when linked to increased resistance to such necrotrophic fungus as *B. cinerea*. This chapter focuses on a group of genes that are found to be co-expressed in response to infection with *B. cinerea*. Genes that are co-expressed are more likely to also be co-regulated by the same TF/s (Allococo et al., 2004). The 1kb promoter sequences of 10 genes selected for their unique characteristics in response to infection with *B. cinerea* were split into 30 fragments and screened using a high-throughput Y1H pooled library screen to uncover a comprehensive GRN. Some of the selected genes also contain strong CNSs within their promoter regions and had a potential binding site for *CCA1*, one of the core circadian clock genes in *A. thaliana*. In order to validate positive interactions identified in the library screen, 126 TFs responsible for 131 interactions within the promoters of the 10 genes were assembled into a mini-library and the CDS was sequence verified against published sequences (TAIR 10). A mini-library was screened against all promoter fragments again in triplicate to confirm previously detected interactions and identify any false negative results from the library screen. In total, 204 interactions were identified from the combined library and pairwise screens. 83 positive interactions identified from the library screens were not reproduced and conversely 48 were confirmed by the pairwise screen. Analysis of the library has revealed a FNR of 76% in terms of the number of new interactions uncovered by the pairwise screen. Reproducibility of the pairwise screen was 83%, in-line with the previously published 90% reproducibility of enhanced Y1H screen (Reece-Hoyes et al., 2011b).

The evidence obtained from the application of a new image-based technique for identification of positive interactions from the combined pairwise and library screens, suggests that five TFs are able to interact and therefore regulate at least 7 (70%) of the selected genes, and three TFs are likely to be regulating 9 (90%) genes. Additionally, the Y1H screen was able to identify individual TFs that are likely to be regulating a small subgroup of the selected genes, e.g. *ORA47*. While many assumptions accompany Y1H screens, e.g. correct folding of the proteins, competition or lack thereof with other proteins and other factors associated with experiments conducted in a foreign organism, taken together, results from the combined screens

present a highly connected GRN with the majority of core TFs that are able to bind up to 70% of the selected genes to be differentially expressed in response to infection with *B. cinerea*, Figure 3.13.

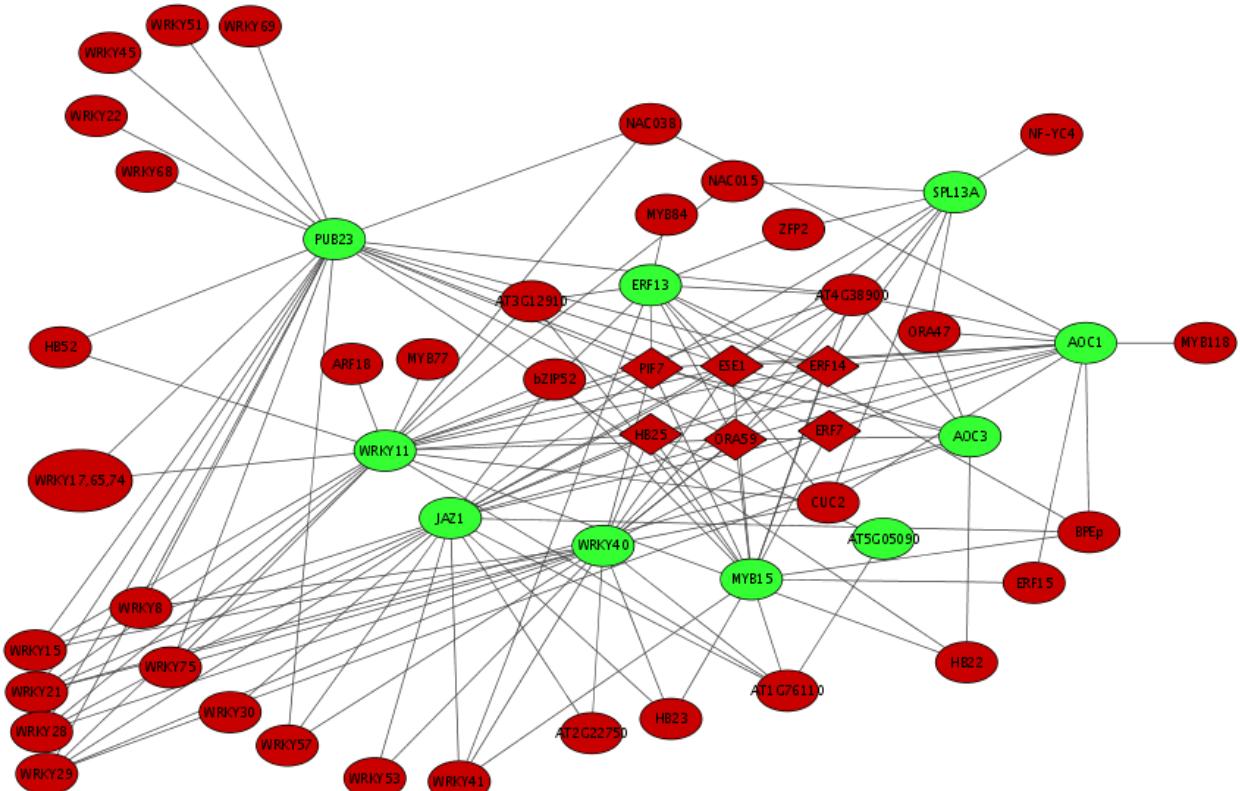


Figure 3.13: Y1H GRN derived from positive interactions in the library screen and validated using three rounds of pairwise screens. Green - promoter fragments of genes used as targets for direct protein-DNA interactions, red - TFs used as sources for protein-DNA interactions, rhombus - TFs regulating over 70% of target genes.

## Chapter 4

# Computational Approaches To Identify Regulatory Elements In Arabidopsis

### 4.1 Introduction

Using predicted direct protein-DNA interactions observed from the Y1H screen obtained in the previous chapter, this chapter aims to further test these interactions. Firstly, by computationally predicting *de novo* binding motifs where interactions could occur using a mixture of bioinformatics and publicly available DNase I sensitivity data (Zhang et al., 2012), then by mutating these motifs and carrying out the Y1H screen once more in order to validate binding of TFs to the predicted motifs. Furthermore, two AP2/ERF domain containing TFs, ESE1 and ERF14 have been shown to interact with 9 and 7, respectively of the 10 promoters tested in the previous chapter. ESE1 has been reported to play a role in the response to salt stress (Zhang et al., 2011) and ERF14 plays a non-redundant role in the Arabidopsis defence response (Oñate-Sánchez et al., 2007), however more detailed information about direct targets of these two TFs is missing. Therefore, constructs overexpressing *ERF14* and *ESE1* will be tested in protoplasts and subjected to microarray analysis in order to identify direct and indirect genome-wide targets. Additionally, a T-DNA knockout line is available for *ERF14*, which will also be used to uncover regulatory targets of this TF. Finally, stable transgenic Arabidopsis lines carrying a T-DNA insertion knocking out the expression of a single TF predicted to be regulating a large proportion of differentially expressed genes, namely *erf14*, *pif7* and *athb25*, will be analysed for altered susceptibility to infection with Botrytis, pro-

viding further evidence of the TFs' role in the plants' defence response network to biotic stress.

#### 4.1.1 TF Binding Site Availability

The main mode of action for TFs is DNA binding through DNA-binding domain(s) contained within the amino acid sequence of the protein or created through the formation of homo- and hetero-dimers (Riechmann et al., 1996). Recent estimates put the number of TFs in *Arabidopsis* at around 1500 (Riechmann et al., 1996), although some reports suggest this figure may be greater than 2000 TFs (Davuluri et al., 2003; Guo et al., 2005; Iida et al., 2005; Riano-Pachon et al., 2007). PLACE database (Higo et al., 1999) reports 508 plant-related regulatory motifs found to be bound by a TF, with only 117 (23%) motifs derived from *Arabidopsis* itself. Rice (*Oryza sativa*), tobacco (*Nicotiana tabacum*) and maize (*Zea mays*) are the other major contributors with 74 (15%), 46 (9%) and 45 (9%) respectively. It is assumed that orthologous genes in other species would bind to similar, if not identical sequences in other species. Additionally, there are only a few motifs associated with whole families of TFs, for example the AP2/ERF domain family of TFs consists of 122 genes in *Arabidopsis* (Nakano et al., 2006), however PLACE only presents 4 different binding motifs for ERF responsive elements. Although some AP2/ERF TFs have been found to be binding the GCC-/AGCC-box (*ERF1* (Cheng et al., 2013) and *ORA59* (Zarei et al., 2011)) it is unlikely that the majority of the AP2/ERF family binds to the same sequence given the large diversity in AP2 domain structure and arrangements within the family itself (Mizoi et al., 2012).

There are two main barriers to uncovering sequence specific binding of a TF family. Firstly, nuclear extracts of the TFs are costly and, more importantly, time-consuming when dealing with approximately 100 different TFs. Secondly, there is a lack of experimental approaches for testing a large number of TFs for potential interactions with all possible k-mer sequences. For example, 122 AP2/ERF TFs tested using the EMSA technique against synthesised arrangements of 6 DNA bases (based on the length of the GCC-box) adds up to a total 158,112 individual experiments. This unrealistically high number means that it is not feasible to undertake such a study given current techniques. Therefore, there is a gap in the knowledge of the sequence specific binding sites associated with individual proteins.

#### 4.1.2 Verification in plants

One of the major challenges of identifying new protein-DNA interactions using *in vitro* or *in vivo* techniques is that the context may be different from that in which these interactions take place *in planta*. For example, interactions identified using Y1H or EMSA techniques are thought to be “context-free”, as they provide evidence for the ability of certain sequences to interact with TFs of interest. Such evidence is assumed to be independent of other factors potentially influencing the ability of TFs to bind DNA, e.g. condition specific histone modifications (Kim et al., 2008). Therefore “context-free” protein-DNA interactions need to be validated *in planta*, in order to establish their functionality under the conditions of interest, as well as demonstrating that observed interactions were not a product of performing the experiment in another species, e.g. yeast in the case of Y1H screens. For example, some bHLH TFs have been suggested to misfold in yeast assays, but bind to corresponding G-box sequences in other experiments (Chow et al., 2008; Xu et al., 2009). Other factors may also include competition between endogenous and TFs of interest for binding to DNA sequence leading to false negatives, as concluded by Dreier et al. (2001) in their study of human zinc finger domain TFs.

There are several strategies for verifying predicted “context-free” protein-DNA interactions and linking them to the specific contexts or conditions under which these interactions take place *in planta*. For example, if an antibody is available for the TF of interest, then ChIP-Seq experiments (Johnson et al., 2007) can be carried out under conditions under which the interactions are thought to take place. As well as uncovering whether or not a TF binds to the promoter of interest, the ChIP-Seq technique can also identify all other genome-wide targets of the TF. All observed interactions also provide information for any DNA sequence dependence between a TF and its ability to bind. A limitation of the ChIP-Seq technique, given that an antibody is available for the TF, is that it is not known whether the gene associated with the promoter binding the TF, is actively transcribed. In order to establish transcriptional events, RNA-Seq can be utilised (Nagalakshmi et al., 2008). In RNA-Seq experiments, mRNA transcripts are sequenced and mapped onto a reference genome. A combination of ChIP-Seq and RNA-Seq provides strong evidence for the TF binding events and the transcription of the associated gene, as indicated by a higher level of mRNA as compared to a control sample. Alternatively, constructs overexpressing the TF of interest can be introduced into host cells using PEG-mediated uptake (Paszkowski et al., 1984) or by electroporation (Deshayes et al., 1985) and after a period of incubation, typically 24 h, mRNA can

be extracted and hybridised to microarray slides, revealing downstream targets associated with the TF. However, because protoplasts are used, it is difficult to link protein-DNA interactions to a specific context and so this technique serves to validate proposed interactions in plant cells only. On the other hand, microinjection (Jaenisch and Mintz, 1974) or biolistic (Kikkert, 1993) techniques can be used to quickly introduce foreign DNA containing the TF of interest directly into plants by bombarding the leaf with high speed gold or tungsten particles coated in the TF DNA sequence. Although both these methods are time consuming and relatively low throughput, they can be applied directly in a condition specific manner. For example, wheat plants overexpressing the *Arabidopsis DREB1A* gene introduced using biolistics were found to be more tolerant to drought stress (Pellegrineschi et al., 2004). Finally, transfection using *Agrobacterium tumefaciens* result in stable transgenic lines harbouring T-DNA insertions in genes of interest, thus reducing or completely abolishing expression of these genes. Homozygous plants with the mutated gene(s) can be directly subjected to a variety of conditions and subsequently analysed using ChIP-Seq, RNA-Seq or microarray technologies to uncover downstream targets of the mutated gene(s).

#### 4.1.3 DNase Assays

Since the early 1970's, it has been known that DNA assembles into nucleosomes, which subsequently form higher order chromatin structures. An immediate question arose, whether or not the chromatin structure was different between active and inactive genes. Weintraub and Groudine investigated if the chromatin structure was equally accessible for DNase I digestion between active and inactive chicken globin genes isolated from different cell types. For their experiment, genomic DNA isolated from the nuclei of different cell types was treated with increasing concentrations of DNase I. After digestion, DNA was purified and cleaved with restriction enzymes to produce known size fragments from globin loci. The DNA fragments were subsequently separated in agarose gel and transferred to nitrocellulose membrane. Specific restriction fragments were detected by hybridisation to radiolabelled probes from globin and control genes. In the samples derived from erythrocyte nuclei, the DNA was digested efficiently by DNase I, in contrast to samples derived from other cell types where it was not digested as efficiently. From this "DNase sensitivity assay" Weintraub and Groudine concluded that the entire chromatin structure of the region is altered in order to allow for active transcription of the globin genes in erythrocytes (Weintraub and Groudine, 1976).

Later, DNase I was used to find the exact location of TFs binding along a DNA fragment using “DNase footprinting” (Galas and Schmitz, 1978). This method was developed as an amalgamation of two earlier methods, Maxam-Gilbert DNA sequencing (Maxam and Gilbert, 1977) and DNase-protected fragment isolation (Schaller et al., 1976) and has subsequently developed a large following in the experimental community due to its simplicity. Conceptually, the DNA sequence is radioactively labelled on one end and then partially digested by DNase I enzyme. The digested DNA produces a ladder of various size fragments whose mobility on a polyacrylamide gel determines the distance from the cleavage site to the radioactively labelled end. Bound TFs protect the DNA sequence from cleavage by DNase I in and around the binding site, generating a “footprint” in the cleavage ladder. The distance between the labelled end and the cleavage site represents the distance to the protein binding site and can be exactly determined by running a standard DNA ladder alongside the footprint.

Originally designed to test protein-DNA interactions of small sequences, recent developments in high-throughput sequencing technologies have lead to a new technique, named DNase-Seq, for assessing open/close chromatin areas of the whole genome (Boyle et al., 2008). Open areas of the chromatin are more accessible for TFs to bind and therefore also more accessible to digestion by the DNase I restriction enzyme. Coupled with deep sequencing technologies, open areas of the chromatin are associated with a large number of sequencing reads, also called DNase I hypersensitive sites, Figure 4.1, (Boyle et al., 2008). Conversely, closed areas of the chromatin lack sequencing reads. This technique has been used to map open chromatin areas in Arabidopsis and human genomes (Zhang et al., 2012; Neph et al., 2012). As an additional byproduct of DNase I hypersensitive sites, a footprint of DNase I cuts can be identified around the TF binding sites. The footprint is characterised by an increased number of cuts immediately prior to and after the TF binding site. Therefore, it is possible to associate the DNA sequence being bound by a TF. For example, SEPALLATA3 (SEP3) TF, is known to bind to MADS box 5'-CC(A/T)(A/T)(A/T)(A/T)(A/T)GG-3' (Kaufmann et al., 2009) and function in Arabidopsis flower development (Mandel et al., 1992; Liljegren et al., 1999; Pelaz et al., 2000a; Vandenbussche et al., 2003). Analysis of the MADS box overlapping with the DNase I hypersensitive areas in leaves and flower buds shows a footprint generated by the TF bound to the sequence motif, Figure 4.2. However, unless the TF interacting with the sequence motif is known *a priori*, it is not pos-

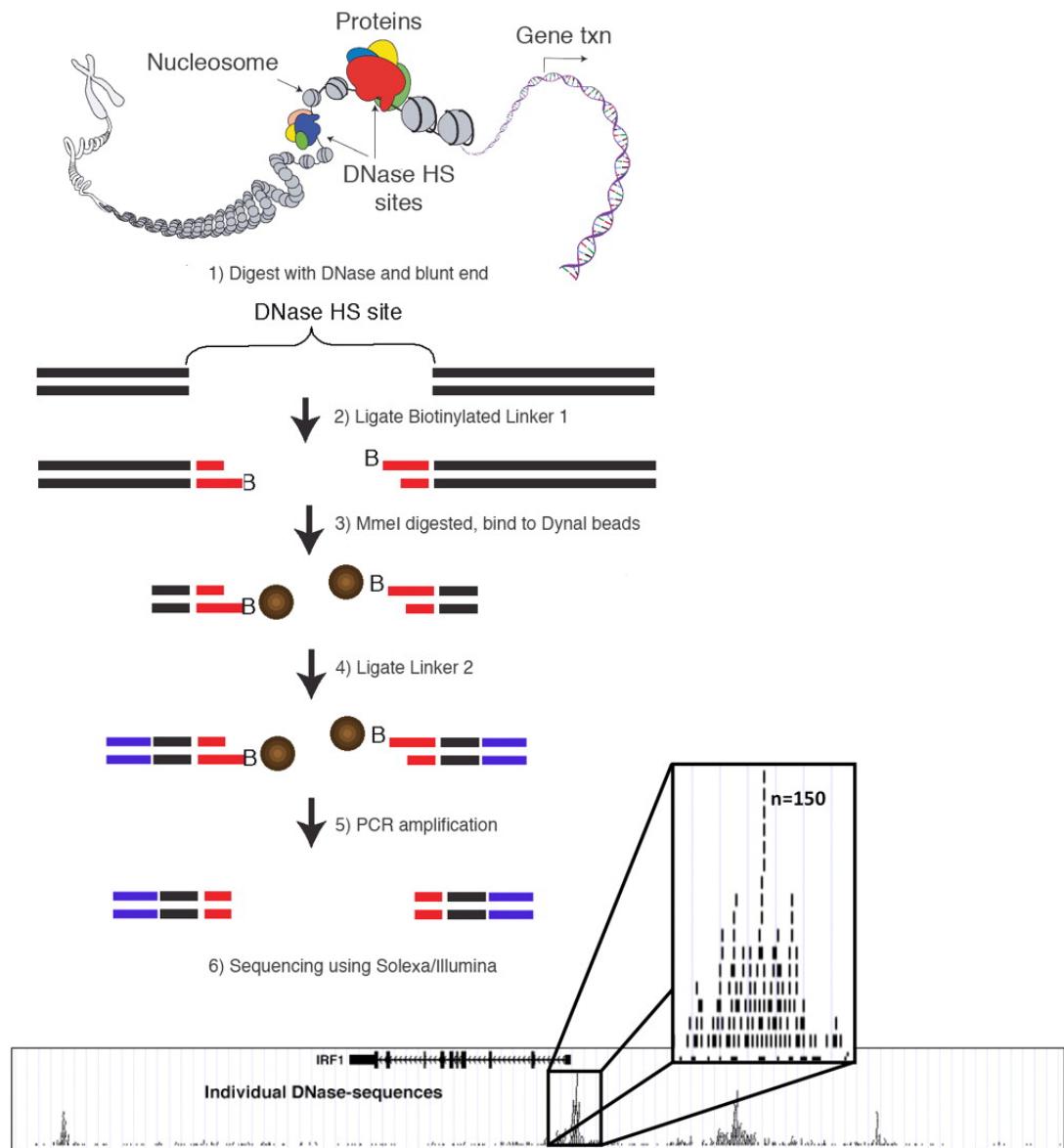


Figure 4.1: Flowchart of DNase-Seq protocol. Briefly, cells are lysed with detergent to release nuclei, and the nuclei are digested with optimal concentrations of DNase I. DNase I-digested DNA is embedded in low-melt gel agarose plugs to reduce additional random shearing. DNA (while still in the plugs) is then blunt-ended, extracted, and ligated to biotinylated linker 1 (red bars). Excess linker is removed by gel purification. Biotinylated fragments (linker 1 plus 20 bases of genomic DNA) are digested with MmeI and captured by streptavidin-coated Dynal beads (brown balls). Linker 2 (blue bars) is ligated to the 2-base overhang generated by MmeI, and the ditagged 20-bp DNAs are amplified by PCR and sequenced by Illumina/Solexa (with permission from Cold Spring Harbour Laboratory Press, (Song and Crawford, 2010)).

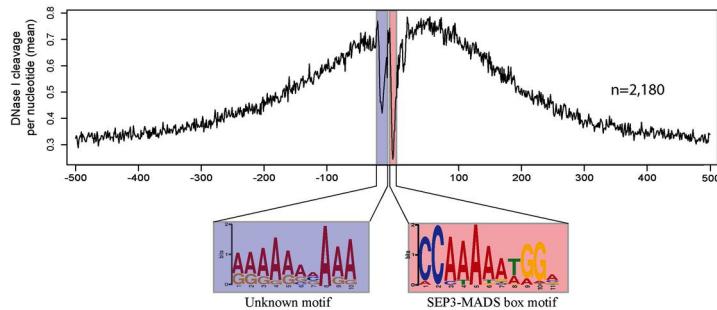


Figure 4.2: SEP3 binding footprints revealed by DNase I hypersensitive sites that overlap with SEP3 binding sites. The x axis represents the distance from the SEP3 motif, and the y axis represents the mean DNase I cut per nucleotide (Zhang et al., 2012).

sible to identify the TF from the footprint data alone.

In summary, there is a gap in the knowledge of the sequence specificity of TFs, such that binding site(s) identified for a single motif are used to represent the whole family of TFs. In order to bridge this gap, this chapter aims to provide additional information on sequence specific binding of some of the TFs used in the Y1H screen described in the previous chapter, by employing a mixture of sequence analysis of promoter fragments to predict where interactions could occur, and publicly available DNase I data (Zhang et al., 2012). Although Y1H screens allow the identification of “context-free” protein-DNA interactions, these need to be validated further *in planta* to provide more information about the conditions under which such interactions may take place. Furthermore, downstream targets of newly reported interactions from the Y1H screen can be obtained using stable transgenic lines harbouring T-DNA insertions in the TFs of interest or using overexpression constructs in protoplasts. mRNA levels from either tissue source can be hybridised to microarrays and analysed to identify direct and indirect targets. Plants with the mutated TFs can also be analysed for altered susceptibility to infection with *Botrytis*, providing further evidence of the TFs functional role in a stress responsive network.

## 4.2 Methods

### 4.2.1 Promoter DNA analysis

Promoter DNA sequences were stored in multiple fasta format and analysed using MeMe v4.9.0 (Bailey et al., 2009). Custom Perl script was used to obtain a set of promoter sequences given an ATG of a TF that had positive interactions with select promoters in Y1H screens. Sequences of promoter fragments that were not found to be interacting with the given TF were bundled together to create a set of negative sequences used as prior in MeMe analysis (“psp-gen” command). 15 conserved motifs were extracted for each set of binding promoters with specified minimum and maximum width of the motif and whether it ought to be palindromic. By default, number of motif occurrences to look for was set at “zero or once per sequence” (zoops) with minimum number of sites set to total number of positive fragments used. However, this option could be manually overridden to “any number” (anr) or “only once per sequence” (oops). Y1H scores were included as individual weights in the motif analysis for each promoter fragment. Scores [0,10] were converted to weights [0,1] by dividing Y1H scores by 10. MeMe was set to output in both HTML and plain text formats.

### 4.2.2 DNase Analysis

Bowtie aligned sequences from Zhang et al. (2012) were extracted from GEO repository (Edgar et al., 2002) (ascension numbers: GSM847326, GSM847327, GSM847329, GSM847329, GSM847330) and converted into BAM format using SAMTools v0.1.18-dev (Li et al., 2009). HOMER v4.1 (Heinz et al., 2010) was used to extract hypersensitive sites in BED format individually for each of the 5 samples using Arabidopsis v1.3 (TAIR10) reference provided by HOMER. A custom parallel application was written in C to scan for presence or absence of a motif in promoters of Arabidopsis genome (TAIR10). A motif was deemed to be present if it matched exactly one of the sequences representing the motif, which is also included as part of the motif output from MeMe. The locations of sites matching MeMe motifs were checked for overlap with hypersensitive sites for each sample using a custom Perl script. Wellington, a novel tool for analysing DNase data (Jason Piper, unpublished) was used to extract information about the starts or end of the sequencing read, where DNase has broken the double stranded DNA, and plot them using “matrix2png” package for Python, as well as plot average cut profile using “averageProfilePlotter” Python script (courtesy of Jason Piper). The colour of the spots in DNase figure represents the strand the cut is located: red - positive strand, green - negative. The

brightness of the spot on a DNase image represents number of cuts, reads starting or ending at that position, brighter the spot, larger number of reads it represents.

#### 4.2.3 Plant growth

Arabidopsis seeds were stratified in 0.1% w/v agarose at 4°C for 72 h in complete darkness. Stratified seeds were sown in pre-watered Arabidopsts soil mix (6:1:1 ratio of Levington F2 compost:sand:vermiculite) containing Intercept (Everris, Ipswich) in 4-cm pots (P24, Plant-pak). Pots were converted with cling film and placed in a growth chamber to germinate. The covering was removed 7 days post sowing and seedlings thinned out to give one plant per pot. Plants were grown in standardised conditions under 16 h light 18 h dark days at 20°C, 350 ppm CO<sub>2</sub> concentration and 120 µmol m<sup>2</sup> s<sup>-1</sup> light.

#### 4.2.4 Fungal growth

*Botrytis cinerea*(Botrytis) strain pepper spores (Denby et al., 2004) were germinated and cultured on sterile tinned apricot halves (Tesco, UK) in petri dishes 4 weeks prior to use. Two weeks prior to use, Botrytis was sub-cultured using the same procedure. Sub-cultures were incubated at 25°C in complete darkness. Spores were harvested in sterile water and filtered through glass wool to remove hyphae. Inoculums were prepared by suspending spores in half strength sterile grape juice (Tesco, UK) at a concentration adjusted to  $1 \times 10^5$  spores/ml. Spores concentration was measured with a hemocytometer.

#### 4.2.5 Phenotype Analysis

Plant leaves inoculated with Botrytis were grown in 3 trays (10-15 leaves per line in each tray) at 90% humidity level in sealed trays in 12 h day/night cycles. Images were taken of the trays with the leaves for phenotypic analysis at 48 h, 57.5 h and 72 h post infection. Area around each fungus in every leaf was manually draw and calculated using ImageJ (Schneider et al., 2012) software. The calculated areas were analysed using one-sample Kolmogorov-Smirnov test for normality, Table C.1. The areas for each samples were normalised using mean and standard deviation within the sample. All samples were approximately derived from a normal distribution and therefore, t-test was used to determine whether there was statistical difference between each sample and control (Col0).

#### **4.2.6 Microarray analysis**

##### **RNA extraction**

Snap frozen *Arabidopsis* leaves were ground in 1 ml Trizol reagent (Invitrogen, Paisley) using a Dremel drill for 1 min until the sample was completely homogenised. The drill-bit was frozen in liquid nitrogen prior to use to prevent thawing of leaf tissue. Samples were incubated at room temperature for 5 min to allow for dissociation of nuclear protein complex before adding 200 µl chloroform. Reaction was shaken vigorously by hand for 15 s and incubated for a further 3 min at room temperature. Samples were centrifuged at  $8000 \times g$  for 15 min at 4°C. The upper aqueous phase (60% of the volume) was transferred to a fresh 1.5 ml Eppendorf tube followed by addition of 0.5 ml of isopropanol to precipitate the RNA. Samples were mixed by inverting tubes several times and incubated at -20°C for 2 h. Samples were centrifuged at  $8,000 \times g$  for 20 min at 4°C. RNA pellets were washed with 1 ml of 75% EtIH followed by centrifugation at  $8,000 \times g$  for 10 min at 4°C. The supernatant was completely removed and pellet allowed to air-dry for 5 min before re-suspension in 100 µl RNase free water. Total RNA was purified using Qiagen RNeasy purification kit (Qiagen, Manchester) according to the manufacturer's instructions, except for the final step where purified RNA was eluted from the column with 2 x 40 µl RNase free water. Total RNA concentration was measured using a Nanodrop ND-1000 spectrophotometer (Thermo-Scientific, Nortumberland) using 1 µl sample. Total RNA quality was determined using a 2100 Bioanalyser with the RNA 6000 Nano LabChip kit according to the manufacturer's instructions (Agilent). The Bioanalyser assesses total RNA integrity by measuring the 18S and 28S rRNA peaks using high-resolution electrophoresis system. Where total RNA samples displayed no rRNA peaks or a poor 18S/28S ratio (< 1), total RNA was isolated from alternative leaf samples.

##### **RNA amplification**

Total RNA was amplified using MessageAmp-II aRNA Amplification Kit (Invitrogen, Paisley) according to the manufacturer's instructions, using a single round of amplification and an *in vitro* transcription step and an incubation time of 14 h. The quality of amplified RNA was determined using a 2100 Bioanalyser with the RNA 6000 Nano LabChip kit according to the manufacturer's instructions (Agilent, Wokingham) and concentration of the purified sample was measured using a Nanodrop ND-1000 spectrophotometer. Good quality amplified RNA should display a size distribution that is approximately a normal distribution (bell shaped) Where

the size distribution was clearly abnormal or the amplified RNA concentration was  $< 300 \text{ ng } \mu\text{l}^{-1}$ , total RNA for that sample was re-amplified.

### **Microarray experimental design**

*Erf* plants infected with Botrytis were compared to Col-0 plants also infected with Botrytis 24 h post infection. RNA from 4 biological replicates were pooled after amplification for mutant and wild type leaves. Comparisons were made using a total of 4 technical replicates: 2 replicates for one set of dyes, then 2 replicates after dyes were swapped. In the protoplast experiment overexpression vectors for *ERF14* and *ESE1* were compared to protoplasts where no vector was added, but otherwise went through the same process. For each sample, RNA from 4 technical replicates were pooled after amplification. Comparisons were made using a total of 4 technical replicates after pooling: 2 replicates for one set of dyes, then 2 replicates after dyes were swapped.

### **Direct labelling of amplified RNA**

Approximately 5  $\mu\text{g}$  of pooled amplified RNA, generated by combining equal amounts of amplified RNA from each of the appropriate biological, or technical replicates in case of protoplasts, was combined with 0.5  $\mu\text{l}$  of random nanomer ( $3 \mu\text{g ml}^{-1}$ ) (Invitrogen, Paisley) and 0.5  $\mu\text{l}$  of RNase inhibitor (RNase OUT; Invitrogen) for total volume of 10.5  $\mu\text{l}$ . Samples were incubated at 70°C in a thermocycler for 10 min. Superscript mastermix was created by combining the following reagents per reaction: 4  $\mu\text{l}$  5  $\times$  Superscript II First Strand Buffer (Invitrogen, Paisley), 2  $\mu\text{l}$  0.1 mol DTT (Invitrogen), 1  $\mu\text{l}$  dNTP mix (10 mmol dATP, 10 mmol dCTP, 10 mmol dGTP, 10 mmol dTTP) and 1  $\mu\text{l}$  Superscript II reverse transcriptase (Invitrogen, Paisley). Samples were labelled by adding 8  $\mu\text{l}$  of superscript mastermix with 1.5  $\mu\text{l}$  of either Cy3- or Cy5-dCTP (GE Healthcare, Chalfont St Giles) followed by incubation in the dark at 42°C for 2.5 h. 2  $\mu\text{l}$  of 2.5 mol NaOH was added to each of the labelled cDNA samples followed by incubation at 37°C for 15 min. Samples were combined with 10  $\mu\text{l}$  of 2 mol MOPS buffer and purified using QiaQuick PCR purification kit (Qiagen, Manchester) according to the manufacturer's instructions. At the end of the procedure, the purified cDNA was eluted with 2  $\times$  30  $\mu\text{l}$  of Buffer EB (Qiagen, Manchester). The concentration of purified sample was measured at 532 nm (Cy3) or 635 nm (Cy5) wavelength using nanodrop ND-1000 spectrophotometer. Samples were kept in the dark throughout labelling steps to minimise light degradation.

### CATMA array hybridisation

CATMAv4 (Hilson et al., 2004) array slides were prepared for hybridisation by incubating them in Coplin jars with Pre-Hybridisation buffer (1.2g Bovine Serum Albumin (Sigma-Aldrich, A9418), 5 × SSC, 0.1% SDS in 120 ml sterile water) (pre-warmed to 42°C in an air incubator) for 1 h. CATMA array slides were washed by submerging in sterile water for a total of 5 washes and a final wash with isopropanol. Slides were dried by centrifugation for 1 min at 2000 × *g*. Combinations of up to 40 pmol of the appropriate Cy3- or Cy5-labelled samples were freeze dried until nearly dry and resuspended in 50 µl of hybridisation buffer (12.5 µl Formidem, 12.5 µl 20 × SSC, 0.35 µl of 14% SDS, 6.25 µl of 4 µg ml<sup>-1</sup> Yeast tRNA (Invitrogen, Paisley) and 18.4 µl sterile water). Resuspended samples were incubated at 95°C for 5 min in a thermocycler followed by centrifugation at 10000 × *g* for 1 min. The hybridisation mix was applied to an array slide located within hybridisation chamber (Corning, Corning) followed by the application of a coverslip (Sigma Aldrich, Gillingham) and chamber cover. Hybridisation chambers were placed in a humid environment at 42°C for 16 h.

Coverslips were removed by submerging array slides in 250 ml of Wash Solution 1 (2 × *SSC*, 0.07% SDS and 250 ml sterile water) (preheated to 42°C in an air incubator) until free. Hybridised slides were then incubated in Wash Solution 1 in a hybridisation rack for 5 min with gentle shaking. Slides were then incubated in 250 ml of Wash Solution 2 (0.1 × *SSC*, 0.07% SDS, 250 ml sterile water) for 10 min with gentle shaking. Then slides were incubated in Wash Solution 3 (0.1 × *SSC* and 995 ml of sterile water) for 1 min with gentle shaking for a total of 4 washes. Finally, slides were briefly immersed in isopropanol and dried by centrifugation at 1000 × *g* for 1 min.

#### 4.2.7 Microarray scanning

Array slides were scanned on Affymetrix slide scanner with default settings. The data extracted from scanned images was quantified using Imagene 7.5 software (BioDiscovery, Inc.) (Médigue et al., 1999).

#### 4.2.8 Expression Analysis

Comparisons between mutant and wild-type samples were analysed using R (Bioconductor) (Gentleman et al., 2004) package limmaGUI (Wettenhall and Smyth, 2004). Raw data was normalised within arrays using PrintTip lowess transformation and then normalised between arrays using aquantile-normalisation. The data

was fitted with a linear model using the least squares method. P-values were adjusted for multiple testing using Benjamini and Hochberg method with threshold of 0.05 to control for false discovery rate.

#### 4.2.9 Gene Ontology (GO) analysis

Overrepresentation within gene lists for Gene Ontology (GO) (Berardini et al., 2004) terms was done using BiNGO (Maere et al., 2005) plugin for Cytoscape 2.8 (Shannon et al., 2003). BiNGO performs hypergeometric tests to determine if a particular GO term associated with a set of genes is expected by chance when considering the number of genes associated with that term in the whole *Arabidopsis* genome.

#### 4.2.10 Pathway Analysis with MapMan

MapMan (version 3.6.0RC1 (Thimm et al., 2004)) was used to visualise changes in gene expression in mutant plants and protoplasts as compared to Col-0 and no vector respectively. Catma probe IDs were mapped to best TAIR IDs using BLAST (provided by Jonathan Moore) and MapMan TAIR9 annotation set was used to map known *Arabidopsis* ATG ids from microarray analysis.

## 4.3 Results

### 4.3.1 New TF specific motifs conserved within the promoter fragments of gene tested in Y1H screen

As described in the introduction to this chapter, there is a gap in the knowledge about the sequence specific interactions of individual TFs. Due to the large number of possible sequence variations,  $n^4$  where  $n$  is the length of the binding site, it is not feasible to test for all potential interactions using current *in vitro* techniques, e.g. EMSA. Instead, when a sequence is identified for a member of a TF family, this sequence is also used as a good proxy for other members of the family. For example, the WRKY TFs are generally thought to recognise the 5'-TTGACY-3' core sequence (Eulgem et al., 2000), however Ciolkowski et al. (2008) have shown that some WRKY TFs have different tolerances for mutations in the core WRKY sequence and in the nucleotides immediately adjacent to it. This suggests that although using a single core sequence maybe a good proxy for interactions, it may not be sufficient to identify specific TFs interacting with the DNA sequence.

The results obtained from the previous chapter, Table 3.12, shows that some groups of TFs preferentially interact with some promoter fragments, whilst others do not. For example, WRKY15/17/21/22/29/41/65/68/69 were found to be interacting with the promoter sequences of the PUB23 gene. However, WRKY15/21/29 were also found to be interacting with the promoter sequences of WRKY11, WRKY40 and JAZ1, whereas the others did not. Given the large number of positive results observed from the Y1H screens, these can be combined together and used for *de novo* motif discovery using the MeMe software (Bailey et al., 2009). Additionally, sequences not binding the proteins of interest are equally as valuable as they shouldn't contain the conserved sequence that the protein binds to and therefore can be used in construction of background models to filter potentially common elements in the promoter fragments that are not directly responsible for protein-DNA interactions. The number of available sequences plays an important part in the confidence of newly found motifs. The strength of interactions seen in the Y1H screen, represented through scoring criteria in the screen, can be directly applied to motif discovery in the form of weighting of individual sequence contributions to the motif, and set programatically for each TF.

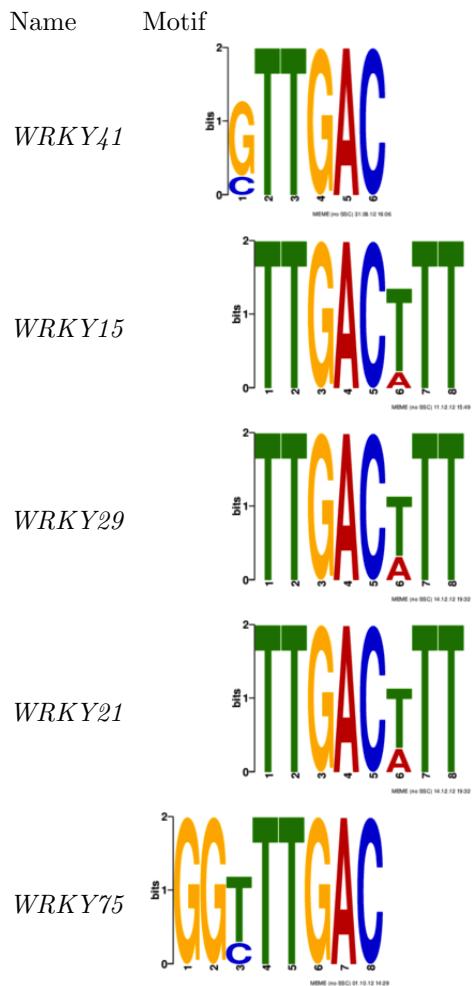
## Conserved motifs found in the promoters interacting with the WRKY TFs

Although MeMe is able to find conserved regions between only two sequences, these are not statistically significant as there is a high chance of short sequences being conserved by chance alone, especially if promoters are adjacent to each other and overlap by 50-100 base pairs. Therefore, only WRKY15/21/29/41/75 were considered for this analysis as they occur in 6, 7, 7, 5 and 4 promoters respectively. For all WRKY TFs the core 5'-TTGACY-3' sequence was present in all motifs discovered by MeMe, Table 4.1. Additionally, different members of the WRKY family were identified as having different nucleotides conserved around the core binding sequence. The promoter fragments interacting with the WRKY41 TF were found to only contain the core binding motif. The fragments interacting with the WRKY15/21/29 had (A/T)TT conserved immediately downstream from the core sequence. Finally, the fragments interacting with the WRKY75 had GG(C/T) conserved immediately upstream from the core sequence.

## Conserved motifs found in promoters interacting with the AP2 domain TFs

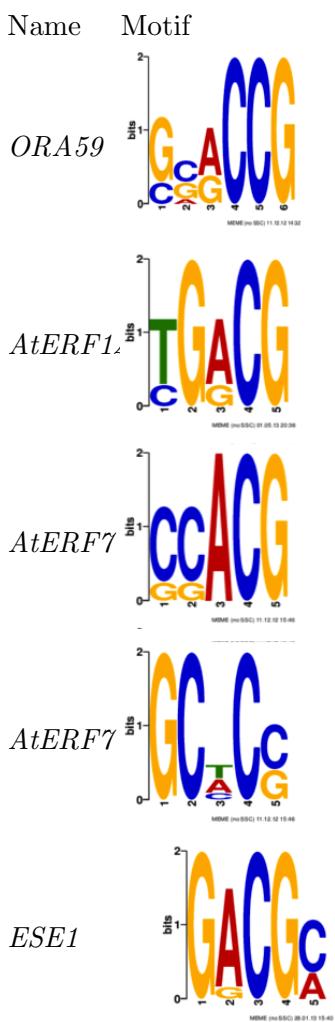
The TFs containing AP2/ERF DNA binding domain, e.g. *ORA59*, *ESE1*, *AtERF7*, *AtERF14* have been previously linked to be regulating components of the stress response pathways in Arabidopsis (Pre et al., 2008; Zhang et al., 2011; Oñate-Sánchez et al., 2007; Fujimoto et al., 2000). They also appear to be

Table 4.1: Motif logos identified by the MeMe software in the fragments found to be interacting with WRKY TFs. The height of each letter represents conservation across different sites where the motif is present.



interacting with a large proportion of the promoters tested in the previous chapter, therefore it would be advantageous to determine any sequence specificity associated with the binding of different AP2 domain proteins in *Arabidopsis*.

Table 4.2: Motif logos identified by the MeMe software in the fragments found to be interacting with AP2 TFs. The height of each letter represents conservation across different sites where the motif is present.



The promoter fragments found to be interacting with the aforementioned AP2 domain TFs have also been analysed using MeMe software to uncover any conserved elements that may function as the binding site for the associated TFs, Table 4.2. The GCC-box motif is known to be interacting with at least some members of the AP2 family TFs, e.g. ERF1 and ORA59 (Brown et al., 2003; Zarei et al., 2011; Cheng et al., 2013), however it was not found to be present in all fragments for each TF. This suggests that the AP2 domain TFs may potentially function through different sequence specific elements within the promoter sequences. ORA59 has been found to bind to 90% of the selected genes described in the previous chapter. MeMe analysis has identified a consensus sequence 5'-(G/C)N(A/G)CCG-3' to be present in all fragments interacting with this TF. This motif is similar to the published GCC-box in terms of its high GC content and relative arrangement of the nucleotides with the motif. The 5'-GACG(A/C)-3' consensus motif was found in the promoter sequences interacting with the ESE1 TF. The new ESE1 motif is also somewhat similar to the published GCC-box motif in having a high GC content. MeMe analysis suggested that AtERF14 interacts with the 5'-(T/C)G(A/G)CG-3' consensus sequence. Finally, AtERF7 has two motifs, 5'-(C/G)(C/G)ACG-3' and 5'-GCNC(C/G)-3', these are relatively similar to each other and to the other newly discovered AP2 binding motifs that are potentially able to bind the TF. All new AP2 domain

within the core sequence, consistent with the GCC-box motif sequence.

### **Conserved motifs found in the promoters interacting with PIF7, bZIP52, AtHB25 and NAC098 TFs**

Finally, this group of TFs was found to be interacting with a large proportion of the promoter fragments tested in the previous chapter, Table 4.3. The homeobox (HB) domain TFs have been previously linked with the response to water stress in *Arabidopsis* (Shin et al., 2004; Park et al., 2011). There is also evidence to suggest that HB TFs bind to 5'-CAAT(A/T)ATTG-3' and to 5'-CAAT(G/C)ATTG-3' (Sessa et al., 1993, 1997). MeMe analysis revealed a very similar motif, 5'-CAANTANTTG-3', conserved in the promoter sequences interacting with the AtHB25 TF. The bHLH TFs are known to interact with the G-box, 5'-CACGTG-3', sequence motifs (Toledo-Ortiz et al., 2003), this is also supported by the MeMe analysis of the promoter fragments interacting with the PIF7 TF, where the G-box, or slight variations of it, were found in all promoter fragments. bZIP TFs were also found to bind similar G-box sequences as bHLH TFs (Menkens et al., 1995), however, this analysis showed a different element, consensus sequence 5'-G(T/A)AACC(C/T)C-3', to be conserved amongst the promoters interacting with the bZIP52 TF. Finally, NAC TFs have been previously shown to interact with the 5'-CATGTG-3' motif (Tran et al., 2004). Analysis of the conserved sequences among the promoter fragments interacting with the NAC098 TF suggests that this TF interacts with the strongly conserved 5'-CNGTGG(G/A)-3' consensus sequence. This new motif is similar to the previously characterised motif in its first base and downstream 5'-GTG-3' sequence, suggesting that these nucleotides are responsible for the binding of the TF.

#### **4.3.2 *De novo* motifs show DNase I footprint in genome-wide locations in *Arabidopsis* leaves and buds tissue**

The binding of TFs to the DNA protect the bound sequence when digested with the DNase I enzyme, identifiable by a footprint in DNase-Seq data, e.g. NRF1 TF in humans (Neph et al., 2012). Recently, a large DNase-Seq dataset became available for the *Arabidopsis* genome from leaf and flower tissues (Zhang et al., 2012). As such, DNase-Seq data can be used to test the hypothesis of whether conserved motifs identified by the MeMe analysis of Y1H promoters are bound by any TFs in flower or bud tissue *in planta* using the newly available data.

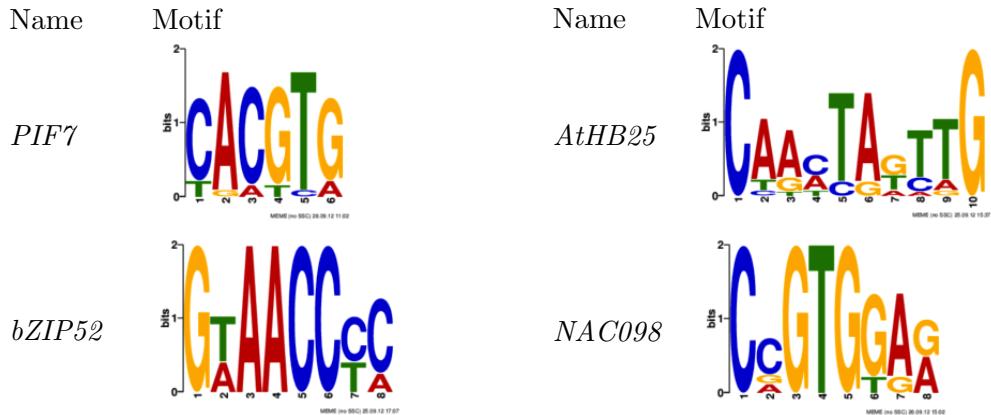


Table 4.3: Motif logos identified by the MeMe software in the fragments found to be interacting with PIF7, bZIP52, AtHB25 and NAC098 TFs. The height of each letter represents conservation across different sites where the motif is present.

The genome wide locations of the AtERF14, ORA59, ESE1 and AtERF7 motifs were identified in the promoters of all Arabidopsis genes using a custom, massively parallel algorithm. At the same time, DNase I hypersensitive sites were separately identified in the Arabidopsis genome for each DNase-Seq sample using Hypergeometric Optimization of Motif EnRichment (HOMER) software (Heinz et al., 2010). DNase profiles were retrieved only for those motifs where locations also overlapped with the hypersensitive site, suggesting transcriptional activity within the immediate vicinity of the motif. A profile of DNase I cuts, one per line, and an average profile of cuts across all sites, graph above, were computed for AtERF14, ORA59, ESE1 and AtERF7 motifs 10 and 14, Figure 4.3a - 4.3h respectively. Although DNase I cut profiles were computed for all motifs identified by the MeMe analysis, only one motif was selected for each TF as being the most likely to be bound by the TF and the most similar to the GCC-box, known to bind other AP2/ERF domain TFs (Cheng et al., 2013; Fujimoto et al., 2000; Zarei et al., 2011). Both motif 10 and motif 14 derived for AtERF7 had similar GC contents and bore a resemblance to the GCC-box and therefore both were chosen for DNase analysis as potential AtERF7 binding sites.

All motifs in Figure 4.3a - 4.3h show a protected site in the middle, where the proposed binding motif is located, as well as showing a characteristic footprint in the average profile of DNase I cuts across all genomic locations overlapping with hypersensitive sites. The profiles are sorted from most to least likely to have a DNase I footprint at the motif location in the middle. Although the footprint in

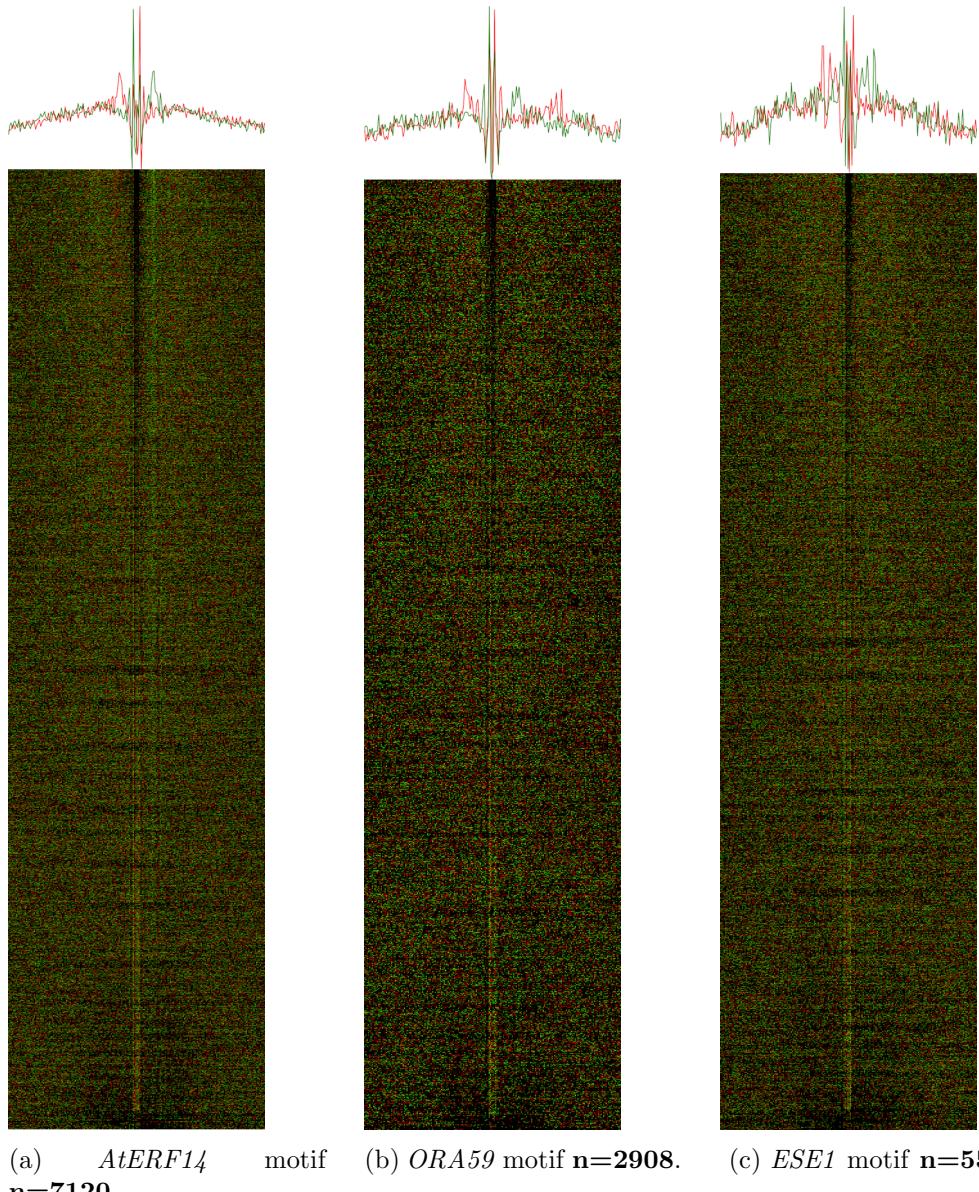


Figure 4.3: Profile of cuts (sequence reads starts) centred on the motif, with 250 bp either side of it, occurring within *Arabidopsis* gene promoters and overlapping with DNase I hypersensitive sites in leaves. Each row is a genomic location where motif occurs. Graph above is average profile for each strand computed from the locations below. Red - cuts on top strand, green - cuts on the bottom strand, colour intensity is directly proportional to number of cuts.

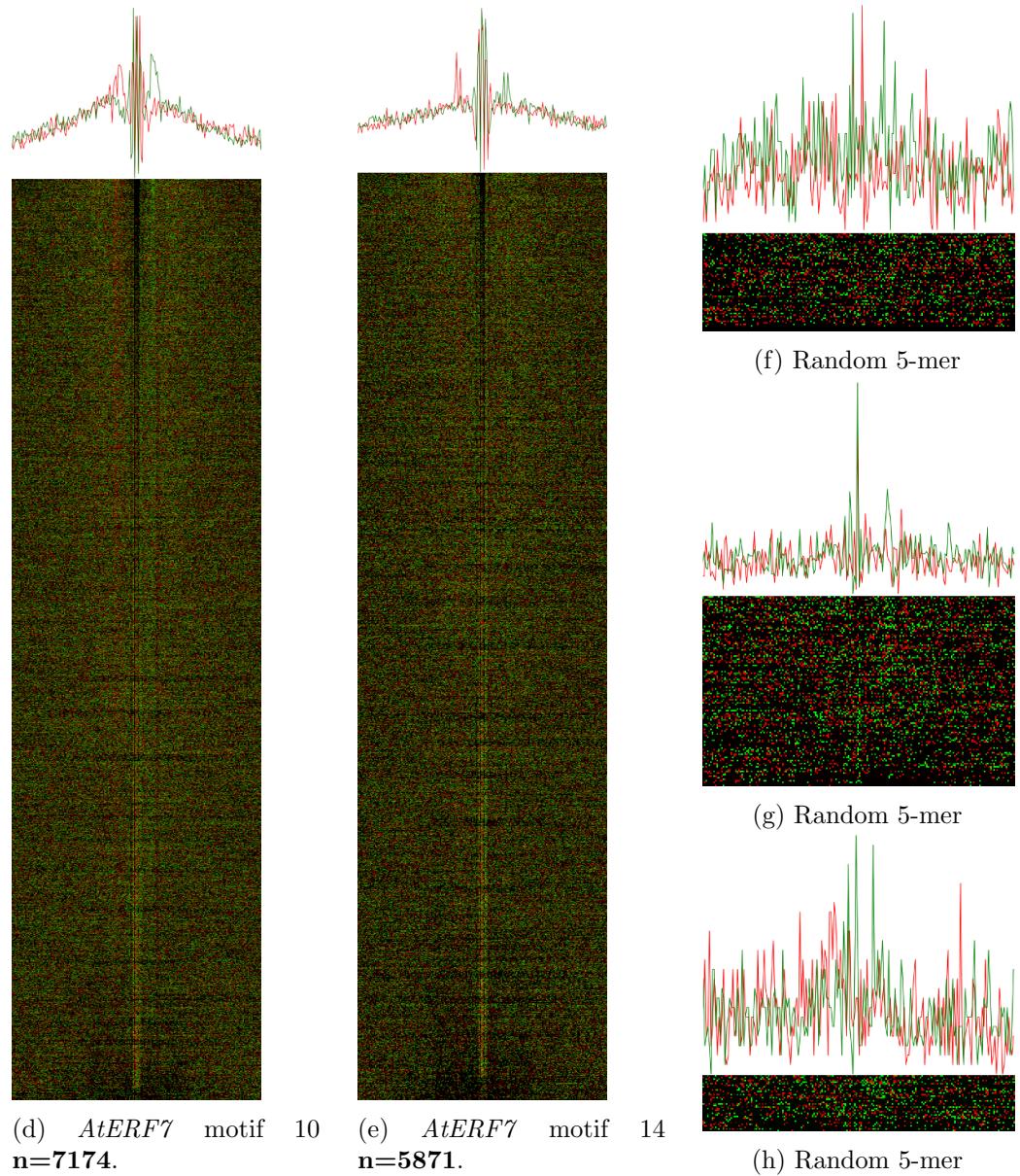


Figure 4.3: Profile of cuts (sequence reads starts) centred on the motif, with 250 bp either side of it, occurring within *Arabidopsis* gene promoters and overlapping with DNase I hypersensitive sites in leaves. Each row is a genomic location where motif occurs. Graph above shows the average profile for each strand computed from the locations below. Red - cuts on the top strand, green - cuts on the bottom strand, colour intensity is directly proportional to number of cuts.

the middle gradually becomes less visible, the average profiles, pictured above the images in Figure 4.3a - 4.3h, show a strong signal present when all sites are considered equally. The presence of the DNase I footprint in the motifs suggests that these sites are bound by a protein in the wild type Arabidopsis leaves, and also in buds (data not shown). However, the identity of the binding protein or complex is not known, unlike in data obtained from ChIP-Seq where the interacting TF is known. However, MeMe analysis suggests AtERF14, ORA59, ESE1 and AtERF7 bind to the corresponding sequence motifs.

#### **4.3.3 Mutations of the new binding sites alter protein-DNA interactions of associated TFs**

So far, 14 motifs have been uncovered as being potentially associated with the specific TFs from MeMe analysis of the promoter fragments interacting with the TFs. Some of the motifs have previously been well characterised as binding to specific TF families, for example, WRKY TFs have been shown to interact with the 5'-TTGACY-3' sequence (Eulgem et al., 2000), which also corresponds to the motifs uncovered by the MeMe analysis of the promoters binding WRKY15/21/29/41/75 TFs, Table 4.1. However, MeMe analysis also suggested new TF specific sequences for AtERF14, AtERF7, ESE1 and ORA59 which have not been previously characterised. Analysis of published DNase-Seq data additionally suggests that the motifs appearing in the promoters of Arabidopsis genes and in DNase hypersensitive areas, are protected from digestion by the enzyme, as indicated by the DNase footprints Figure 4.3. In order to test whether these motifs are able to interact with the suggested TFs, predicted sites were mutated in the promoter fragments of the JAZ1 gene (Y1H\_174), which interacts with these TFs in the Y1H screens detailed in the previous chapter. A total of five sets of mutations were made to alter sequences across four different sites, Figure 4.4.

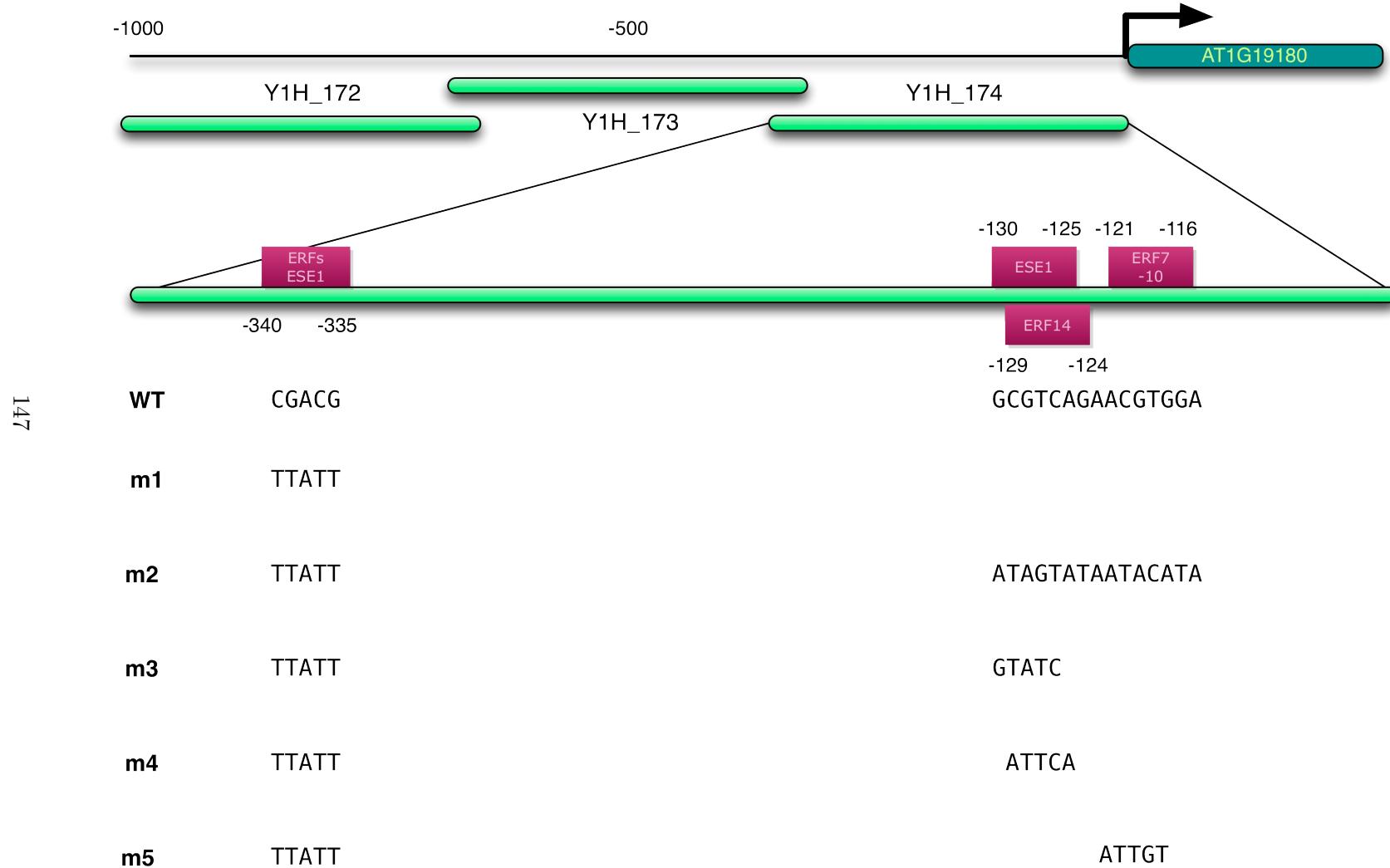


Figure 4.4: Overview of mutagenesis for different binding sites across *JAZ1* promoter fragment (Y1H\_174).

	1	2	3	4	5	6
<b>A</b>	AT3G27785	AT1G76110	AT3G01970	AT3G62340	AT5G24110	AT5G65410
<b>B</b>	AT1G06160	AT2G22750	AT3G12910	AT4G01250	AT5G53950	AT5G28650
<b>C</b>	AT1G06850	AT2G23320	AT3G20310	EMPTY	AT5G39760	AT5G24930
<b>D</b>	AT1G29280	AT2G24430	AT3G23220	AT4G18170	AT5G46350	AT1G21960
<b>E</b>	EMPTY	AT2G24570	AT1G04370	AT4G23550	AT2G46830	AT4G39250
<b>F</b>	AT1G59640	AT2G30590	AT3G50060	AT4G23810	AT5G53980	AT2G17600
<b>G</b>	AT1G69310	AT2G31230	AT3G58710	AT4G38900	AT5G61270	AT4G24060
<b>H</b>	AT1G74930	AT2G36610	AT3G61890	AT5G13080	AT5G64810	CONTROL

Table 4.4: Mutagenesis library arrangement.

Wild-type (WT) and mutated Y1H\_174 were screened first by mating using the mini-library, Table 4.4, to determine if any of the mutations altered the previously observed positive protein-DNA interactions, Figure 4.5. No significant change was observed between WT and **m1** mutations, suggesting that the motifs at the end of the fragment do not play a role in the previously observed interactions, Figure 4.5b. The **m2** fragment showed a slight increase in binding for *AtHB23* (grid location: C05), *ESE1* (D03) and *AtERF14* (E03), but a markedly weaker interaction for *BPE* (F01), Figure 4.5c. Both the **m3** and **m4** mutation sets have a very similar set of positive interactions as the WT fragment, Figure 4.5d and 4.5e respectively. Finally, **m5** mutations show the strongest reduction in positive interactions for *ORA59* (B01), bZIP52 (C01), *BIGPETAL* (*BPE*) (F01), *HMG* (A02), *AT2g22750* (B02), *ESE1* (D03) and *AtERF14* (E03), Figure 4.5f. The *AT4g38900* and *PIF7* TFs, located at G04 and G05 respectively, show no change in the level of yeast growth as compared to the WT fragment. Although positive interactions are much weaker than observed in the WT, interactions are not completely abolished, suggesting that these TFs still bind to the promoter sequences and are able to activate transcription of the reporter gene. As the **m5** promoter showed the largest reduction in positive protein-DNA interaction it was tested further by co-transforming with *ORA59*, *AtERF14*, *ESE1*, *BPE* and *HMG* to quantitatively measure the observed reduction in positive interaction, Figure 4.6. All of the tested TFs showed reduced interactions in **m5** fragment as compared to the WT. Additionally, *BPE* showed a complete reduction in expression at the highest concentration of cells -  $1 \times 10^8$  cells/ml.

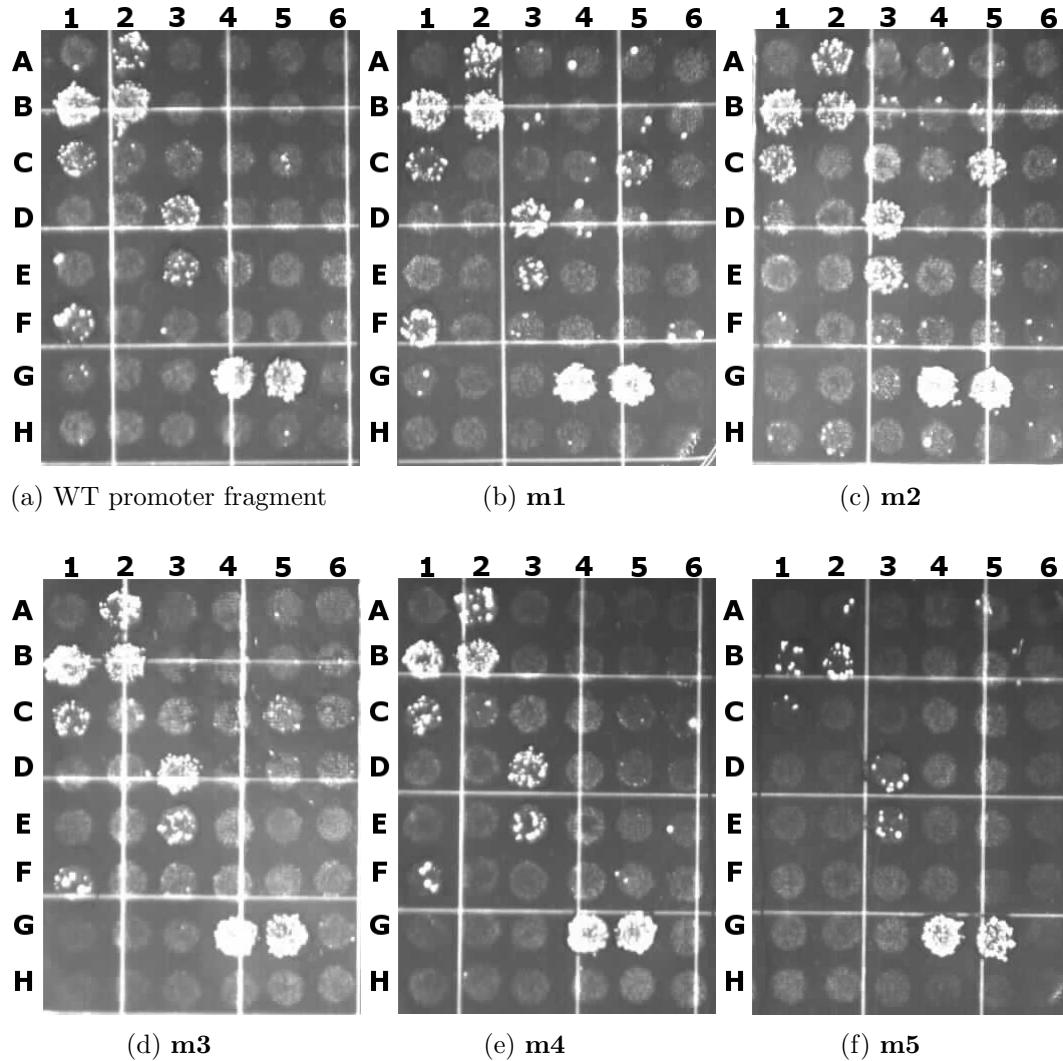
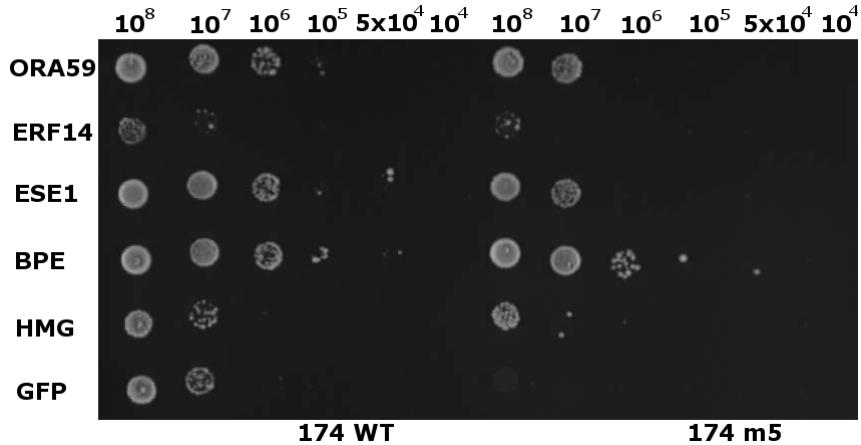


Figure 4.5: Images of WT and mutagenised promoters on SD-LTH agar plates vs mini-library arrangements (Table 4.4). Bright spots indicate growing yeast colonies.

Figure 4.6: Pairwise Y1H screen (by co-transformation) screen of serial dilutions of Y1H-174 WT and **m5** promoter fragments vs select TFs.



In summary, the fragments used in the Y1H screen in the previous chapter were analysed for the presence of potentially new binding sites using MeMe software. The fragments interacting with the WRKY TFs showed conservation of previously characterised TTGACY binding site. AP2 TFs have been shown to interact with the GCC-box, however it is not present in the promoter fragments of the genes analysed using the Y1H screen, suggesting that the AP2 TF can potentially interact through a different sequence specific element. MeMe analysis has outlined new binding motifs for ORA59, ESE1, AtERF14 and two sites for AtERF7 TFs. To increase the confidence in the newly found motifs, DNase I footprints were extracted from genome wide locations overlapping with the hypersensitive sites. The new motifs have a typical DNase I footprint associated with them, suggesting that a TF is bound to the sequences in *Arabidopsis* leaf and bud tissues. Five sets of sequence mutations spanning four different sites where protein-DNA interaction could potentially occur were tested for altered protein-DNA interactions. Interestingly, mutations in AtERF7 motif 10 (**m5**) exhibited the strongest reduction in previously seen protein-DNA interactions affecting seven different TFs simultaneously.

#### 4.3.4 Phenotype screen of *erf14*, *pif7* and *athb25* KO plants show increased susceptibility to infection with Botrytis

The promoters of the genes tested in the previous chapter have specific temporal expression profiles associated with them, suggesting that the genes have a role in the stress response network. Some of the selected genes are known to be involved in

Arabidopsis stress response from previous studies, e.g. WRKY40 (Xu et al., 2006) and JAZ1 (Pauwels and Goossens, 2008). Therefore, TFs found to be regulating a large proportion of these genes may play an important role in the plants' response to infection with Botrytis and may also regulate other genes. Publicly available (NASC), stable transgenic lines bearing a T-DNA insertion in the coding regions of *AtERF14* (SALK\_118494), *PIF7* (SALK\_062756C) and *AtHB25* (SALK\_133857C) were obtained in order to test if the loss-of-function mutants have altered susceptibility to infection with Botrytis. The lines were genotyped using PCR with primers specific for the mutated TFs to check for homozygosity. The homozygous lines were grown for 28 days and detached leaves from WT (Col-4) and transgenic plants were infected with Botrytis spores in grape juice (see Methods). Lesion sizes were measured 48, 57.5 and 72 hours post infection (hpi), Figure 4.7. Statistically significant (at 5% level) smaller lesion sizes were observed in *pif7* and *athb25* lines at all three time points. *erf14* displayed a weak phenotype at 57.5 hpi and was reduced at 72 hpi, just above the 5% significance level. These results suggest that both PIF7 and AtHB25 play a significant role in Arabidopsis stress response and act as repressors of defence genes, since lesions grow slower on the leaves of the plants not expressing these TFs. On the other hand, AtERF14 also appears to function as a repressor in the early stages of the infection process, however lesion spread recovers to the levels seen in the WT leaves 72 h post infection.

#### 4.3.5 Microarray analysis of *erf14* KO plants reveal new targets of the TF

*AtERF14* TF has been suggested to play a non-redundant role in plant defence (Oñate-Sánchez et al., 2007). The results from the Y1H screen have revealed *AtERF14* to be interacting with the promoters of seven of the ten genes selected for screening, including JAZ1 and WRKY40, which have previously been shown to be involved in Arabidopsis stress response (Xu et al., 2006; Pauwels and Goossens, 2008). Additionally, *AtERF14* has a distinct temporal expression profile in response to infection with Botrytis, this gave rise to hypothesis that *AtERF14* may play a unique role in regulating the defence response in Arabidopsis. To test this hypothesis, stable transgenic lines harbouring a T-DNA insertion in the *AtERF14* coding region were grown for 28 days after which leaves were infected with Botrytis spores (see Methods). Two time points, 24 and 28 hpi were chosen to capture the peak expression of *AtERF14* which is known to occur in the wild type leaves. qPCR analysis of WT and *erf14* showed much lower expression of the *AtERF14* in the

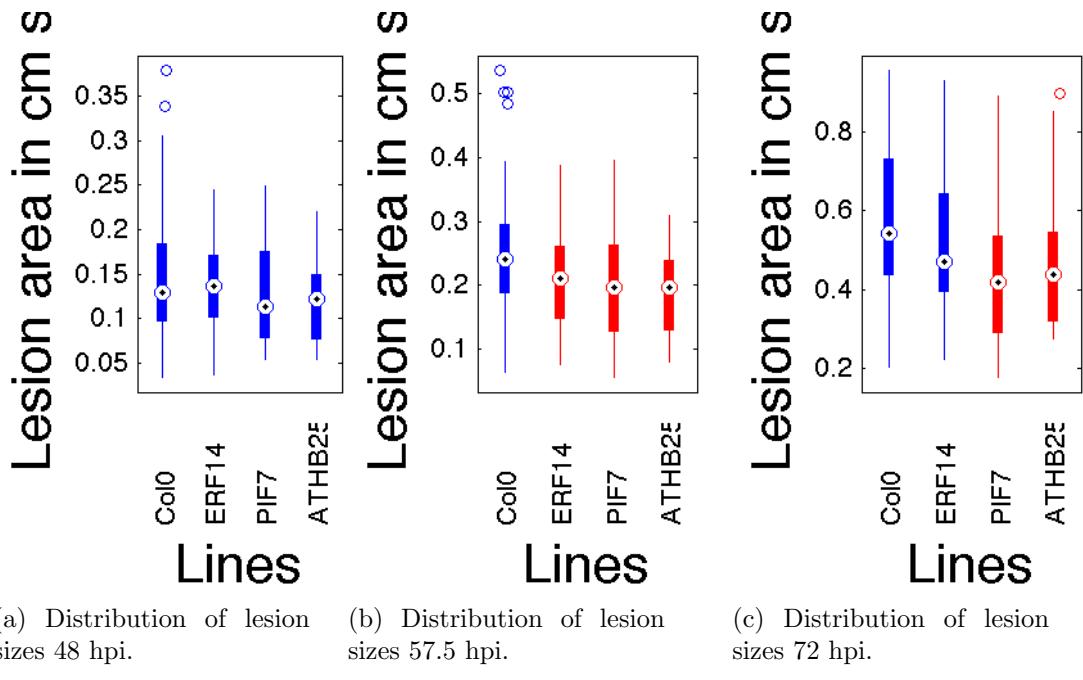


Figure 4.7: Botrytis susceptibility screen of T-DNA knockout lines of predicted regulators of defence response. Box plots are coloured according to the outcomes of two hypotheses tests, if the alternative hypothesis was accepted at the 5% significance level with the t-test, the plots are coloured in red. Otherwise, the plots are coloured in blue.

transgenic plant compared to WT, relative to the expression of the housekeeping gene  $\beta$ -actin, Figure 4.8, suggesting expression of *AtERF14* is reduced, but not completely knocked out. mRNA from the *erf14* leaves infected with Botrytis 24 hpi was extracted, labelled and hybridised to CATMA (v4) microarrays in order to identify direct or indirect downstream targets of *AtERF14* TF. The targets of *AtERF14* would be expected to be differentially expressed from that found in WT leaves infected with Botrytis 24 hpi, Table 4.5. A number of defence related genes are differentially expressed in the *erf14* line e.g. defence response: *LCR67*; salt stress: *KIN2* and *AT5g43060*; dehydration: *ATDR4*. Surprisingly, *AtERF14* targets from the Y1H screen were up regulated in the *erf14* line Table 4.9a.

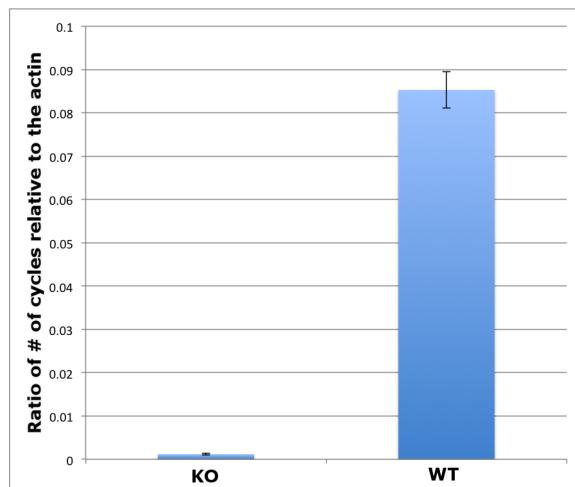


Figure 4.8: Ratio of # of cycles required to reach threshold in qPCR of mRNA of *AtERF14* in *erf14* (KO) and wild type (WT) plants relative to the # of cycles required to reach threshold of actin mRNA.

ATG	Synonym	Log <sub>2</sub> FC	Adjusted P	ATG	Synonym	Log <sub>2</sub> FC	Adjusted P
AT5G57655		-1.0	0.018	AT2G27840	HDT4	0.51	0.039
AT4G15530	PPDK	-0.95	0.035	AT2G29460	ATGSTU4	0.54	0.037
AT4G33150		-0.94	0.035	AT3G60245		0.57	0.035
AT1G15380		-0.94	0.038	AT4G11360	RHA1B	0.59	0.038
AT4G32810	CCD8	-0.90	0.039	AT2G44670		0.62	0.027
AT3G55610	P5CS2	-0.90	0.034	AT5G64140	RPS28	0.63	0.035
AT4G36760	ATAPP1	-0.78	0.034	AT5G15970	KIN2	0.67	0.034
AT1G79600		-0.78	0.023	AT1G73330	ATDR4	0.68	0.035
AT1G58030	CAT2	-0.78	0.025	AT4G21840		0.68	0.025
AT5G37510	EMB1467	-0.75	0.025	AT5G05060		0.72	0.026
AT1G50480	THFS	-0.72	0.032	AT4G06746	RAP2.9	0.78	0.026
AT5G49360	BXL1	-0.69	0.042	AT1G22890		0.78	0.025
AT1G27980	DPL1	-0.68	0.035	AT1G75830	PDF1.1	0.80	0.023
AT5G46180	DELTA-OAT	-0.68	0.041	AT1G54010		0.80	0.037
AT5G03900		-0.67	0.038	AT3G11340		0.92	0.023
AT5G26210	AL4	-0.67	0.026	AT1G38630		1.5	0.025
AT4G20110		-0.65	0.043	AT3G28899		2.4	0.025
AT1G55020	LOX1	-0.63	0.043				
AT5G43060	RD21b	-0.63	0.027				
AT1G15740		-0.63	0.035				
AT2G40670	ARR16	-0.62	0.035				
AT4G38220		-0.62	0.039				
AT3G26380		-0.60	0.035				
AT5G65110	ACX2	-0.59	0.040				
AT5G63800	MUM2	-0.58	0.035				
AT4G19860		-0.56	0.035				
AT1G69480		-0.54	0.035				
AT1G70070	EMB25	-0.53	0.038				
AT5G53450	ORG1	-0.48	0.043				
AT2G42790	CSY3	-0.48	0.043				

Table 4.5: Analysis of genes differentially expressed at 24 h post infection reveals a large number of potential targets of AtERF14 involved in stressed response pathways in Arabbidopsis. These targets include previously known players in stress response network , for example PDF1.1, RAP2.9 and CCD8.

Table 4.6: Most overrepresented GO terms *erf14* differentially expressed genes.

(a) “Biological Processes” GO terms.

P-value	Description
$3.2 \times 10^{-6}$	cellular lipid catabolic process
$4.9 \times 10^{-6}$	lipid catabolic process
$9.1 \times 10^{-6}$	carboxylic acid catabolic process
$9.1 \times 10^{-6}$	organic acid catabolic process
$1.7 \times 10^{-5}$	oxoacid metabolic process
$1.7 \times 10^{-5}$	carboxylic acid metabolic process
$1.8 \times 10^{-5}$	organic acid metabolic process
$2.0 \times 10^{-5}$	cellular ketone metabolic process
$4.4 \times 10^{-5}$	cellular catabolic process
$5.2 \times 10^{-5}$	proline biosynthetic process

(b) “Molecular Function” GO terms.

P-value	Description
$2.6 \times 10^{-5}$	catalytic activity
$8.5 \times 10^{-4}$	oxidoreductase activity
$1.0 \times 10^{-3}$	dioxygenase activity
$1.3 \times 10^{-3}$	oxidoreductase activity
$1.9 \times 10^{-3}$	N-1-naphthylphthalamic acid binding
$1.9 \times 10^{-3}$	xylose isomerase activity
$1.9 \times 10^{-3}$	pyruvate, phosphate dikinase activity
$1.9 \times 10^{-3}$	saccharopine dehydrogenase activity
$1.9 \times 10^{-3}$	glutamate-5-semialdehyde dehydrogenase activity
$1.9 \times 10^{-3}$	glutamate 5-kinase activity

Table 4.7: AtERF14 targets genes involved in catabolic and reductase activities, as revealed by the GO Terms of genes determined to be differentially expressed in *erf14* compared to WT 24 h post infection.

### GO Term analysis of *AtERF14* targets in protoplasts

In order to identify whether *AtERF14* regulates genes involved in certain cell processes and functions, GO terms of differentially expressed genes (P-value < 0.05) were tested for overrepresentation in “Biological Processes” and “Molecular Function” categories. GO analysis of “Biological Processes” terms revealed a range of catabolic activities: lipid catabolic process (P-value  $4.9 \times 10^{-6}$ ), carboxylic acid (P-value  $9.1 \times 10^{-6}$ ) and organic acid catabolic process (P-value  $9.1 \times 10^{-6}$ ), Table 4.6a, as well as a number of metabolic processes, such as oxoacid metabolic process (P-value  $1.7 \times 10^{-5}$ ), carboxylic acid metabolic process (P-value  $1.7 \times 10^{-5}$ ), cellular ketone metabolic process ( $2.0 \times 10^{-5}$ ) and proline metabolic process ( $1.2 \times 10^{-4}$ ). “Molecular Functions” GO terms revealed “catalytic activity” as the most significant term (P-value  $2.6 \times 10^{-5}$ ), Table 4.6b.

### Hundreds of genes are differentially expressed in protoplasts overexpressing *AtERF14*, and *ESE1* TFs

Overexpression of TFs in protoplasts allows the levels of the TF to rapidly increase within a 24 h incubation period and therefore enables the assessment of changes to transcription of genes targeted by the TF of interest. Whereas stable KO lines do not always guarantee a complete reduction in the TF expression and so downstream targets may not be immediately obvious. Furthermore, stable KO lines require a fully grown plant, whereas overexpression in protoplasts can be done in a much shorter time frame. To further verify the direct targets of *AtERF14* and *ESE1*, plasmids overexpressing these TFs were introduced into protoplasts. After 24hr incubation

total RNA was extracted, labelled and hybridised to CATMA (v4) microarrays (see Methods). Subsequent analysis have identified 878 and 1718 differentially expressed genes ( $P$ -value  $< 0.05$ ) for *AtERF14* and *ESE1* respectively, compared to protoplasts without overexpression plasmids.

Assays performed in protoplasts have been shown to have innate differential expression associated with the stress of the detaching leaf and digesting the cell walls. As such, previously published genes associated with the innate protoplast response to stress (Birnbaum et al., 2003; Gifford et al., 2008) were removed from lists of differentially expressed genes in *AtERF14* (38 genes removed) and *ESE1* (59 genes removed). The overlap between differentially expressed genes in protoplasts overexpressing *AtERF14* and *ESE1* (at  $P$ -value  $< 0.05$ ) was 465 genes, suggesting that both TFs regulate very similar sets of target genes. In order to find specific direct targets for each TF, the set of overlapping 465 genes was excluded from differentially expressed genes for each TF, leaving 394 and 1215 significantly differentially expressed genes for *AtERF14* and *ESE1* respectively.

The GO Terms associated with the differentially expressed genes were analysed for overrepresentation in order to establish biological processes associated with the targets of *AtERF14* and *ESE1* TFs. The set of differently expressed genes, excluding overlaps and innate protoplast response genes, have revealed “response to stress” ( $1.7 \times 10^{-4}$  in *AtERF14* and  $5.5 \times 10^{-9}$  in *ESE1*) as significantly overrepresented category among “Biological Processes”, Table 4.8a and 4.8c respectively. No significant “Molecular Function” terms were overrepresented in the *AtERF14* set, whereas “catalytic activity” and “nucleotide binding” were two main categories overrepresented in the *ESE1* target set of genes ( $10^{-9}$  and  $1.4 \times 10^{-6}$  respectively), Table 4.8b.

P-value	Description	P-value	Description
$4.9 \times 10^{-6}$	response to metal ion	$5.5 \times 10^{-9}$	response to stress
$1.8 \times 10^{-5}$	response to cadmium ion	$7.1 \times 10^{-8}$	response to stimulus
$2.4 \times 10^{-5}$	response to chemical stimulus	$4.9 \times 10^{-7}$	nitrogen compound metabolic process
$2.6 \times 10^{-5}$	response to inorganic substance	$1.2 \times 10^{-6}$	cellular nitrogen compound metabolic process
$1.3 \times 10^{-4}$	response to stimulus	$2.1 \times 10^{-6}$	response to temperature stimulus
$1.3 \times 10^{-4}$	catabolic process	$2.5 \times 10^{-6}$	small molecule metabolic process
$1.7 \times 10^{-4}$	response to stress	$3.6 \times 10^{-6}$	response to abiotic stimulus
$1.8 \times 10^{-4}$	cellular process	$2.0 \times 10^{-5}$	response to other organism
$4.1 \times 10^{-4}$	small molecule metabolic process	$2.8 \times 10^{-5}$	response to osmotic stress
(a) Overrepresentation of “Biological Processes” GO terms of 394 differentially expressed genes in protoplasts overexpressing <i>AtERF14</i>		$3.4 \times 10^{-5}$	multi-organism process
		$4.9 \times 10^{-5}$	nucleic acid metabolic process
		$5.4 \times 10^{-5}$	response to biotic stimulus
		$5.6 \times 10^{-5}$	response to inorganic substance
		$1.0 \times 10^{-4}$	response to chemical stimulus
		$1.2 \times 10^{-4}$	small molecule biosynthetic process
		$1.7 \times 10^{-4}$	response to cadmium ion
		$1.9 \times 10^{-4}$	cofactor metabolic process
		$2.1 \times 10^{-4}$	defense response
		$2.9 \times 10^{-4}$	response to cold
		$2.9 \times 10^{-4}$	metabolic process
		$3.2 \times 10^{-4}$	vitamin metabolic process
		$3.4 \times 10^{-4}$	heterocycle metabolic process
		$4.0 \times 10^{-4}$	response to salt stress
		$4.1 \times 10^{-4}$	protein import
		$4.7 \times 10^{-4}$	secondary metabolic process
		$4.7 \times 10^{-4}$	defense response to fungus
		$5.3 \times 10^{-4}$	protein targeting to chloroplast
		$5.5 \times 10^{-4}$	response to metal ion
		$7.0 \times 10^{-4}$	response to bacterium
		$7.3 \times 10^{-4}$	mRNA 3'-end processing
		$7.4 \times 10^{-4}$	vitamin biosynthetic process
		$7.6 \times 10^{-4}$	cellular homeostasis
		$8.4 \times 10^{-4}$	nucleic acid metabolic process
		$9.5 \times 10^{-4}$	porphyrin metabolic process
		$9.8 \times 10^{-4}$	homeostatic process
		$1.0 \times 10^{-3}$	lignin biosynthetic process
		$1.0 \times 10^{-3}$	cellular metabolic process
		$1.2 \times 10^{-3}$	water-soluble vitamin metabolic process
		$1.2 \times 10^{-3}$	tetrapyrrole metabolic process
(b) Overrepresentation of “Molecular Function” GO terms of 1215 differentially expressed genes in protoplasts overexpressing <i>ESE1</i>		(c) Overrepresentation of “Biological Processes” GO terms of 1215 differentially expressed genes in protoplasts overexpressing <i>ESE1</i>	

Table 4.8: *erf14* targets genes involved in stress response, chemical stimulus and variety of defence response pathways as revealed by the GO Terms associated with the genes found to be differentially expressed in protoplasts overexpressing *AtERF14*, compared to protoplasts without any plasmids.

#### 4.3.6 Reported Y1H interactions for *AtERF14* and *ESE1* are confirmed in protoplasts

The genes corresponding to the promoters which were seen to interact with *AtERF14*, and *ESE1* in the Y1H experiment, were found to be differentially expressed in the protoplasts overexpressing the respective TFs as compared to protoplast cells alone, as well as being differentially expressed in the stable *erf14* KO leaves infected with Botrytis as compared to wildtype leaves infected with Botrytis. This data is consistent with the predictions from Y1H experiments that these TFs interact with the promoters of these genes and therefore potentially regulate their transcription.

Gene Name	Log <sub>2</sub> FC	P-value	Adjusted P-Value
JAZ1	0.214222306	0.013	0.23
WRKY40	0.435416218	0.001	0.093
AOC3	0.388000316	0.003	0.138
ERF13	0.086413724	0.249	0.661
PUB23	0.368320765	0.058	0.406
MYB15	0.189807243	0.115	0.518
WRKY11	0.10639487	0.278	0.683

(a) Expression levels of predicted targets of *AtERF14* from Y1H screen in the *erf14* KO.

(b) Expression levels of predicted targets of *AtERF14* in protoplasts overexpressing *AtERF14*.

(c) Expression levels of predicted targets of *ESE1* in protoplasts overexpressing *ESE1*.

Name	Log <sub>2</sub> FC	P	Adjusted P	Name	Log <sub>2</sub> FC	P	Adjusted P
JAZ1	0.54	$1.4 \times 10^{-4}$	0.010	JAZ1	0.077	0.44	0.75
WRKY40	0.17	0.075	0.40	WRKY40	0.10	0.23	0.57
AOC3	-0.063	0.33	0.75	AOC1	-0.16	0.26	0.60
ERF13	0.064	0.55	0.88	AOC3	-0.19	0.022	0.16
PUB23	0.015	0.96	0.99	ERF13	-0.21	0.12	0.42
MYB15	0.32	$9.8 \times 10^{-4}$	0.033	PUB23	0.65	0.088	0.35
WRKY11	0.17	0.19	0.60	MYB15	0.0093	0.90	0.97
				WRKY11	0.064	0.65	0.87
				AT5G50570	-0.62	$2.1 \times 10^{-4}$	0.0095

Table 4.9: JAZ1 and MYB15, that were found to be interacting with *AtERF14* in Y1H, are also significantly differentially expressed in *erf14* KO lines compared to WT and in protoplasts overexpressing *AtERF14*.

The *erf14* KO analysis of the predicted Y1H targets shows all targets to be up

regulated, suggesting that *AtERF14* functions as transcriptional repressor of those genes. In the protoplasts overexpressing *AtERF14*, only *JAZ1* and *MYB15* were found to be significantly up regulated (adjusted P-value < 0.05), indicative of a positive relationship between levels of *AtERF14* and the two target genes. AT5G50570 is the only gene from the set of predicted Y1H targets regulated by *ESE1* that has been found to be differentially expressed in the protoplasts overexpressing the TF, as compared to the protoplasts without the overexpression plasmid. Additionally, AT5G50570 is down regulated in the protoplasts suggesting that *ESE1* may function as a transcriptional repressor of the gene.

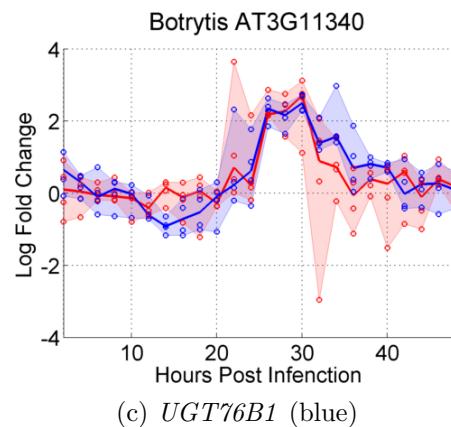
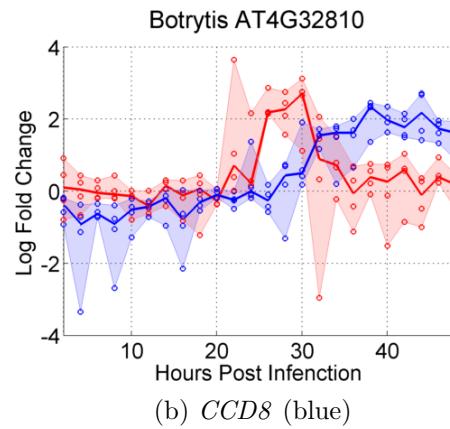
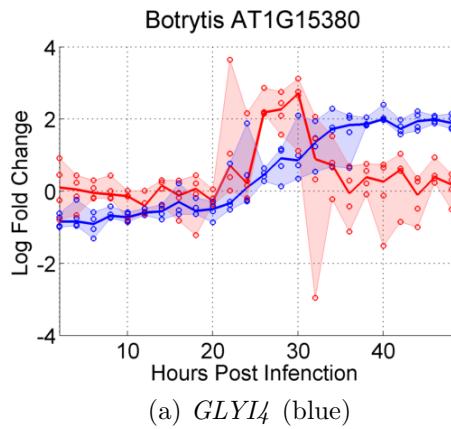
### New stress responsive targets of *AtERF14* TF

Analysis of the differentially expressed genes (at 5% significance level) from *erf14* transgenic lines and protoplasts overexpressing *AtERF14* TF have identified 4 genes that significantly overlap (hypergeometric P-value: 0.02558626) between the two experiments: *AT3g11340 (UGT76B1)*, *AT1G54010*, *AT1g15380 (GLYI4)* and *AT4g32810 (CDD8)*, Table 4.10. Of these 4 genes *UGT76B1* have been previously shown to be modulating plant defence and senescence through the SA and JA signalling pathways (von Saint Paul et al., 2011). The overlap suggests that these 4 genes act as direct or indirect targets of *AtERF14*. The data also suggests that *AtERF14* acts as a transcriptional repressor of *UGT76B1* and *AT1G54010* as both are up regulated in the *erf14* KO lines, whilst down regulated in the overexpression lines. *UGT76B1* loss-of-function plants have been shown to be more susceptible to necrotrophic *Alternaria brassicicola* (von Saint Paul et al., 2011), consistent with the observation that *AtERF14* is a repressor of *UGT76B1*, and *erf14* plants are less susceptible to necrotrophic Botrytis. Analysis of the mRNA expression level in response to infection with Botrytis available from the PRESTA project shows that *UGT76B1* expression closely follows that of *AtERF14* further strengthening the link between these two genes, Figure 4.9c. On the other hand *CDD8* and *GLYI4* are both expressed late in response to infection with Botrytis, > 30 hpi, Figure 4.9b and 4.9a respectively.

Name	$\log_2$ FC	Adjusted P
<i>UGT76B1</i>	0.92	0.023
AT1G54010	0.80	0.037
<i>GLYI4</i>	-0.94	0.038
<i>CCD8</i>	-0.90	0.039

(a) *erf14* KO

Name	$\log_2$ FC	Adjusted P
<i>UGT76B1</i>	-0.29	0.027
AT1G54010	-0.56	0.010
<i>GLYI4</i>	-0.84	0.019
<i>CCD8</i>	-0.67	0.0069

(b) *AtERF14* overexpression.Table 4.10: Changes in expression levels of select genes in *erf14* and protoplasts overexpressing *AtERF14*.Figure 4.9: mRNA expression levels of *AtERF14* (red) and associated targets (blue) during infection with Botrytis.

## 4.4 Discussion

### 4.4.1 Characterisation of novel *cis*-acting elements in the promoters of gene screened in the Y1H experiments

Successful predictions of functional binding site motifs is an open problem in the field of biology. No high-throughput and cost-effective screen has been devised to test for all possible binding sequences that protein containing a DNA binding domain can interact with. This is mostly due to the unfeasably large number of sequences required to be tested, arising from combinatorial variations of nucleotides in the binding sites. This leads to *in vitro* techniques like EMSA being used to validate predicted high confidence protein-DNA interactions. Results from ChIP-Seq experiments have been shown to successfully identify genome wide binding events and motifs associated with them, given that an antibody exists for the TF of interest (Robertson et al., 2007; Yant et al., 2010). However, antibodies are only available for a small number of DNA binding proteins and therefore there is a substantial gap in the knowledge of sequence specific interactions of other proteins. Computational techniques have been developed for the prediction of potential protein binding sites, when provided with a number of DNA sequences thought to be interacting with the protein, such as MeMe (Bailey and Elkan, 1994), AlignACE (Neuwald et al., 1995) and SCOPE (Chakravarty et al., 2007). Of these, MeMe has been widely used for *de novo* motif discovery and offers a rich set of adjustable parameters including creation of custom negative control sets. Subsequently, MeMe software was used to find any conserved binding site motifs present in the promoters of genes found to be interacting in the Y1H screens presented in the previous chapter. Identification, through a Y1H screen, of TF specific sites where direct protein-DNA interactions occur, also presents another facet of the data rarely explored, namely sequences used for promoter fragment constructions are not used for bioinformatic analysis beyond scanning for the presence of existing motifs. However, as mentioned above, only a limited number of motifs is available for each family of TFs, which may not be responsible for the interactions of other family members.

The results obtained in the previous chapter, Table 3.12, show that some TFs were found to be positively interacting with the selected promoter fragments. For example, WRKY15/17/21/22/29/41/65/68/69 were found to be interacting with the promoter sequences of PUB23 gene. However, WRKY15/21/29 were also found to be interacting with the promoter sequences of WRKY11, WRKY40, JAZ1, and WRKY21/29 also interacted with the promoter fragments of AOC3. This suggests

that there are features that are recognised by WRKY17/22/65/68/69 in the promoter of PUB23 that are not present in the promoters of WRKY11, WRKY40, JAZ1 and AOC3. We hypothesised that these features are sequence specific around the core binding motif, or are an altogether new and previously uncharacterised motif conserved across multiple promoter fragments. However, the new motif may be similar to the previously described one and differs by mutations and/or substitutions of one or more bases. The tolerance of different WRKY genes to mutations within the core binding sequence or nucleotides immediately adjacent to it has previously been described by Ciolkowski et al. (2008). Therefore, MeMe software (Bailey et al., 2009) is used to identify the short sequences conserved between the different promoters interacting with the same WRKY and other TFs.

Limited data was available for WRKY TFs, forming 3 groups of sequences: those binding WRKY15/21/29, WRKY75, and WRKY41, other WRKY TFs showed positive results in one promoter fragment making identification of a new WRKY motif not plausible. WRKY15 and WRKY21 are part of the same II<sup>d</sup> group and WRKY29 is part of II<sup>e</sup> group, which is closely related to II<sup>d</sup> phylogenetically (Eulgem et al., 2000). Therefore, the 5'-TTGAC(A/T)TT-3' sequence may be indicative of clades II<sup>d</sup> and II<sup>e</sup>. On the other hand, WRKY41 is part of clade III of the WRKY superfamily, which lacks a number of extra domains conserved among the II<sup>e</sup> and II<sup>d</sup> clades (Eulgem et al., 2000), and therefore WRKYS of that clade are able to recognise only the core sequence of the motif. Finally WRKY75 belongs to clade II<sup>c</sup> (Zhang and Wang, 2005), a different phylogenetic branch to the other three clades (Eulgem et al., 2000), suggesting that the GG(T/C) sequence immediately upstream from the WRKY binding site may be specific for that clade alone.

The highly conserved binding motif 5'-CACGTG-3' (G-Box, (Toledo-Ortiz et al., 2003)) for bHLH TFs such as *PIF7* was also found to be present at the appropriate locations within the promoter found to be interacting with *PIF7* TF. Although the motif is present as a whole, not all bases were unanimously conserved in all the sequences. These point mutations may be indicative of a tolerance to base substitutions by the TF at a cost of weaker protein-DNA interactions. Slight variations of motifs known to be associated with a family of TFs may represent specificity for individual TFs, e.g. homeobox domain proteins have been shown to bind 5'-CAAT(A/T)ATTG-3' (BS-1) and 5'-CAAT(G/C)ATTG-3' (BS-2) sequences in *Arabidopsis* (Sessa et al., 1997). Using the sequences found to be interacting with the *AtHB25* TF, MeMe analysis uncovered a 5'-CAANTANTTG-3' binding motif

that preserves the palindromic nature of the previously found binding site. The variation from the BS-1 and BS-2 sequences may represent a *AtHB25* specific binding site.

#### 4.4.2 Sequence specificity of AP2 domain proteins

AP2 domain TFs, for example *ORA59*, *AtERF14* and *AtERF7*, represent a large proportion of TFs interacting with the selected subset of genes in the Y1H screen. AP2 domain proteins have also been shown to play nonredundant roles in the stress response (Pre et al., 2008; Oñate-Sánchez et al., 2007; Moffat et al., 2012), therefore it would be advantageous to determine the sequence specificity of individual AP2 TFs. Previous studies have broadly separated AP2 proteins into 2 classes, A and B, which were further separated into six subgroups for each class, giving rise to 12 groups in total (Nakano et al., 2006). Different members of these subgroups have been found to have preferences associated with different binding motifs; some were found to bind to the ethylene responsive element (EREBP) also known as GCC-box, e.g. *ERF1* (Ohme-Takagi and Shinshi, 1995) whereas temperature and cold responsive elements containing an AP2 domain bound to the C-repeat core sequence 5'-CCGAC-3' (Baker et al., 1994), e.g. *CBF1* (Stockinger et al., 1997), *CBF2* and *CBF3* (Gilmour et al., 1998) as well as *DREB1* and *DREB2* (Liu et al., 1998). Finally, some AP2 domain proteins function using bipartite sequence recognition, e.g. *RAV1* and *RAV2* TFs recognise 5'-CAACA-3' and 5'-CACCTG-3' by their AP2 and B3-like domains respectively (Kagaya et al., 1999). The naming convention is confusing in the case of AP2 domain proteins, for example members of the B3 group of AP2 domain proteins do not contain a characterised B3 binding domain, unlike *RAV1* and *RAV2* TFs that contain AP2/B3 domains but are not characterised under the AP2 family of TFs. This suggests that the AP2 DB domain is able to interact with a variety of sequence motifs with a particular preference for C/G richness.

The binding motif for AtERF14, ESE1 and AtERF7 are not known, however, AtERF14 and ESE1 are in the same phylogenetic clade (IXc) as ERF1 and ORA59 TFs (Nakano et al., 2006), which have been shown to interact with the GCC-box sequence 5'-GCCGCC-3' (Ohme-Takagi and Shinshi, 1995). However, the GCC box was not present in the promoters tested in the Y1H screen, suggesting the presence of a different *cis*-acting element. Furthermore, both AtERF14 and ESE1 are comparatively smaller proteins than ERF1 and ORA59 14.6 kDa, 15.7 kDa, 24.7 kDa and

27.1 kDa respectively. Smaller TFs may interact with shorter binding sequences and therefore the full GCC motif may not be required for the protein-DNA interaction to take place. Instead, a shorter motif may be adequate for the TFs to interact with the major or minor grooves of the DNA whilst anchored by stronger interactions with guanine and cytosine residues, which are preferential for AP2 domain TFs. In contrast, AtERF7 forms the VIIa clade together with ERF3/4/8/9/10/11 and 12, which have been found to contain an EAR motif. ERF4 has also been reported to act as a transcriptional repressor of the ABA responsive genes ABI2, RD29B and RAB18 through the GCC element present in the promoters of these genes (Yang et al., 2005). However, AtERF7 was not tested for direct interaction with the GCC box. The *de novo* motifs identified for the AP2 TFs using MeMe analysis were found to be shorter than the GCC motif and have an anchoring GC sequence within them, supporting the hypothesis of new *cis*-acting elements potentially more suitable for smaller TFs.

#### 4.4.3 DNase I analysis

Open regions of the genomic DNA have been found to be more accessible to the DNase I enzyme allowing it to be cut at random nucleotides (Boyle et al., 2008). DNase I hypersensitive assays are increasingly used for the identification of open chromatin regions over 500bp long (Zhang et al., 2012; Neph et al., 2012), as well as for identification of individual binding sites of 5-10bp (Neph et al., 2012). In order to increase the confidence of the motifs predicted by MeMe, genomic locations of the new motifs were combined with previously identified hypersensitive sites (Zhang et al., 2012) and overlapping locations were analysed for a typical DNase I cutting profile (protected in the middle and cuts either side of the binding site). In general, motifs with a likely DNase I profile in hypersensitive sites were found more often, in the order of 10 times more likely to be present, than random motifs. In comparison with DNase I profiles previously published for different motifs (Neph et al., 2012), the profiles obtained here contain a lot more random cuts uniformly distributed around the binding sites. This may be explained by advancements in the technique used to generate DNase I digested fragments, whereby prior to sequencing, only fragments of a certain size, ~ 50bp, were used instead of all digested fragments as used by Zhang et al. (2012). Even though data is noisy, strong profiles are visible as compared to random k-mer sequences, Figure 4.3f, 4.3g and 4.3h.

Sequences found to be conserved in the promoter fragments that interact with

certain TFs also show stronger DNase I profiles than previously reported motifs. For example, the new motif associated with the *ORA59* TF, 5'-(C/G)(C/G)(A/G)CCG-3', has characteristic features of the DNase footprint, Figure 4.10b. On the other hand, the GCC-boc motif previously found to be interacting with the *ORA59* TF (Zarei et al., 2011), lacks an increased number of cuts either side of the motif and has no identifiable protected site where the motif is located. However, the DNase I data is from unstressed, *Arabidopsis* wildtype leaf and bud tissue, whereas the GCC-box may be located in the closed areas, which become open in response to stress with the help from chromatin modification factors (Sokol et al., 2007).

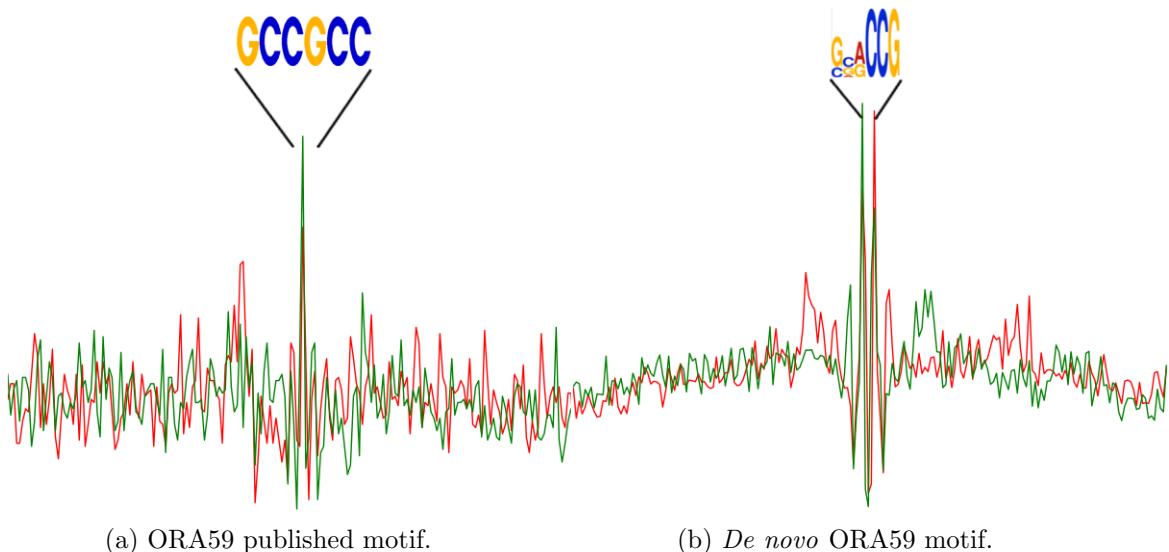


Figure 4.10: DNase I cutting profile of published and *de novo* motifs in *Arabidopsis* leaves for *ORA59* TF.

Interestingly, when analysing DNase I footprints for *AtERF14*, it was identified that the proposed motif appears as part of a larger, palindromic motif, which is also associated with a very strong DNase I footprint. This longer binding site may be indicative of TFs other than *AtERF14* binding to the palindromic sequence in the wildtype leaves and buds.

#### 4.4.4 The *de novo* motifs interact with specific TFs

Mutational analysis of the motifs predicted to be in the promoter fragments of the genes interacting with *ORA59*, *AtERF14*, *ESE1* and *AtERF7* TFs that were also found to contain DNase I footprints, was carried out in order to test whether the

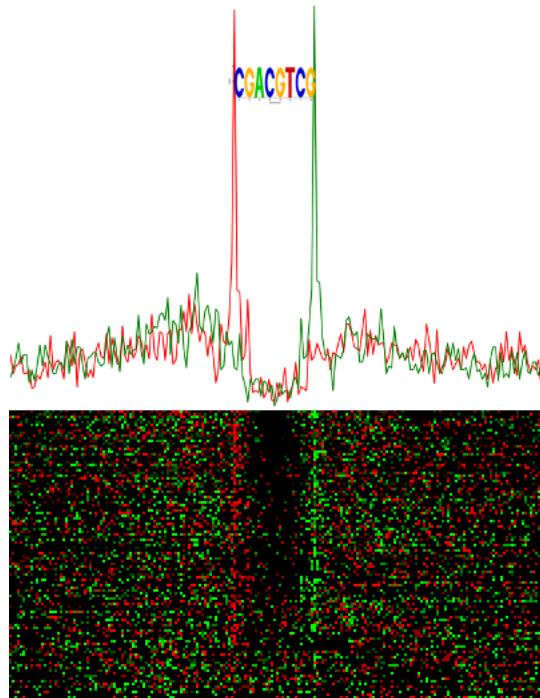


Figure 4.11: DNase I profile of extended palindromic sequence derived from *AtERF14* predicted motif.

TFs were interacting with the associated motifs in a Y1H screen. Surprisingly, only mutations of AtERF7 (motif 10) showed reduced interactions not only with the *AtERF7* TF, but also with a number of other motifs. Additional TFs affected by the mutations of the new AtERF7 motif include those with an AP2 domain and some without, e.g. BPE and HMG. The reasons for the ineffectiveness of mutations in other motifs to alter specific protein-DNA interactions seen in the Y1H screens are not clear. The strict MeMe analysis identified those motifs as being located in the appropriate locations of the promoter fragments tested and conservation was proportional to the strength of interaction, represented by the Y1H score, seen in the Y1H screen. Therefore, these motifs were prime candidates through which direct protein-DNA interactions could take place. Theoretically, it is possible that the mutated motifs could serve as a protein-DNA interaction of host TFs, and therefore validation of these motifs through a Y1H screen may not be appropriate. Alternative, ChIP-Seq experiments would provide genome-wide locations for the TF binding and therefore the motifs could be validated *in planta*.

Another alternative explanation for this result may be that some TFs, instead of recognising a specific sequence, recognise DNA shape, which may instead be created by multiple sequences. It has previously been hypothesised that the

backbone structure itself may be responsible for the binding of certain TFs, like *lac* operon, where mutations to the overall structure of the DNA backbone had significant implications on operon activity (Klug et al., 1979). Furthermore, correlation between the ability of the DNA to twist and roll has been found to be significant when considered in the context of direct protein-DNA interactions (Gorin et al., 1995). Although little is known about the molecular basis of protein-DNA interactions, the structure conferred by certain DNA sequences must be more favourable for some proteins than others, serving as a template for specific DNA interactions. Currently, most protein-DNA interaction predictions are based on scoring previously established binding motif sequences, usually represented in the form of PSSMs, along the DNA sequence, considering the DNA as a “static rod”. However, some databases are starting to include more information about structural properties of the DNA sequence (Gardiner et al., 2003). This information may help to develop more sensitive methods of identifying strong protein-DNA interactions than a binary presence or absence of the binding motif.

#### 4.4.5 Summary of new motif interaction patterns

In summary, in order to provide much needed TF-specific binding site motifs, information on which is currently lacking, a bioinformatics approach using the MeMe software suite was first adopted to find potentially conserved motifs within the previously tested promoter fragments. This analysis has identified unique motifs for a number of TFs including AtERF14, AtERF7, ESE1 and ORA59. Publicly available DNase-Seq data (Zhang et al., 2012) was used to further test if the predicted motifs were bound by a TF in wildtype *Arabidopsis* leaf and bud tissues, showing a DNase I footprint for the selected motifs. However, amongst the selected motifs only the AtERF7 (motif 10) site had an effect on protein-DNA interactions, altering positive protein-DNA interactions for other TFs as well as for AtERF7. Figures 4.12 - 4.19 summarize the results from all chapters and include information about Y1H fragments, discovered CNSs and potential binding sites of the TFs found to be interacting in the Y1H screen described in this chapter. Additionally, the profile of the chromatin in the wild type is superimposed along the promoter, as described in Zhang et al. (2012).

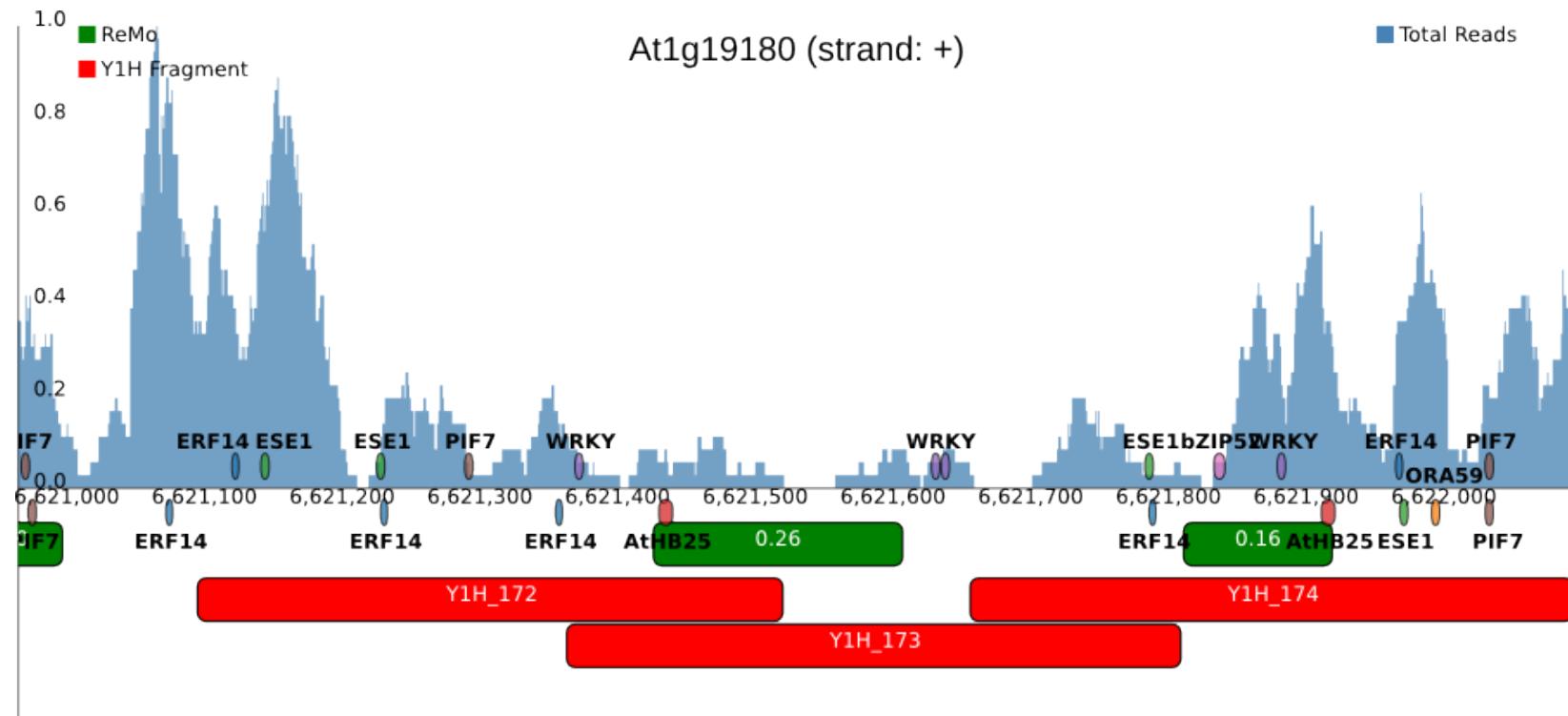


Figure 4.12: Summary of all results obtained in this thesis for *At1g19180*: Y1H fragments (red), predicted CNSs (green) locations of the predicted motifs and DNase profile (blue) showing open chromatin areas. Height of the DNase peaks is proportional to the DNase-Seq reads in the Arabidopsis leaf tissue but normalised by maximum read depth along the whole promoter. Red line at the edge is the annotated TSS.

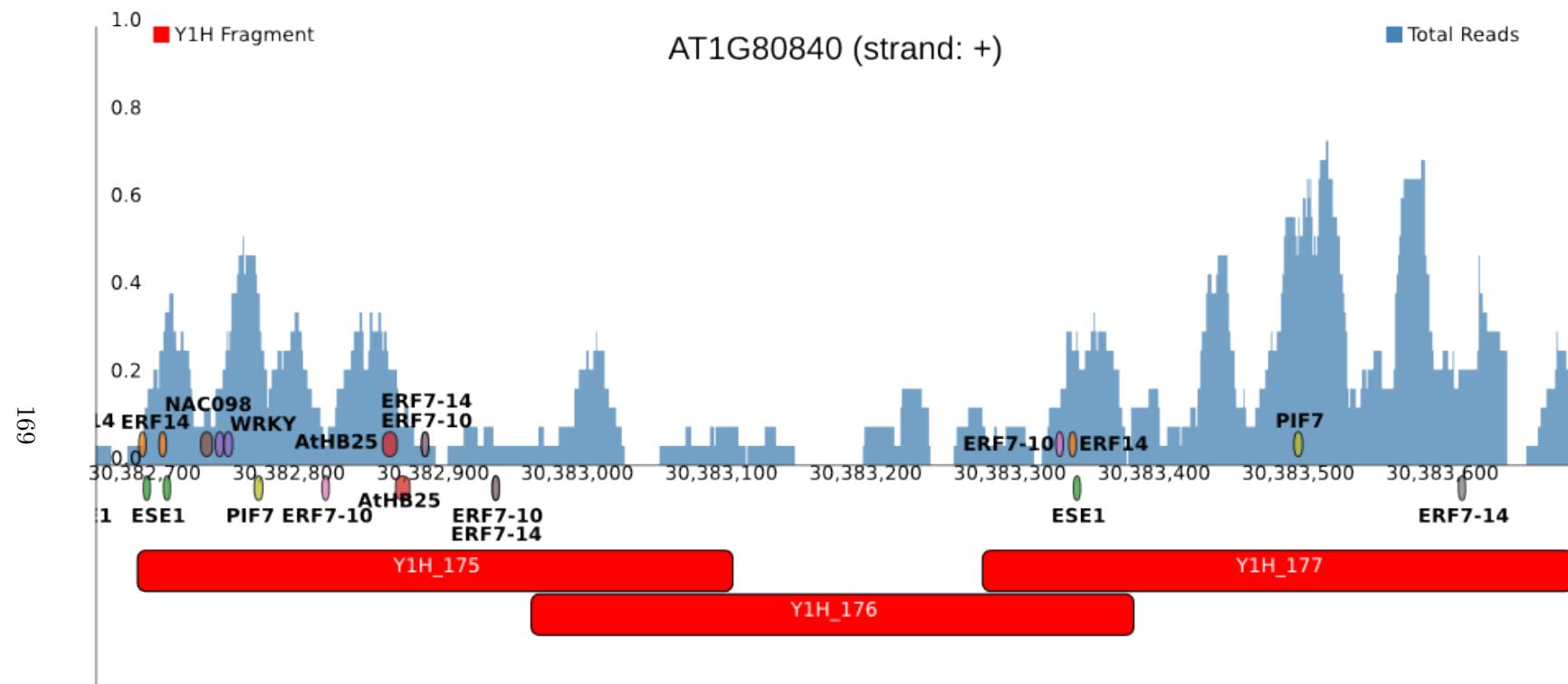


Figure 4.13: Summary of all results obtained in this thesis for *At1g80840*: Y1H fragments (red), predicted CNSs (green) locations of the predicted motifs and DNase profile (blue) showing open chromatin areas. Height of the DNase peaks is proportional to the DNase-Seq reads in the *Arabidopsis* leaf tissue but normalised by maximum read depth along the whole promoter. Red line at the edge is the annotated TSS.

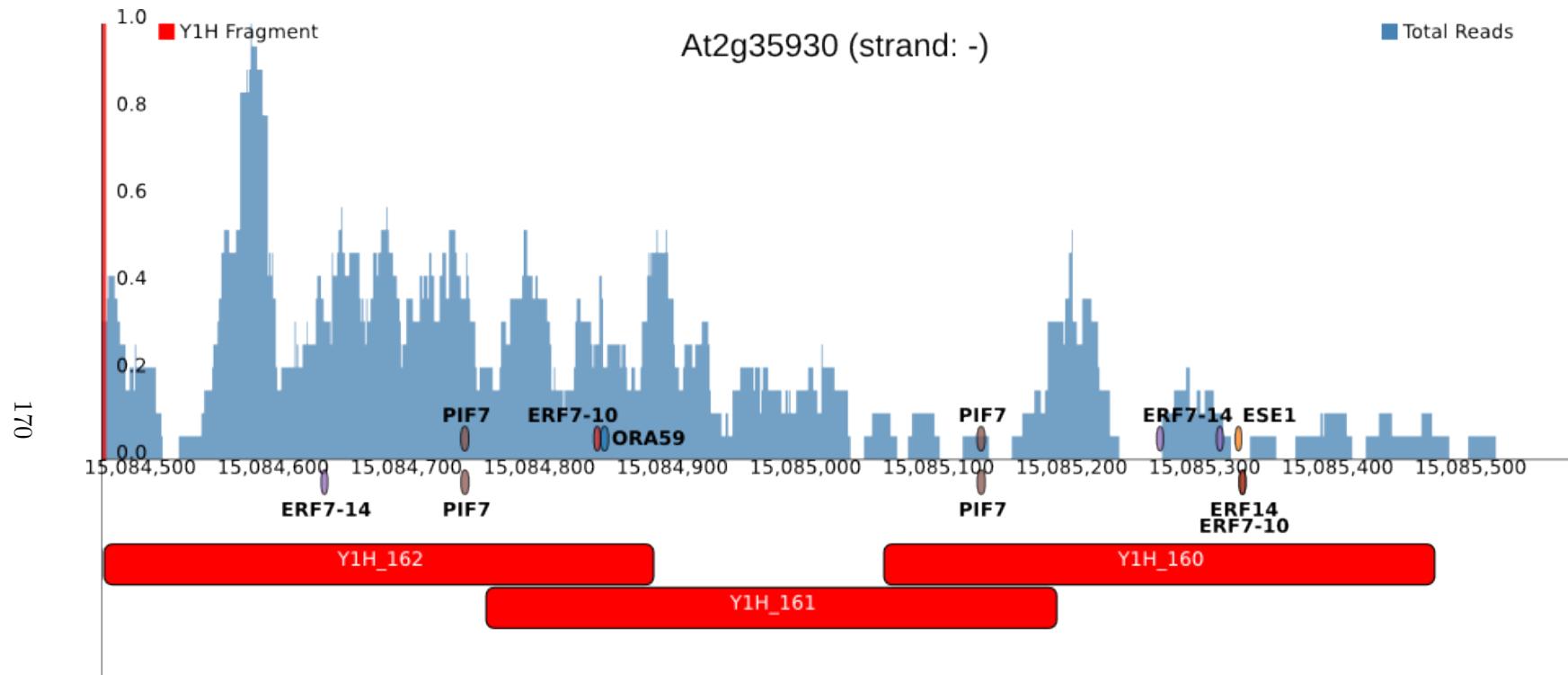


Figure 4.14: Summary of all results obtained in this thesis for *At2g35930*: Y1H fragments (red), predicted CNSs (green) locations of the predicted motifs and DNase profile (blue) showing open chromatin areas. Height of the DNase peaks is proportional to the DNase-Seq reads in the Arabidopsis leaf tissue but normalised by maximum read depth along the whole promoter. Red line at the edge is the annotated TSS.

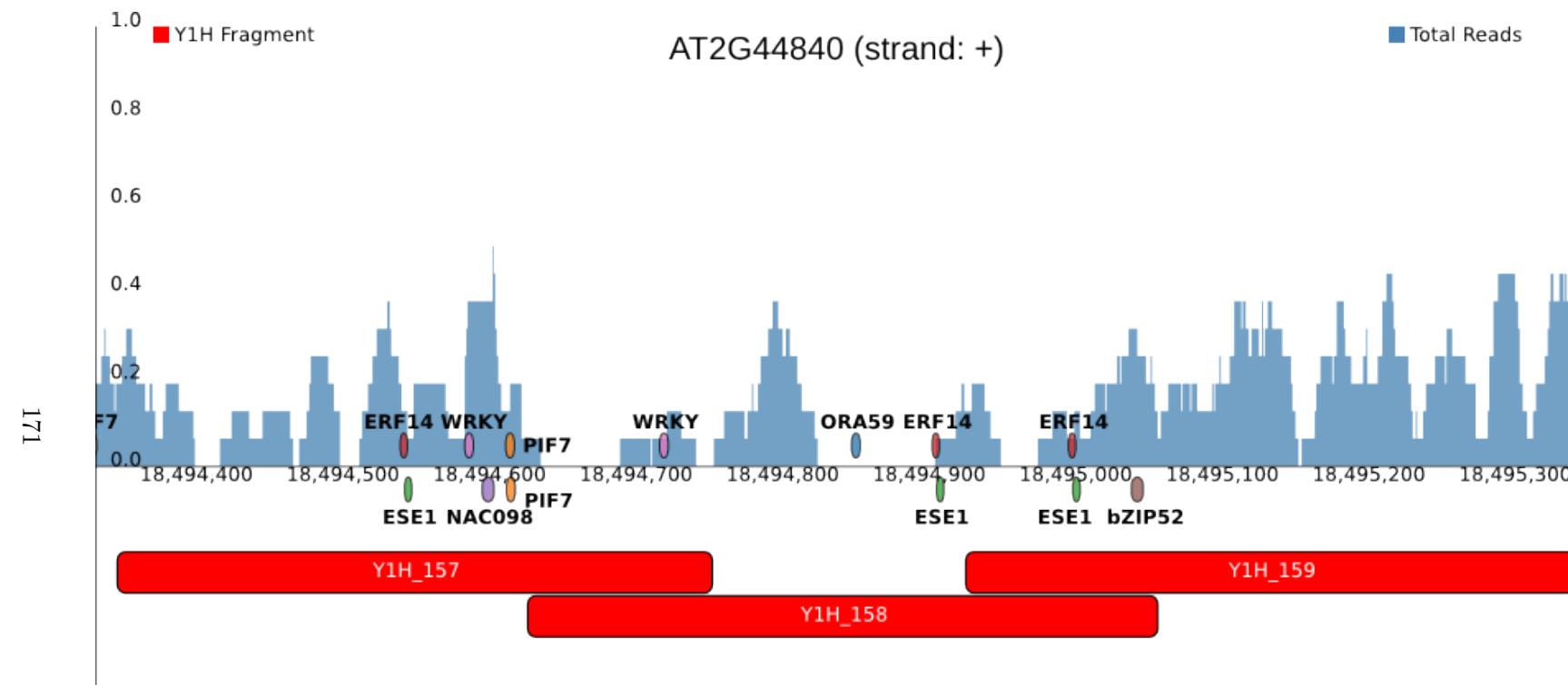


Figure 4.15: Summary of all results obtained in this thesis for *At2g44840*: Y1H fragments (red), predicted CNSs (green) locations of the predicted motifs and DNase profile (blue) showing open chromatin areas. Height of the DNase peaks is proportional to the DNase-Seq reads in the *Arabidopsis* leaf tissue but normalised by maximum read depth along the whole promoter. Red line at the edge is the annotated TSS.

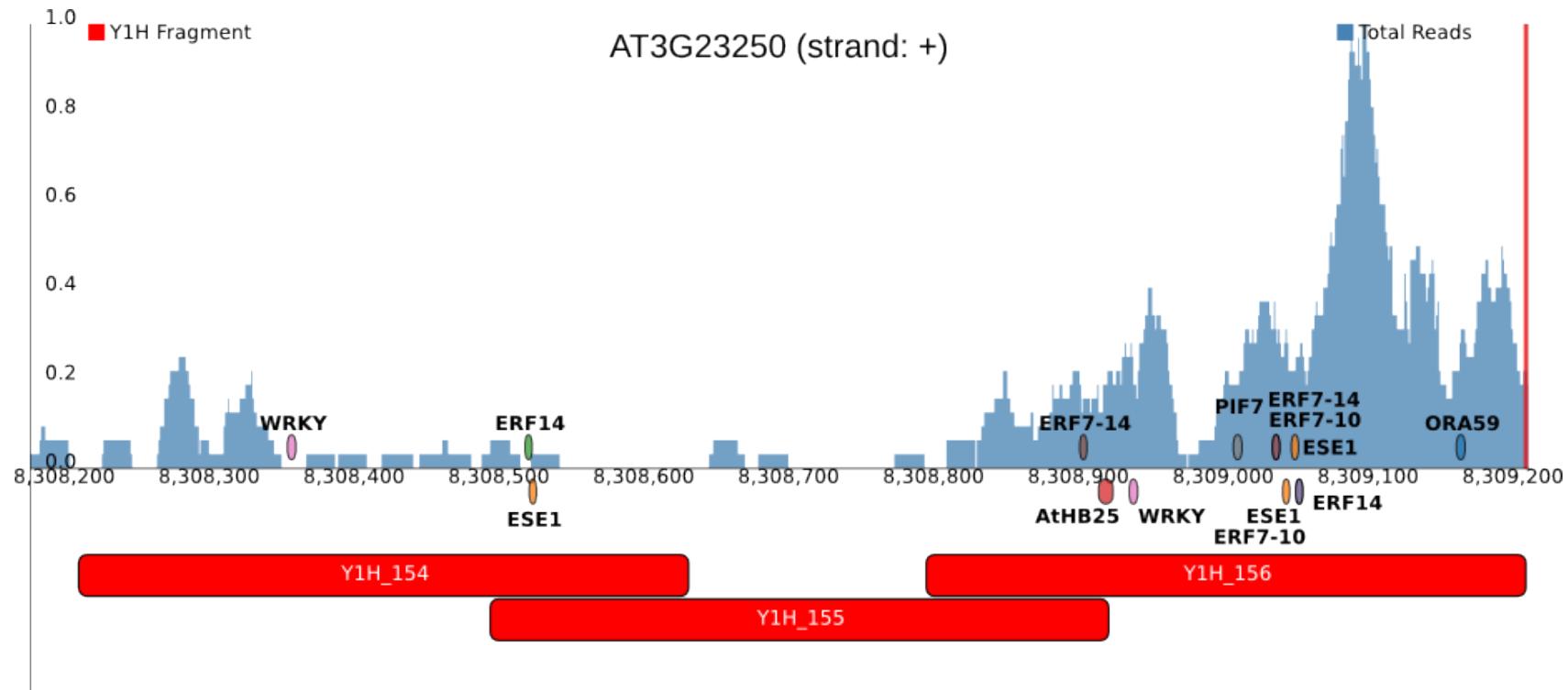


Figure 4.16: Summary of all results obtained in this thesis for *At3g23250*: Y1H fragments (red), predicted CNSs (green) locations of the predicted motifs and DNase profile (blue) showing open chromatin areas. Height of the DNase peaks is proportional to the DNase-Seq reads in the Arabidopsis leaf tissue but normalised by maximum read depth along the whole promoter. Red line at the edge is the annotated TSS.

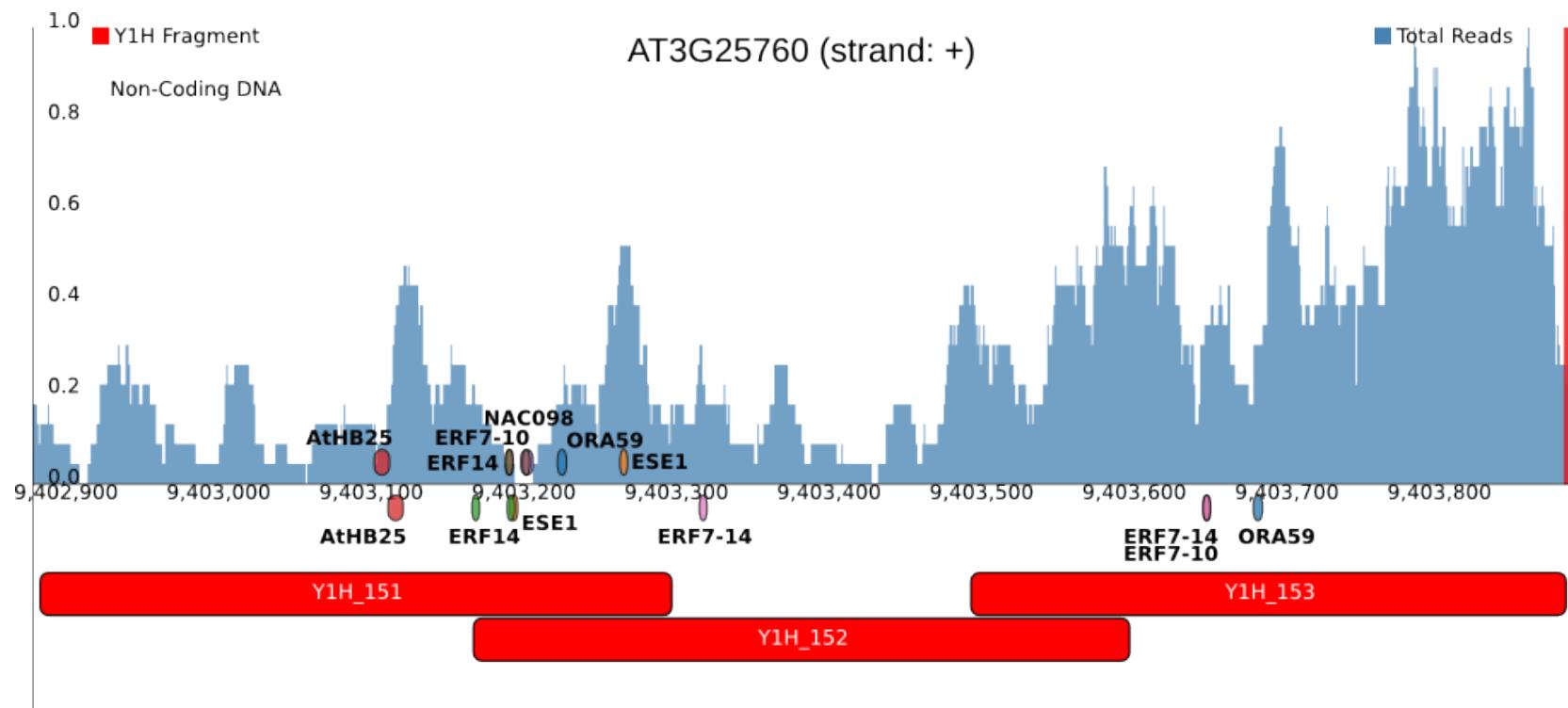


Figure 4.17: Summary of all results obtained in this thesis for *At3g25760*: Y1H fragments (red), predicted CNSs (green) locations of the predicted motifs and DNase profile (blue) showing open chromatin areas. Height of the DNase peaks is proportional to the DNase-Seq reads in the *Arabidopsis* leaf tissue but normalised by maximum read depth along the whole promoter. Red line at the edge is the annotated TSS.

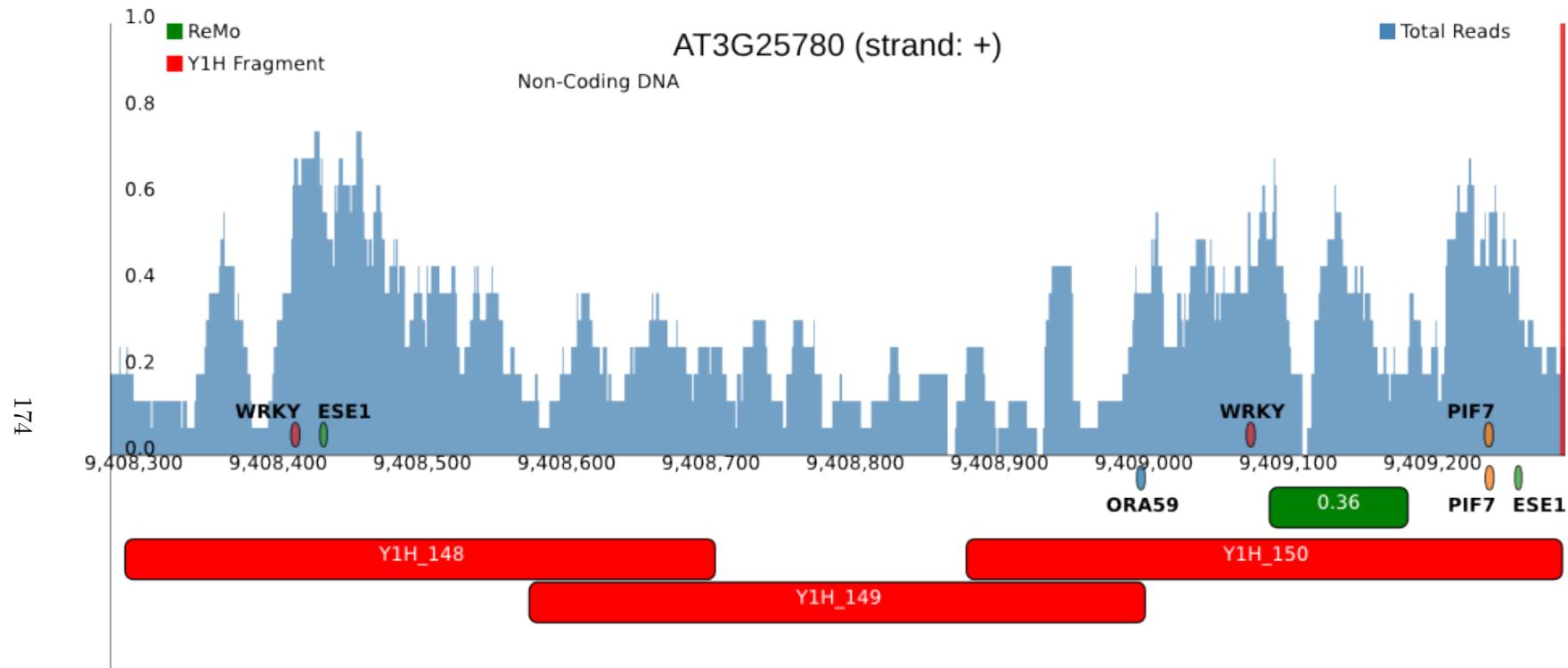


Figure 4.18: Summary of all results obtained in this thesis for *At3g25780*: Y1H fragments (red), predicted CNSs (green) locations of the predicted motifs and DNase profile (blue) showing open chromatin areas. Height of the DNase peaks is proportional to the DNase-Seq reads in the *Arabidopsis* leaf tissue but normalised by maximum read depth along the whole promoter. Red line at the edge is the annotated TSS.

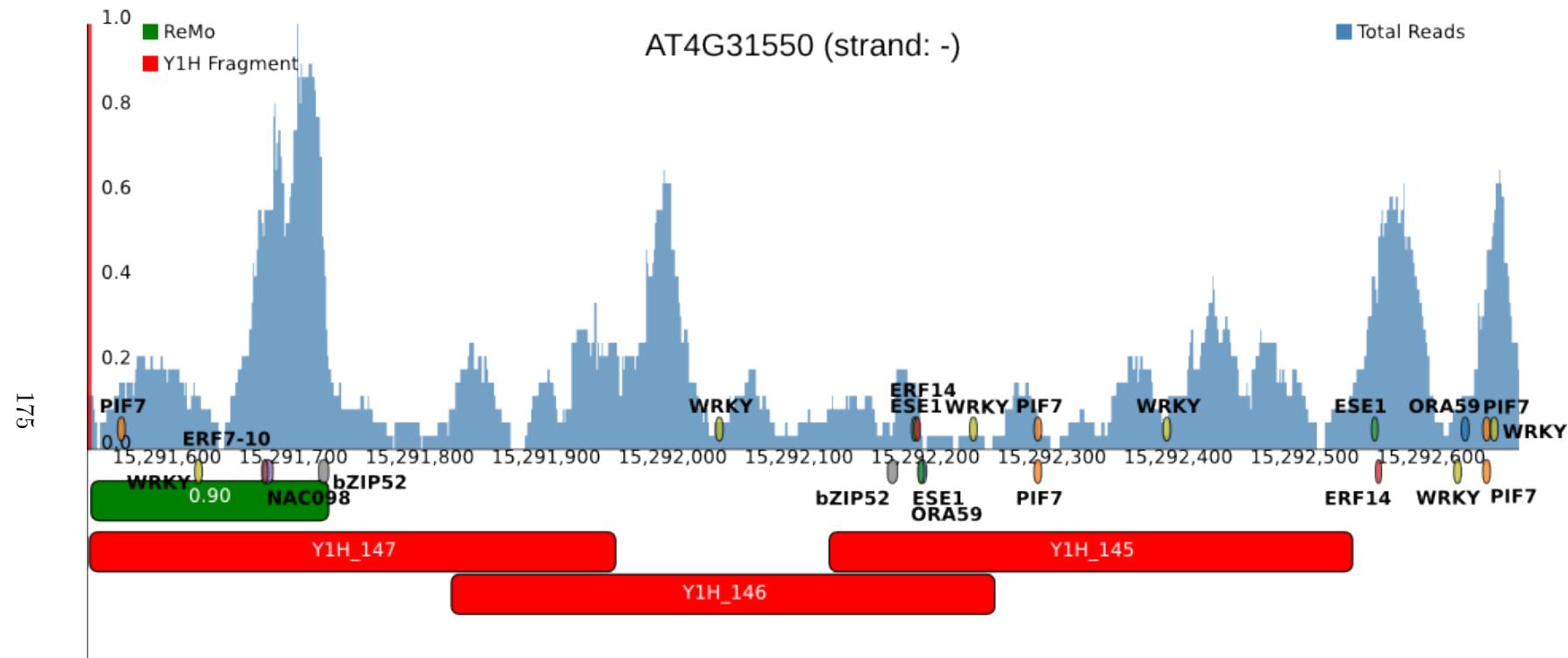


Figure 4.19: Summary of all results obtained in this thesis for *At4g31550*: Y1H fragments (red), predicted CNSs (green) locations of the predicted motifs and DNase profile (blue) showing open chromatin areas. Height of the DNase peaks is proportional to the DNase-Seq reads in the *Arabidopsis* leaf tissue but normalised by maximum read depth along the whole promoter. Red line at the edge is the annotated TSS.

#### **4.4.6 Role of *PIF7* and *AtHB25* in Botrytis infection**

A phenotype screen of the *pif7* and *athb25* stable transgenic plants has uncovered improved resistance to infection with Botrytis as well as accelerated growth as compared to Col-0 plants (data not shown). *PIF7* has been shown to be involved in the repression of known low-temperature stress responsive *DREB1* in a circadian manner, through *TOC1* (Kidokoro et al., 2009), so it may also function through analogous mechanisms to repress other defence genes in *A. thaliana*. Y1H results suggest that *PIF7* TF interacts with a number of genes differentially expressed in response to infection with Botrytis, including JAZ1, which is known to function downstream of the JA signalling pathway in response to wounding (Chung et al., 2008). JAZ1 functions as a repressor by interacting through its ZIM domain with stress responsive genes, e.g. MYC2 in a COI dependent manner (Lorenzo et al., 2004). This suggests that *PIF7* may repress genes on its own or by regulating the expression of JAZ1. Moreover, *PIF7* also interacts with AOC1 and AOC3, the intermediates of the JA biosynthetic pathway, providing an additional layer of regulation in a JA dependent manner.

A literature search suggests that functions of the *AtHB25* TF remain largely uncategorised. Similarly to *PIF7*, *AtHB25* appears to be a repressor of its target genes, as T-DNA insertions knocking out the gene have improved the plant's tolerance to Botrytis and have accelerated the plant's growth. Previous reports indicate *AtHB25* involvement in improving drought tolerance (Tran et al., 2007), whereas results obtained here indicate a negative role in response to biotic stresses, such as Botrytis. It would be interesting to determine the effect in dual stress situations when both drought and Botrytis are affecting the plant. Moreover, MeMe analysis suggested the 5'-CAANTANTTG-3' consensus motif to be specific for interaction with *AtHB25* TF. In order to verify this hypothesis, mutations of the motif could be tested *in vitro* using EMSA, *in vivo* in a Y1H screen together with the TF or using promoters, with and without the binding site for *AtHB25*, fused to a reporter construct, such as GFP, together with the plasmid overexpressing the TF *in planta*.

#### **4.4.7 Role of *AtERF14* in Botrytis infection**

*AtERF14* has been shown to play a non redundant role in the plant's defence response (Oñate-Sánchez et al., 2007). A mRNA time-series expression profile in response to infection with Botrytis also suggests it may be a potential candidate to be specifically expressed in response to infection. Y1H screens have revealed a

potentially large number of downstream targets, Figure 4.20. Therefore, a gene knockout line from NASC was obtained and an overexpression construct was made to be tested in protoplasts to probe for new targets of the TF as well as to validate interactions predicted from the Y1H screens. In qPCR analysis of the knockout, a large number of cycles was required to reach the threshold, suggesting that the gene is knocked down and expressed in very low amounts as compared to the wildtype, instead of being completely knocked out. Therefore likely downstream targets would also only show partial changes in the mRNA expression levels. Moreover, the timing post-infection is difficult to ascertain with high precision as infection rates vary between leaves and lines, making it difficult to capture the peak of the gene expression.

### Predicted direct targets of *AtERF14* TF in response to infection with Botrytis

Analysis of the genes differentially expressed in response to overexpression of the TF in protoplasts, and in the leaves with knocked down expression of *AtERF14* show a number of the genes are also differentially expressed in the PRESTA time-series data in response to infection with Botrytis. Firstly, expression of the *THFS*, *EMB25*, *P5SC2*, *At1g19860*, *AQI* and *RD21b* genes is inversely proportional to the expression of the *AtERF14* TF, suggesting that *AtERF14* directly represses their expression in response to Botrytis, Figure 4.20. *P5SC2* and *RD21b* have previously been shown to be involved in the osmotic stress response in Arabidopsis (Székely et al., 2008), additionally the *rd21* mutants are significantly more susceptible to Botrytis (Shindo et al., 2012), further supporting the hypothesis that *AtERF14* acts as a repressor of these genes key defence gene and *erf14* mutants are more resistant to infection with Botrytis.

Conversely, expression of some predicted targets of *AtERF14* in response to Botrytis infection is directly correlated with the expression of the ERF TF. For example, *At1G22890*, *PDF1.1*, *RAP2.9*, *RHAIB*, *UGT76B1*, *At3g28899*, *MSRB8* and *RPS28* are expressed in unison with *AtERF14* in the PRESTA time-series experiments. Notably, *UGT76B1*, KO lines of which have been shown to be more susceptible to infection with *P. syringae*, whereas *UGT76B1* overexpression lines improved the plant's resilience to the same infection (von Saint Paul et al., 2011). Additionally, expression of *UGT76B1* in *erf14* KO and overexpression experiments suggests that *ERF14* acts as a repressor of *UGT76B1*, Table 4.10. Repression of defence genes in Arabidopsis by *AtERF14* has been proposed in the past (Camehl

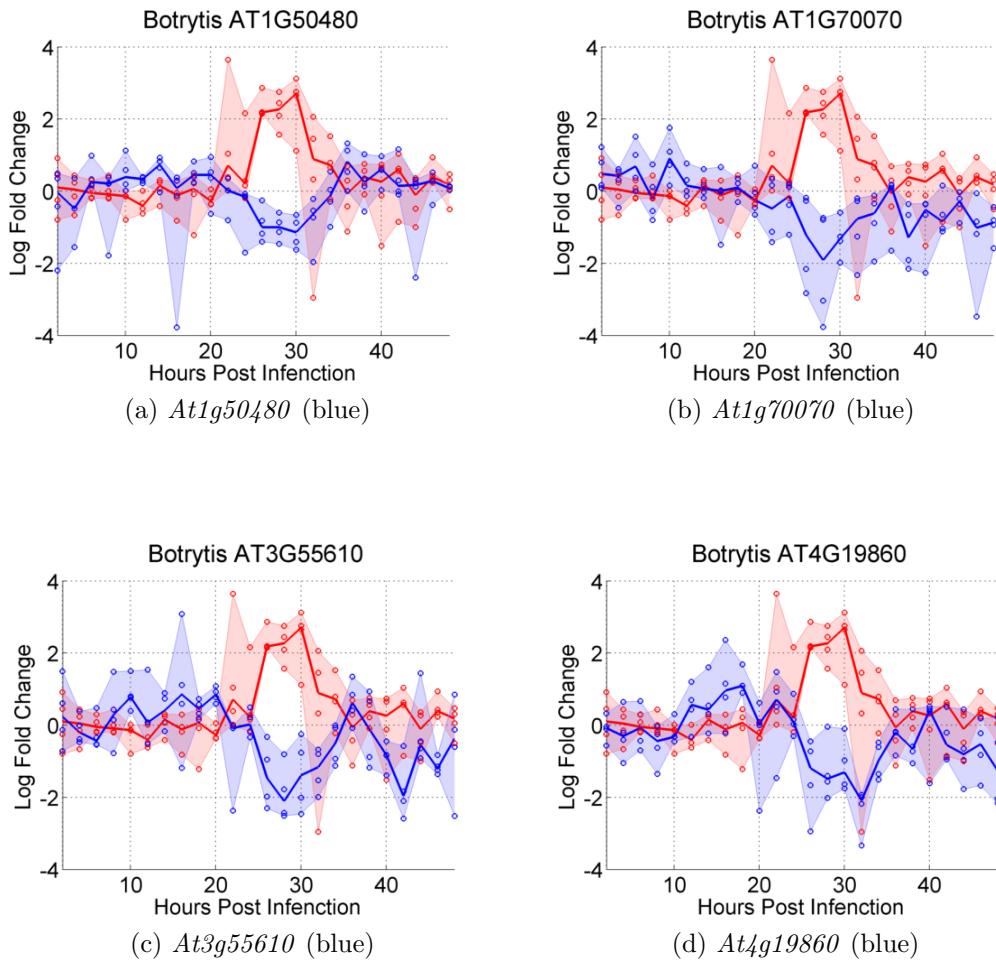


Figure 4.20: mRNA expression profiles of *AtERF14* (red) and associated targets (blue), found to be differentially expressed in *erf14* KO line, and have correlated expression during the infection with *Botrytis* supporting regulatory link.

and Oelmüller, 2010), and strengthened further with improved, although not significantly, resistance to *Botrytis* in the *erf14* mutants, Figure 4.7. *AtERF14* is a relatively small protein (133 aa) of which 60 aa form the AP2 DNA-binding domain with no other conserved domains identified using the CDD tool (Marchler-Bauer et al., 2011). Although not much is known about transcription activation domains in plants, small areas outside of the AP2 domain in *AtERF14* seem unlikely to contain any additional interaction domains, further supporting the repressive nature of the TF.

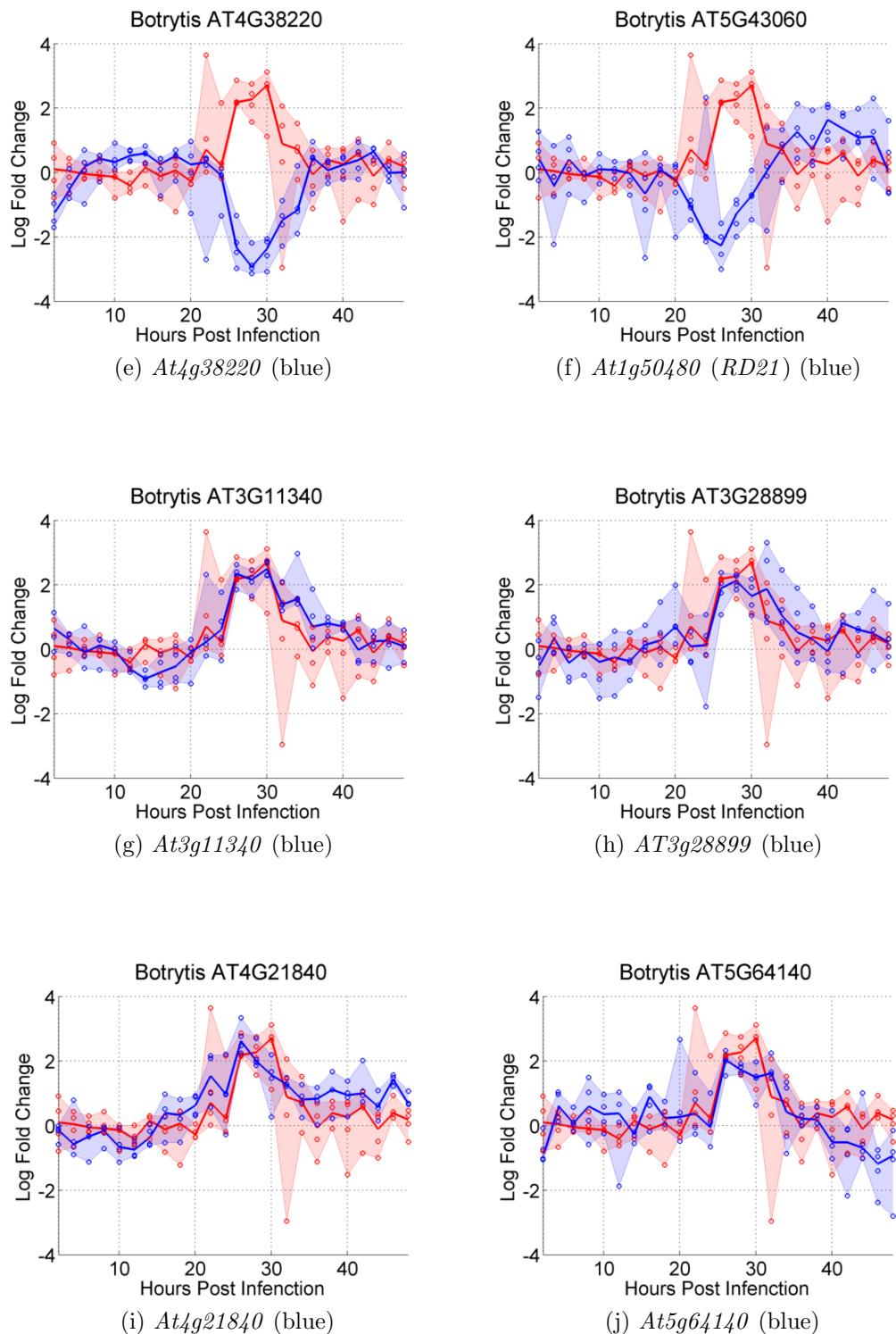


Figure 4.20: mRNA expression profiles of *AtERF14* (red) and associated targets (blue), found to be differentially expressed in *erf14* KO line, and have correlated expression during the infection with Botrytis supporting regulatory link.

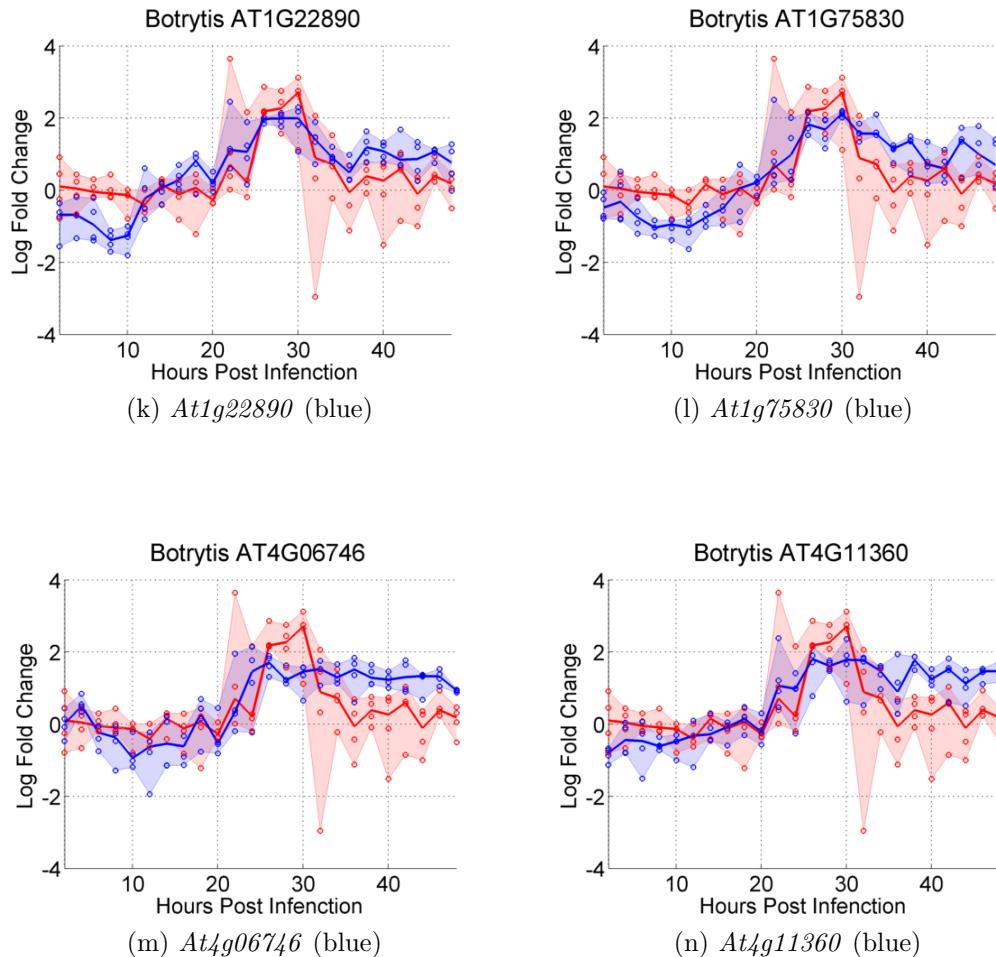


Figure 4.20: mRNA expression profiles of *AtERF14* (red) and associated targets (blue), found to be differentially expressed in *erf14* KO line, and have correlated expression during the infection with Botrytis supporting regulatory link.

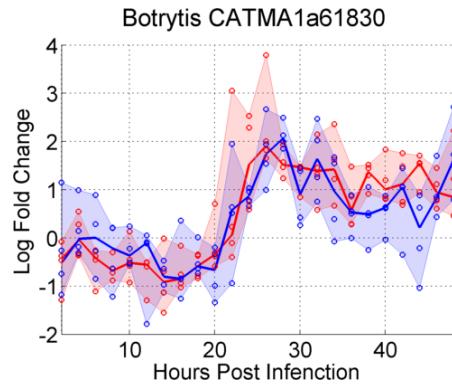
#### 4.4.8 Role of *ESE1* in Botrytis infection

*ESE1* has been shown to play an important role in the response to salt stress in Arabidopsis (Zhang et al., 2011). However, *ESE1*'s function in infection with Botrytis has not been established. The mRNA expression pattern from the PRESTA time-series dataset suggests that the TF is actively transcribed in response to the infection with the necrotroph Botrytis. The Y1H screen has revealed the potential regulation of the JA pathway, through *JAZ1* and intermediate JA-biosynthesis genes *AOC1* and *AOC3*. Overexpression of the TF in protoplasts did not result in any significant up- or down-regulation of JA genes. The only gene found to be significantly down-regulated in the protoplast analysis was *AT5g50570*, suggesting a repressive role *ESE1* plays on gene expression. MapMan analysis of all differentially expressed genes have shown that the biotic and abiotic stress pathways have the largest number of genes changing in expression and that these genes also have the most significant changes of expression. Other pathways have relatively few genes affected by overexpression of *ESE1* (data not shown).

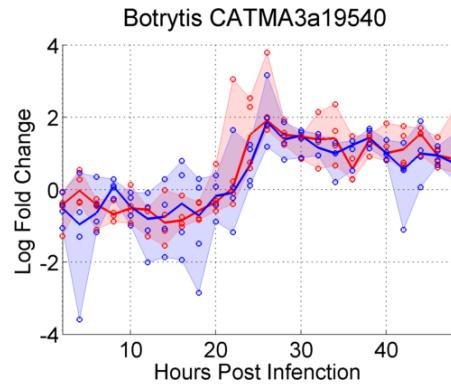
Analysis of the genes with the largest positive fold changes in response to *ESE1* overexpression in protoplasts, suggesting a positive effect of the *ESE1* TF, reveal that many are associated with unknown functions, and expression analysis of these genes in the PRESTA time-series shows that some are also differentially expressed in response to Botrytis in the same manner as *ESE1*, Figure 4.21a - 4.21c. Both lines of evidence suggest that *ESE1* helps to activate transcription of these genes either by itself or together with other genes. Conversely, some genes that are found to be up regulated in the overexpression dataset are down regulated in response to Botrytis, Figure 4.21d - 4.21f suggesting that although *ESE1* can activate their expression in protoplasts, it functions as a repressor in the context of infection with Botrytis.

Similarly, genes found to be significantly down regulated in protoplasts are also differentially expressed in response to Botrytis. Although down regulated in protoplasts, most are up regulated in the Botrytis time-series, with the exception of *ATGSR2*, suggesting that *ESE1* represses their expression in some contexts, but helps to activate transcription potentially with other factors, in response to infection with Botrytis. Alternatively, *ESE1* could be acting in a certain *cis*-regulatory location when expressed in protoplasts, but the changes in chromatin structure associated with the response to infection, could be changing the location where *ESE1* binds to the DNA turning it from an activator into a repressor. For example *HDA6* and *HDA19* are histone modification factors that act as transcriptional repressors

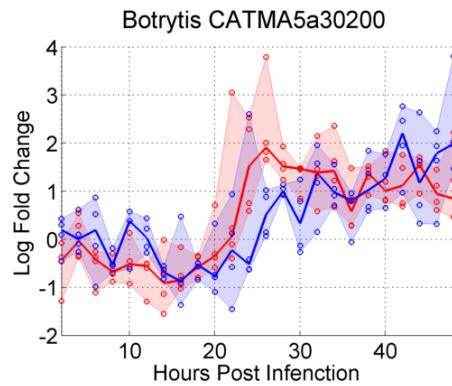
and are known to have an important role in the expression of some key defence response genes, e.g. *ERF1* and *PR* (Zhou et al., 2005).



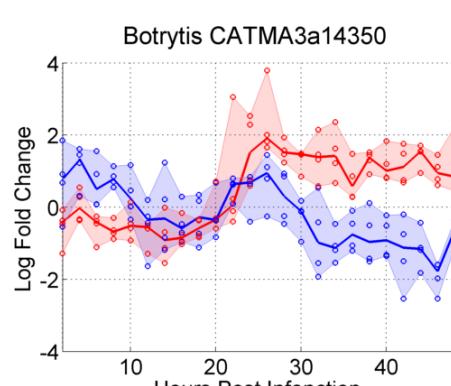
(a) *At1g72600* (blue) potentially directly positively regulated by *ESE1* during Botrytis infection



(b) *At3g19920* (blue) potentially directly positively regulated by *ESE1* during Botrytis infection



(c) *At5g35090* (blue) potentially directly positively regulated by *ESE1* during Botrytis infection



(d) *At3g15000* (blue) down regulated by *ESE1* during Botrytis infection

Figure 4.21: mRNA expression profiles of *ESE1* (red) and associated targets (blue), found to be differentially expressed in protoplasts overexpressing *ESE1* TF, and have correlated expression during the infection with Botrytis supporting regulatory link.

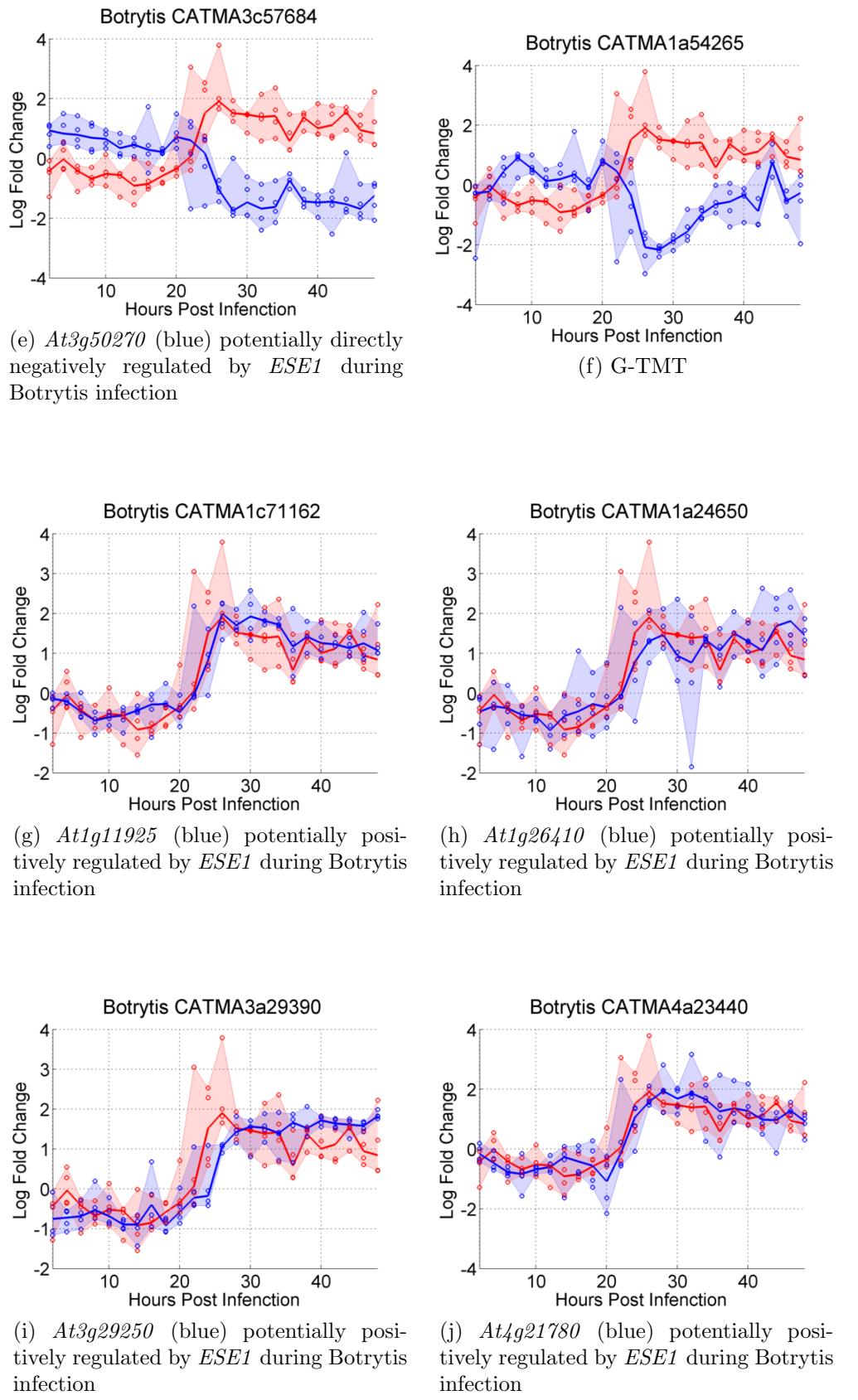
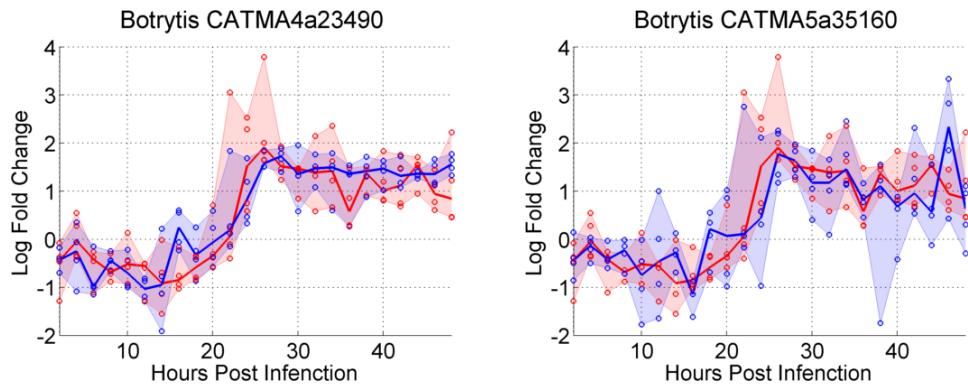
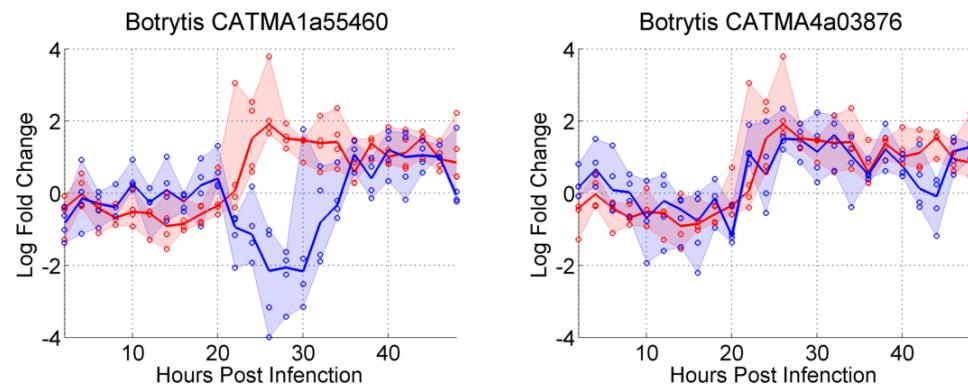


Figure 4.21: mRNA expression profiles of *ESE1* (red) and associated targets (blue), found to be differentially expressed in protoplasts overexpressing *ESE1* TF, and have correlated expression during the infection with Botrytis supporting regulatory link.



(k) *At4g21830* (blue) potentially positively regulated by *ESE1* during Botrytis infection

(l) *At5g39580* (blue) potentially directly negatively regulated by *ESE1* during Botrytis infection



(m) *AtGSR2* (blue) potentially negatively regulated by *ESE1* during Botrytis infection

(n) *RMA1* (blue) potentially directly negatively regulated by *ESE1* during Botrytis infection

Figure 4.21: mRNA expression profiles of *ESE1* (red) and associated targets (blue), found to be differentially expressed in protoplasts overexpressing *ESE1* TF, and have correlated expression during the infection with Botrytis supporting regulatory link.

#### **4.4.9 Context of regulation by *AtERF14* and *ESE1* TFs and their respected motifs**

Positive results from the Y1H experiments provide a “context-free” perspective on the potential direct protein-DNA interactions between the TFs and the promoters of the associated genes. However, it is not known whether this interaction takes place, if at all, during a plants’ growth, development and/or stress response. It is also very difficult to prove that an observed “context-free” interaction can never take place *in planta* under any condition. Conversely, if an interaction does not take place in a “context-free” environment, that could be due to the limitations of the experiment. For example, a TF interacting with promoter elements distal from the TSS may not be able to activate the expression of the reporter gene in Y1H, therefore producing a false negative result. One way of validating if the proposed interactions could occur is by overexpressing the TF in protoplasts and assessing the effect on the expression of the putative target gene. Although this provides a strategy to test interactions in an environment closely resembling that of a whole plant, some conditions cannot be tested adequately in protoplasts. For example, if the TF is proposed to be regulating under stress conditions, these cannot be easily duplicated in the protoplast system. Thus, the context of the interaction is limited to those conditions which can be created for protoplasts. ChIP-Seq can be used instead of protoplasts to test for binding locations of the TF and if coupled with RT-PCR, can test whether the target genes’ expression is altered in plant’s constitutively overexpressing the TF of interest.

Similarly to the context of protein-DNA interaction as a whole, the location of this interaction can also be context dependent. The whole transcriptome is known to undergo massive reprogramming in response to stress (Kreps et al., 2002; Morcuende et al., 2007; Pauwels et al., 2008; Kusano et al., 2011). These changes also affect the state of open and closed chromatin (Kim et al., 2012), in turn allowing or restricting access to the binding motifs for the TF(s) of interest. The dependency of the genomic locations and their effect on the expression of the associated genes on the context is much harder to test as that would require stable lines with mutations of the binding site in the appropriate locations.

## 4.5 Conclusions

In conclusion, this chapter presents a bioinformatics approach to elucidating new binding motifs that are also TF specific. 14 motifs have been identified as being present in the promoter fragments of those genes tested in the Y1H screens. Seven of the motifs are predicted to be specific to *ORA59*, *AtERF14*, *AtERF7*, *ESE1*, *AtHB25*, *NAC098* and *bZIP52* respectively, Table 4.1, 4.2 and 4.3. A consensus 5'-CCACG-3' motif showed the strongest effect on altering the predicted protein-DNA interactions in the context of Y1H screens. These motifs tie the expression of the TF known to play a significant role in the response to Botrytis infection with that of their target genes. In particular, the mutants of *erf14*, *pif7*, and *athb25* are shown to be more resistant to infection with the necrotroph. In the case of AtERF14, a number of genes regulated by the TF are uncategorised and thus have an unknown function, e.g. *At1g22890* and *At2g44670*, but may still play an important role in the plant's resistance to Botrytis.

## Chapter 5

# General Discussion

This thesis presents a body of interdisciplinary work utilising bioinformatic approaches together with experimental techniques to identify potentially functional ncDNA regions. From this an extended, “context-free” GRN was built using a Y1H library screen on the promoters of selected genes found to be differentially expressed in response to infection with *Botrytis B. cinerea* as compared to a mock infection. Some of the TFs found to be interacting with a large proportion of the promoter fragments have produced a detectable increase and significant resistance to infection with *B. cinerea*. Using bioinformatics tools on the promoter fragments found to be interacting with the TFs tested in the Y1H screen, potential TF specific motifs responsible for the observed interactions have been identified. In turn, TF specific motifs can serve as prior information for modelling future GRNs with higher precision.

Firstly, this thesis has focused on bioinformatic approaches to identifying new conserved non-coding sequences (CNSs). CNSs occur in the intergenic regions of the genome where regulation of gene expression takes place through the binding of TFs. The continuous fall in the cost of sequencing has led to an ever increasing pool of genomic sequences available for study. Evolutionary theory has been successfully applied to the coding regions in the genomes in order to track changes within ecotypes and across different species. However, the rules governing evolution in the non-coding sequence of the genome appears to differ to the rules directing the flow of evolution in the coding sequences. Point mutations, insertions or deletions in the protein coding sequence can render the gene non-functional. On the other hand, the same alterations in the ncDNA can abolish or boost expression in certain tissues or in certain conditions. This has lead to an increased interest in tracking evolu-

tionary changes across multiple genomes, in the hope that new functional areas can be uncovered using these techniques. The evidence on the CNSs identified between *Arabidopsis*, papaya, poplar and grape suggest them to be functional and involved in a variety of developmental processes, as well as harbouring a much larger number of motifs than randomly chosen promoter sequences resembling CNSs in their genomic locations.

Sequencing of new plant species and the 1001 genomes project (Ossowski et al., 2008) is paving the way for the identification and functional annotation of new orthologous sequences in plants. Improved genomic annotations will increase the scope and confidence of methods relying on validated gene boundaries, as they would yield higher confidence predictions of sequences found in the promoter regions of genes. A question only lightly touched on by the work presented here on the CNSs is that of neo-, sub- and non-functionalisation of the orthologous and/or paralogous genes that are caused, in part, by whole genome duplication events. The CNSs present in the promoter region of a gene in one species which is orthologous to another two genes in a second species may provide clues to the functionality of the gene in each species, and whether neo-, sub- and non-functionalisation of the gene is caused by the changes in the promoter elements or by mutations in the coding sequence of the gene. If an ecotype is found to be more resilient to stress or to produce higher biomass plants, identified CNSs may also provide an explanation for such features. However, without more genome wide data available for analysis it is very difficult to draw conclusions from information contained in the CNSs alone.

Secondly, this thesis focuses on the identification of direct protein-DNA interactions in the promoters of genes, some of which contain strong CNSs. The functional binding motifs within the promoters of genes are able to integrate complex signals present during development and stress conditions. The cues are read and integrated together to form gene regulatory networks through the binding of TFs to the promoter regions of genes, thus altering the transcription rates of the associated genes. The library Y1H approach chosen for elucidating positive direct protein-DNA interactions offers the ability to test an entire complement of TFs from a chosen genome in a single experiment. Unlike many other techniques, e.g. ChIP-Seq, the gene regulatory network is built from the bottom up, which reveals the complexity of the single gene regulation and can serve as good prior information to build better models for the transcriptional regulation of genes. The results presented here show a highly connected and complex network of TFs predicted to

be regulating their associated target genes in a “context-free” setting. Moreover, phenotypic analysis of some TFs, found to be interacting with a large proportion of the promoter fragments, have identified them as being functionally important in the biotic stress response, since plants with mutated TFs were observed to have altered susceptibility to infection with Botrytis.

The Y1H screen techniques can be further enhanced in two ways. Firstly, although 65% of TFs, in this case from *Arabidopsis*, are present in the Y1H library format, they are represented as a single transcript variant. Previous reports suggest that 42% to 60% of intron-containing genes in *Arabidopsis* are alternatively spliced (Filichkin et al., 2010; Marquez et al., 2012), implying that the current gene regulatory networks may be missing complexity associated with multiplicity of alternatively spliced TFs. Inclusion of all isoforms, although labour intensive, may lead to a better understanding of the changes in the nucleous environment and the subsequent effect on direct protein-DNA interactions. Secondly, progress to fully automate the screening process may allow further increases in the throughput rate of the screen.

Finally, mounting evidence suggests a Kuhnian paradigm shift in the way TF binding is viewed and interpreted. So far, research has had limited success in identifying TF binding sites due to the large number of potential nucleotide combinations. The main focus has been on the identification of a handful of motifs for a limited number of members of a family of TFs, and then using these motifs for all remaining members of the family due to the homology of the DNA binding domain in multigene TF families. Moreover, the presence of the binding site alone is not indicative that a TF will bind to this sequence *in vivo*. The reasons for this discrepancy still remain unknown, but clues for the difference in binding potential may lie in the structure and sequence immediately adjacent to the binding sites. There are parallels to be drawn from protein folding research, in which an understanding of protein structure allows us to better understand protein function. Similarly, understanding the folding of the TFs themselves and the DNA sequence around the binding site and the way it changes depending on other proteins present in the immediate vicinity, may provide an explanation for present and absent binding given the presence of the same binding motif. However, unlike protein folding, structural analysis of the DNA sequence itself has been slow and only a handful of molecular dynamic models exist for DNA folding alone, or in the presence of a TF. For example, HADDOCK tools allow for *ab initio* protein-DNA interactions (van Dijk et al., 2006). Advances in

the area of modelling techniques and computational power availability means that protein-DNA interactions could be predicted with confidence *in silico* leading to only a handful of true positive interactions having to be experimentally verified.

## **Appendix A**

# **Conserved Noncoding Sequences Highlight Shared Components of Regulatory Networks in Dicotyledonous Plants**

Gene Identifier	Gene Name
AT3G54320	WRI1
AT3G24650	ABI3
AT1G69120	AP1 (APETALA1)
AT4G36920	AP2
AT5G20240	PI (PISTILLATA)
AT3G26790	FUS3
AT1G64280	NPR1
AT1G51190	PLT2 (PLETHORA2)
AT1G14350	FOUR LIPS (AtMYB124)
AT2G17950	WUSCHEL (WUS) (PGA6)
AT5G61850	LFY (LEAFY)
AT3G54340	AP3
AT4G18960	AG (AGAMOUS)
AT1G21970	LEC1
AT1G28300	LEC2
AT5G61960	AML1 (arabidopsis MEI-like 1)
AT3G26744	ICE1
AT3G24140	FAMA
AT3G60460	DUO1 (R2R3 myb)
AT1G71930	VND7
AT5G18830	SPL7
AT1G32330	A1d
AT3G06120	MUTE
AT5G53210	SPEECHLESS (SPCH)
AT1G62360	SHOOT MERISTEMLESS (STM)
AT4G29860	TAN
AT5G47670	L1L (LEC1-like)
AT1G04370	AtERF14 (Ethylene-responsive element binding factor 14)
AT4G17750	HsfA1a (HSF1)
AT5G16820	A1b
AT1G52740	H2A.Z
AT4G25470	CBF2
AT2G02820	MYB88

Table A.1: Manually curated list by Laura Baxter of plant “Master Regulators” derived from current literature.

Conservation Score Threshold	Paralogs		Random Gene Pairs		False Positive Rate
	No. of Genes	No. of Aligned Regions	No. of Genes	No. of Aligned Regions	
1	224	243	1	1	0.0003
0.9	479	564	1	2	0.0003
0.8	719	882	1	3	0.0003
0.7	771	964	1	3	0.0003
0.6	952	1247	1	3	0.0003
0.5	971	1289	1	3	0.0003
0.4	1005	1335	2	4	0.0006
0.3	1149	1573	4	6	0.0012

Table A.2: Numbers of aligned regions and associated genes from paralogous promoters and from promoters of randomly paired paralog genes at different thresholds of conservation score (compiled by Laura Baxter).

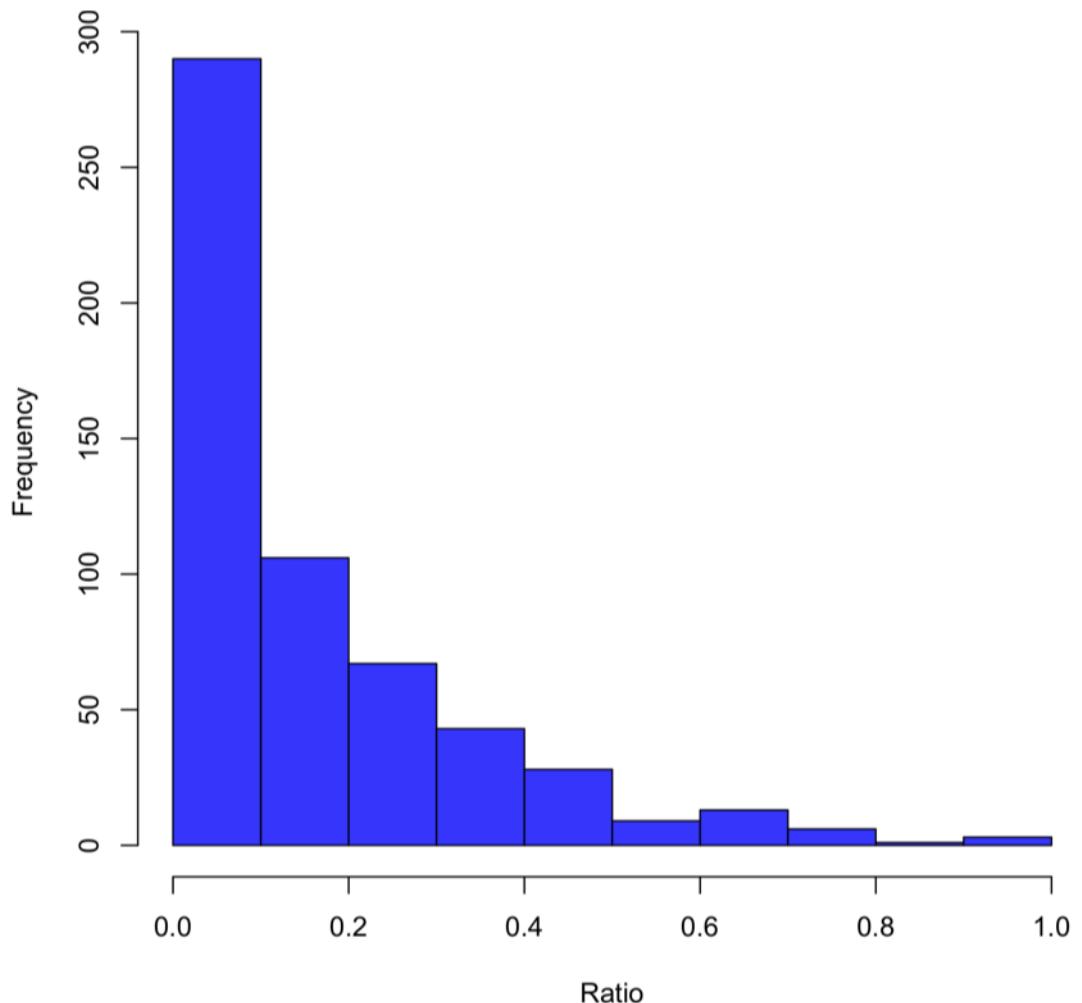


Figure A.1: Distribution of distances between CNS and TSSs in Arabidopsis, normalized by intergenic length.

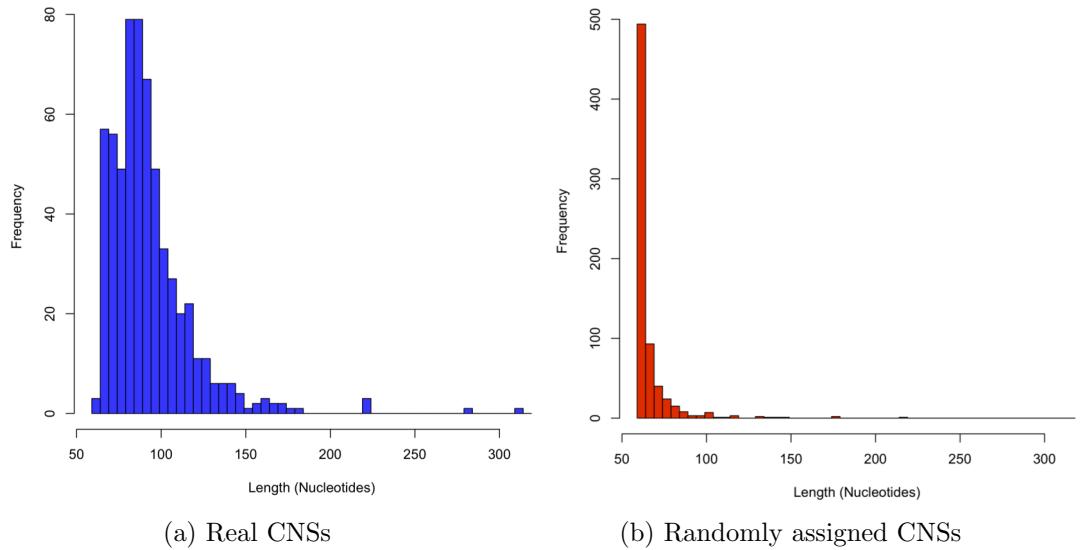


Figure A.2: The distribution of alignment lengths in orthologs and randomly assigned gene pairs. CNS lengths identified in orthologous promoters (above left), applying the 0.7 threshold, randomly assigned gene pair promoters (control, above right), applying the 0.3 threshold. The different thresholds were applied between the two datasets to provide comparable numbers in each distribution (602 conserved regions from orthologous promoters, and 700 alignments from random gene pair promoters).

## **Appendix B**

# **Elucidating Functional Elements and Gene Regulatory Network Using Yeast One-Hybrid Screens**

Table B.1: Positive interactions between the library of TFs and promoter 29 promoter constructs. Number in brackets indicates number of positive colonies recorded for associated interaction ranging for 1 to 5.

197

<i>At5g50570</i>		<i>At5g05090</i>			<i>At3g25780</i>			
Y1H-139	Y1H-140	Y1H-141	Y1H-142	Y1H-143	Y1H-144	Y1H-148	Y1H-149	Y1H-150
<i>At1g61110</i> (1)	<i>At5g63790</i> (1)			<i>AT5g39610</i> (4)		<i>At1g64000</i> (2)	<i>At1g74930</i> (2)	<i>At1g06160</i> (1)
<i>At5g63790</i> (5)						<i>At1g75390</i> (1)	<i>At5g21120</i> (5)	
						<i>At4g31660</i> (1)		
						<i>At5g42630</i> (1)		
						<i>At5g60200</i> (2)		

Continued on next page

**Table B.1 – continued from previous page**

Positive interactions between the library of TFs and promoter 29 promoter constructs.

Y1H-145	<i>At4g31550</i>			<i>At3g25760</i>			<i>At3g23250</i>		
Y1H-146	Y1H-147	Y1H-151	Y1H-152	Y1H-153	Y1H-154	Y1H-155	Y1H-156		
<i>At5g43410</i> (1)	<i>At2g36610</i> (1)	<i>At1g06850</i> (1)	<i>At2g24430</i> (5)	<i>At5g47220</i> (1)	<i>At1g74930</i> (4)		<i>At1g76110</i> (5)	<i>At3g20310</i> (1)	
<i>At1g67970</i> (1)	<i>At5g53980</i> (2)	<i>At1g33280</i> (3)	<i>At3g27785</i> (5)	<i>At4g25470</i> (1)	<i>At1g04370</i> (1)		<i>At1g73360</i> (2)	<i>At4g11070</i> (1)	
		<i>At2g24430</i> (4)	<i>At4g13040</i> (2)		<i>At2g17600</i> (1)			<i>At2g46270</i> (1)	
		<i>At2g36610</i> (1)	<i>At5g63790</i> (5)		<i>At2g31230</i> (1)			<i>At5g49300</i> (1)	
		<i>At2g46830</i> (2)	<i>At5g57520</i> (1)		<i>At2g33310</i> (1)				
		<i>At3g12910</i> (3)			<i>At2g33860</i> (1)				
		<i>At3g61890</i> (2)			<i>At3g13540</i> (1)				
		<i>At2g28340</i> (1)			<i>At3g61830</i> (1)				
		<i>At2g46160</i> (1)			<i>At4g17920</i> (1)				
		<i>At3g61910</i> (5)			<i>At4g24060</i> (1)				
		<i>At4g37790</i> (1)			<i>At5g39610</i> (1)				
		<i>At5g17300</i> (1)							
		<i>At5g39610</i> (1)							
		<i>At5g53980</i> (1)							
		<i>At5g66770</i> (1)							

Continued on next page

**Table B.1 – continued from previous page**

Positive interactions between the library of TFs and promoter 29 promoter constructs.

<i>At1g80840</i>			<i>At2g44840</i>		
Y1H-175	Y1H-176	Y1H-177	Y1H-157	Y1H-158	Y1H-159
<i>AT4G11070</i> (5)	<i>AT5G06500</i> (1)	<i>AT1G50640</i> (1)	<i>At5g53950</i> (2)	<i>At2g28240</i> (5)	<i>At3g49690</i> (1)
<i>AT4G23810</i> (5)			<i>At2g24430</i> (5)		<i>At5g57520</i> (2)
<i>AT5G61270</i> (5)			<i>At3g58120</i> (1)		<i>At2g24430</i> (1)
<i>AT5G65410</i> (1)			<i>At5g57520</i> (2)		<i>At3g27920</i> (1)
					<i>At5g63470</i> (1)
					<i>At3g47600</i> (1)
					<i>At5g62320</i> (1)
					<i>At1g66140</i> (2)

**Table B.1 – continued from previous page**  
Positive interactions between the library of TFs and promoter 29 promoter constructs.

<i>At2g35930</i>			<i>At1g19180</i>		
Y1H-160	Y1H-161	Y1H-162	Y1H-172	Y1H-173	Y1H-174
<i>AT4G14410</i> (3)	<i>AT1G69310</i> (3)	<i>AT1G29280</i> (4)	<i>AT4G11070</i> (5)	<i>AT5G39760</i> (1)	<i>AT1G50640</i> (1)
<i>AT2G23320</i> (5)	<i>AT2G23320</i> (3)		<i>AT4G23810</i> (5)	<i>AT5G65410</i> (2)	<i>AT1G55520</i> (2)
<i>AT2G30590</i> (5)	<i>AT2G24570</i> (1)		<i>AT5G24110</i> (1)		<i>AT1G57560</i> (1)
<i>AT4G18170</i> (4)	<i>AT2G30590</i> (3)		<i>AT3G09370</i> (1)		<i>AT1G64000</i> (1)
<i>AT5G13080</i> (3)	<i>AT3G62340</i> (2)				<i>AT2G18300</i> (1)
<i>AT5G46350</i> (1)	<i>AT4G01250</i> (1)				<i>AT2G34000</i> (1)
<i>AT1G02680</i> (1)	<i>AT4G18170</i> (2)				<i>AT2G41710</i> (1)
<i>AT1G12860</i> (1)	<i>AT4G23550</i> (1)				<i>AT3G01140</i> (1)
<i>AT1G24625</i> (1)	<i>AT5G46350</i> (4)				<i>AT3G13445</i> (2)
<i>AT1G29280</i> (3)					<i>AT3G15540</i> (1)
<i>AT2G24570</i> (3)					<i>AT4G05100</i> (1)
<i>AT2G41070</i> (1)					<i>AT4G29080</i> (1)
<i>AT3G19290</i> (1)					<i>AT4G32890</i> (1)
<i>AT3G50060</i> (1)					<i>AT4G36780</i> (1)
<i>AT3G58710</i> (1)					<i>AT5G17810</i> (1)
<i>AT3G62340</i> (2)					<i>AT5G28650</i> (2)
<i>AT4G01250</i> (1)					<i>AT5G63790</i> (2)
<i>AT4G32040</i> (1)					<i>AT5G64810</i> (1)
<i>AT5G11260</i> (1)					
<i>AT5G17490</i> (1)					
<i>AT5G67190</i> (1)					

## **Appendix C**

# **Computational Approaches To Identify Regulatory Elements In Arabidopsis**

Hours Post Infection	Probability that data was derived from Normal Distribution			
	Col	<i>AtERF14</i>	<i>PIF7</i>	<i>AtHB25</i>
48 h	0.5407(30)	0.6169(39)	0.5775(31)	0.3429(21)
57.7 h	0.1582(30)	0.4248(39)	0.2364(31)	0.1583(21)
72 h	0.7372(30)	0.45849(37)	0.9283(31)	0.3993(21)

Table C.1: All of the samples are derived from the normal distribution as determined by one-sample Kolmogorov-Smirnoff test. The values in the brackets show number of samples in each category.

# Bibliography

- Abe H., Urao T., Ito T., Seki M., Shinozaki K., and Yamaguchi-Shinozaki K. Arabidopsis atmyc2 (bhlh) and atmyb2 (myb) function as transcriptional activators in abscisic acid signaling. *Plant Cell*, 15(1):63–78, Jan 2003.
- Abramovitch R. B., Kim Y.-J., Chen S., Dickman M. B., and Martin G. B. Pseudomonas type iii effector avrptob induces plant disease susceptibility by inhibition of host programmed cell death. *EMBO J*, 22(1):60–9, Jan 2003. doi: 10.1093/emboj/cdg006.
- AbuQamar S., Chen X., Dhawan R., Bluhm B., Salmeron J., Lam S., Dietrich R. A., and Mengiste T. Expression profiling and mutant analysis reveals complex regulatory networks involved in arabidopsis response to botrytis infection. *Plant J*, 48(1):28–44, Oct 2006. doi: 10.1111/j.1365-313X.2006.02849.x.
- Aikawa S., Kobayashi M. J., Satake A., Shimizu K. K., and Kudoh H. Robust control of the seasonal expression of the arabidopsis flc gene in a fluctuating environment. *Proc Natl Acad Sci U S A*, 107(25):11632–7, Jun 2010. doi: 10.1073/pnas.0914293107.
- Alabadí D., Oyama T., Yanovsky M. J., Harmon F. G., Más P., and Kay S. A. Reciprocal regulation between toc1 and lhy/ccal within the arabidopsis circadian clock. *Science*, 293(5531):880–3, Aug 2001. doi: 10.1126/science.1061320.
- Albert I., Mavrich T. N., Tomsho L. P., Qi J., Zanton S. J., Schuster S. C., and Pugh B. F. Translational and rotational settings of h2a.z nucleosomes across the saccharomyces cerevisiae genome. *Nature*, 446(7135):572–6, Mar 2007. doi: 10.1038/nature05632.
- Albert R. and Barabási A.-L. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 01 2002. URL <http://link.aps.org/doi/10.1103/RevModPhys.74.47>.

Albright S. R. and Tjian R. Tafs revisited: more data reveal new twists and confirm old ideas. *Gene*, 242(1–2):1–13, 1 2000. doi: [http://dx.doi.org/10.1016/S0378-1119\(99\)00495-3](http://dx.doi.org/10.1016/S0378-1119(99)00495-3). URL <http://www.sciencedirect.com/science/article/pii/S0378111999004953>.

Allococo D. J., Kohane I. S., and Butte A. J. Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics*, 5:18, Feb 2004. ISSN 1471-2105 (Electronic); 1471-2105 (Linking). doi: 10.1186/1471-2105-5-18.

Altschul S., Gish W., Miller W., Myers E., and Lipman D. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

Andersson L. S., Larhammar M., Memic F., Wootz H., Schwochow D., Rubin C.-J., Patra K., Arnason T., Wellbring L., Hjälm G., Imsland F., Petersen J. L., McCue M. E., Mickelson J. R., Cothran G., Ahituv N., Roepstorff L., Mikko S., Vallstedt A., Lindgren G., Andersson L., and Kullander K. Mutations in dmrt3 affect locomotion in horses and spinal circuit function in mice. *Nature*, 488(7413):642–6, Aug 2012. doi: 10.1038/nature11399.

Andronis C., Barak S., Knowles S. M., Sugano S., and Tobin E. M. The clock protein CCA1 and the bZIP transcription factor HY5 physically interact to regulate gene expression in Arabidopsis. *Mol Plant*, 1(1):58–67, Jan 2008. doi: 10.1093/mp.

Arabidopsis Genome Initiative . Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*, 408(6814):796–815, Dec 2000. doi: 10.1038/35048692.

Arabidopsis Interactome Mapping Consortium . Evidence for network evolution in an arabidopsis interactome map. *Science*, 333(6042):601–7, Jul 2011. doi: 10.1126/science.1203877.

Asselbergh B., De Vleesschauwer D., and Höfte M. Global switches and fine-tuning-aba modulates plant pathogen defense. *Mol Plant Microbe Interact*, 21(6):709–19, Jun 2008. doi: 10.1094/MPMI-21-6-0709.

Bailey T. L. and Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2:28–36, 1994.

Bailey T. L., Boden M., Buske F. A., Frith M., Grant C. E., Clementi L., Ren J., Li W. W., and Noble W. S. Meme suite: tools for motif discovery and searching. *Nucleic Acids Res*, 37(Web Server issue):W202–8, Jul 2009. doi: 10.1093/nar/gkp335.

- Baker S. S., Wilhelm K. S., and Thomashow M. F. The 5'-region of arabidopsis thaliana cor15a has cis-acting elements that confer cold-, drought- and abaregulated gene expression. *Plant Mol Biol*, 24(5):701–13, Mar 1994.
- Baranowskij N., Frohberg C., Prat S., and Willmitzer L. A novel dna binding protein with homology to myb oncoproteins containing only one repeat can function as a transcriptional activator. *EMBO J*, 13(22):5383–92, Nov 1994.
- Bari R. and Jones J. D. G. Role of plant hormones in plant defence responses. *Plant Mol Biol*, 69(4):473–88, Mar 2009. doi: 10.1007/s11103-008-9435-0.
- Barrero J. M., Piqueras P., González-Guzmán M., Serrano R., Rodríguez P. L., Ponce M. R., and Micol J. L. A mutational analysis of the aba1 gene of arabidopsis thaliana highlights the involvement of aba in vegetative development. *J Exp Bot*, 56(418):2071–83, Aug 2005. doi: 10.1093/jxb/eri206.
- Baxter L., Jironkin A., Hickman R., Moore J., Barrington C., Krusche P., Dyer N. P., Buchanan-Wollaston V., Tiskin A., Beynon J., Denby K., and Ott S. Conserved noncoding sequences highlight shared components of regulatory networks in dicotyledonous plants. *Plant Cell*, 24(10):3949–65, Oct 2012. doi: 10.1105/tpc.112.103010.
- Berardini T. Z., Mundodi S., Reiser L., Huala E., Garcia-Hernandez M., Zhang P., Mueller L. A., Yoon J., Doyle A., Lander G., Moseyko N., Yoo D., Xu I., Zoeckler B., Montoya M., Miller N., Weems D., and Rhee S. Y. Functional annotation of the arabidopsis genome using controlled vocabularies. *Plant Physiol*, 135(2):745–55, Jun 2004. doi: 10.1104/pp.104.040071.
- Berrocal-Lobo M., Molina A., and Solano R. Constitutive expression of ethylene-response-factor1 in arabidopsis confers resistance to several necrotrophic fungi. *Plant J*, 29(1):23–32, Jan 2002.
- Bessman M. J., Lehman I. R., Simms E. S., and Kornberg A. Enzymatic synthesis of deoxyribonucleic acid. ii. general properties of the reaction. *J Biol Chem*, 233 (1):171–7, Jul 1958.
- Birnbaum K., Shasha D. E., Wang J. Y., Jung J. W., Lambert G. M., Galbraith D. W., and Benfey P. N. A gene expression map of the arabidopsis root. *Science*, 302(5652):1956–60, Dec 2003. doi: 10.1126/science.1090022.
- Blais A. and Dynlacht B. D. Constructing transcriptional regulatory networks. *Genes Dev*, 19(13):1499–511, Jul 2005. doi: 10.1101/gad.1325605.

Boyle A. P., Davis S., Shulha H. P., Meltzer P., Margulies E. H., Weng Z., Furey T. S., and Crawford G. E. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–22, Jan 2008. doi: 10.1016/j.cell.2007.12.014.

Boyle A. P., Song L., Lee B.-K., London D., Keefe D., Birney E., Iyer V. R., Crawford G. E., and Furey T. S. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res*, 21(3):456–64, Mar 2011. doi: 10.1101/gr.112656.110.

Bray N., Dubchak I., and Pachter L. Avid: A global alignment program. *Genome Res*, 13(1):97–102, Jan 2003. doi: 10.1101/gr.789803.

Breathnach R. and Chambon P. Organization and expression of eucaryotic split genes coding for proteins. *Annu Rev Biochem*, 50:349–83, 1981. doi: 10.1146/annurev.bi.50.070181.002025.

Breeze E., Harrison E., McHattie S., Hughes L., Hickman R., Hill C., Kiddle S., Kim Y.-S., Penfold C. A., Jenkins D., Zhang C., Morris K., Jenner C., Jackson S., Thomas B., Tabrett A., Legaie R., Moore J. D., Wild D. L., Ott S., Rand D., Beynon J., Denby K., Mead A., and Buchanan-Wollaston V. High-resolution temporal profiling of transcripts during arabidopsis leaf senescence reveals a distinct chronology of processes and regulation. *Plant Cell*, 23(3):873–94, Mar 2011. doi: 10.1105/tpc.111.083345.

Brown R. L., Kazan K., McGrath K. C., Maclean D. J., and Manners J. M. A role for the gcc-box in jasmonate-mediated activation of the pdf1.2 gene of arabidopsis. *Plant Physiol*, 132(2):1020–32, Jun 2003. doi: 10.1104/pp.102.017814.

Bryne J. C., Valen E., Tang M.-H. E., Marstrand T., Winther O., da Piedade I., Krogh A., Lenhard B., and Sandelin A. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res*, 36(Database issue):D102–6, Jan 2008. doi: 10.1093/nar.

Bu Q., Jiang H., Li C.-B., Zhai Q., Zhang J., Wu X., Sun J., Xie Q., and Li C. Role of the arabidopsis thaliana nac transcription factors anac019 and anac055 in regulating jasmonic acid-signaled defense responses. *Cell Res*, 18(7):756–67, Jul 2008. doi: 10.1038/cr.2008.53.

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. 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, May 2005. doi: 10.1111/j.1365-313X.2005.02399.x.

Burland T. G. Dnastar's lasergene sequence analysis software. *Methods Mol Biol*, 132:71–91, 2000. ISSN 1064-3745 (Print); 1064-3745 (Linking).

Camehl I. and Oelmüller R. Do ethylene response factors9 and -14 repress pr gene expression in the interaction between piriformospora indica and arabidopsis? *Plant Signal Behav*, 5(8):932–6, Aug 2010.

Cao H., Bowling S. A., Gordon A. S., and Dong X. Characterization of an arabidopsis mutant that is nonresponsive to inducers of systemic acquired resistance. *Plant Cell*, 6(11):1583–1592, Nov 1994. doi: 10.1105/tpc.6.11.1583.

Cavallini B., Faus I., Matthes H., Chipoulet J. M., Winsor B., Egly J. M., and Chambon P. Cloning of the gene encoding the yeast protein btfly, which can substitute for the human tata box-binding factor. *Proc Natl Acad Sci U S A*, 86(24):9803–7, Dec 1989.

Cawley S., Bekiranov S., Ng H. H., Kapranov P., Sekinger E. A., Kampa D., Piccolboni A., Sementchenko V., Cheng J., Williams A. J., Wheeler R., Wong B., Drenkow J., Yamanaka M., Patel S., Brubaker S., Tammana H., Helt G., Struhl K., and Gingeras T. R. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding rnas. *Cell*, 116(4):499–509, Feb 2004.

Chakravarty A., Carlson J. M., Khetani R. S., and Gross R. H. A novel ensemble learning method for de novo computational identification of dna binding sites. *BMC Bioinformatics*, 8:249, 2007. doi: 10.1186/1471-2105-8-249.

Chen D., Xu G., Tang W., Jing Y., Ji Q., Fei Z., and Lin R. Antagonistic basic helix-loop-helix/bzip transcription factors form transcriptional modules that integrate light and reactive oxygen species signaling in arabidopsis. *Plant Cell*, 25(5):1657–73, May 2013. doi: 10.1105/tpc.112.104869.

Chen H., Hwang J. E., Lim C. J., Kim D. Y., Lee S. Y., and Lim C. O. Arabidopsis dreb2c functions as a transcriptional activator of hsfa3 during the heat stress response. *Biochem Biophys Res Commun*, 401(2):238–44, Oct 2010. doi: 10.1016/j.bbrc.2010.09.038.

Chen K. and Rajewsky N. The evolution of gene regulation by transcription factors and micrornas. *Nat Rev Genet*, 8(2):93–103, Feb 2007. doi: 10.1038/nrg1990.

Chen W., Provart N. J., Glazebrook J., Katagiri F., Chang H.-S., Eulgem T., Mauch F., Luan S., Zou G., Whitham S. A., Budworth P. R., Tao Y., Xie Z., Chen X., Lam S., Kreps J. A., Harper J. F., Si-Ammour A., Mauch-Mani B., Heinlein M., Kobayashi K., Hohn T., Dangl J. L., Wang X., and Zhu T. Expression profile matrix of arabidopsis transcription factor genes suggests their putative functions in response to environmental stresses. *Plant Cell*, 14(3):559–74, Mar 2002.

Cheng M.-C., Liao P.-M., Kuo W.-W., and Lin T.-P. The arabidopsis ethylene response factor1 regulates abiotic stress-responsive gene expression by binding to different cis-acting elements in response to different stress signals. *Plant Physiol*, 162(3):1566–1582, Jul 2013. ISSN 1532-2548 (Electronic); 0032-0889 (Linking). doi: 10.1104/pp.113.221911.

Cheong Y. H., Chang H.-S., Gupta R., Wang X., Zhu T., and Luan S. Transcriptional profiling reveals novel interactions between wounding, pathogen, abiotic stress, and hormonal responses in arabidopsis. *Plant Physiol*, 129(2):661–77, Jun 2002. doi: 10.1104/pp.002857.

Chilton M. D., Drummond M. H., Merio D. J., Sciaky D., Montoya A. L., Gordon M. P., and Nester E. W. Stable incorporation of plasmid dna into higher plant cells: the molecular basis of crown gall tumorigenesis. *Cell*, 11(2):263–71, Jun 1977.

Chini A., Fonseca S., Fernandez G., Adie B., Chico J. M., Lorenzo O., Garcia-Casado G., Lopez-Vidriero I., Lozano F. M., Ponce M. R., Micol J. L., and Solano R. The jaz family of repressors is the missing link in jasmonate signalling. *Nature*, 448 (7154):666–671, 08 2007. URL <http://dx.doi.org/10.1038/nature06006>.

Choi H., Hong J., Ha J., Kang J., and Kim S. Y. Abfs, a family of aba-responsive element binding factors. *J Biol Chem*, 275(3):1723–30, Jan 2000.

Chow H.-K., Xu J., Shahravan S. H., De Jong A. T., Chen G., and Shin J. A. Hybrids of the bhlh and bzip protein motifs display different dna-binding activities in vivo vs. in vitro. *PLoS One*, 3(10):e3514, 2008. doi: 10.1371/journal.pone.0003514.

Chung B. Y. W., Simons C., Firth A. E., Brown C. M., and Hellens R. P. Effect of 5'utr introns on gene expression in arabidopsis thaliana. *BMC Genomics*, 7:120, 2006. doi: 10.1186/1471-2164-7-120.

- Chung H. S., Koo A. J. K., Gao X., Jayanty S., Thines B., Jones A. D., and Howe G. A. Regulation and function of arabidopsis jasmonate zim-domain genes in response to wounding and herbivory. *Plant Physiol*, 146(3):952–64, Mar 2008. doi: 10.1104/pp.107.115691.
- Ciolkowski I., Wanke D., Birkenbihl R. P., and Somssich I. E. Studies on dna-binding selectivity of wrky transcription factors lend structural clues into wrky-domain function. *Plant Mol Biol*, 68(1-2):81–92, Sep 2008. doi: 10.1007/s11103-008-9353-1.
- Clarke J. D., Aarts N., Feys B. J., Dong X., and Parker J. E. Constitutive disease resistance requires eds1 in the arabidopsis mutants cpr1 and cpr6 and is partially eds1-dependent in cpr5. *Plant J*, 26(4):409–20, May 2001.
- Clough S. J. and Bent A. F. Floral dip: a simplified method for agrobacterium-mediated transformation of arabidopsis thaliana. *Plant J*, 16(6):735–43, Dec 1998.
- Colinas J., Birnbaum K., and Benfey P. N. Using cauliflower to find conserved non-coding regions in arabidopsis. *Plant Physiol*, 129(2):451–4, Jun 2002. doi: 10.1104/pp.002501.
- Crick F. Central dogma of molecular biology. *Nature*, 227(5258):561–3, Aug 1970.
- Crombach A. and Hogeweg P. Evolution of evolvability in gene regulatory networks. *PLoS Comput Biol*, 4(7):e1000112, 2008. doi: 10.1371/journal.pcbi.1000112.
- Dangl J. L. and Jones J. D. Plant pathogens and integrated defence responses to infection. *Nature*, 411(6839):826–33, Jun 2001. doi: 10.1038/35081161.
- Davidson E. H. and Erwin D. H. Gene regulatory networks and the evolution of animal body plans. *Science*, 311(5762):796–800, Feb 2006. doi: 10.1126/science.1113832.
- Davuluri R. V., Sun H., Palaniswamy S. K., Matthews N., Molina C., Kurtz M., and Grotewold E. Agris: Arabidopsis gene regulatory information server, an information resource of arabidopsis cis-regulatory elements and transcription factors. *BMC Bioinformatics*, 4:25, Jun 2003. doi: 10.1186/1471-2105-4-25.
- de Torres-Zabala M., Truman W., Bennett M. H., Lafforgue G., Mansfield J. W., Rodriguez Egea P., Bögre L., and Grant M. *Pseudomonas syringae* pv. tomato hijacks the arabidopsis abscisic acid signalling pathway to cause disease. *EMBO J*, 26(5):1434–43, Mar 2007. doi: 10.1038/sj.emboj.7601575.

Deckmann K., Rörsch F., Geisslinger G., and Grösch S. Identification of dna-protein complexes using an improved, combined western blotting-electrophoretic mobility shift assay (wemsa) with a fluorescence imaging system. *Mol Biosyst*, 8(5):1389–95, Apr 2012. doi: 10.1039/c2mb05500g.

Denby K. J., Kumar P., and Kliebenstein D. J. Identification of botrytis cinerea susceptibility loci in arabidopsis thaliana. *Plant J*, 38(3):473–86, May 2004. doi: 10.1111/j.0960-7412.2004.02059.x.

Deplancke B., Dupuy D., Vidal M., and Walhout A. J. M. A gateway-compatible yeast one-hybrid system. *Genome Res*, 14(10B):2093–101, Oct 2004. doi: 10.1101/gr.2445504.

Deshayes A., Herrera-Estrella L., and Caboche M. Liposome-mediated transformation of tobacco mesophyll protoplasts by an escherichia coli plasmid. *EMBO J*, 4 (11):2731–7, Nov 1985.

Dong X. Npr1, all things considered. *Curr Opin Plant Biol*, 7(5):547–52, Oct 2004. doi: 10.1016/j.pbi.2004.07.005.

Dreier B., Beerli R. R., Segal D. J., Flippin J. D., and Barbas C. F., 3rd. Development of zinc finger domains for recognition of the 5'-ann-3' family of dna sequences and their use in the construction of artificial transcription factors. *J Biol Chem*, 276(31):29466–78, Aug 2001. doi: 10.1074/jbc.M102604200.

Duret L. and Bucher P. Searching for regulatory elements in human noncoding sequences. *Curr Opin Struct Biol*, 7(3):399–406, Jun 1997.

Economou A. D., Ohazama A., Porntaveetus T., Sharpe P. T., Kondo S., Basson M. A., Gritli-Linde A., Cobourne M. T., and Green J. B. A. Periodic stripe formation by a turing mechanism operating at growth zones in the mammalian palate. *Nat Genet*, 44(3):348–51, Mar 2012. doi: 10.1038/ng.1090.

Edgar R., Domrachev M., and Lash A. E. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Res*, 30(1): 207–10, Jan 2002.

Eisenmann D. M., Dollard C., and Winston F. Spt15, the gene encoding the yeast tata binding factor tfid, is required for normal transcription initiation in vivo. *Cell*, 58(6):1183–91, Sep 1989.

- el Deiry W. S., Kern S. E., Pietenpol J. A., Kinzler K. W., and Vogelstein B. Definition of a consensus binding site for p53. *Nat Genet*, 1(1):45–9, Apr 1992. doi: 10.1038/ng0492-45.
- Elad Y., Williamson B., Tudzynski P., and Delen N., editors. *Botrytis: Biology, Pathology and Control*. Springer Netherlands, 2007a.
- Elad Y., Williamson B., Tudzynski P., Delen N., Droby S., and Licherter A. *Post-Harvest Botrytis Infection: Etiology, Development and Management*, pages 349–367. Springer Netherlands, 2007b. ISBN 978-1-4020-2624-9. doi: 10.1007/978-1-4020-2626-3{\\_}19. URL [http://dx.doi.org/10.1007/978-1-4020-2626-3\\_19](http://dx.doi.org/10.1007/978-1-4020-2626-3_19).
- Elgar G. and Vavouri T. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends Genet*, 24(7):344–52, Jul 2008. doi: 10.1016/j.tig.2008.04.005.
- ENCODE Project Consortium , Bernstein B. E., Birney E., Dunham I., Green E. D., Gunter C., and Snyder M. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012. doi: 10.1038/nature11247.
- Espinosa-Soto C., Padilla-Longoria P., and Alvarez-Buylla E. R. A gene regulatory network model for cell-fate determination during arabidopsis thaliana flower development that is robust and recovers experimental gene expression profiles. *Plant Cell*, 16(11):2923–39, Nov 2004. doi: 10.1105/tpc.104.021725.
- Eulgem T., Rushton P. J., Robatzek S., and Somssich I. E. The wrky superfamily of plant transcription factors. *Trends Plant Sci*, 5(5):199–206, May 2000.
- Eun S. O. and Lee Y. Actin filaments of guard cells are reorganized in response to light and abscisic acid. *Plant Physiol*, 115(4):1491–8, Dec 1997.
- Felix G., Duran J. D., Volko S., and Boller T. Plants have a sensitive perception system for the most conserved domain of bacterial flagellin. *Plant J*, 18(3):265–76, May 1999.
- Ferrari S., Galletti R., Denoux C., De Lorenzo G., Ausubel F. M., and Dewdney J. Resistance to botrytis cinerea induced in arabidopsis by elicitors is independent of salicylic acid, ethylene, or jasmonate signaling but requires phytoalexin deficient3. *Plant Physiol*, 144(1):367–79, May 2007. doi: 10.1104/pp.107.095596.
- Ferré-D’Amaré A. R., Pognonec P., Roeder R. G., and Burley S. K. Structure and function of the b/hlh/z domain of usf. *EMBO J*, 13(1):180–9, Jan 1994.

Fields S. and Song O. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–6, Jul 1989. doi: 10.1038/340245a0.

Filichkin S. A., Priest H. D., Givan S. A., Shen R., Bryant D. W., Fox S. E., Wong W.-K., and Mockler T. C. Genome-wide mapping of alternative splicing in arabidopsis thaliana. *Genome Res*, 20(1):45–58, Jan 2010. doi: 10.1101/gr.093302.109.

Finkelstein R. R., Wang M. L., Lynch T. J., Rao S., and Goodman H. M. The arabidopsis abscisic acid response locus abi4 encodes an apetala 2 domain protein. *Plant Cell*, 10(6):1043–54, Jun 1998.

Frazer K. A., Elnitski L., Church D. M., Dubchak I., and Hardison R. C. Cross-species sequence comparisons: a review of methods and available resources. *Genome Res*, 13(1):1–12, Jan 2003. doi: 10.1101/gr.222003.

Freeling M., Rapaka L., Lyons E., Pedersen B., and Thomas B. C. G-boxes, bigfoot genes, and environmental response: characterization of intragenomic conserved noncoding sequences in arabidopsis. *Plant Cell*, 19(5):1441–57, May 2007. doi: 10.1105/tpc.107.050419.

Fried M. and Crothers D. M. Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Res*, 9(23):6505–25, Dec 1981.

Fujimoto S. Y., Ohta M., Usui A., Shinshi H., and Ohme-Takagi M. Arabidopsis ethylene-responsive element binding factors act as transcriptional activators or repressors of gcc box-mediated gene expression. *Plant Cell*, 12(3):393–404, Mar 2000.

Galas D. J. and Schmitz A. DNase footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res*, 5(9):3157–70, Sep 1978.

Galon Y., Nave R., Boyce J. M., Nachmias D., Knight M. R., and Fromm H. Calmodulin-binding transcription activator (camta) 3 mediates biotic defense responses in arabidopsis. *FEBS Lett*, 582(6):943–8, Mar 2008. doi: 10.1016/j.febslet.2008.02.037.

Gannon F., O'Hare K., Perrin F., LePennec J. P., Benoist C., Cochet M., Breathnach R., Royal A., Garapin A., Cami B., and Chambon P. Organisation and sequences at the 5' end of a cloned complete ovalbumin gene. *Nature*, 278(5703):428–34, Mar 1979.

Gardiner E. J., Hunter C. A., Packer M. J., Palmer D. S., and Willett P. Sequence-dependent dna structure: a database of octamer structural parameters. *J Mol Biol*, 332(5):1025–35, Oct 2003.

Garner M. M. and Revzin A. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic Acids Res*, 9(13):3047–60, Jul 1981.

Gentleman R. C., Carey V. J., Bates D. M., Bolstad B., Dettling M., Dudoit S., Ellis B., Gautier L., Ge Y., Gentry J., Hornik K., Hothorn T., Huber W., Iacus S., Irizarry R., Leisch F., Li C., Maechler M., Rossini A. J., Sawitzki G., Smith C., Smyth G., Tierney L., Yang J. Y. H., and Zhang J. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80, 2004. doi: 10.1186/gb-2004-5-10-r80.

Gepstein S. and Thimann K. V. Changes in the abscisic acid content of oat leaves during senescence. *Proc Natl Acad Sci U S A*, 77(4):2050–3, Apr 1980.

Gifford M. L., Dean A., Gutierrez R. A., Coruzzi G. M., and Birnbaum K. D. Cell-specific nitrogen responses mediate developmental plasticity. *Proc Natl Acad Sci U S A*, 105(2):803–8, Jan 2008. doi: 10.1073/pnas.0709559105.

Gilmour D. S. and Lis J. T. Detecting protein-dna interactions in vivo: distribution of rna polymerase on specific bacterial genes. *Proc Natl Acad Sci U S A*, 81(14):4275–9, Jul 1984.

Gilmour S. J., Zarka D. G., Stockinger E. J., Salazar M. P., Houghton J. M., and Thomashow M. F. Low temperature regulation of the arabidopsis cbf family of ap2 transcriptional activators as an early step in cold-induced cor gene expression. *Plant J*, 16(4):433–42, Nov 1998.

Godoy M., Franco-Zorrilla J. M., Pérez-Pérez J., Oliveros J. C., Lorenzo O., and Solano R. Improved protein-binding microarrays for the identification of dna-binding specificities of transcription factors. *Plant J*, 66(4):700–11, May 2011. doi: 10.1111/j.1365-313X.2011.04519.x.

Gómez-Gómez L. and Boller T. Fls2: an lrr receptor-like kinase involved in the perception of the bacterial elicitor flagellin in arabidopsis. *Mol Cell*, 5(6):1003–11, Jun 2000.

- Goodspeed D., Chehab E. W., Min-Venditti A., Braam J., and Covington M. F. Arabidopsis synchronizes jasmonate-mediated defense with insect circadian behavior. *Proceedings of the National Academy of Sciences*, 109(12):4674–4677, 03 2012. URL <http://www.pnas.org/content/109/12/4674>.
- Gorin A. A., Zhurkin V. B., and Olson W. K. B-dna twisting correlates with base-pair morphology. *J Mol Biol*, 247(1):34–48, Mar 1995.
- Govrin E. M. and Levine A. The hypersensitive response facilitates plant infection by the necrotrophic pathogen *botrytis cinerea*. *Curr Biol*, 10(13):751–7, Jun 2000.
- Greenberg J. T. Programmed cell death in plant-pathogen interactions. *Annu Rev Plant Physiol Plant Mol Biol*, 48:525–545, Jun 1997. doi: 10.1146/annurev.arplant.48.1.525.
- Gross D. S. and Garrard W. T. Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem*, 57:159–97, 1988. doi: 10.1146/annurev.bi.57.070188.001111.
- Gu Y. Q., Yang C., Thara V. K., Zhou J., and Martin G. B. Pt14 is induced by ethylene and salicylic acid, and its product is phosphorylated by the pto kinase. *Plant Cell*, 12(5):771–86, May 2000.
- Guelzim N., Bottani S., Bourgine P., and Képès F. Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet*, 31(1):60–3, May 2002. doi: 10.1038/ng873.
- Guerrero F. and Mullet J. E. Increased abscisic acid biosynthesis during plant dehydration requires transcription. *Plant Physiol*, 80(2):588–91, Feb 1986.
- Guo A., He K., Liu D., Bai S., Gu X., Wei L., and Luo J. Datf: a database of arabidopsis transcription factors. *Bioinformatics*, 21(10):2568–9, May 2005. doi: 10.1093/bioinformatics/bti334.
- Guo H. and Moose S. P. Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution. *Plant Cell*, 15(5):1143–58, May 2003.
- Haberer G., Hindemitt T., Meyers B. C., and Mayer K. F. X. Transcriptional similarities, dissimilarities, and conservation of cis-elements in duplicated genes of arabidopsis. *Plant Physiol*, 136(2):3009–22, Oct 2004. doi: 10.1104/pp.104.046466.

- Hahn S., Buratowski S., Sharp P. A., and Guarente L. Isolation of the gene encoding the yeast tata binding protein tfiid: a gene identical to the spt15 suppressor of ty element insertions. *Cell*, 58(6):1173–81, Sep 1989.
- Hahn S. Structure and mechanism of the rna polymerase ii transcription machinery. *Nat Struct Mol Biol*, 11(5):394–403, May 2004. doi: 10.1038/nsmb763.
- Hancock J. G. and Lorbeer J. W. Pathogenesis of botrytis cinerea, b. squamosa and b. allii on onion leaves. *Phytopathology*, 53:669–673, 1963.
- Hawker L. E. and Hendy R. J. An eelectron-microscope study of germination of conidia of botrytis cinerea. *J Gen Microbiol*, 33:43–6, Oct 1963.
- Hedges S. B., Dudley J., and Kumar S. Timetree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, 22(23):2971–2, Dec 2006. doi: 10.1093/bioinformatics/btl505.
- Heinz S., Benner C., Spann N., Bertolino E., Lin Y. C., Laslo P., Cheng J. X., Murre C., Singh H., and Glass C. K. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Mol Cell*, 38(4):576–89, May 2010. doi: 10.1016/j.molcel.2010.05.004.
- Hellinger E. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 136:210–271, 1909. URL <http://eudml.org/doc/149313>.
- Herms D. and Mattson W. The dilemma of plants: To grow or defend. *The Quarterly Review of Biology*, 67(3):283–335, Sep 1992.
- Herr A. J., Jensen M. B., Dalmay T., and Baulcombe D. C. Rna polymerase iv directs silencing of endogenous dna. *Science*, 308(5718):118–20, Apr 2005. doi: 10.1126/science.1106910.
- Hesselberth J. R., Chen X., Zhang Z., Sabo P. J., Sandstrom R., Reynolds A. P., Thurman R. E., Neph S., Kuehn M. S., Noble W. S., Fields S., and Stamatoyannopoulos J. A. Global mapping of protein-dna interactions in vivo by digital genomic footprinting. *Nat Methods*, 6(4):283–9, Apr 2009. doi: 10.1038/nmeth.1313.
- Heyndrickx K. S. and Vandepoele K. Systematic identification of functional plant modules through the integration of complementary data sources. *Plant Physiol*, 159(3):884–901, Jul 2012. doi: 10.1104/pp.112.196725.

Higo K., Ugawa Y., Iwamoto M., and Korenaga T. Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res*, 27(1):297–300, Jan 1999.

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., Lundberg 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., Vercruyse S., Vuylsteke M., Weingartner M., Weisbeek P. J., Wirta V., Wittink F. R. A., Zabeau M., and Small I. Versatile gene-specific sequence tags for arabidopsis functional genomics: transcript profiling and reverse genetics applications. *Genome Res*, 14(10B):2176–89, Oct 2004. doi: 10.1101/gr.2544504.

Hoffman A., Sinn E., Yamamoto T., Wang J., Roy A., Horikoshi M., and Roeder R. G. Highly conserved core domain and unique n terminus with presumptive regulatory motifs in a human tata factor (tfiid). *Nature*, 346(6282):387–90, Jul 1990. doi: 10.1038/346387a0.

Horikoshi M., Wang C. K., Fujii H., Cromlish J. A., Weil P. A., and Roeder R. G. Cloning and structure of a yeast gene encoding a general transcription initiation factor tfiid that binds to the tata box. *Nature*, 341(6240):299–303, Sep 1989. doi: 10.1038/341299a0.

Hua J., Grisafi P., Cheng S. H., and Fink G. R. Plant growth homeostasis is controlled by the arabidopsis bon1 and bap1 genes. *Genes Dev*, 15(17):2263–72, Sep 2001. doi: 10.1101/gad.918101.

Huang W., Pérez-García P., Pokhilko A., Millar A. J., Antoshechkin I., Riechmann J. L., and Mas P. Mapping the core of the arabidopsis circadian clock defines the network structure of the oscillator. *Science*, 336(6077):75–9, Apr 2012. doi: 10.1126/science.1219075.

Huq E. and Quail P. H. Pif4, a phytochrome-interacting blhh factor, functions as a negative regulator of phytochrome b signaling in arabidopsis. *EMBO J*, 21(10): 2441–50, May 2002. doi: 10.1093/emboj/21.10.2441.

Iida K., Seki M., Sakurai T., Satou M., Akiyama K., Toyoda T., Konagaya A., and Shinozaki K. Rartf: database and tools for complete sets of arabidopsis tran-

scription factors. *DNA Res*, 12(4):247–256, 2005. ISSN 1756-1663 (Electronic); 1340-2838 (Linking). doi: 10.1093/dnares/dsi011.

Inada D. C., Bashir A., Lee C., Thomas B. C., Ko C., Goff S. A., and Freeling M. Conserved noncoding sequences in the grasses. *Genome Res*, 13(9):2030–41, Sep 2003. doi: 10.1101/gr.1280703.

Initiative T. A. G. Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*, 408(6814):796–815, 12 2000. URL <http://dx.doi.org/10.1038/35048692>.

International Food Policy Research Institute . Global hunger index 2011 by severity — international food policy research institute (ifpri), 2011. URL <http://www.ifpri.org/publication/global-hunger-index-2011-severity>.

Jackson R. J. and Standart N. How do micrornas regulate gene expression? *Sci STKE*, 2007(367):re1, Jan 2007. doi: 10.1126/stke.3672007re1.

Jaenisch R. and Mintz B. Simian virus 40 dna sequences in dna of healthy adult mice derived from preimplantation blastocysts injected with viral dna. *Proc Natl Acad Sci U S A*, 71(4):1250–4, Apr 1974.

Jakoby M., Weisshaar B., Dröge-Laser W., Vicente-Carbajosa J., Tiedemann J., Kroj T., Parcy F., and bZIP Research Group . bzip transcription factors in arabidopsis. *Trends Plant Sci*, 7(3):106–11, Mar 2002.

Jambunathan N., Siani J. M., and McNellis T. W. A humidity-sensitive arabidopsis copine mutant exhibits precocious cell death and increased disease resistance. *Plant Cell*, 13(10):2225–40, Oct 2001.

Jarvis W. R. *Botryotinia and Botrytis species : taxonomy, physiology, and pathogenicity : a guide to the literature / W. R. Jarvis*. Research Branch, Canada Dept. of Agriculture: obtainable from Information Division, Canada Dept. of Agriculture, Ottawa, 1977.

Jiang C. and Pugh B. F. Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet*, 10(3):161–72, Mar 2009. doi: 10.1038/nrg2522.

Jin H. and Martin C. Multifunctionality and diversity within the plant myb-gene family. *Plant Mol Biol*, 41(5):577–85, Nov 1999.

Jofuku K. D., den Boer B. G., Van Montagu M., and Okamuro J. K. Control of arabidopsis flower and seed development by the homeotic gene *apetala2*. *Plant Cell*, 6(9):1211–25, Sep 1994.

Johnson C., Boden E., and Arias J. Salicylic acid and npr1 induce the recruitment of trans-activating tga factors to a defense gene promoter in arabidopsis. *Plant Cell*, 15(8):1846–58, Aug 2003.

Johnson D. S., Mortazavi A., Myers R. M., and Wold B. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–502, Jun 2007. doi: 10.1126/science.1141319.

Kagaya Y., Ohmiya K., and Hattori T. Rav1, a novel dna-binding protein, binds to bipartite recognition sequence through two distinct dna-binding domains uniquely found in higher plants. *Nucleic Acids Res*, 27(2):470–8, Jan 1999.

Kao C. C., Lieberman P. M., Schmidt M. C., Zhou Q., Pei R., and Berk A. J. Cloning of a transcriptionally active human tata binding factor. *Science*, 248 (4963):1646–50, Jun 1990.

Kaplan N., Moore I. K., Fondufe-Mittendorf Y., Gossett A. J., Tillo D., Field Y., LeProust E. M., Hughes T. R., Lieb J. D., Widom J., and Segal E. The dna-encoded nucleosome organization of a eukaryotic genome. *Nature*, 458(7236): 362–6, Mar 2009. doi: 10.1038/nature07667.

Kaplinsky N. J., Braun D. M., Penterman J., Goff S. A., and Freeling M. Utility and distribution of conserved noncoding sequences in the grasses. *Proc Natl Acad Sci U S A*, 99(9):6147–51, Apr 2002. doi: 10.1073/pnas.052139599.

Katagiri F. A global view of defense gene expression regulation—a highly interconnected signaling network. *Curr Opin Plant Biol*, 7(5):506–11, Oct 2004. doi: 10.1016/j.pbi.2004.07.013.

Kaufmann K., Muiño J. M., Jauregui R., Airoldi C. A., Smaczniak C., Krajewski P., and Angenent G. C. Target genes of the mads transcription factor *sepallata3*: integration of developmental and hormonal pathways in the arabidopsis flower. *PLoS Biol*, 7(4):e1000090, Apr 2009. doi: 10.1371/journal.pbio.1000090.

Kaufmann K., Muiño J. M., Østerås M., Farinelli L., Krajewski P., and Angenent G. C. Chromatin immunoprecipitation (chip) of plant transcription factors followed by sequencing (chip-seq) or hybridization to whole genome arrays (chip-chip). *Nat Protoc*, 5(3):457–72, Mar 2010. doi: 10.1038/nprot.2009.244.

- Kesarwani M., Yoo J., and Dong X. Genetic interactions of tga transcription factors in the regulation of pathogenesis-related genes and disease resistance in arabidopsis. *Plant Physiol*, 144(1):336–46, May 2007. doi: 10.1104/pp.106.095299.
- Kidokoro S., Maruyama K., Nakashima K., Imura Y., Narusaka Y., Shinwari Z. K., Osakabe Y., Fujita Y., Mizoi J., Shinozaki K., and Yamaguchi-Shinozaki K. The phytochrome-interacting factor pif7 negatively regulates dreb1 expression under circadian control in arabidopsis. *Plant Physiol*, 151(4):2046–2057, Dec 2009. ISSN 1532-2548 (Electronic); 0032-0889 (Linking). doi: 10.1104/pp.109.147033.
- Kikkert J. The biolistic®pds-1000/he device. 33(3):221–226, 1993. doi: 10.1007/BF02319005. URL <http://dx.doi.org/10.1007/BF02319005>.
- Kim H.-S., Desveaux D., Singer A. U., Patel P., Sondek J., and Dangl J. L. The pseudomonas syringae effector avrrpt2 cleaves its c-terminally acylated target, rin4, from arabidopsis membranes to block rpm1 activation. *Proc Natl Acad Sci U S A*, 102(18):6496–501, May 2005a. doi: 10.1073/pnas.0500792102.
- Kim J.-M., To T. K., Ishida J., Morosawa T., Kawashima M., Matsui A., Toyoda T., Kimura H., Shinozaki K., and Seki M. Alterations of lysine modifications on the histone h3 n-tail under drought stress conditions in arabidopsis thaliana. *Plant Cell Physiol*, 49(10):1580–8, Oct 2008. doi: 10.1093/pcp/pcn133.
- Kim J.-M., To T. K., Ishida J., Matsui A., Kimura H., and Seki M. Transition of chromatin status during the process of recovery from drought stress in arabidopsis thaliana. *Plant Cell Physiol*, 53(5):847–56, May 2012. doi: 10.1093/pcp/pcs053.
- Kim M. G., da Cunha L., McFall A. J., Belkhadir Y., DebRoy S., Dangl J. L., and Mackey D. Two pseudomonas syringae type iii effectors inhibit rin4-regulated basal defense in arabidopsis. *Cell*, 121(5):749–59, Jun 2005b. doi: 10.1016/j.cell.2005.03.025.
- Kim T. H., Barrera L. O., Zheng M., Qu C., Singer M. A., Richmond T. A., Wu Y., Green R. D., and Ren B. A high-resolution map of active promoters in the human genome. *Nature*, 436(7052):876–80, Aug 2005c. doi: 10.1038/nature03877.
- Klug A., Jack A., Viswamitra M. A., Kennard O., Shakked Z., and Steitz T. A. A hypothesis on a specific sequence-dependent conformation of dna and its relation to the binding of the lac-repressor protein. *J Mol Biol*, 131(4):669–80, Jul 1979.
- Koch E. and Slusarenko A. Arabidopsis is susceptible to infection by a downy mildew fungus. *Plant Cell*, 2(5):437–45, May 1990. doi: 10.1105/tpc.2.5.437.

Koohy H., Dyer N. P., Reid J. E., Koentges G., and Ott S. An alignment-free model for comparison of regulatory sequences. *Bioinformatics*, 26(19):2391–7, Oct 2010. doi: 10.1093/bioinformatics/btq453.

Koyama T., Furutani M., Tasaka M., and Ohme-Takagi M. Tcp transcription factors control the morphology of shoot lateral organs via negative regulation of the expression of boundary-specific genes in arabidopsis. *Plant Cell*, 19(2):473–84, Feb 2007. doi: 10.1105/tpc.106.044792.

Koyama T., Mitsuda N., Seki M., Shinozaki K., and Ohme-Takagi M. Tcp transcription factors regulate the activities of asymmetric leaves1 and mir164, as well as the auxin response, during differentiation of leaves in arabidopsis. *Plant Cell*, 22(11):3574–88, Nov 2010. doi: 10.1105/tpc.110.075598.

Kreps J. A., Wu Y., Chang H.-S., Zhu T., Wang X., and Harper J. F. Transcriptome changes for arabidopsis in response to salt, osmotic, and cold stress. *Plant Physiol*, 130(4):2129–41, Dec 2002. doi: 10.1104/pp.008532.

Krusche P. and Tiskin A. Computing alignment plots efficiently. In *Advances in Parallel Computing*, volume 19, pages 158–165. IOS Press, 2010. doi: <http://dx.doi.org/10.3233/978-1-60750-530-3-158>.

Kusano M., Tohge T., Fukushima A., Kobayashi M., Hayashi N., Otsuki H., Kondou Y., Goto H., Kawashima M., Matsuda F., Niida R., Matsui M., Saito K., and Fernie A. R. Metabolomics reveals comprehensive reprogramming involving two independent metabolic responses of arabidopsis to uv-b light. *Plant J*, 67(2):354–69, Jul 2011. doi: 10.1111/j.1365-313X.2011.04599.x.

Laluk K., Luo H., Chai M., Dhawan R., Lai Z., and Mengiste T. Biochemical and genetic requirements for function of the immune response regulator botrytis-induced kinase1 in plant growth, ethylene signaling, and pamp-triggered immunity in arabidopsis. *Plant Cell*, 23(8):2831–49, Aug 2011. doi: 10.1105/tpc.111.087122.

Lee M. M. and Schiefelbein J. Developmentally distinct myb genes encode functionally equivalent proteins in arabidopsis. *Development*, 128(9):1539–46, May 2001.

Lee S.-j., Park J. H., Lee M. H., Yu J.-h., and Kim S. Y. Isolation and functional characterization of ce1 binding proteins. *BMC Plant Biol*, 10:277, 2010. doi: 10.1186/1471-2229-10-277.

- Lehman I. R., Bessman M. J., Simms E. S., and Kornberg A. Enzymatic synthesis of deoxyribonucleic acid. i. preparation of substrates and partial purification of an enzyme from escherichia coli. *J Biol Chem*, 233(1):163–70, Jul 1958.
- Lehmann J., Atzorn R., Brückner C., Reinbothe S., Leopold J., Wasternack C., and Parthier B. Accumulation of jasmonate, abscisic acid, specific transcripts and proteins in osmotically stressed barley leaf segments. 197(1):156–162, 1995. doi: 10.1007/BF00239952. URL <http://dx.doi.org/10.1007/BF00239952>.
- Leivar P., Monte E., Al-Sady B., Carle C., Storer A., Alonso J. M., Ecker J. R., and Quail P. H. 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):337–352, Feb 2008. ISSN 1040-4651 (Print); 1040-4651 (Linking). doi: 10.1105/tpc.107.052142.
- Leslie A. G., Arnott S., Chandrasekaran R., and Ratliff R. L. Polymorphism of dna double helices. *J Mol Biol*, 143(1):49–72, Oct 1980.
- Letovsky J. and Dynan W. S. Measurement of the binding of transcription factor sp1 to a single gc box recognition sequence. *Nucleic Acids Res*, 17(7):2639–53, Apr 1989.
- Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., and 1000 Genome Project Data Processing Subgroup . The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–9, Aug 2009. doi: 10.1093/bioinformatics/btp352.
- Li J., Brader G., and Palva E. T. The wrky70 transcription factor: a node of convergence for jasmonate-mediated and salicylate-mediated signals in plant defense. *Plant Cell*, 16(2):319–31, Feb 2004. doi: 10.1105/tpc.016980.
- Liljegren S. J., Gustafson-Brown C., Pinyopich A., Ditta G. S., and Yanofsky M. F. Interactions among APETALA1, LEAFY, and TERMINAL FLOWER1 specify meristem fate. *Plant Cell*, 11(6):1007–18, Jun 1999.
- Liu H., Sachidanandam R., and Stein L. Comparative genomics between rice and arabidopsis shows scant collinearity in gene order. *Genome Res*, 11(12):2020–6, Dec 2001. doi: 10.1101/gr.194501.
- Liu Q., Kasuga M., Sakuma Y., Abe H., Miura S., Yamaguchi-Shinozaki K., and Shinozaki K. Two transcription factors, dreb1 and dreb2, with an erebp/ap2 dna binding domain separate two cellular signal transduction pathways in drought-

and low-temperature-responsive gene expression, respectively, in arabidopsis. *Plant Cell*, 10(8):1391–406, Aug 1998.

Liu W. X., Liu H. L., Chai Z. J., Xu X. P., Song Y. R., and Qu L. Q. Evaluation of seed storage-protein gene 5' untranslated regions in enhancing gene expression in transgenic rice seed. *Theor Appl Genet*, 121(7):1267–74, Nov 2010. doi: 10.1007/s00122-010-1386-6.

Loake G. and Grant M. Salicylic acid in plant defence—the players and protagonists. *Curr Opin Plant Biol*, 10(5):466–72, Oct 2007. doi: 10.1016/j.pbi.2007.08.008.

Locke J. C. W., Kozma-Bognár L., Gould P. D., Fehér B., Kevei E., Nagy F., Turner M. S., Hall A., and Millar A. J. Experimental validation of a predicted feedback loop in the multi-oscillator clock of arabidopsis thaliana. *Mol Syst Biol*, 2:59, 2006. doi: 10.1038/msb4100102.

Loots G. G., Locksley R. M., Blankspoor C. M., Wang Z. E., Miller W., Rubin E. M., and Frazer K. A. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*, 288(5463):136–40, Apr 2000.

Lorenzo O., Piqueras R., Sánchez-Serrano J. J., and Solano R. Ethylene response factor1 integrates signals from ethylene and jasmonate pathways in plant defense. *Plant Cell*, 15(1):165–78, Jan 2003.

Lorenzo O., Chico J. M., Sánchez-Serrano J. J., and Solano R. Jasmonate-insensitive1 encodes a myc transcription factor essential to discriminate between different jasmonate-regulated defense responses in arabidopsis. *Plant Cell*, 16(7):1938–50, Jul 2004. doi: 10.1105/tpc.022319.

Lyons E., Pedersen B., Kane J., Alam M., Ming R., Tang H., Wang X., Bowers J., Paterson A., Lisch D., and Freeling M. Finding and comparing syntenic regions among arabidopsis and the outgroups papaya, poplar, and grape: Coge with rosids. *Plant Physiol*, 148(4):1772–81, Dec 2008. doi: 10.1104/pp.108.124867.

Lysenko E. A. and Kuznetsov V. V. [plastid rna polymerases]. *Mol Biol (Mosk)*, 39 (5):762–75, 2005.

Ma P. C., Rould M. A., Weintraub H., and Pabo C. O. Crystal structure of myod bhlh domain-dna complex: perspectives on dna recognition and implications for transcriptional activation. *Cell*, 77(3):451–9, May 1994.

- Mackey D., Holt B. F., 3rd, Wiig A., and Dangl J. L. Rin4 interacts with *pseudomonas syringae* type iii effector molecules and is required for rpm1-mediated resistance in arabidopsis. *Cell*, 108(6):743–54, Mar 2002.
- Maere S., Heymans K., and Kuiper M. Bingo: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16):3448–9, Aug 2005. doi: 10.1093/bioinformatics/bti551.
- Malys N. and McCarthy J. E. G. Translation initiation: variations in the mechanism can be anticipated. *Cell Mol Life Sci*, 68(6):991–1003, Mar 2011. doi: 10.1007/s00018-010-0588-z.
- Mandel M. A., Gustafson-Brown C., Savidge B., and Yanofsky M. F. Molecular characterization of the Arabidopsis floral homeotic gene APETALA1. *Nature*, 360(6401):273–7, Nov 1992. doi: 10.1038/360273a0.
- Manners J. M., Penninckx I. A., Vermaere K., Kazan K., Brown R. L., Morgan A., Maclean D. J., Curtis M. D., Cammue B. P., and Broekaert W. F. The promoter of the plant defensin gene pdf1.2 from arabidopsis is systemically activated by fungal pathogens and responds to methyl jasmonate but not to salicylic acid. *Plant Mol Biol*, 38(6):1071–80, Dec 1998.
- Mao G., Meng X., Liu Y., Zheng Z., Chen Z., and Zhang S. Phosphorylation of a wrky transcription factor by two pathogen-responsive mapks drives phytoalexin biosynthesis in arabidopsis. *Plant Cell*, 23(4):1639–53, Apr 2011. doi: 10.1105/tpc.111.084996.
- Marchler-Bauer A., Lu S., Anderson J. B., Chitsaz F., Derbyshire M. K., DeWeese-Scott C., Fong J. H., Geer L. Y., Geer R. C., Gonzales N. R., Gwadz M., Hurwitz D. I., Jackson J. D., Ke Z., Lanczycki C. J., Lu F., Marchler G. H., Mullokan-dov M., Omelchenko M. V., Robertson C. L., Song J. S., Thanki N., Yamashita R. A., Zhang D., Zhang N., Zheng C., and Bryant S. H. Cdd: a conserved domain database for the functional annotation of proteins. *Nucleic Acids Res*, 39(Database issue):D225–9, Jan 2011. doi: 10.1093/nar/gkq1189.
- Marquez Y., Brown J. W. S., Simpson C., Barta A., and Kalyna M. Transcriptome survey reveals increased complexity of the alternative splicing landscape in arabidopsis. *Genome Res*, 22(6):1184–95, Jun 2012. doi: 10.1101/gr.134106.111.
- Martínez-García J. F., Moyano E., Alcocer M. J., and Martin C. Two bzip proteins from antirrhinum flowers preferentially bind a hybrid c-box/g-box motif and help

to define a new sub-family of bzip transcription factors. *Plant J*, 13(4):489–505, Feb 1998.

Martínez-García J. F., Huq E., and Quail P. H. Direct targeting of light signals to a promoter element-bound transcription factor. *Science*, 288(5467):859–63, May 2000.

MATLAB . version 8.0.0.783 (R2012b). The MathWorks Inc., Natick, Massachusetts, 2012.

Maxam A. M. and Gilbert W. A new method for sequencing DNA. *Proc Natl Acad Sci U S A*, 74(2):560–4, Feb 1977.

McGrath K. C., Dombrecht B., Manners J. M., Schenk P. M., Edgar C. I., Maclean D. J., Scheible W.-R., Udvardi M. K., and Kazan K. Repressor- and activator-type ethylene response factors functioning in jasmonate signaling and disease resistance identified via a genome-wide screen of arabidopsis transcription factor gene expression. *Plant Physiol*, 139(2):949–959, Oct 2005. ISSN 0032-0889 (Print); 0032-0889 (Linking). doi: 10.1104/pp.105.068544.

Médigue C., Rechenmann F., Danchin A., and Viari A. Imagene: an integrated computer environment for sequence annotation and analysis. *Bioinformatics*, 15 (1):2–15, Jan 1999.

Meijer A. H., Ouwerkerk P. B., and Hoge J. H. Vectors for transcription factor cloning and target site identification by means of genetic selection in yeast. *Yeast*, 14(15):1407–15, Nov 1998. doi: 10.1002/(SICI)1097-0061(199811)14:15<1407::AID-YEA325>3.0.CO;2-M.

Melotto M., Underwood W., Koczan J., Nomura K., and He S. Y. Plant stomata function in innate immunity against bacterial invasion. *Cell*, 126(5):969–80, Sep 2006. doi: 10.1016/j.cell.2006.06.054.

Mengiste T., Chen X., Salmeron J., and Dietrich R. The botrytis susceptible1 gene encodes an r2r3myb transcription factor protein that is required for biotic and abiotic stress responses in arabidopsis. *Plant Cell*, 15(11):2551–65, Nov 2003. doi: 10.1105/tpc.014167.

Menkens A. E., Schindler U., and Cashmore A. R. The g-box: a ubiquitous regulatory dna element in plants bound by the gbf family of bzip proteins. *Trends Biochem Sci*, 20(12):506–10, Dec 1995.

Mertz E. T., Bates L. S., and Nelson O. E. Mutant gene that changes protein composition and increases lysine content of maize endosperm. *Science*, 145(3629): 279–80, Jul 1964.

Ming R., Hou S., Feng Y., Yu Q., Dionne-Laporte A., Saw J. H., Senin P., Wang W., Ly B. V., Lewis K. L. T., Salzberg S. L., Feng L., Jones M. R., Skelton R. L., Murray J. E., Chen C., Qian W., Shen J., Du P., Eustice M., Tong E., Tang H., Lyons E., Paull R. E., Michael T. P., Wall K., Rice D. W., Albert H., Wang M.-L., Zhu Y. J., Schatz M., Nagarajan N., Acob R. A., Guan P., Blas A., Wai C. M., Ackerman C. M., Ren Y., Liu C., Wang J., Wang J., Na J.-K., Shakirov E. V., Haas B., Thimmapuram J., Nelson D., Wang X., Bowers J. E., Gschwend A. R., Delcher A. L., Singh R., Suzuki J. Y., Tripathi S., Neupane K., Wei H., Irikura B., Paidi M., Jiang N., Zhang W., Presting G., Windsor A., Navajas-Pérez R., Torres M. J., Feltus F. A., Porter B., Li Y., Burroughs A. M., Luo M.-C., Liu L., Christopher D. A., Mount S. M., Moore P. H., Sugimura T., Jiang J., Schuler M. A., Friedman V., Mitchell-Olds T., Shippen D. E., dePamphilis C. W., Palmer J. D., Freeling M., Paterson A. H., Gonsalves D., Wang L., and Alam M. The draft genome of the transgenic tropical fruit tree papaya (*carica papaya linnaeus*). *Nature*, 452(7190):991–6, Apr 2008. doi: 10.1038/nature06856.

Mirny L. A. Nucleosome-mediated cooperativity between transcription factors. *Proc Natl Acad Sci U S A*, 107(52):22534–9, Dec 2010. doi: 10.1073/pnas.0913805107.

Mizoi J., Shinozaki K., and Yamaguchi-Shinozaki K. Ap2/erf family transcription factors in plant abiotic stress responses. *Biochim Biophys Acta*, 1819(2):86–96, Feb 2012. ISSN 0006-3002 (Print); 0006-3002 (Linking). doi: 10.1016/j.bbagr.2011.08.004.

Moffat C. S., Ingle R. A., Wathugala D. L., Saunders N. J., Knight H., and Knight M. R. Erf5 and erf6 play redundant roles as positive regulators of ja/et-mediated defense against *botrytis cinerea* in arabidopsis. *PLoS One*, 7(4):e35995, 2012. doi: 10.1371/journal.pone.0035995.

Molina C. and Grotewold E. Genome wide analysis of arabidopsis core promoters. *BMC Genomics*, 6:25, 2005. doi: 10.1186/1471-2164-6-25.

Morcuende R., Bari R., Gibon Y., Zheng W., Pant B. D., Bläsing O., Usadel B., Czechowski T., Udvardi M. K., Stitt M., and Scheible W.-R. Genome-wide reprogramming of metabolism and regulatory networks of arabidopsis in response to phosphorus. *Plant Cell Environ*, 30(1):85–112, Jan 2007. doi: 10.1111/j.1365-3040.2006.01608.x.

Moreno-Hagelsieb G. and Latimer K. Choosing blast options for better detection of orthologs as reciprocal best hits. *Bioinformatics*, 24(3):319–24, Feb 2008. doi: 10.1093/bioinformatics/btm585.

Morgan P. W. and Drew M. C. Ethylene and plant responses to stress. *Physiologia Plantarum*, 100(3):620–630, 1997. ISSN 1399-3054. doi: 10.1111/j.1399-3054.1997.tb03068.x. URL <http://dx.doi.org/10.1111/j.1399-3054.1997.tb03068.x>.

Morohashi K., Xie Z., and Grotewold E. Gene-specific and genome-wide chip approaches to study plant transcriptional networks. *Methods Mol Biol*, 553:3–12, 2009. doi: 10.1007/978-1-60327-563-7\_1.

Morris K., MacKerness S. A., Page T., John C. F., Murphy A. M., Carr J. P., and Buchanan-Wollaston V. Salicylic acid has a role in regulating gene expression during leaf senescence. *Plant J*, 23(5):677–85, Sep 2000.

Moses A. M., Pollard D. A., Nix D. A., Iyer V. N., Li X.-Y., Biggin M. D., and Eisen M. B. Large-scale turnover of functional transcription factor binding sites in drosophila. *PLoS Comput Biol*, 2(10):e130, Oct 2006. doi: 10.1371/journal.pcbi.0020130.

Mueller S., Hilbert B., Dueckershoff K., Roitsch T., Krischke M., Mueller M. J., and Berger S. General detoxification and stress responses are mediated by oxidized lipids through tga transcription factors in arabidopsis. *Plant Cell*, 20(3):768–85, Mar 2008. doi: 10.1105/tpc.107.054809.

Mukumoto F., Hirose S., Imaseki H., and Yamazaki K. Dna sequence requirement of a tata element-binding protein from arabidopsis for transcription in vitro. *Plant Mol Biol*, 23(5):995–1003, Dec 1993.

Murat F., Xu J.-H., Tannier E., Abrouk M., Guilhot N., Pont C., Messing J., and Salse J. Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Res*, 20(11):1545–57, Nov 2010. doi: 10.1101/gr.109744.110.

Murre C., McCaw P. S., and Baltimore D. A new dna binding and dimerization motif in immunoglobulin enhancer binding, daughterless, myod, and myc proteins. *Cell*, 56(5):777–83, Mar 1989.

- Nagalakshmi U., Wang Z., Waern K., Shou C., Raha D., Gerstein M., and Snyder M. The transcriptional landscape of the yeast genome defined by rna sequencing. *Science*, 320(5881):1344–9, Jun 2008. doi: 10.1126/science.1158441.
- Nair S. K. and Burley S. K. Recognizing dna in the library. *Nature*, 404(6779):715, 717–8, Apr 2000. doi: 10.1038/35008182.
- Nakano T., Suzuki K., Fujimura T., and Shinshi H. Genome-wide analysis of the erf gene family in arabidopsis and rice. *Plant Physiol*, 140(2):411–32, Feb 2006. doi: 10.1104/pp.105.073783.
- Narusaka Y., Nakashima K., Shinwari Z. K., Sakuma Y., Furihata T., Abe H., Narusaka M., Shinozaki K., and Yamaguchi-Shinozaki K. Interaction between two cis-acting elements, abre and dre, in aba-dependent expression of arabidopsis rd29a gene in response to dehydration and high-salinity stresses. *Plant J*, 34(2):137–48, Apr 2003.
- Ndamukong I., Abdallat A. A., Thurow C., Fode B., Zander M., Weigel R., and Gatz C. Sa-inducible arabidopsis glutaredoxin interacts with tga factors and suppresses ja-responsive pdf1.2 transcription. *Plant J*, 50(1):128–39, Apr 2007. doi: 10.1111/j.1365-313X.2007.03039.x.
- Needleman S. B. and Wunsch C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–53, Mar 1970.
- Neph S., Vierstra J., Stergachis A. B., Reynolds A. P., Haugen E., Vernot B., Thurman R. E., John S., Sandstrom R., Johnson A. K., Maurano M. T., Humbert R., Rynes E., Wang H., Vong S., Lee K., Bates D., Diegel M., Roach V., Dunn D., Neri J., Schafer A., Hansen R. S., Kutyavin T., Giste E., Weaver M., Canfield T., Sabo P., Zhang M., Balasundaram G., Byron R., MacCoss M. J., Akey J. M., Bender M. A., Groudine M., Kaul R., and Stamatoyannopoulos J. A. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414):83–90, Sep 2012. doi: 10.1038/nature11212.
- Neuwald A. F., Liu J. S., and Lawrence C. E. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci*, 4(8):1618–32, Aug 1995. doi: 10.1002/pro.5560040820.
- Ni M., Tepperman J. M., and Quail P. H. Pif3, a phytochrome-interacting factor necessary for normal photoinduced signal transduction, is a novel basic helix-loop-helix protein. *Cell*, 95(5):657–67, Nov 1998.

Nilsen T. W. Mechanisms of microRNA-mediated gene regulation in animal cells. *Trends Genet*, 23(5):243–9, May 2007. doi: 10.1016/j.tig.2007.02.011.

Oh S. A., Park J. H., Lee G. I., Paek K. H., Park S. K., and Nam H. G. Identification of three genetic loci controlling leaf senescence in *arabidopsis thaliana*. *Plant J*, 12(3):527–35, Sep 1997.

Ohme-Takagi M. and Shinshi H. Ethylene-inducible dna binding proteins that interact with an ethylene-responsive element. *Plant Cell*, 7(2):173–82, Feb 1995. doi: 10.1105/tpc.7.2.173.

Ohta M., Matsui K., Hiratsu K., Shinshi H., and Ohme-Takagi M. Repression domains of class ii erf transcriptional repressors share an essential motif for active repression. *Plant Cell*, 13(8):1959–68, Aug 2001.

Oñate-Sánchez L. and Singh K. B. Identification of *arabidopsis* ethylene-responsive element binding factors with distinct induction kinetics after pathogen infection. *Plant Physiol*, 128(4):1313–22, Apr 2002. doi: 10.1104/pp.010862.

Oñate-Sánchez L., Anderson J. P., Young J., and Singh K. B. Aterf14, a member of the erf family of transcription factors, plays a nonredundant role in plant defense. *Plant Physiol*, 143(1):400–9, Jan 2007. doi: 10.1104/pp.106.086637.

Onodera Y., Haag J. R., Ream T., Costa Nunes P., Pontes O., and Pikaard C. S. Plant nuclear rna polymerase iv mediates sirna and dna methylation-dependent heterochromatin formation. *Cell*, 120(5):613–22, Mar 2005. doi: 10.1016/j.cell.2005.02.007.

Osborne T. B. and Mendel L. B. Amino acids in nutrition and growth. *The journal of Biological Chemistry*, 17:325–349, 1914.

Ossowski S., Schneeberger K., Clark R. M., Lanz C., Warthmann N., and Weigel D. Sequencing of natural strains of *arabidopsis thaliana* with short reads. *Genome Res*, 18(12):2024–33, Dec 2008. doi: 10.1101/gr.080200.108.

Ou B., Yin K.-Q., Liu S.-N., Yang Y., Gu T., Wing Hui J. M., Zhang L., Miao J., Kondou Y., Matsui M., Gu H.-Y., and Qu L.-J. A high-throughput screening system for *arabidopsis* transcription factors and its application to med25-dependent transcriptional regulation. *Mol Plant*, 4(3):546–555, May 2011. ISSN 1752-9867 (Electronic); 1674-2052 (Linking). doi: 10.1093/mp/ssr002.

- Ozsolak F., Song J. S., Liu X. S., and Fisher D. E. High-throughput mapping of the chromatin structure of human promoters. *Nat Biotechnol*, 25(2):244–8, Feb 2007. doi: 10.1038/nbt1279.
- Parcy F., Nilsson O., Busch M. A., Lee I., and Weigel D. A genetic framework for floral patterning. *Nature*, 395(6702):561–6, Oct 1998. doi: 10.1038/26903.
- Park J., Lee H.-J., Cheon C.-I., Kim S.-H., Hur Y.-S., Auh C.-K., Im K.-H., Yun D.-J., Lee S., and Davis K. R. The arabidopsis thaliana homeobox gene athb12 is involved in symptom development caused by geminivirus infection. *PLoS One*, 6(5):e20054, 2011. doi: 10.1371/journal.pone.0020054.
- Paszkowski J., Shillito R. D., Saul M., Mandák V., Hohn T., Hohn B., and Potrykus I. Direct gene transfer to plants. *EMBO J*, 3(12):2717–22, Dec 1984.
- Pauwels L. and Goossens A. Fine-tuning of early events in the jasmonate response. *Plant Signal Behav*, 3(10):846–7, Oct 2008.
- Pauwels L., Morreel K., De Witte E., Lammertyn F., Van Montagu M., Boerjan W., Inzé D., and Goossens A. Mapping methyl jasmonate-mediated transcriptional reprogramming of metabolism and cell cycle progression in cultured arabidopsis cells. *Proc Natl Acad Sci U S A*, 105(4):1380–5, Jan 2008. doi: 10.1073/pnas.0711203105.
- Paxton J. D. Phytoalexins — a working redefinition. *Journal of Phytopathology*, 101(2):106–109, 1981. ISSN 1439-0434. doi: 10.1111/j.1439-0434.1981.tb03327.x. URL <http://dx.doi.org/10.1111/j.1439-0434.1981.tb03327.x>.
- Pelaz S., Ditta G. S., Baumann E., Wisman E., and Yanofsky M. F. B and C floral organ identity functions require SEPALLATA MADS-box genes. *Nature*, 405(6783):200–3, May 2000a. doi: 10.1038/35012103.
- Pelaz S., Ditta G. S., Baumann E., Wisman E., and Yanofsky M. F. B and c floral organ identity functions require sepallata mads-box genes. *Nature*, 405(6783): 200–3, May 2000b. doi: 10.1038/35012103.
- Pellegrineschi A., Reynolds M., Pacheco M., Brito R. M., Almeraya R., Yamaguchi-Shinozaki K., and Hoisington D. Stress-induced expression in wheat of the arabidopsis thaliana dreb1a gene delays water stress symptoms under greenhouse conditions. *Genome*, 47(3):493–500, Jun 2004. doi: 10.1139/g03-140.

Penninckx I. A., Eggermont K., Terras F. R., Thomma B. P., De Samblanx G. W., Buchala A., Métraux J. P., Manners J. M., and Broekaert W. F. Pathogen-induced systemic activation of a plant defensin gene in arabidopsis follows a salicylic acid-independent pathway. *Plant Cell*, 8(12):2309–23, Dec 1996.

Peterson M. G., Tanese N., Pugh B. F., and Tjian R. Functional domains and upstream activation properties of cloned human tata binding protein. *Science*, 248(4963):1625–30, Jun 1990.

Picot E., Krusche P., Tiskin A., Carré I., and Ott S. Evolutionary analysis of regulatory sequences (ears) in plants. *Plant J*, 64(1):165–76, Oct 2010. doi: 10.1111/j.1365-313X.2010.04314.x.

Pieterse C. M. J., Leon-Reyes A., Van der Ent S., and Van Wees S. C. M. Networking by small-molecule hormones in plant immunity. *Nat Chem Biol*, 5(5):308–16, May 2009. doi: 10.1038/nchembio.164.

Pillai R. S., Bhattacharyya S. N., and Filipowicz W. Repression of protein synthesis by mirnas: how many mechanisms? *Trends Cell Biol*, 17(3):118–26, Mar 2007. doi: 10.1016/j.tcb.2006.12.007.

Pokhilko A., Fernández A. P., Edwards K. D., Southern M. M., Halliday K. J., and Millar A. J. The clock gene circuit in arabidopsis includes a repressilator with additional feedback loops. *Mol Syst Biol*, 8:574, 2012. doi: 10.1038/msb.2012.6.

Pozo M., Loon L. C., and Pieterse C. J. Jasmonates - signals in plant-microbe interactions. 23(3):211–222, 2004. doi: 10.1007/s00344-004-0031-5. URL <http://dx.doi.org/10.1007/s00344-004-0031-5>.

Pre M., Atallah M., Champion A., De Vos M., Pieterse C. M. J., and Memelink J. The ap2/erf domain transcription factor ora59 integrates jasmonic acid and ethylene signals in plant defense. *Plant Physiol*, 147(3):1347–1357, Jul 2008. ISSN 0032-0889 (Print); 0032-0889 (Linking). doi: 10.1104/pp.108.117523.

Pruneda-Paz J. L., Breton G., Para A., and Kay S. A. A functional genomics approach reveals che as a component of the arabidopsis circadian clock. *Science*, 323(5920):1481–5, Mar 2009. doi: 10.1126/science.1167206.

Qiu J.-L., Fiil B. K., Petersen K., Nielsen H. B., Botanga C. J., Thorgrimsen S., Palma K., Suarez-Rodriguez M. C., Sandbech-Clausen S., Lichota J., Brodersen P., Grasser K. D., Mattsson O., Glazebrook J., Mundy J., and Petersen

M. Arabidopsis map kinase 4 regulates gene expression through transcription factor release in the nucleus. *EMBO J*, 27(16):2214–21, Aug 2008. doi: 10.1038/emboj.2008.147.

Ramírez V., Agorio A., Coego A., García-Andrade J., Hernández M. J., Balaguer B., Ouwerkerk P. B. F., Zarra I., and Vera P. Myb46 modulates disease susceptibility to *botrytis cinerea* in arabidopsis. *Plant Physiol*, 155(4):1920–35, Apr 2011. doi: 10.1104/pp.110.171843.

Reece-Hoyes J. S., Barutcu A. R., McCord R. P., Jeong J. S., Jiang L., MacWilliams A., Yang X., Salehi-Ashtiani K., Hill D. E., Blackshaw S., Zhu H., Dekker J., and Walhout A. J. M. Yeast one-hybrid assays for gene-centered human gene regulatory network mapping. *Nat Methods*, 8(12):1050–2, Dec 2011a. doi: 10.1038/nmeth.1764.

Reece-Hoyes J. S., Diallo A., Lajoie B., Kent A., Shrestha S., Kadreppa S., Pesyna C., Dekker J., Myers C. L., and Walhout A. J. M. Enhanced yeast one-hybrid assays for high-throughput gene-centered regulatory network mapping. *Nat Methods*, 8(12):1059–64, Dec 2011b. doi: 10.1038/nmeth.1748.

Reineke A. R., Bornberg-Bauer E., and Gu J. Evolutionary divergence and limits of conserved non-coding sequence detection in plant genomes. *Nucleic Acids Res*, 39(14):6029–43, Aug 2011. doi: 10.1093/nar/gkr179.

Ren D., Liu Y., Yang K.-Y., Han L., Mao G., Glazebrook J., and Zhang S. A fungal-responsive mapk cascade regulates phytoalexin biosynthesis in arabidopsis. *Proc Natl Acad Sci U S A*, 105(14):5638–43, Apr 2008. doi: 10.1073/pnas.0711301105.

Reymond P., Weber H., Damond M., and Farmer E. E. Differential gene expression in response to mechanical wounding and insect feeding in arabidopsis. *Plant Cell*, 12(5):707–20, May 2000.

Riano-Pachon D. M., Ruzicic S., Dreyer I., and Mueller-Roeber B. Plntfdb: an integrative plant transcription factor database. *BMC Bioinformatics*, 8:42, 2007. ISSN 1471-2105 (Electronic); 1471-2105 (Linking). doi: 10.1186/1471-2105-8-42.

Riechmann J. L., Krizek B. A., and Meyerowitz E. M. Dimerization specificity of arabidopsis mads domain homeotic proteins apetala1, apetala3, pistillata, and agamous. *Proc Natl Acad Sci U S A*, 93(10):4793–8, May 1996.

Riechmann J. L., Heard J., Martin G., Reuber L., Jiang C., Keddie J., Adam L., Pineda O., Ratcliffe O. J., Samaha R. R., Creelman R., Pilgrim M., Broun P.,

Zhang J. Z., Ghandehari D., Sherman B. K., and Yu G. Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, 290(5499): 2105–10, Dec 2000.

Rivas-San Vicente M. and Plasencia J. Salicylic acid beyond defence: its role in plant growth and development. *J Exp Bot*, 62(10):3321–38, Jun 2011. doi: 10.1093/jxb/err031.

Robertson G., Hirst M., Bainbridge M., Bilenky M., Zhao Y., Zeng T., Euskirchen G., Bernier B., Varhol R., Delaney A., Thiessen N., Griffith O. L., He A., Marra M., Snyder M., and Jones S. Genome-wide profiles of stat1 dna association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods*, 4(8):651–7, Aug 2007. doi: 10.1038/nmeth1068.

Robinson K. A., Koepke J. I., Kharodawala M., and Lopes J. M. A network of yeast basic helix-loop-helix interactions. *Nucleic Acids Res*, 28(22):4460–6, Nov 2000.

Rose W. C. The amino acid requirements of adult man. *Nutr Abstr Rev Ser Hum Exp*, 27(3):631–47, Jul 1957.

Rosinski J. A. and Atchley W. R. Molecular evolution of the myb family of transcription factors: evidence for polyphyletic origin. *J Mol Evol*, 46(1):74–83, Jan 1998.

Ross-Innes C. S., Stark R., Holmes K. A., Schmidt D., Spyrou C., Russell R., Massie C. E., Vowler S. L., Eldridge M., and Carroll J. S. Cooperative interaction between retinoic acid receptor-alpha and estrogen receptor in breast cancer. *Genes Dev*, 24(2):171–82, Jan 2010. doi: 10.1101/gad.552910.

Rounsley S. D., Ditta G. S., and Yanofsky M. F. Diverse roles for mads box genes in arabidopsis development. *Plant Cell*, 7(8):1259–69, Aug 1995. doi: 10.1105/tpc.7.8.1259.

Rushton P. J. and Somssich I. E. Transcriptional control of plant genes responsive to pathogens. *Curr Opin Plant Biol*, 1(4):311–5, Aug 1998.

Rushton P. J., Torres J. T., Parniske M., Wernert P., Hahlbrock K., and Somssich I. E. Interaction of elicitor-induced dna-binding proteins with elicitor response elements in the promoters of parsley pr1 genes. *EMBO J*, 15(20):5690–700, Oct 1996.

- Sakuma Y., Liu Q., Dubouzet J. G., Abe H., Shinozaki K., and Yamaguchi-Shinozaki K. Dna-binding specificity of the erf/ap2 domain of arabidopsis drebs, transcription factors involved in dehydration- and cold-inducible gene expression. *Biochem Biophys Res Commun*, 290(3):998–1009, Jan 2002. doi: 10.1006/bbrc.2001.6299.
- Salama R. A. and Stekel D. J. Inclusion of neighboring base interdependencies substantially improves genome-wide prokaryotic transcription factor binding site prediction. *Nucleic Acids Res*, 38(12):e135, Jul 2010. doi: 10.1093/nar/gkq274.
- Sasaki-Sekimoto Y., Taki N., Obayashi T., Aono M., Matsumoto F., Sakurai N., Suzuki H., Hirai M. Y., Noji M., Saito K., Masuda T., Takamiya K.-i., Shibata D., and Ohta H. Coordinated activation of metabolic pathways for antioxidants and defence compounds by jasmonates and their roles in stress tolerance in arabidopsis. *Plant J*, 44(4):653–68, Nov 2005. doi: 10.1111/j.1365-313X.2005.02560.x.
- Schaffer R., Ramsay N., Samach A., Corden S., Putterill J., Carré I. A., and Coupland G. The late elongated hypocotyl mutation of arabidopsis disrupts circadian rhythms and the photoperiodic control of flowering. *Cell*, 93(7):1219–29, Jun 1998.
- Schaffer R., Landgraf J., Accerbi M., Simon V., Larson M., and Wisman E. Microarray analysis of diurnal and circadian-regulated genes in arabidopsis. *Plant Cell*, 13(1):113–23, Jan 2001.
- Schaller F. Enzymes of the biosynthesis of octadecanoid-derived signalling molecules. *J Exp Bot*, 52(354):11–23, Jan 2001. ISSN 0022-0957 (Print); 0022-0957 (Linking).
- Schaller H., Uhlmann A., and Geider K. A DNA fragment from the origin of single-strand to double-strand DNA replication of bacteriophage fd. *Proc Natl Acad Sci U S A*, 73(1):49–53, Jan 1976.
- Schauer S. E., Schlüter P. M., Baskar R., Gheyselinck J., Bolaños A., Curtis M. D., and Grossniklaus U. Intronic regulatory elements determine the divergent expression patterns of agamous-like6 subfamily members in arabidopsis. *Plant J*, 59(6):987–1000, Sep 2009. doi: 10.1111/j.1365-313X.2009.03928.x.
- Schmidt M. C., Kao C. C., Pei R., and Berk A. J. Yeast tata-box transcription factor gene. *Proc Natl Acad Sci U S A*, 86(20):7785–9, Oct 1989.
- Schneider C. A., Rasband W. S., and Eliceiri K. W. NIH image to imagej: 25 years of image analysis. *Nat Methods*, 9(7):671–5, Jul 2012.

Schneider T. D. and Stephens R. M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, 18(20):6097–100, Oct 1990.

Schroeder J. I., Kwak J. M., and Allen G. J. Guard cell abscisic acid signalling and engineering drought hardiness in plants. *Nature*, 410(6826):327–30, Mar 2001. doi: 10.1038/35066500.

Sengupta D. J., Wickens M., and Fields S. Identification of rnas that bind to a specific protein using the yeast three-hybrid system. *RNA*, 5(4):596–601, Apr 1999.

Sessa G., Morelli G., and Ruberti I. The athb-1 and -2 hd-zip domains homodimerize forming complexes of different dna binding specificities. *EMBO J*, 12(9):3507–17, Sep 1993.

Sessa G., Morelli G., and Ruberti I. Dna-binding specificity of the homeodomain-leucine zipper domain. *J Mol Biol*, 274(3):303–9, Dec 1997. doi: 10.1006/jmbi.1997.1408.

Shannon P., Markiel A., Ozier O., Baliga N. S., Wang J. T., Ramage D., Amin N., Schwikowski B., and Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–504, Nov 2003. doi: 10.1101/gr.1239303.

Shen Q. and Ho T. H. Functional dissection of an abscisic acid (aba)-inducible gene reveals two independent aba-responsive complexes each containing a g-box and a novel cis-acting element. *Plant Cell*, 7(3):295–307, Mar 1995.

Shimizu T., Toumoto A., Ihara K., Shimizu M., Kyogoku Y., Ogawa N., Oshima Y., and Hakoshima T. Crystal structure of pho4 bhlh domain-dna complex: flanking base recognition. *EMBO J*, 16(15):4689–97, Aug 1997. doi: 10.1093/emboj/16.15.4689.

Shin D., Koo Y. D., Lee J., Lee H.-J., Baek D., Lee S., Cheon C.-I., Kwak S.-S., Lee S. Y., and Yun D.-J. Athb-12, a homeobox-leucine zipper domain protein from arabidopsis thaliana, increases salt tolerance in yeast by regulating sodium exclusion. *Biochem Biophys Res Commun*, 323(2):534–40, Oct 2004. doi: 10.1016/j.bbrc.2004.08.127.

Shindo T., Misas-Villamil J. C., Hörger A. C., Song J., and van der Hoorn R. A. L. A role in immunity for arabidopsis cysteine protease rd21, the ortholog of the tomato

immune protease c14. *PLoS One*, 7(1):e29317, 2012. doi: 10.1371/journal.pone.0029317.

Shirano Y., Kachroo P., Shah J., and Klessig D. F. A gain-of-function mutation in an arabidopsis toll interleukin1 receptor-nucleotide binding site-leucine-rich repeat type r gene triggers defense responses and results in enhanced disease resistance. *Plant Cell*, 14(12):3149–62, Dec 2002.

Smale S. T. and Kadonaga J. T. The rna polymerase ii core promoter. *Annu Rev Biochem*, 72:449–79, 2003. doi: 10.1146/annurev.biochem.72.121801.161520.

Smedley D., Haider S., Ballester B., Holland R., London D., Thorisson G., and Kasprzyk A. Biomart—biological queries made easy. *BMC Genomics*, 10:22, 2009. ISSN 1471-2164 (Electronic); 1471-2164 (Linking). doi: 10.1186/1471-2164-10-22.

Smolen G. A., Pawlowski L., Wilensky S. E., and Bender J. Dominant alleles of the basic helix-loop-helix transcription factor atr2 activate stress-responsive genes in arabidopsis. *Genetics*, 161(3):1235–46, Jul 2002.

Sokol A., Kwiatkowska A., Jerzmanowski A., and Prymakowska-Bosak M. 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(1):245–54, Dec 2007. doi: 10.1007/s00425-007-0612-1.

Solano R., Stepanova A., Chao Q., and Ecker J. R. Nuclear events in ethylene signaling: a transcriptional cascade mediated by ethylene-insensitive3 and ethylene-response-factor1. *Genes Dev*, 12(23):3703–14, Dec 1998.

Song L. and Crawford G. E. Dnase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc*, 2010(2):pdb.prot5384, Feb 2010. doi: 10.1101/pdb.prot5384.

Spangler J. B., Subramaniam S., Freeling M., and Feltus F. A. Evidence of function for conserved noncoding sequences in arabidopsis thaliana. *New Phytol*, 193(1): 241–52, Jan 2012. doi: 10.1111/j.1469-8137.2011.03916.x.

Spensley M., Kim J.-Y., Picot E., Reid J., Ott S., Helliwell C., and Carré I. A. Evolutionarily conserved regulatory motifs in the promoter of the arabidopsis clock gene late elongated hypocotyl. *Plant Cell*, 21(9):2606–23, Sep 2009. doi: 10.1105/tpc.109.069898.

Spoel S. H., Mou Z., Tada Y., Spivey N. W., Genschik P., and Dong X. Proteasome-mediated turnover of the transcription coactivator npr1 plays dual roles in regulating plant immunity. *Cell*, 137(5):860–72, May 2009. doi: 10.1016/j.cell.2009.03.038.

Standart N. and Jackson R. J. Micrornas repress translation of m7gppp-capped target mrnas in vitro by inhibiting initiation and promoting deadenylation. *Genes Dev*, 21(16):1975–82, Aug 2007. doi: 10.1101/gad.1591507.

Stein A., Takasuka T. E., and Collings C. K. Are nucleosome positions in vivo primarily determined by histone-dna sequence preferences? *Nucleic Acids Res*, 38(3):709–19, Jan 2010. doi: 10.1093/nar/gkp1043.

Stewart A. J., Hannenhalli S., and Plotkin J. B. Why transcription factor binding sites are ten nucleotides long. *Genetics*, 192(3):973–85, Nov 2012. doi: 10.1534/genetics.112.143370.

Stockinger E. J., Gilmour S. J., and Thomashow M. F. Arabidopsis thaliana cbf1 encodes an ap2 domain-containing transcriptional activator that binds to the c-repeat/dre, a cis-acting dna regulatory element that stimulates transcription in response to low temperature and water deficit. *Proc Natl Acad Sci U S A*, 94(3):1035–40, Feb 1997.

Stuart L. M., Paquette N., and Boyer L. Effector-triggered versus pattern-triggered immunity: how animals sense pathogens. *Nat Rev Immunol*, 13(3):199–206, Mar 2013. doi: 10.1038/nri3398.

Székely G., Abrahám E., Csépló A., Rigó G., Zsigmond L., Csiszár J., Ayaydin F., Strizhov N., Jásik J., Schmelzer E., Koncz C., and Szabados L. Duplicated p5cs genes of arabidopsis play distinct roles in stress regulation and developmental control of proline biosynthesis. *Plant J*, 53(1):11–28, Jan 2008. doi: 10.1111/j.1365-313X.2007.03318.x.

Tagle D. A., Koop B. F., Goodman M., Slightom J. L., Hess D. L., and Jones R. T. Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J Mol Biol*, 203(2):439–55, Sep 1988.

Tang H., Lyons E., Pedersen B., Schnable J. C., Paterson A. H., and Freeling M. Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC Bioinformatics*, 12:102, 2011. doi: 10.1186/1471-2105-12-102.

- Tanimoto M., Roberts K., and Dolan L. Ethylene is a positive regulator of root hair development in *arabidopsis thaliana*. *Plant J*, 8(6):943–8, Dec 1995.
- Tao Y., Xie Z., Chen W., Glazebrook J., Chang H.-S., Han B., Zhu T., Zou G., and Katagiri F. Quantitative nature of *arabidopsis* responses during compatible and incompatible interactions with the bacterial pathogen *pseudomonas syringae*. *Plant Cell*, 15(2):317–30, Feb 2003.
- Teichmann S. A. and Babu M. M. Gene regulatory network growth by duplication. *Nat Genet*, 36(5):492–6, May 2004. doi: 10.1038/ng1340.
- Thimm O., Bläsing O., Gibon Y., Nagel A., Meyer S., Krüger P., Selbig J., Müller L. A., Rhee S. Y., and Stitt M. Mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J*, 37(6):914–39, Mar 2004.
- Thines B., Katsir L., Melotto M., Niu Y., Mandaokar A., Liu G., Nomura K., He S. Y., Howe G. A., and Browse J. Jaz repressor proteins are targets of the scfcoi1 complex during jasmonate signalling. *Nature*, 448(7154):661–665, 08 2007. URL <http://dx.doi.org/10.1038/nature05960>.
- Thomas B. C., Rapaka L., Lyons E., Pedersen B., and Freeling M. Arabidopsis intragenomic conserved noncoding sequence. *Proc Natl Acad Sci U S A*, 104(9): 3348–53, Feb 2007. doi: 10.1073/pnas.0611574104.
- Tiskin A. Semi-local string comparison: Algorithmic techniques and applications. 1(4):571–603, 2008. doi: 10.1007/s11786-007-0033-3. URL <http://dx.doi.org/10.1007/s11786-007-0033-3>.
- Toledo-Ortiz G., Huq E., and Quail P. H. The *arabidopsis* basic/helix-loop-helix transcription factor family. *Plant Cell*, 15(8):1749–70, Aug 2003.
- Ton J., Flors V., and Mauch-Mani B. The multifaceted role of aba in disease resistance. *Trends Plant Sci*, 14(6):310–7, Jun 2009. doi: 10.1016/j.tplants.2009.03.006.
- Tran L.-S. P., Nakashima K., Sakuma Y., Simpson S. D., Fujita Y., Maruyama K., Fujita M., Seki M., Shinozaki K., and Yamaguchi-Shinozaki K. 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(9):2481–98, Sep 2004. doi: 10.1105/tpc.104.022699.

Tran L.-S. P., Nakashima K., Sakuma Y., Osakabe Y., Qin F., Simpson S. D., Maruyama K., Fujita Y., Shinozaki K., and Yamaguchi-Shinozaki K. Co-expression of the stress-inducible zinc finger homeodomain zfhd1 and nac transcription factors enhances expression of the erd1 gene in arabidopsis. *Plant J*, 49 (1):46–63, Jan 2007. doi: 10.1111/j.1365-313X.2006.02932.x.

Turing A. M. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 237(641):37–72, 08 1952. URL <http://rstb.royalsocietypublishing.org/content/237/641/37>.

Tuskan G. A., Difazio S., Jansson S., Bohlmann J., Grigoriev I., Hellsten U., Putnam N., Ralph S., Rombauts S., Salamov A., Schein J., Sterck L., Aerts A., Bhalerao R. R., Bhalerao R. P., Blaudez D., Boerjan W., Brun A., Brunner A., Busov V., Campbell M., Carlson J., Chalot M., Chapman J., Chen G.-L., Cooper D., Coutinho P. M., Couturier J., Covert S., Cronk Q., Cunningham R., Davis J., Degroeve S., Déjardin A., Depamphilis C., Detter J., Dirks B., Dubchak I., Duplessis S., Ehlting J., Ellis B., Gendler K., Goodstein D., Gribskov M., Grimwood J., Groover A., Gunter L., Hamberger B., Heinze B., Helariutta Y., Henrissat B., Holligan D., Holt R., Huang W., Islam-Faridi N., Jones S., Jones-Rhoades M., Jorgensen R., Joshi C., Kangasjärvi J., Karlsson J., Kelleher C., Kirkpatrick R., Kirst M., Kohler A., Kalluri U., Larimer F., Leebens-Mack J., Leplé J.-C., Locascio P., Lou Y., Lucas S., Martin F., Montanini B., Napoli C., Nelson D. R., Nelson C., Nieminen K., Nilsson O., Pereda V., Peter G., Philippe R., Pilate G., Poliakov A., Razumovskaya J., Richardson P., Rinaldi C., Ritland K., Rouzé P., Ryaboy D., Schmutz J., Schrader J., Segerman B., Shin H., Siddiqui A., Sterky F., Terry A., Tsai C.-J., Uberbacher E., Unneberg P., Vahala J., Wall K., Wessler S., Yang G., Yin T., Douglas C., Marra M., Sandberg G., Van de Peer Y., and Rokhsar D. The genome of black cottonwood, populus trichocarpa (torr. & gray). *Science*, 313(5793):1596–604, Sep 2006. doi: 10.1126/science.1128691.

United Nations . World population prospects: The 2012 revision, 2012. URL [http://esa.un.org/wpp/unpp/panel\\_population.htm](http://esa.un.org/wpp/unpp/panel_population.htm).

United Nations . United nations millennium development goals, 2013. URL <http://www.un.org/millenniumgoals/>.

Uno Y., Furihata T., Abe H., Yoshida R., Shinozaki K., and Yamaguchi-Shinozaki K. Arabidopsis basic leucine zipper transcription factors involved in an abscisic acid-dependent signal transduction pathway under drought and high-

salinity conditions. *Proc Natl Acad Sci U S A*, 97(21):11632–7, Oct 2000. doi: 10.1073/pnas.190309197.

Urnov F. D. Chromatin remodeling as a guide to transcriptional regulatory networks in mammals. *J Cell Biochem*, 88(4):684–94, Mar 2003. doi: 10.1002/jcb.10397.

Valencia-Sanchez M. A., Liu J., Hannon G. J., and Parker R. Control of translation and mrna degradation by mirnas and sirnas. *Genes Dev*, 20(5):515–24, Mar 2006. doi: 10.1101/gad.1399806.

van der Geer P., Hunter T., and Lindberg R. A. Receptor protein-tyrosine kinases and their signal transduction pathways. *Annu Rev Cell Biol*, 10:251–337, 1994. doi: 10.1146/annurev.cb.10.110194.001343.

van der Graaff E., Schwacke R., Schneider A., Desimone M., Flügge U.-I., and Kunze R. Transcription analysis of arabidopsis membrane transporters and hormone pathways during developmental and induced leaf senescence. *Plant Physiol*, 141(2):776–92, Jun 2006. doi: 10.1104/pp.106.079293.

van Dijk M., van Dijk A. D. J., Hsu V., Boelens R., and Bonvin A. M. J. J. Information-driven protein-dna docking using haddock: it is a matter of flexibility. *Nucleic Acids Res*, 34(11):3317–25, 2006. doi: 10.1093/nar/gkl412.

van Kan J. A. L. Licensed to kill: the lifestyle of a necrotrophic plant pathogen. *Trends Plant Sci*, 11(5):247–53, May 2006. doi: 10.1016/j.tplants.2006.03.005.

Van Loon L. and Van Strien E. {The families of pathogenesis-related proteins, their activities, and comparative analysis of PR-1 type proteins}. *Physiological and Molecular Plant Pathology*, 55(2):85–97, 08 1999. doi: doi:10.1006/pmpp.1999.0213. URL <http://dx.doi.org/10.1006/pmpp.1999.0213>.

van Loon L. C., Geraats B. P. J., and Linthorst H. J. M. Ethylene as a modulator of disease resistance in plants. *Trends Plant Sci*, 11(4):184–91, Apr 2006. doi: 10.1016/j.tplants.2006.02.005.

Vandenbussche M., Zethof J., Souer E., Koes R., Tornielli G. B., Pezzotti M., Ferrario S., Angenent G. C., and Gerats T. Toward the analysis of the petunia MADS box gene family by reverse and forward transposon insertion mutagenesis approaches: B, C, and D floral organ identity functions require SEPALLATA-like MADS box genes in petunia. *Plant Cell*, 15(11):2680–93, Nov 2003. doi: 10.1105/tpc.017376.

Vandepoele K., Casneuf T., and Van de Peer Y. Identification of novel regulatory modules in dicotyledonous plants using expression data and comparative genomics. *Genome Biol*, 7(11):R103, 2006. doi: 10.1186/gb-2006-7-11-r103.

Velasco R., Zharkikh A., Troggio M., Cartwright D. A., Cestaro A., Pruss D., Pindo M., Fitzgerald L. M., Vezzulli S., Reid J., Malacarne G., Iliev D., Coppola G., Wardell B., Micheletti D., Macalma T., Facci M., Mitchell J. T., Perazzolli M., Eldredge G., Gatto P., Oyzerski R., Moretto M., Gutin N., Stefanini M., Chen Y., Segala C., Davenport C., Demattè L., Mraz A., Battilana J., Stormo K., Costa F., Tao Q., Si-Ammour A., Harkins T., Lackey A., Perbost C., Taillon B., Stella A., Solovyev V., Fawcett J. A., Sterck L., Vandepoele K., Grando S. M., Toppo S., Moser C., Lanchbury J., Bogden R., Skolnick M., Sgaramella V., Bhatnagar S. K., Fontana P., Gutin A., Van de Peer Y., Salamini F., and Viola R. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS One*, 2(12):e1326, 2007. doi: 10.1371/journal.pone.0001326.

Veronese P., Chen X., Bluhm B., Salmeron J., Dietrich R., and Mengiste T. The bos loci of arabidopsis are required for resistance to botrytis cinerea infection. *Plant J*, 40(4):558–74, Nov 2004. doi: 10.1111/j.1365-313X.2004.02232.x.

Vert G., Walcher C. L., Chory J., and Nemhauser J. L. Integration of auxin and brassinosteroid pathways by auxin response factor 2. *Proc Natl Acad Sci U S A*, 105(28):9829–34, Jul 2008. doi: 10.1073/pnas.0803996105.

Vick B. A. and Zimmerman D. C. Biosynthesis of jasmonic acid by several plant species. *Plant Physiol*, 75(2):458–61, Jun 1984.

Vidal M. *The Yeast Two-Hybrid*, pages 109–147. Oxford University Press, New York, NY, 1997.

von Saint Paul V., Zhang W., Kanawati B., Geist B., Faus-Kessler T., Schmitt-Kopplin P., and Schäffner A. R. The arabidopsis glucosyltransferase ugt76b1 conjugates isoleucic acid and modulates plant defense and senescence. *Plant Cell*, 23(11):4124–45, Nov 2011. doi: 10.1105/tpc.111.088443.

Walhout A. J. M. Unraveling transcription regulatory networks by protein-dna and protein-protein interaction mapping. *Genome Res*, 16(12):1445–54, Dec 2006. doi: 10.1101/gr.5321506.

Wang C.-T. and Xu Y.-N. The 5untranslated region of the fad3 mrna is required for its translational enhancement at low temperature in arabidopsis

roots. *Plant Science*, 179(3):234–240, 9 2010. doi: <http://dx.doi.org/10.1016/j.plantsci.2010.05.008>. URL <http://www.sciencedirect.com/science/article/pii/S0168945210001470>.

Wang W., Barnaby J. Y., Tada Y., Li H., Tör M., Caldelari D., Lee D.-u., Fu X.-D., and Dong X. Timing of plant immune responses by a central circadian regulator. *Nature*, 470(7332):110–4, Feb 2011. doi: 10.1038/nature09766.

Wang X., Basnayake B. M. V. S., Zhang H., Li G., Li W., Virk N., Mengiste T., and Song F. The arabidopsis ataf1, a nac transcription factor, is a negative regulator of defense responses against necrotrophic fungal and bacterial pathogens. *Mol Plant Microbe Interact*, 22(10):1227–38, Oct 2009. doi: 10.1094/MPMI-22-10-1227.

Wang Z. Y., Kenigsbuch D., Sun L., Harel E., Ong M. S., and Tobin E. M. A myb-related transcription factor is involved in the phytochrome regulation of an arabidopsis lhcb gene. *Plant Cell*, 9(4):491–507, Apr 1997. doi: 10.1105/tpc.9.4.491.

Weintraub H. and Groudine M. Chromosomal subunits in active genes have an altered conformation. *Science*, 193(4256):848–56, Sep 1976.

Wettenhall J. M. and Smyth G. K. limmagui: a graphical user interface for linear modeling of microarray data. *Bioinformatics*, 20(18):3705–6, Dec 2004. doi: 10.1093/bioinformatics/bth449.

Willcock E. and Hopkins F. The importance of individual amino acids in metabolism; observations on the effect of adding tryptophan to a diet in which zein is the sole nitrogenous constituent. *The Journal of Physiology*, 3:88–102, 1906.

Williamson B., Tudzynski B., Tudzynski P., and van Kan J. A. L. Botrytis cinerea: the cause of grey mould disease. *Mol Plant Pathol*, 8(5):561–80, Sep 2007. doi: 10.1111/j.1364-3703.2007.00417.x.

Windram O., Madhou P., McHattie S., Hill C., Hickman R., Cooke E., Jenkins D. J., Penfold C. A., Baxter L., Breeze E., Kiddie S. J., Rhodes J., Atwell S., Kliebenstein D. J., Kim Y.-S., Stegle O., Borgwardt K., Zhang C., Tabrett A., Legaie R., Moore J., Finkenstadt B., Wild D. L., Mead A., Rand D., Beynon J., Ott S., Buchanan-Wollaston V., and Denby K. J. Arabidopsis defense against botrytis cinerea: chronology and regulation deciphered by high-resolution temporal transcriptomic analysis. *Plant Cell*, 24(9):3530–57, Sep 2012. doi: 10.1105/tpc.112.102046.

- Wingender E., Chen X., Hehl R., Karas H., Liebich I., Matys V., Meinhardt T., Prüss M., Reuter I., and Schacherer F. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res*, 28(1):316–9, Jan 2000.
- Wloch D. M., Szafraniec K., Borts R. H., and Korona R. Direct estimate of the mutation rate and the distribution of fitness effects in the yeast *Saccharomyces cerevisiae*. *Genetics*, 159(2):441–52, Oct 2001.
- Xiong Y., Liu T., Tian C., Sun S., Li J., and Chen M. Transcription factors in rice: a genome-wide comparative analysis between monocots and eudicots. *Plant Mol Biol*, 59(1):191–203, Sep 2005. doi: 10.1007/s11103-005-6503-6.
- Xu J., Chen G., De Jong A. T., Shahravan S. H., and Shin J. A. Max-e47, a designed minimalist protein that targets the e-box dna site in vivo and in vitro. *J Am Chem Soc*, 131(22):7839–48, Jun 2009. doi: 10.1021/ja901306q.
- Xu L. C., Thali M., and Schaffner W. Upstream box/tata box order is the major determinant of the direction of transcription. *Nucleic Acids Res*, 19(24):6699–704, Dec 1991.
- Xu X., Chen C., Fan B., and Chen Z. Physical and functional interactions between pathogen-induced *Arabidopsis* wrky18, wrky40, and wrky60 transcription factors. *Plant Cell*, 18(5):1310–26, May 2006. doi: 10.1105/tpc.105.037523.
- Xu Y., Chang P., Liu D., Narasimhan M. L., Raghothama K. G., Hasegawa P. M., and Bressan R. A. Plant defense genes are synergistically induced by ethylene and methyl jasmonate. *Plant Cell*, 6(8):1077–1085, Aug 1994. doi: 10.1105/tpc.6.8.1077.
- Yang P., Chen C., Wang Z., Fan B., and Chen Z. A pathogen- and salicylic acid-induced wrky dna-binding activity recognizes the elicitor response element of the tobacco class i chitinase gene promoter. *The Plant Journal*, 18(2):141–149, 1999. ISSN 1365-313X. doi: 10.1046/j.1365-313X.1999.00437.x. URL <http://dx.doi.org/10.1046/j.1365-313X.1999.00437.x>.
- Yang Y., Costa A., Leonhardt N., Siegel R. S., and Schroeder J. I. Isolation of a strong *Arabidopsis* guard cell promoter and its potential as a research tool. *Plant Methods*, 4:6, 2008. doi: 10.1186/1746-4811-4-6.
- Yang Z., Tian L., Latoszek-Green M., Brown D., and Wu K. *Arabidopsis* erf4 is a transcriptional repressor capable of modulating ethylene and abscisic acid responses. *Plant Mol Biol*, 58(4):585–96, Jul 2005. doi: 10.1007/s11103-005-7294-5.

- Yant L., Mathieu J., Dinh T. T., Ott F., Lanz C., Wollmann H., Chen X., and Schmid M. Orchestration of the floral transition and floral development in arabidopsis by the bifunctional transcription factor apetala2. *Plant Cell*, 22(7):2156–70, Jul 2010. doi: 10.1105/tpc.110.075606.
- Yoshida T., Fujita Y., Sayama H., Kidokoro S., Maruyama K., Mizoi J., Shinozaki K., and Yamaguchi-Shinozaki K. Areb1, areb2, and abf3 are master transcription factors that cooperatively regulate abre-dependent aba signaling involved in drought stress tolerance and require aba for full activation. *Plant J*, 61(4):672–85, Feb 2010. doi: 10.1111/j.1365-313X.2009.04092.x.
- Young J. M., Kuykendall L. D., Martínez-Romero E., Kerr A., and Sawada H. A revision of rhizobium frank 1889, with an emended description of the genus, and the inclusion of all species of agrobacterium conn 1942 and allorhizobium undicola de lajudie et al. 1998 as new combinations: Rhizobium radiobacter, r. rhizogenes, r. rubi, r. undicola and r. vitis. *Int J Syst Evol Microbiol*, 51(Pt 1):89–103, Jan 2001.
- Yu E. Y., Kim S. E., Kim J. H., Ko J. H., Cho M. H., and Chung I. K. Sequence-specific dna recognition by the myb-like domain of plant telomeric protein rtbp1. *J Biol Chem*, 275(31):24208–14, Aug 2000. doi: 10.1074/jbc.M003250200.
- Zarei A., Korbes A. P., Younessi P., Montiel G., Champion A., and Memelink J. Two gcc boxes and ap2/erf-domain transcription factor ora59 in jasmonate/ethylene-mediated activation of the pdf1.2 promoter in arabidopsis. *Plant Mol Biol*, 75 (4-5):321–331, Mar 2011. ISSN 1573-5028 (Electronic); 0167-4412 (Linking). doi: 10.1007/s11103-010-9728-y.
- Zentella R., Zhang Z.-L., Park M., Thomas S. G., Endo A., Murase K., Fleet C. M., Jikumaru Y., Nambara E., Kamiya Y., and Sun T.-P. Global analysis of della direct targets in early gibberellin signaling in arabidopsis. *Plant Cell*, 19(10): 3037–57, Oct 2007. doi: 10.1105/tpc.107.054999.
- Zhang L., Li Z., Quan R., Li G., Wang R., and Huang R. An ap2 domain-containing gene, ese1, targeted by the ethylene signaling component ein3 is important for the salt response in arabidopsis. *Plant Physiol*, 157(2):854–865, Oct 2011. ISSN 1532-2548 (Electronic); 0032-0889 (Linking). doi: 10.1104/pp.111.179028.
- Zhang W., Zhang T., Wu Y., and Jiang J. Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in Arabidopsis. *Plant Cell*, 24(7):2719–31, Jul 2012. doi: 10.1105/tpc.112.098061.

Zhang Y. and Wang L. The wrky transcription factor superfamily: its origin in eukaryotes and expansion in plants. *BMC Evol Biol*, 5:1, 2005. doi: 10.1186/1471-2148-5-1.

Zhang Y., Goritschnig S., Dong X., and Li X. A gain-of-function mutation in a plant disease resistance gene leads to constitutive activation of downstream signal transduction pathways in suppressor of npr1-1, constitutive 1. *Plant Cell*, 15(11): 2636–46, Nov 2003. doi: 10.1105/tpc.015842.

Zheng Z., Qamar S. A., Chen Z., and Mengiste T. Arabidopsis wrky33 transcription factor is required for resistance to necrotrophic fungal pathogens. *Plant J*, 48(4): 592–605, Nov 2006. doi: 10.1111/j.1365-313X.2006.02901.x.

Zhou C., Zhang L., Duan J., Miki B., and Wu K. Histone deacetylase19 is involved in jasmonic acid and ethylene signaling of pathogen response in arabidopsis. *Plant Cell*, 17(4):1196–204, Apr 2005. doi: 10.1105/tpc.104.028514.

Zhou N., Tootle T. L., and Glazebrook J. Arabidopsis pad3, a gene required for camalexin biosynthesis, encodes a putative cytochrome p450 monooxygenase. *Plant Cell*, 11(12):2419–28, Dec 1999.

Zhu Q., Zhang J., Gao X., Tong J., Xiao L., Li W., and Zhang H. The arabidopsis ap2/erf transcription factor rap2.6 participates in aba, salt and osmotic stress responses. *Gene*, 457(1-2):1–12, Jun 2010. doi: 10.1016/j.gene.2010.02.011.

Zhu Z., An F., Feng Y., Li P., Xue L., A M., Jiang Z., Kim J.-M., To T. K., Li W., Zhang X., Yu Q., Dong Z., Chen W.-Q., Seki M., Zhou J.-M., and Guo H. Derepression of ethylene-stabilized transcription factors (ein3/eil1) mediates jasmonate and ethylene signaling synergy in arabidopsis. *Proc Natl Acad Sci U S A*, 108(30):12539–44, Jul 2011. doi: 10.1073/pnas.1103959108.

Zimmerli L., Metraux J. P., and Mauch-Mani B. beta-aminobutyric acid-induced protection of arabidopsis against the necrotrophic fungus botrytis cinerea. *Plant Physiol*, 126(2):517–523, Jun 2001. ISSN 0032-0889 (Print); 0032-0889 (Linking).

Zipfel C. and Felix G. Plants and animals: a different taste for microbes? *Curr Opin Plant Biol*, 8(4):353–60, Aug 2005. doi: 10.1016/j.pbi.2005.05.004.