A Bioinformatic Analysis of Genes Involved in Stress Responses in *Arabidopsis thaliana*

by Linda Karen Hughes

Thesis submitted in fulfillment of the requirements

for the degree of Doctor of Philosophy

University of Warwick, Warwick HRI

September 2009

Project supervisors:

Prof. Jim Beynon (Warwick HRI)

This project was carried out under a BBSRC funded studentship.

Length of thesis 39394 excl. tables and figures

Table of Contents

Acknowledgements

I would like to express my thanks to the following people for their help and support in the work towards this thesis. I have to thank Prof. Jim Beynon as my academic supervisor and to Dr. Vicky Buchanan-Wollaston for her expert advice and guidance in relation to *A. thaliana* matters. In particular, I am indebted to Dr. Laura Baxter for all her generous help, advice and guidance in both general scientific and bioinformatics matters throughout this PhD process. I would also like to thank staff at Warwick HRI, in particular Neale Grant, Kerry-Sue Peplow and Sarah Stevenson for all their moral support and in keeping my spirits up. Many thanks to Stuart McHattie and Kanishka Bhattacharya for their help in practical and theoretical modelling matters respectively. Thanks to the BBSRC for funding this investigation. Finally, I would like to express my deep thanks and gratitude to my partner David Pashute for all his help, support and tolerance during what has arguably one of the most difficult times in my career.

Declaration

The work referred to in this thesis is my own work and has not been submitted for a degree at another university.

Summary

Hyaloperonospora arabidopsidis is an obligate biotrophic oomycete shown to cause downy mildew in *Arabidopsis thaliana*. The main focus of this project is examining plant stress response and the strategies employed by *H. arabidopsidis* to infect Arabidopsis and evade plant stress responses. Two regions of the *H. arabidopsidis* genome containing genes expressed in planta during infection were bioinformatically annotated. The results indicated the genes were involved in regulatory processes associated with the pathogenicity of *H. arabidopsidis* but not a direct role in pathogenicity. *H. arabidopsidis* infects its host by secreting effector proteins into the cytoplasm and apoplastic space of the host. The secretome of *H. arabidopsidis* was analysed to identify classes of cysteine rich apoplastic effectors. This identified 15 candidate elicitin (ELI) and elicitin-like (ELL) sequences, three Kazal-like serine protease inhibitors and four candidates similar to the protein sequences of *Ppats* 14 and 24, expressed during infection.

A second set of aims was to identify potential signalling networks up activated during plant defence responses to infection by *H. arabidopsidis* using a new model developed by Beal et al (Beal, Falciani et al. 2005) to eventually engineer transcriptional networks. Unfortunately this failed due to problems with the experiment. However, it was still possible to identify signalling networks from a second microarray time course experimental data set centred on signalling networks up regulated in response to the onset of senescence, as they share overlapping signalling pathways. The modelling methodology was used to model the anthocyanin biosynthesis pathway. The model predicted the presence of AtMYB15 as a positive regulator of anthocyanin biosynthesis along with AtMYB90. Research carried out by Nichola Warner (Warner 2008) suggested that AtMYB90 was not essential for anthocyanin biosynthesis during senescence based on by comparing the phenotype of the MYB90 knock out, IM28, with the wild type (WT) Col-0 using a time course microarray. Models of networks of transcriptional regulation of the anthocyanin biosynthesis pathway for IM28 and WT implicate AtMYB29 as a positive regulator of anthocyanin biosynthesis.

List of Tables

Table 5.3 Predicted regulatory relationships between the 48 genes chosen for modelling for the IM28 dataset.

List of Figures

Chapter 1: Introduction

Food production depends on the growth of agricultural plants adapted to local environments. In recent times environments have been stable and, consequently, crop yields have been high. However, the levels of global carbon dioxide in the atmosphere have been growing steadily since the middle of the last century. This will cause changes in global weather and seasonal patterns within 20 years. These environmental changes will cause stress on current crop types. Hence, it is essential that we understand in detail how crop plants respond to these stresses to enable us to build more robust plant types.

To work on such complex systems as stress responses in a crop plant is impractical and, therefore, model organisms are used. *Arabidopsis thaliana*, otherwise known as Thale Cress and Arabidopsis in common use, belongs to the genus *Brassicaceae*. It is found in naturally occurring populations in regions of Europe, Asia and North America (Meinke, Cherry et al. 1998). In recent years Arabidopsis has become the model organism of choice for studying the essential principles of plant biology. Arabidopsis was considered the most appropriate model organism for plant molecular biology as;

1. It has a relatively short life cycle, the process of development from germinating seed through to senescence and subsequent first seed maturation can occur within 6 weeks,

2. The relatively small size of Arabidopsis (plants are usually between 3-15 cm in length) enables it to be grown at high density in a small space,

3. Arabidopsis mutant lines produce phenotypes that are defective in almost all areas of plant development including plant growth, flowering and senescence (Mur, Bi et al. 1997; Feys and Parker 2000; Mur, Brown et al. 2000; Devoto and Turner 2005).

4. It has a small (140Mb) genome that has been completely sequenced,

5. It is readily transformable,

6. A wide range of molecular tools are available, such as microarrays, collections of knockout lines and RNA interference (RNAi) constructs.

Consequently, a large community of researchers has been built up that contribute to a large and rapidly growing dataset of information and physical resources.

The main aim of this thesis was to model altered transcriptional activity in Arabidopsis in response to stress, with the long-term goal of identifying key genes that regulate global responses to stress. The two stresses used in this study were responses to senescence and infection by the oomycete pathogen *Hyaloperonospora arabidopsidis*. As the content of this thesis is diverse only a brief introduction is given here with more detailed and specific information described in the appropriate chapters.

1.1 Senescence

Senescence is a highly regulated process of resource remobilisation that represents the final stage of the leaf growth and development cycle. During senescence leaf cells undergo both structural and metabolic changes. One of the earliest cell structure changes is the breakdown of organelles such as the chloroplast. On the metabolic level there is an increase in catabolism of cellular components such as lipids, proteins and nucleic acids, into nutrients that can be reallocated to new growth areas of the plant such as the developing seeds. The process results in the gradual degradation of plant leaf tissue, leading to the death of the leaf (Lim, Kim et al. 2007). The onset of senescence is affected by numerous endogenous factors occurring within the Arabidopsis cell and also environmental factors. The environmental factors include starvation, inadequate amount of light, drought, pathogen attack and responses to extreme changes in temperature (He and Gan 2002). All of these factors will lead to an earlier onset of senescence so as to recycle the nutrients within the plant into storage organs, such as seed.. Examples of this have been seen in plant responses to drought and pathogen attack where early onset of senescence has been shown to occur (Mathews, Carroll et al. 1990).

1.2 Internal regulation of senescence

Naturally occurring age dependent senescence is regulated by a number of phytohormones including ethylene (ET), salicylic acid (SA) and jasmonic acid (JA). The current literature also suggests that the levels of these phytohormones have roles in governing plant responses to environmental stress indicating a significant overlap between the signalling pathways governing the two types of senescence. It is thought that the environmental stresses outlined earlier impact upon the synthesis of these hormones, which in turn affects the signalling pathways they are involved in, leading to increased expression of stress responsive genes (Devoto and Turner 2005). The individual roles of each of the hormones are outlined below along with the environmental stress that impact upon them.

1.3 Ethylene

ET is a gaseous hormone, which has been shown to play significant roles in plant growth and development, fruit ripening and flowering (Jones and Woodson 1997; Smalle and VanderStraeten 1997; Suzuki, Kikuchi et al. 1997). It has also been shown to be a positive regulator of age dependent senescence (Jing, Schippers et al. 2005). In fact studies using the ethylene resistant mutant *etr1-1* and the ethylene insensitive mutant *ein-2* resulted in a significant delay in the onset of senescence (Grbic and Bleecker 1995; Oh, Park et al. 1997). In terms of responses to stress, ET has lately been implicated in salt induced senescence. A recent study into the effects of increased salinity on the expression of several hormones showed a correlation between an increase in the ET precursor 1-aminocyclopropane-1-carboxylic acid (ACC) and the early onset of senescence in tomato (Ghanem, Albacete et al. 2008).

1.4 Salicylic Acid

SA has been shown to be involved in age dependent senescence (Grbic and Bleecker 1995). Studies have shown that SA levels increased four-fold in senescing leaves (Morris, Mackerness et al. 2000). A study comparing gene expression in transgenic plants expressing *NahG*, resulting in the loss of the SA induced signalling pathway, to the wild type Col-0 strain showed a two-fold decrease in expression of around 19 percent of genes previously up regulated during senescence in the wild type. The study also showed a delay in age-dependent senescence in the phenotype of the transgenic *NahG* plant (Buchanan-Wollaston, Page et al. 2005). In terms of responses to stress, SA has been shown to be involved in plant responses to pathogen attack through the production of Reactive Oxygen Species (ROS) resulting in programmed cell death at the site of infection (Mur, Bi et al. 1997; Mur, Brown et al. 2000) and as part of systemic acquired resistance (SAR) in which pathogen attack in one part of the plant induces resistance to the pathogen in the rest of the plant (Penninckx, Eggermont et al. 1996).

1.5 Methyl Jasmonate and Jasmonic Acid

JA signalling pathways and those of its precursor methyl jasmonate (MeJA) are implicated in promoting senescence. A study in which JA was exogenously applied to Arabidopsis leaves resulted in an early onset of senescence. Also application of JA to JA insensitive mutant *coi1* plants resulted in a failure to promote the early onset of senescence. Furthermore a four-fold increase in both MeJA and JA levels was seen during senescence (He, Fukushige et al. 2002). These results imply that the JA signalling pathway is necessary in order for JA to encourage the onset of senescence. However, it should be noted that studies on mutants defective in the JA signalling pathway resulted in a delay in senescence, which implies that although it is necessary to have the JA signalling pathway to enable JA promotion of senescence it is not essential for senescence to occur (He, Fukushige et al. 2002). As well as promoting the onset of senescence the JA signalling pathway has been implicated in governing plant responses to numerous stresses resulting in premature senescence. Among these are dark induced senescence, wounding, pathogen defence, and temperature changes (Thomma, Eggermont et al. 1998; Buchanan-Wollaston, Page et al. 2005; Devoto and Turner 2005; Fung, Wang et al. 2006; Wang, Cao et al. 2008).

1.6 Interlinking pathways in response to stress

The current model is that ET, SA and JA signalling pathways do not function independently but rather function as part of complex signalling network whereby the different pathways influence one another either positively or negatively in response to different stresses. Studies investigating the effects on Arabidopsis in response to infection with *Alternaria brassicicola* showed that both ET and JA are required for the activation of the defence related gene PDF1.2 (Penninckx, Eggermont et al. 1996). There is also further evidence to suggest that JA and ET co-ordinately regulate many plant defence response genes. A microarray experiment monitoring Arabidopsis gene expression to various stimuli showed that up to 50 percent of the genes induced by ET stimulation were also up regulated in response to treatment with JA (Schenk, Kazan et al. 2000). In contrast to the relationship between ET and JA, there is evidence that the interactions between JA and SA are mutually antagonistic. Studies using the *eds4* and *pad4* Arabidopsis mutant plants, deficient in SA accumulation, resulted in an enhanced response to JA dependent gene expression (Gupta, Willits et al. 2000). Further to this a study into the effects of SA on JA levels in tomato showed SA to be inhibitory (Doherty, Selvendran et al. 1988). JA has also been shown to have an inhibitory effect on SA signalling. The characterisation of three JA signalling mutants, mitogen-activated protein kinase4 (*mpk4*), suppressor of SA insensitivity2 (*ssi2*) and *coi1*, resulted in enhanced SA mediated defence gene expression. This implies that JA signalling impairs SA signalling (Kloek, Verbsky et al. 2001).

The complex nature of the interactions between the pathways could be due to the wide range of pathogens and stresses. There is evidence to suggest that the signalling pathway Arabidopsis uses in response to pathogen attack is dependent on the invading pathogen. Studies have shown that SA signalling is used as part of a basal resistance to infection by bacterial, *Pseudomonas syringae,* and oomycete, *H. arabidopsidis,* pathogens (Feys and Parker 2000; Kachroo, Yoshioka et al. 2000). In contrast, JA signalling has been used to mediate resistance to the fungal pathogen *Botrytis cinerea* (Thomma, Eggermont et al. 1998).

1.7 Mechanisms of pathogenicity

Pathogenicity can be defined as the ability of a pathogen to produce an infectious disease in an organism. Understanding the mechanisms by which a pathogen infects its host can leads to better ways of treatment and prevention of infection of valuable crops. Plant pathogens use a range of strategies to infect the host. Pathogenic bacteria multiply in the apoplastic spaces between the cells after entering through the stomata or through wounds. Some pathogenic fungi and oomycetes can develop feeding structures called haustoria, into the host cell plasma membrane. The pathogens can then deliver effector proteins into the plant cell to enhance pathogenicity (Jones and Dangl 2006). The microbial pathogens however, have different methods of transporting effector proteins into their respective host cell targets; phytopathogenic bacteria such as *Pseudomonas, Erwinia* and *Xanthomonas* species use a type III secretion system (TSS) to inject effector proteins into the cytoplasm of the host cell in plants (Alfano and Collmer 2004). The widespread use of the TSS in many bacterial species suggests this mechanism is conserved amongst many prokaryotes (Bhattacharjee, Hiller et al. 2006). However, much less is known about the method of targeting and transporting eukaryotic effectors to their host targets.

The plant response to pathogen infection is as follows; there are two branches of the plant immune system, in the first type transmembrane pattern recognition receptors (PRR) detect slow evolving pathogenic elicitors/PAMPs (pathogen-associated molecular patterns) leading to PAMP triggered immunity (PTI). The second type of defence is by resistance (R) receptor proteins, often membrane-bound, with distinct leucine rich repeat regions (LRR). The R receptor proteins are thought to act as a second line of defence against specific pathogenic effectors released by the pathogen after successfully avoiding PTI. R receptor proteins either directly recognise and interact with effectors, (Ellis, Dodds et al. 2007) or indirectly recognise effectors through monitoring the integrity of the sites of effector targets largely within the plant cell, any interference will result in effector recognition (Jones and Dangl 2006).

H. arabidopsidis is an obligate biotrophic oomycete that has been shown to cause downy mildew in *Arabidopsis thaliana* (Holub and Beynon 1997). Downy mildews are obligate pathogens that are incapable of surviving apart from their hosts. Because *H. arabidopsidis* is the most commonly occurring eukaryotic pathogen of Arabidopsis, it has become one of the two most widely used model pathogens (along with *P. syringae*) for investigating Arabidopsis defence networks. The *H. arabidopsidis* and Arabidopsis interaction system is currently also being developed as a model to determine the mechanisms by which biotrophs manipulate their hosts.

The present concept of pathogenicity between oomycetes and their plant hosts is that the pathogen secretes an array of different effector proteins that interfere with plant defence responses enabling invasion of the host tissue (Huitema, Bos et al. 2004).

This group of secreted effector proteins form part of a general set of secreted proteins that perform different functions both related and unrelated to pathogenicity. The secreted proteins have been defined as the ‘secretome’ of the pathogen. In terms of pathogenicity, the effectors play a vital role in aiding the infection of the host via the targeting of a varied range of effector proteins towards two distinct host cell sites; the cytoplasm and the apoplastic space. These effectors manipulate and interfere with the physiological and biochemical functioning of the host cell not only to disrupt and suppress the host immune response, but in so doing, promote the pathogen’s own ability to infect the host. The effector families can be loosely divided into two groups targeted to either the cytoplasm or in the apoplast. The effectors can be further subdivided into those that promote virulence and those, the avirulence (AVR) proteins, that trigger resistance (*R*) gene mediated host defence responses.

1.8 Cytoplasmic effectors

Many effectors were originally cloned as their presence triggered *R* mediated resistance. (Feys and Parker 2000; Tyler 2002). It is thought that receptors encoded by *R* genes recognise and indirectly interact with the protein products of *Avr* genes through monitoring the integrity of effector targets, to trigger a localised cell death called a Hypersensitive Response (HR) at the point of invasion, thereby, containing the infection.. Many *Avr* genes from the genus *Phytophthora* have been cloned, including *Avrs 1, 2, 3, 4, 10* and 11 in *P. infestans* (van der Lee, Robold et al. 2001), *Avrs 1a, 1b, 1k, 3a, 3b, 3c, 4, 5* and *6* in *P. sojae* (Whisson, Drenth et al. 1995; Gijzen, Forster et al. 1996; Shan, Cao et al. 2004) and *Arabidopsis thaliana* Recognised (*ATR*) 1 and *ATR13* genes from *H. arabidopsidis* (Allen, Bittner-Eddy et al. 2004; Rehmany, Gordon et al. 2005). As all of these AVR proteins are recognised by cytoplasmically located R proteins they are thought to be targeted to the host cytoplasm.

Sequence comparisons were made between the amino acid sequences of *Avr1b-1* from *P. sojae*, *Avr3a* from *P. infestans* and *ATR1* and *ATR13* from *H. arabidopsidis*. Alignment of the sequences revealed a conserved motif at the N terminus region of the 4 genes found within 32 amino acids of a predicted signal peptide cleavage site, and the motif had a consensus sequence of RXLR. This motif was followed by a varying number of amino acids and then a short motif of enriched acidic amino acid residues with a consensus sequence of DEER (Rehmany, Gordon et al. 2005; Tyler, Tripathy et al. 2006). Screening of genomic sequences for any sequences containing the RXLR-DEER returned approximately 350 candidate *Avr*s in each of the *P. sojae* and *P. ramorum* species (Tyler, Tripathy et al. 2006). The RXLR motif is similar to the RXLX (E/Q) motif found in a host targeting signals of many species of malarial parasites, including *Plasmodium falciparum*, and has also been shown to be important in mediating the translocation of *Plasmodium* virulence proteins into the host erythrocytes (Hiller, Bhattacharjee et al. 2004; Bhattacharjee, Hiller et al. 2006). Hence the theory is that RXLRs mediate the translocation of oomycete cytoplasmic effector proteins into the host cell (Armstrong, Whisson et al. 2005; Birch, Rehmany et al. 2006) (Morgan and Kamoun 2007). This theory was later confirmed via an experiment conducted by Whisson and associates (2007) using the *Phytophthora infestans* RXLR-EER-containing protein Avr3a. The protein was used as a reporter for translocation because it triggers RXLR-EER-independent HR upon recognition by plant cells containing the R3a resistance protein. Replacement of Avr3a RXLR-

EER motifs with alanine residues, or with residues KMIK-DDK, which conserves the physicochemical properties of the protein, resulted in the failure of *P. infestans* to deliver either Avr3a or an Avr3a–GUS fusion protein into plant cells. This result demonstrates that these motifs are required for translocation of the secreted cytoplasmic effector proteins into the host cell (Whisson, Boevink et al. 2007). Thus the oomycete *Avr* family of cytoplasmic effectors is defined by their detection by host plant R receptors and by the characteristic RXLR motif in their amino acid sequences.

1.9 Apoplastic Effectors

As many of the classes of apoplastic effector are discussed in more depth in subsequent chapters they will only be discussed briefly here. The Cellulose Binding Elicitor and Lectin-like (CBEL) glycoprotein was shown to be crucial to the *Phytophthora Parasitica var. Nicotiana* zoospore's ability to attach itself to leaf surfaces (Gaulin, Jauneau et al. 2002). CBEL also has a pathogen-associated molecular pattern (PAMP) and thus is recognised by the host, which triggers an HR at the site of infection. GP42, is a glycoprotein found in the cell walls of most *Phytophthora* species and has been shown to trigger HR responses in the leaves of potato and parsley (Nurnberger, Nennstiel et al. 1994; Scheel, Hahlbrock et al. 1995). Part of the host response includes the secretion of an array of enzymes such as glucan, serine and cysteine proteases, chitinases and other hydrolytic enzymes, which act to decompose the protein compounds of the pathogen. Three distinct families of effectors have been identified in *Phytophthora* which specifically act to counteract the degradation caused by these enzymes; firstly, the Glucanase Inhibitor protein

(GIP) family of effectors, which include GIP1 and GIP2 from *P. sojae* and PiGIP1 to 4 from *P. infestans*, that act to inhibit endo-β-1,3 glucanases (Rose, Ham et al. 2002; Bishop, Ripoll et al. 2005; Damasceno, Bishop et al. 2008). Secondly, the Serine protease inhibitor family, which include EPI10 and EPI1, identified in *P. infestans* and both have been shown to interact with and target the same P69B subtilisin-like serine protease from tomato (Tian, Huitema et al. 2004; Tian, Benedetti et al. 2005). Thirdly, there is the cysteine protease family of inhibitors, to which EPIC1 and EPIC2 from *P. infestans* belong (Tian, Benedetti et al. 2005). A family of effector proteins called the Necrosis and Ethylene inducing Protein 1 (NEP1) like effector family, have been shown to induce a necrotic response in their hosts (Pemberton and Salmond 2004; Bae, Bowers et al. 2005; Pemberton, Whitehead et al. 2005).

One of the most easily recognised of all the apoplastic effector families are the small cysteine-rich class of effectors, these effectors are identifiable by the number of cysteines present relative to the sequence length, which is usually less than 150 amino acids (aa) and also by the number of aa’s between each cysteine. Many eukaryotic *Avr* genes, such as *Cladosporium fulvum Avr2*, *Avr4*, and *Avr9*, encode small (<150 amino acids) secreted proteins with a large number of cysteine residues, which can induce defence responses from the host (van't Slot and Knogge 2002). One of the most well studied families of cysteine- rich effectors is the Elicitins (ELI) and Elicitin-like (ELL) effectors. These effectors are pumped into the apoplastic space between the plant cells from the haustoria and are thought to promote infection of the plant through interaction with host cell receptors on the cell surface The ELI domain can be characterised by a highly conserved 96 amino acid domain, containing the six cysteine residue pattern C_1-23-C_2-23-C_3-4-C_4-14-C_5-23-C_6 that forms 3 disulphide bonds (Fefeu, Bouaziz et al. 1997). ELLs, although sharing the six cysteine spacing pattern, are more diverse in the size of the domain and spacing between the cysteine residues, particularly at the C-terminus. To date, other than *Hyaloperonospora arabidopsidis*, the ELI and ELL gene products have only been found in *Phytophthora* and *Pythium* species.

1.10 Using a systems biology approach to model transcriptional networks

Although the analysis of the *H. arabidopsidis* genome has identified numerous potential pathogenicity effectors, still little is known about how these effectors act as a group to infect Arabidopsis and why some isolates of *H. arabidopsidis* can elicit a plant defence response from some accessions of Arabidopsis and others will not be detected at all. The current theory is that different subsets of *H. arabidopsidis* effectors are recognised by different accessions of Arabidopsis. However it is difficult to investigate this, as there are so many *H. arabidopsidis* effectors.

The major challenge in understanding cellular functions from the Arabidopsis genomic sequence is in attempting to explain the principles and mechanisms that govern the observed organism. Although all cells in the organism may contain the same genomic material, they may differ drastically at the protein level due to regulation at the mRNA level. Such mechanisms include RNA transcription, splicing, post-translational modifications and mRNA degradation (Friedman, Linial et al. 2000). One of the major intersections of regulation occurs during transcription where proteins themselves can bind to regulatory sites in the DNA and thus alter the transcription of genes they regulate. Advances in DNA microarray technology, whereby the entire genome of an organism can be represented by either cDNA or oligonucleotide probes, have allowed a quantitative analysis of the abundance of mRNA targets simultaneously (Chang, Li et al. 2005). Further to this the ability of microarrays to capture mRNA expression across the genome at numerous time points has enabled the analysis of the gene expression patterns in order to determine the regulatory relationships between the protein products of the genes and their targets (Chang, Li et al. 2005). Systems biology attempts to do this by combining both experimental and theoretical approaches, with mathematical modelling performing a key role. Mathematics has already been used at many levels from genome sequencing to bioinformatics, to extract statistically relevant patterns from experimental data i.e. BLAST and HMMER (Altschul, Madden et al. 1998; R. Durbin 1998). Thus a mathematical modelling approach may contribute towards understanding the complex signalling networks governing Arabidopsis responses to stress caused by infection with *H. arabidopsidis*.

The main strength of a mathematical modelling approach is the ability to generate experimentally testable hypotheses on the underlying mechanisms of the signalling network as well as providing predictions of novel ones. These predictions can then be tested using perturbation methods such RNA interference (RNAi) of expression of wild type and mutant genes (Breeze, Harrison et al. 2008). The results can then fed back into the mathematical model thereby iteratively producing more refined transcriptional network models that give an insight into the system. A validated model of a signalling network from an organism under perceived normal behavioural conditions will allow one to track changes in the signalling network due to perturbations. Another added benefit of this type of approach is that it is both easier and cheaper to model perturbations that may impact the signalling network (by removing them from the model) than to carry out comparable experiments on the living organism i.e. where many smaller perturbations produce significant effects on the network when combined (Albert 2007).

There have been numerous approaches taken to model the data such as Boolean Networks whereby a genes state can be described by a Boolean flag as being active with a 1 or inactive with a 0. Thus in Boolean networks a genes state can be predicted based upon whether other genes are also active or inactive (Akutsu T 1999). Another approach is to use a dynamic framework to model the data. The use of Dynamic Bayesian Networks (DBN) to model microarray data was first explored by Friedman et al (Friedman, Linial et al. 2000) by using *S. cerevisiae* cell cycle measurements at a single time point (Spellman, Sherlock et al. 1998) using discretised expression data and returned casual relationships between genes thought to initiate cell cycle and its control and demonstrated the usefulness of using a dynamic framework to model the data. These can be categorised as either discrete or continuous. Discrete frameworks treat the data as separate values at each time point whilst continuous frameworks treat the data at each time point as part of a continuum (Friedman, Linial et al. 2000). Thus because there is a lot of variability and noise in most biological systems and because the expression levels at each time point are related, a continuous dynamic framework has the most potential to accurately describe the data (Albert 2007).

One such Dynamic continuous mathematical model designed to reverse engineer transcriptional networks from microarray time course gene expression data has been developed by Beal (Beal, Falciani et al. 2005). Their approach uses a class of DBN known as Linear Dynamical Systems (LDS) otherwise known as Kalman Filters (Kalman 1960) to model the data. The Beal model has extended upon the principles of the original Kalman filter model so that it can accommodate continuous gene expression data using an "output to input" feed forward loop and also model unknown factors that contribute towards explaining the expressed data. The Beal model also uses new sampling techniques to try and improve the accuracy of generated networks. The ability to model unknown factors is crucial as mRNA levels are a complex mix of a variety of events including the rate of transcription and mRNA degradation. Thus the Beal method will be used to try and elucidate the signalling networks governing Arabidopsis responses to infection with *H. arabidopsidis*. The principles behind this model are explained in more depth later in the thesis.

1.11 Aims of the thesis

1. Determine the effector content of *H. arabidopsidis*. In this bioinformatics approach I analysed regions of the genome known to contain pathogen genes up regulated on infection of Arabidopsis. Once the full genome had become available I developed a bioinformatics analysis to identify the RXLR class of effectors in the pathogen genome.
2. Analyse the diversity of the RXLR effector class to determine the extent of diversifying selection occurring at each locus. This enabled possible identification of the effectors that are recognised by host R proteins.

3. Use systems biology network inference techniques to identify the changes in networks of host gene transcription as a consequence of infection by isolates of *H. arabidopsidis.* Studies in aim two assessed the levels of allelic diversity amongst two isolates of *H. arabidopsidis*. Hence, differences in host gene transcription between the isolates will be due primarily to the variation in effector gene content.
4. Use systems biology network inference to model gene networks mediating responses during senescence of Arabidopsis. In this work I predict the involvement of novel transcription factors in the anthocyanin pathways elicited during senescence.

Chapter 2: Annotation of two *H. arabidopsidis* pathogenic regions

2.1 Introduction

H. arabidopsidis is an obligate biotrophic oomycete that has been shown to cause downy mildew in *Arabidopsis thaliana* (Holub and Beynon 1997). Infection occurs after a *H. arabidopsidis* conidium comes into contact with the host leaf and then germinates to produce an appressorium. The appressorium then produces a hypha which penetrates between two of the leaf's epidermal cells. The hyphae then produce nodule-like feeding structures called haustoria, which grow into the epidermal cells as the penetration hypha grows down between them. The further the hyphae grow down into the mesophyll layer of the leaf the more haustoria are produced. Later, sporangia-like structures called conidiophores containing conidiospores then grow out of the stomata of the leaf. The conidia are then released from the conidiophore to commence new rounds of infection (Slusarenko and Schlaich 2003). *H. arabidopsidis* is of interest because it is closely related to *Phytophthora infestans*, a hemi-biotrophic oomycete which is the principal contributory agent of potato late blight (Slusarenko and Schlaich 2003; Kamoun and Smart 2005), *Phytophthora sojae*, the main cause of root and stem rot in soybean (Tyler 2007) and *Phytophthora ramorum*, the causal agent of sudden oak death (Rizzo, Garbelotto et al. 2005). Although *H. arabidopsidis* infection of Arabidopsis does not cause the destruction of a valuable cash crop it is nevertheless considered a useful model for studying the molecular interaction mechanisms that govern the infection of the host by the pathogen because *H. arabidopsidis* is thought to utilise the same methods and 'weaponry' to infect its host.

The present theory on the mechanisms of pathogenicity of oomycetes and their hosts are that the pathogen secretes an array of different effector proteins that interfere with plant defence responses (Huitema, Bos et al. 2004). This has in turn resulted in the evolution of a number of plant defence mechanisms to detect and respond to the effectors. One such mechanism of immunity, in which the plant establishes a heightened state of resistance to pathogen attack after previous exposure, is referred to as Systemic Acquired Resistance (SAR) and acts as a defence against invasion by a broad range of pathogens. In addition to SAR, genetic analyses of disease resistance by researchers has identified a class of genes within Arabidopsis that are dominant and are able to confer resistance (*R*) of the plant to the pathogen. These *R* genes have been shown to confer resistance against a dominant subset of effectors called Avirulence (*Avr*) genes, which are mainly secreted into the cytoplasm of the host cell and are ultimately responsible for whether a particular *R gene* is able to initiate a defence response against a specific pathogen isolate (Feys and Parker 2000; Tyler 2002). It was proposed that receptors encoded by *R genes* recognise and indirectly interact with the protein products of *Avr* genes through the monitoring of their cell targets to trigger a localised cell death called a Hypersensitive Response at the point of invasion, thereby containing the infection. This type of R mediated resistance is also termed ‘gene for gene’ resistance. The theory was subsequently confirmed when it was shown that the *Pi-ta* R gene in rice directly interacted with the secreted protein product of the *Avr-pita*$_{176}$ avirulence gene from the fungal pathogen *Magnaporthe grisea* causing a resistance response (Jia, McAdams et al. 2000). This has consequently lead to the cloning of many R genes from potato (Gebhardt and Valkonen 2001), soybean (Buzzell and Anderson 1992) (Weng, Yu. et al. 2001), tomato (Moreau, Thoquet et al. 1998) and Arabidopsis (Parker, Coleman et al. 1997; Botella, Parker et al. 1998; McDowell, Dhandaydham et al. 1998; Bittner-Eddy, Can et al. 1999). *Avr* genes from species in the related *Phytophthora* genus were also cloned using map based cloning methods including *Avrs 1, 2, 3, 4, 10* and 11 in *P. infestans* (van der Lee, Robold et al. 2001), *Avrs 1a, 1b, 1k, 3a, 3b, 3c, 4, 5* and *6* in *P. sojae* (Whisson, Drenth et al. 1995; Gijzen, Forster et al. 1996; Shan, Cao et al. 2004) and Arabidopsis Thaliana Recognised (ATR) 1 and *ATR13* genes from *H. arabidopsidis* (Allen, Bittner-Eddy et al. 2004; Rehmany, Gordon et al. 2005). The isolation of the *Avr*s in the *Phytophthora* species has revealed that some *Avr*s are clustered together at the same loci such as *Avrs 3, 10* and *11* in *P. infestans* and *Avrs 1a, 1k,* and *4, 6* in *P. sojae*, this raises an interesting question as to whether there are similarly linked ATRs in the *H. arabidopsidis* regions of the already identified *ATR13* gene, this region has yet to be analysed bioinformatically for the presence of other *Avr* genes.

The cloning of several *Avr* genes has allowed researchers the opportunity to determine whether the AVR proteins from related oomycetes share a conservation of their amino acid sequence. Sequence comparisons were made between the amino acid sequences of *Avr1b-1* from *P. sojae, Avr3a* from *P. infestans* and *ATR1* and *ATR13* from *H. arabidopsidis*. Alignment of the sequences revealed no overall sequence similarity between the sequences. However, a conserved motif at the N terminal region of the 4 genes was found within 32 amino acids of a predicted signal peptide cleavage site, and the motif had a consensus sequence of RXLR (single letter amino acid code). This motif was followed by a varying number of amino acids and then a short motif of enriched acidic amino acid residues with a consensus sequence of DEER (Rehmany, Gordon et al. 2005; Tyler, Tripathy et al. 2006). The discovery of a conserved motif common to the avirulence genes of related oomycete species has led to the screening of genomic sequences from *P. sojae* and *P. ramorum* for any sequences containing the RXLR-DEER motif that might indicate the presence of an *Avr* gene. The screening has returned approximately 350 candidate *Avr*s in each of the *P. sojae* and *P. ramorum* species. The conservation of the RXLR-DEER motif across numerous avirulence proteins suggests it is important to the function of these effector proteins (Tyler, Tripathy et al. 2006). The RXLR motif is similar to the RXLX (E/Q) motif found in a host targeting signals of many species of malarial parasites including *Plasmodium falciparum* and has also been shown to be important in mediating the translocation of plasmodium virulence proteins into the host erythrocytes (Hiller, Bhattacharjee et al. 2004). Hence the theory is that RXLRs mediate the translocation of oomycete cytoplasmic effector proteins into the host cell (Birch, Rehmany et al. 2006) (Morgan and Kamoun 2007). The theory was confirmed by Whisson and associates (2007). Replacement of Avr3a RXLR-EER motifs with alanine residues, or with residues KMIK-DDK, resulted in the failure of *P. infestans* to deliver either Avr3a or an Avr3a–GUS fusion protein into plant cells thus demonstrating that the motifs are essential for the translocation of the secreted cytoplasmic effector proteins into the host cell (Whisson, Boevink et al. 2007).

Although a lot of attention has been paid towards the isolation and cloning of *Avr* genes in different oomycete species, efforts to identify other genes associated with pathogenicity within the oomycetes have been undertaken including the use of random EST sequencing to identify genes that are expressed during *P. sojae* and *P.*

infestans infection of their respective hosts (Kamoun, Hraber et al. 1999; Qutob, Hraber et al. 2000). Efforts by Bittner-Eddy and associates (Bittner-Eddy, Allen et al. 2003) to identify genes expressed by *H. arabidopsidis* during infection of Arabidopsis through the use of a Suppression Subtractive Hybridisation (SSH) method yielded a cDNA library of 57 putative *H. arabidopsidis* genes. BAC P1202 contains two of these genes; *Peronospora Parasitica Arabidopsis thaliana* (*Ppat*) *3* and *8* which have similarity to an H+-translocating inorganic pyrophosphatase and a serine/threonine protein kinase, respectively.

The availability of clone sequences in regions enriched for pathogenicity related sequences at this time prior to the release of the first sequenced genome provided an opportunity to study these regions. The focus of this study was to bioinformatically annotate the BAC P1202 region and the locus of the *ATR13* gene in an attempt to locate the presence of any *Avr* genes in these regions based upon the indicative RXLR-DEER motifs, and whether any *Avr* genes found are clustered as has been the case with *Avr*s found in *P. infestans* and *P. sojae*. Based on the clusters of *Avrs* found in both *P. sojae* and *P. ramorum* one would expect to find a similar conservation and clustered distribution of *Avr*s in *H. arabidopsidis*. The aim was also to identify the presence of any other genes related to the pathogenicity of *H. arabidopsidis* in the regions. The study focused on whether the annotated regions of both BAC P1202 and the *ATR13* locus are syntenic with the genomes of *P. sojae*, *P. infestans* and *P. ramorum* and whether there is any conservation of sequence between the regions and the syntenic loci in the 3 *Phytophthora* species.

2.2 Methods

2.2.1 EST and Genome databases

The EST and genomic sequences of *H. arabidopsidis* (version 6), *P. sojae* (version1) and *P. ramorum* (version 1) were obtained from the website of the VBI (Virginia Bioinformatics Institute) http://vmd.vbi.vt.edu/. The Genomic sequence of *P. infestans* (version1) was obtained from the Broad Institute:

http://www.broad.mit.edu/annotation/genome/*Phytophthora*_infestans/.

2.2.2 Annotation of BAC P1202 and the *ATR13* locus

Contigs P12M10-1 and P12M10-2, BACP1417and BAC P1202 were obtained from Peter Bittner-Eddy and Rebecca Allen, Personal communication.. Contigs P12M10-1 and P12M10-2 were assembled with BAC P14P17 into a single read that spanned the *ATR13* region using the SeqMan program from the Lasergene sequence analysis package version 7.0 available at www.dnastar.com. BAC P1202 and *ATR13* locus sequences were then annotated as follows: -

Identification of all open reading frames (ORFs) in the *H. arabidopsidis* genome and subsequent translation to amino acid sequence was performed by EMBOSS version 3.0 application getorf (Rice, Longden et al. 2000). Predicted genes were discovered by running all ORFs through the HMM (Hidden Markov Model) based gene structure prediction programs Fgenesh www.softberry.com, GeneMark (Besemer and Borodovsky 2005), and Glimmer (Delcher, Harmon et al. 1999).

To determine possible functions of the genes, all ORFs and predicted genes were analysed with the Standalone NETBLAST version 2.2.18 BLASTP and TBLASTX programs (Altschul, Madden et al. 1998; Altschul 1999) against the National Centre for Biotechnology Information (NCBI) non redundant databases http://www.ncbi.nlm.nih.gov, and the web based TBLASTX program against the VBI Microbial database containing the annotated genomes of related pathogens *P. sojae* and *P. ramorum* and the Broad Institute containing the annotated genome of *P. infestans*.

To identify any potential secreted proteins, the ORFs and predicted genes were translated into protein sequences and run through the standalone signal peptide predictor program SignalP version 3.0 (Klee and Ellis 2005). The presence of any RXLR, DEER motifs and other effector domains in the ORF dataset were detected using a combination of the customised RXLR identification Perl script gettingRXLRs.pl (Linda Hughes) (Appendix A) written using Perl version 5.6.1 http://www.perl.org, Pfam (Bateman, Coin et al. 2004), Prosite (Falquet, Pagni et al. 2002) and Swissprot (Boeckmann, Bairoch et al. 2003) protein databases. All ORFs and predicted genes were analysed using BLASTN and TBLASTX analysis against *H. arabidopsidis* EST data to determine whether genes were being expressed. The sequences of BAC P1202 and the *ATR13* locus were physically annotated using the ARTEMIS visualisation and sequence analysis tool (Rutherford, Parkhill et al. 2000)

.

2.2.4 Identification of Orthologues and determination of syntenic regions

The regions syntenic to BAC P1202 and the *ATR13* locus were identified through BLASTP analysis of the predicted protein sequences against the whole genomes of *P. sojae*, *P. infestans* and *P. ramorum* using the Standalone BLAST version 2.2.18 (Altschul, Madden et al. 1998; Altschul 1999). The sequences were visualised using the Artemis Comparison Tool (ACT) (Carver, Rutherford et al. 2005).

BLAST was used to find orthologues between the genes of *H. arabidopsidis* regions being investigated and the three *Phytophthora* genomes. The genes from the corresponding syntenic regions of the *Phytophthora* genomes were screened against the genes from BAC P1202 and the *ATR13* locus using the reciprocal BLASTP program BL2SEQ version 2.2.18 (Altschul, Madden et al. 1998; Altschul 1999). Those gene pairs from the syntenic regions of the three genomes that shared reciprocal best BLAST hits, with those from the corresponding BAC P1202 and *ATR13* locus regions, were assigned as being orthologues.

2.3 Results

2.3.1 Annotation of BAC P1202

BAC (Bacterial Artificial Chromosome) P1202 contains the DNA sequence of a region of the *H. arabidopsidis* genome that contains two cDNA clones, *Ppat 3* (Peronospora Parasitica recognition of Arabidopsis Thaliana) and *Ppat 8* identified in the SSH (Suppression Subtractive Hybridisation) generated cDNA Library of genes expressed in *planta* during infection of the host (Bittner-Eddy, Allen et al. 2003). The study of the BAC P1202 region initially focused on determining the presence of any novel genes that are associated with pathogenicity in the region. The study also focused on whether any of the genes identified are also conserved organisationally in other closely related *Phytophthora* species. The region encompassed by BAC P1202 is 143kb in length and a nucleotide sequence BLAST against version 6 of the *H. arabidopsidis* genome shows a 99% match with Scaffold 198 of the assembly.

Analysis of the BAC P1202 region using GeneMark, Fgenesh, and GlimmerM gene predictor programs revealed 16 predicted genes. Further study and identification of the predicted genes via a manual analysis using a BLASTN (nucleotide Basic Local Alignment Search Tool versus a nucleotide database) analysis of the nucleotide sequences against the unigenes in the VBI (Virginia Bioinformatics Institute) microbial database identified ESTs (Expressed Sequence Tag) for 10 of the 16 putative genes (Table 2.1). ESTs were identified on the basis that they had at least a 90% similarity to the predicted gene sequence. The remaining six predicted genes were classified as hypothetical genes as they had no EST hit and had no predicted function. Nine predicted transposable elements were also found in the region.

Table. 2.1. Identification of ESTs for predicted genes on BAC P1202 using BLASTN[b]

ORF, size (aa)	Location on BAC P1202	EST hits, size (aa)	Bit score, *E*-value	Degree of similarity[a]
Protein Kinase, 652	37312-39270	CL5046Contig1, 274,	661, 0.0,	99%, 273,1504-686:824-6,
Glycoprotein, 123	59053-59421	CL22Contig3, 732	306, 3e-84	100%, 122, 1-366:866-1231
Inorganic H+ pyrophosphatase, 773	121470-123791	CL2968Contig1	1387, 0.0	98%,543, 6-1634:1630-2
NAD- Dependent Epimerase, 234	127026-127728	CL2957Contig1, 416	1386, 0.0	99%, 233, 3-701:852-154
Phosphatidylinositol-4-phosphate 5-kinase, 704	129372-131820	Hp_ENSC_34o14, 251	197, e-139	93%, 86, 504-247:502-245
Unknown protein 1, 384	109578-110729	CL4403 contig1, 273	446, e-126	100%, 180, 1150-611:114-653
Unknown protein 2, 313	39212:40153	CL1613Contig1	571, e-163	100%, 98, 648-941:164-457
Unknown protein 3, 223	55900-56568	CL705Contig1	554, e-158	100% 222, 667-2:102-767
Unknown protein 4, 221	114699-115361	CL914Contig1	544, e-155	100%, 221, 661-2:56-715
Unknown protein 5, 425	118046-119320	Hp_ENSC_25c07	603, e-173	100%, 248, 119-862:27-770

BLASTX (translated nucleotide query versus a protein database) and BLASTP (protein query versus protein database) analysis of the 10 *H. arabidopsidis* ESTs that represent the predicted genes against the NCBI (National Centre for Biotechnology Information) non-redundant nucleotide database returned significant similarity scores for five of the predicted genes to genes of known functions in other species (Table 2.2). The scores were deemed significant if their Expect value (E-value) was less than 10, which equates to a probability that one would expect to see 10 matches with a similar score by chance within the entire NCBI database. The molecular functions of the five genes have been defined as follows; *Ppat 3* encodes an H+ translocating inorganic pyrophosphatase, *Ppat 8* a serine/threonine protein kinase, a Phosphatidylinositol-4-phosphate 5-kinase, a Glycoprotein and a NAD- Dependent Epimerase were also identified. The remaining 5 predicted genes were categorised as unknown proteins as their validity was confirmed by ESTs but no predicted functions could be defined. The unknown proteins were assigned a number from 1 to 5 based upon their relative position on the BAC as shown in Figure 2.1 and will subsequently be referred to individually by that number.

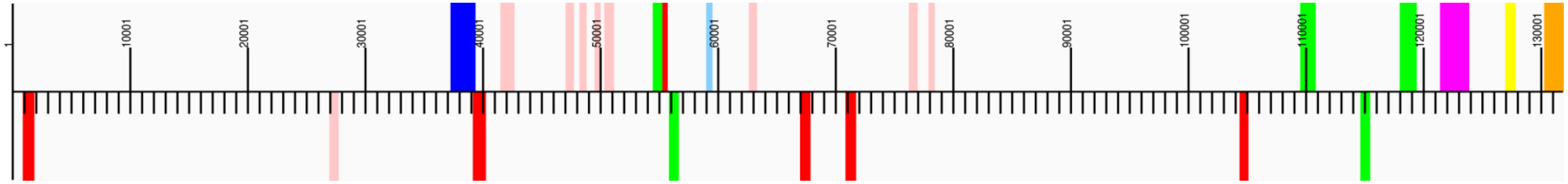
Analysis of the 16 predicted genes using the SignalP program to detect the presence of any signal peptide containing sequences showed no significant hits. These results are consistent with a study of BAC P1202 region for genes whose protein products had predicted signal peptides and an RXLR and DEER motif characteristic of an avirulence protein, using the SignalP program and the Perl script gettingRXLRs.pl. The analysis did not return any sequences containing these motifs. This suggests there are no avirulence proteins present in this region of the *H. arabidopsidis* genome.

Table. 2.2. Similarities of putative ORF products from BAC P1202 to protein sequences in NCBI non redundant database detected using TBLASTX[b]

ORF, size (aa)	ORF location (bp)	Protein Homologue, size (aa)	Degree of similarity[a]	*E*-value	Homologue Function	Organism
Protein Kinase, 652	37312-39270	XP_001506792, 375	51%, 296, 1036-1932:25-276	162 bits (409), Expect = 7e-38	Calcium/calmodulin-dependent protein kinase	*Ornithorhynchus anatinus*
		XP_701499, 476	53%, 307, 1036-1932:10-270	162 bits (409), Expect = 1e-37	Calcium/calmodulin-dependent protein kinase	*Danio rerio*
		XP_683698, 394	51%, 306, 1033-1932:17-276	162 bits (409), Expect = 1e-37	Calcium/calmodulin-dependent protein kinase	*Danio rerio*
		XP_001146086, 490	52%, 307, 1036-1932:11-271	161 bits (408), Expect = 1e-37	Calcium/calmodulin-dependent protein kinase	*Pan troglodytes*
		XP_001146252, 478	52%, 307, 1036-1932:11-271	161 bits (408), Expect = 1e-37	Calcium/calmodulin-dependent protein kinase	*Pan troglodytes*
Glycoprotein endopeptidase, 123	59053-59421	XP_001270107, 377	80%, 118, 19-369:1-118	155 bits (393), Expect = 7e-37	O-sialoglycoprotein endopeptidase	*Aspergillus clavatus*
		NP_194003, 353	80%, 115, 19-369:5-115	155 bits (393), Expect = 7e-37	Glycoprotease M22 family protein	*Arabidopsis thaliana*
		XP_747627, 352	80%, 117, 19-366:1-117	153 bits (386), Expect = 4e-36	O-sialoglycoprotein endopeptidase	*Aspergillus fumigatus*
		XP_001257656, 352	79%, 117, 19-366:1-117	150 bits (380), Expect = 2e-35	O-sialoglycoprotein endopeptidase	*Neosartorya fischeri*
		ABC01899, 346	76%, 115, 19-363:6-116	149 bits (375), Expect = 8e-35	Glycoprotein endopeptidase-like protein	*Solanum tuberosum*
Pyrophosphatase, 773	121470-123791	XP_001415754, 713	76%, 703, 175-2232:18-708	793 bits (2049), Expect = 0.0	H+-PPase family transporter: proton	*Ostreococcus lucimarinus*
		Q06572, 762	70%, 750, 43-2205:77-751	701 bits (1808), Expect = 0.0	Pyrophosphate-energized vacuolar membrane proton pump	*Hordeum vulgare*
		ACA63883, 762	70%, 750, 43-2205:17-751	698 bits (1801), Expect = 0.0	Vacuolar proton-inorganic pyrophosphatase	*Hordeum vulgare*
		XP_001694682, 763	69%, 764, 19-2235:8-756	697 bits (1799), Expect = 0.0	Inorganic pyrophosphatase	*Chlamydomonas reinhardtii*
		CAC44451, 762	69%, 764, 19-2235:8-755	691 bits (1783), Expect = 0.0	Proton-translocating inorganic pyrophosphatase	*Chlamydomonas reinhardtii*
NAD- Dependent Epimerase, 234	127026-127728	NP_579517, 307	57%, 233, 4-702:68-294	161 bits (408), Expect = 3e-38	NDP-sugar dehydratase or epimerase	*Pyrococcus furiosus*
		NP_143580, 306	58%, 233, 4-702:68-294	158 bits (399), Expect = 4e-37	UDP-glucose 4-epimerase	*Pyrococcus horikoshii*
		ABI15605, 306	56%, 233, 4-702: 63-287	155 bits (392), Expect = 2e-36	UDP-glucose 4-epimerase	*Spironucleus barkhanus*
		NP_125996, 307	57%, 233, 4-702:68-294	151 bits (381), Expect = 4e-35	UDP-glucose 4-epimerase	*Pyrococcus abyssi*
		NP_266367, 313	57%, 233, 4-702:73-304	148 bits (374), Expect = 3e-34	UDP-glucose 4-epimerase	*Lactococcus lactis*
Phosphatidylinositol-4-phosphate 5-kinase, 704	129372-131820	XP_001890185, 832	44%, 392, 341-728:341-683	128 bits (322), Expect = 2e-27	Phosphatidylinositol-4-phosphate 5-kinase PIPK5	*Laccaria bicolor*
		XP_001885811, 1119	42%, 408, 341-738:765-1117	127 bits (319), Expect = 3e-27	Phosphatidylinositol-4-phosphate 5-kinase PIPK5	*Laccaria bicolor*
		NP_001051025, 731	44%, 454, 288-721:329-720	122 bits (306), Expect = 1e-25	Phosphatidylinositol-4-phosphate 5-kinase	*Oryza sativa*
		AAC50912, 500	42%, 385, 341-724:126-432	119 bits (299), Expect = 7e-25	Phosphatidylinositol-4-phosphate 5-kinase,	*Homo sapiens*
		Q99755, 562	42%, 385, 341-724:139-445	119 bits (299), Expect = 8e-25	Phosphatidylinositol-4-phosphate 5-kinase	*Homo sapiens*

a The values can be represented as follows: %similarity, length of overlap(aa), length of query (bp), length of subject (bp)
b ORFS were TBLASTX against the NCBI non redundant nucleotide database http://blast.ncbi.nlm.nih.gov/

H. arabidopsidis BAC P1202 (143 kb)



Key

🟦 *Ppat 8*: Serine protein Kinase

🟪 *Ppat 3*: H+ translocating inorganic pyrophosphatase

🟨 NAD- dependent epimerase

🟧 Phosphatidylinositol-4-phosphate 5 kinase

🟦 Glycoprotein

🟥 Hypothetical protein

🟩 Unknown protein

🟥 Transposable element

Figure. 2.1. A diagram representing the 143 kb region of BAC P1202 and the genes found within it. The different types of genes are characterised by the coloured bars along the length of the diagram and can be identified using the key. The region is not drawn to scale; each scaling unit below the horizontal line of the diagram represents 1 kb and above the horizontal line represents 10 kb.

2.3.2 BAC P1202 synteny with *P. sojae*, *P. infestans* and *P. ramorum*

TBLASTX (Translated nucleotide Basic Local Alignment Search Tool versus a translated nucleotide database) and BLASTP analysis of BAC P1202 against version 3 of the *P. sojae* assembly, version 1 of the *P. infestans* assembly and version 1 of the *P. ramorum* assembly show there is synteny with scaffold 16 of. *P. sojae*, super contig 7 of *P. infestans* and scaffold 2 of *P. ramorum*. Comparative analysis of the BAC with the 3 *Phytophthora* scaffolds revealed that the high degree of similarity between the sequences corresponds to the positions of many of the genes found on BAC P1202. The high level of similarity indicates possible orthology between the genes. An analysis to identify possible orthologous genes in the 3 *Phytophthora* scaffolds to those found in BAC P1202 was performed using a reciprocal TBLASTX and BLASTP. The syntenic regions of all 3 scaffolds were annotated for function using the same method used to annotate BAC P1202 as described earlier.

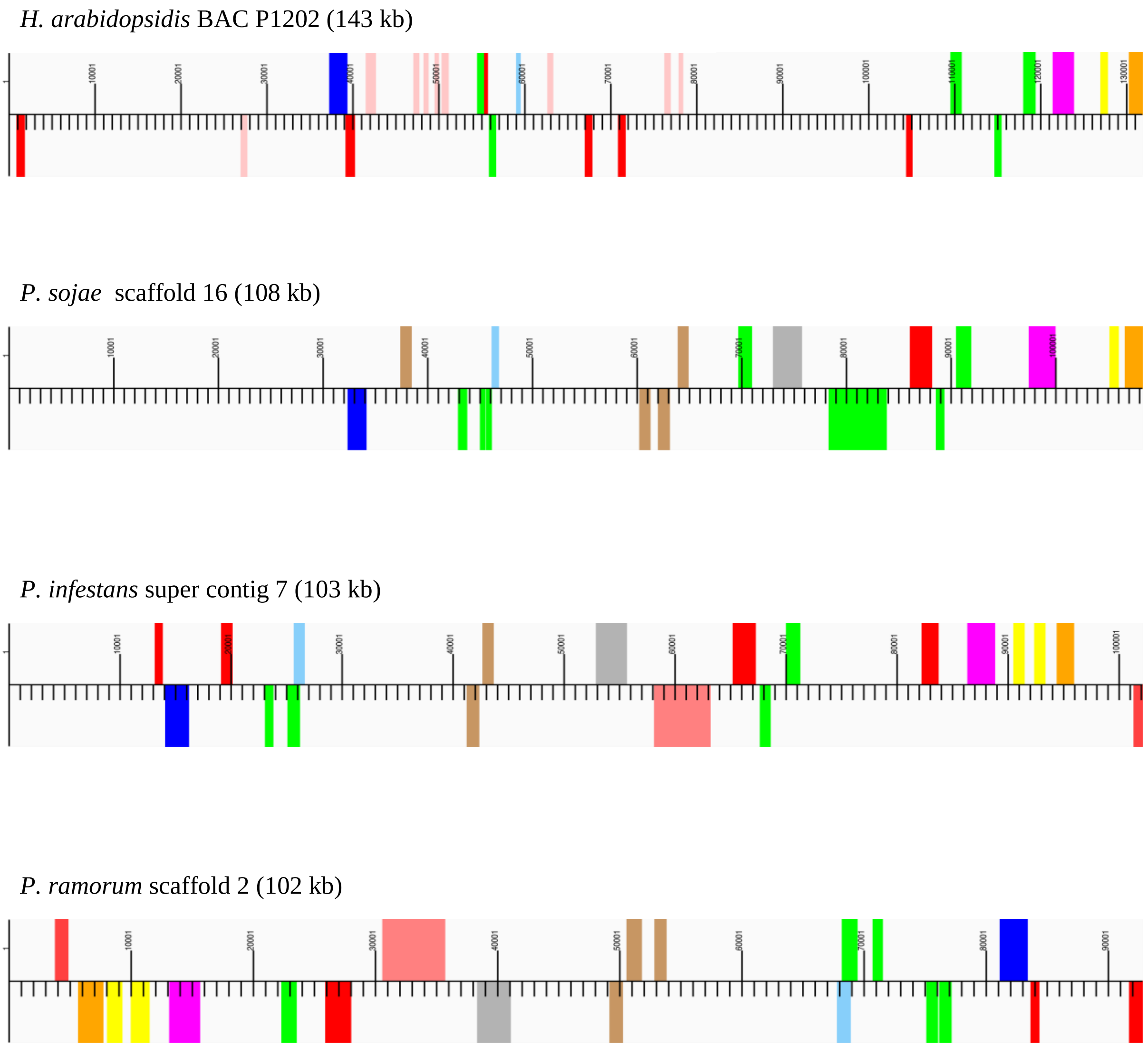
The syntenic locale of *P. sojae* scaffold 16 spans approximately 108kb region of genomic sequence from positions 1450240 to 1558601, and contains 19 annotated genes. On scaffold 2 of *P. ramorum* the syntenic region is around 102kb in length, from positions 504701 to 604542 and contains 20 annotated genes. Super contig 7 of *P. infestans* contains a syntenic region that is approximately 103kb in length, from positions 3946033 to 4049468 and contains 18 predicted genes. Of the genes found on BAC P1202, 10 can be classified as having orthologues in all 3 syntenic regions on the scaffolds of *P. sojae*, *P. infestans* and *P. ramorum* based on reciprocal best BLAST hits (Table 2.3). The 10 genes that have orthologues include *Ppats 3* and *8*, the NAD- Dependent Epimerase, the Phosphotidylinositol-4-phosphate 5-kinase, a Glycoprotein and 5 genes that code for proteins of unknown function. Interestingly, annotation of the 3 scaffolds did not reveal the presence of any transposable elements in these regions. Annotation of the syntenic regions of the 3 *Phytophthora* scaffolds did not return any RXLR containing sequences that would indicate the presence of an RXLR effector.

Analysis of the gene order and orientation between BAC P1202 and the syntenic regions of the 3 scaffolds reveal differences between the species. The primary difference between the syntenic regions and BAC P1202 was the absence of an extradiol ring-cleavage dioxygenase, a membrane occupation and recognition nexus MORN repeat protein and a Formin like protein 20 from the BAC region. For the syntenic region on scaffold 16 of *P. sojae*, eight of the 10 orthologue pairs show the same gene order and gene orientation however 2 genes, the orthologue to unknown protein 2 and the orthologue to *Ppat 8*, show a reverse orientation to that of BAC P1202. Similarly, for the syntenic region on super contig 7 of *P. infestans,* 8 of the 10-orthologue pairs showed the same gene order and orientation but again 2 genes, the orthologue to unknown protein 2 and the orthologue to *Ppat 8,* show a reversed orientation to that of BAC P1202. In contrast, the syntenic region on scaffold 2 of *P. ramorum* shows that the orientation of 8 of the 10 orthologue pairs has also been reversed when compared to BAC P1202. The 2 genes that did not show a reversed orientation were unknown protein 2 and the orthologue to *Ppat8.*

Table. 2.3. Similarities of BAC P1202 ORF products to potential orthologous genes in *P. sojae, P. infestans* and *P. ramorum* detected using TBLASTX[b]

ORF, size (aa)	**Location on BAC P1202**	**Orthologue transcript ID**	**Orthologue location**	**Degree of similarity[a]**	**Organism**
Protein Kinase, 652	37312-39270	129911	Scaffold 16:1482652-1484328	85%, 224, 1457-786:1586-915	*P. sojae*
		PITT_05587	Supercontig 7: 3960153-3962439	87%, 653, 1-1956:48-680	*P. infestans*
		73132	Scaffold_2:588946-591126	93%, 696, 745-1959:874-2088	*P. ramorum*
Glycoprotein, 123	59053-59421	108428	Scaffold_16:1496410-1496982	92%, 113, 28-366:1-339	*P. sojae*
		PITG_05594	Supercontig 7: 3972215-3977380	90%, 122, 1-366:1-366	*P. infestans*
		71077	Scaffold_2:574497-576065	92%, 113, 28-366:1-339	*P. ramorum*
Inorganic H+ pyrophosphatase, 773	121470-123791	108429	Scaffold 16:1547735-1550149	98%, 730, 1-2190:1-2190	*P. sojae*
		PITG_05615.1	Supercontig 7: 4032454-4035014	90%, 748, 1- 2244:1-2244	*P. infestans*
		71076	Scaffold_2:520864-523253	98%, 785, 1-2229:1-2229	*P. ramorum*
NAD- Dependent Epimerase, 234	127026-127728	129935	Scaffold 16:1553365-1556195	96%, 234, 1-702:379-1080	*P. sojae*
		PITG_05617	Supercontig 7: 4038045-4039444	96%, 234, 1-702, 127-360	*P. infestans*
		73107	Scaffold_2:515857-517139	96%, 1-702:379-1080	*P. ramorum*
Phosphatidylinositol-4-phosphate 5-kinase, 704	129372-131820	129936	Scaffold 16:1556906-1558513	95%, 202, 1-606:949-1554	*P. sojae*
		PITG_05618	Supercontig 7: 4040484-4042893	85%, 205, 1-1320:951-2265	*P. infestans*
		73106	Scaffold_2:513174-515114	94%, 206, 1-618:925-1542	*P. ramorum*
Unknown protein 1, 384	109578-110729	-	Scaffold_16:1523286-1525934	78%, 306, 13-930:43-960	*P. sojae*
		PITG_05605	Supercontig 7: 3998982-4001630	74%, 374, 7-1116:33-1152	*P. infestans*
		93495	Scaffold_2:548943-552135	81%, 238, 220-1140:79-969	*P. ramorum*
Unknown protein 2, 313	39212:40153	155906	Scaffold_16:1481731-1482424	68%, 136, 9-939:1-694	*P. sojae*
		PITG_05586	Supercontig 7:3959213-3959815	61%, 122, 9-588:51-600	*P. infestans*
		93499	Scaffold_2:591402-592127	68%, 140, 9-588:42-723	*P. ramorum*
Unknown protein 3, 223	55900-56568	108426	Scaffold_16:1493198-1493947	71%, 70, 662-453:668-459	*P. sojae*
		PITG_05591.1	Supercontig 7: 3969146-3969775	60%, 194, 82-663:5-188	*P. infestans*
		71079	Scaffold_2:578621-579262	83%, 92, 388-663:364-639	*P. ramorum*
Unknown protein 4, 221	114699-115361	129929	Scaffold_16:1538854-1539534	76%, 170, 1-660:4-669	*P. sojae*
		PITG_05609	Supercontig 7: 4013729-4014577	75%, 222, 4-660:426-1077	*P. infestans*
		73114	Scaffold_2:532567-533259	72%, 161, 97-522:529-954	*P. ramorum*
Unknown protein 5, 425	118046-119320	155912	Scaffold_16:1540776-1542098	84%, 307 91-1011:1-921	*P. sojae*
		PITG_05610.1	Supercontig 7: 4016119-4017250	66%, 353, 91-1149:1-1011	*P. infestans*
		73113	Scaffold_2:530134-531270	78%, 333, 91-1089:1-999	*P. ramorum*

a The values can be represented as follows: %similarity, length of overlap(aa), length of query (bp), length of subject (bp).
b ORFS were TBLASTX against Transcripts of P. sojae, P. infestans and P. ramorum from the VBI .
Note: *P = Phytophthora*.



🟦 *Ppat 8*: Serine protein Kinase

🟪 *Ppat 3*: H+ transocating inorganic pyrophosphatase

🟨 NAD- dependent epimerase

🟧 Phosphatidylinositol-4-phosphate 5 kinase

⬜ MORN repeat protein

🟥 Formin like protein 20

🟦 Glycoprotein

🟥 Hypothetical protein

🟩 Unknown protein

🟥 Transposable element

🟫 Extradiol ring-cleavage dioxygenase

Figure. 2.2. A diagram representing the genes found in *H. arabidopsidis* BAC P1202 and its syntenic regions in *P. sojae, P. infestans* and *P. ramorum*. The different types of genes are characterised by the coloured bars along the length of the diagram and can be identified using the key. The regions are not drawn to scale; each scaling unit below the horizontal line of the diagram represents 1 kb and above the horizontal line represents 10 kb.

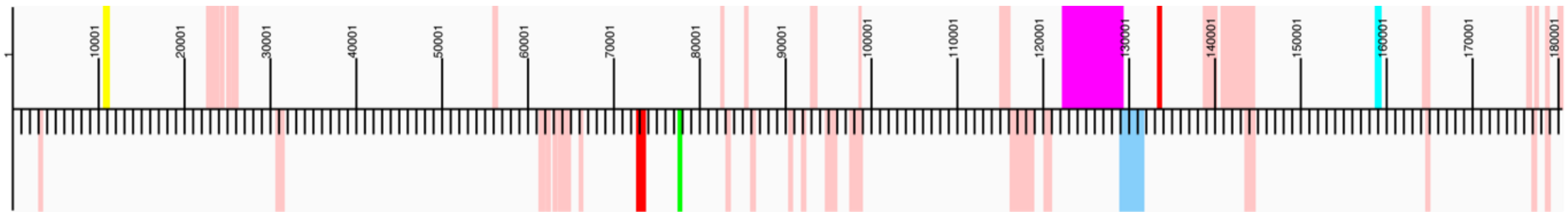
2.3.3 Annotation of *ATR13* region

Two contigs, P12M10-1 and 2, spanning the *ATR13* avirulence gene region were assembled with BAC P14P17 which covers the downstream or 3' region of the gene, into a single read that spans the *ATR13* locus. The assembled locus is 190 kb in length. Annotation of the region revealed 7 predicted genes (Figure 2.3). Of the genes found, three genes were annotated for function, an acetyl CoA carboxylase at positions 122293 to 129240, an Acyl transferase between positions 128979 and 131663, and a GTP-binding protein at positions 158727 to 159287 (Table 2.4). *ATR13* was located between positions 10615 to 11181. Of the remaining 3 predicted genes, 2 were classified as hypothetical genes as they had no known function and no associated ESTs and one was defined as an unknown protein as it had an EST but no known function (Table 2.5). Some 56 predicted transposable elements have also been identified in this region; interestingly, the genes are grouped into large clusters along the entire length of the locus. The process of annotation revealed the presence of 2 inverted identical 61 kb repeat regions; the first between positions 15287 and 72357 on the forward stand and the second between 15287 and 72214 on the reverse strand. It was thought that the repeat regions could have been artefacts as the result of a sequence assembly error. To test this, primers were designed in the gap between the repeat regions and a PCR reaction was carried out to determine if the region returned a single PCR product 2 kb in length. The product obtained matched the size expected, and the repeat was therefore deemed real and not part of an assembly error.

Analysis of the *ATR13* locus was carried out using the same method used to analyse the BAC P1202 region for any candidate effector sequences containing the distinctive RXLR pattern and signal peptide as described earlier. The analysis returned no other candidate genes in this region. Subsequent analysis to find other classes of effector genes also returned no candidate genes in this region.

A nucleotide BLAST analysis with version 6 of the *H. arabidopsidis* genome revealed the locus had strong hits to both scaffolds 43 and 118. Further analysis using ACT (Artemis Comparison Tool) showed there was strong similarity between scaffold 43 and the downstream region of the *ATR13* locus from nucleotide positions 58571 to 189288 of the locus as shown in Figure 2.4. Conversely, scaffold 118 has strong similarity toward the upstream region of the locus containing the *ATR13* gene from nucleotide positions 1 to 72214. Although the *ATR13* locus has similarity to the 2 different scaffolds, there is a 4.7 kb region, from nucleotide positions 58571 to 63275, where the similarity between the 2 scaffolds and the *ATR13* locus overlap. The notable lack of similarity between the locus and the upstream region of scaffold 43 from nucleotide positions 291722 to 433615, suggests there has been a miss-assembly in scaffold 43 at these positions and that scaffolds 43 and 118 should in fact belong to the same scaffold.

H. arabidopsidis ATR13 locus (190 kb)



Key

- 🟨 *ATR13*
- ■ Transposable element
- 🟪 Acetyl coA carboxylase
- 🟦 Acyl transferase
- ■ GTP Binding protein
- 🟥 Hypothetical protein
- 🟩 Unknown protein

Figure. 2.3. A diagram representing the 190 kb region of the ATR13 locus and the genes found within it. The different types of genes are represented by the coloured bars along the length of the diagram and can be identified using the key. The region is not drawn to scale, each scaling unit below the horizontal line of the diagram represents 1 kb and above the horizontal line 10 kb.

Table. 2.4. Identification of ESTs for predicted genes From the *ATR13 locus* using BLASTN[b]

ORF, size (aa)	Location on BAC P1202	EST hits, size (aa)	Bit score, *E*-value	Degree of similarity
ATR13, 189	10615-11181	HpRXLR132*	NA	NA
Acetyl coA carboxylase, 2316	122293-129240	Hp_MCc_01H01**	875, 0.0	99%, 876, 5450-6324:1-875
Acyl transferase, 895	128979-131663	Hp_ENSC_20l02	747, 0.0	100%, 301, 1005-1907:2-904
GTP binding protein, 187	158727-159287	CL4779Contig1	404, e-113	98%, 167, 22-522:766-266

* Sequenced at HRI, amplified from cDNA using gene specific primers
** EST obtained from Mary Coates EST library
a The values can be represented as follows: %similarity, length of overlap(aa), length of query (bp), length of subject (bp)
b ORFS were BLASTN against Hp Unigenes from the VBI http://vmd.vbi.vt.edu/

Table. 2.5. Similarities of ORF products from the *ATR13* locus to protein sequences in NCBI non-redundant database detected using TBLASTX[b]

ORF, size (aa)	**ORF location (bp)**	**Protein Homologue, size (aa)**	**Degree of similarity[a]**	***E*-value**	**Homologue Function**	**Organism**
Acetyl-Coenzyme A carboxylase, 2316	122293:129240	NP_990836, 2324	59%, 2311, 121-6921:101-2315	1647 bits (4264), Expect = 0.0	Acetyl-Coenzyme A carboxylase	*Gallus gallus*
		S41121, 2339	59%, 2311, 121-6921:101-2330	1642 bits (4253), Expect = 0.0	Acetyl-CoA carboxylase	*Homo sapiens*
		XP_001371374, 2391	58%, 2312, 121-6921: 146-2383	1639 bits (4245), Expect = 0.0	Similar to Acetyl-Coenzyme A carboxylase	*Monodelphis domestica*
		XP_867576, 2323	58%, 2315, 121-6921: 101-2315	1632 bits (4227), Expect = 0.0	Similar to Acetyl-Coenzyme A carboxylase	*Canis familiaris*
		EDL15735, 2379	59%, 2294, 121-6906:101-2290	1632 bits (4226), Expect = 0.0	Acetyl-Coenzyme A carboxylase	*Mus musculus*
Glycerol-3-phosphate O-acyltransferase, 895	128979:131663	XP_569487 , 755	52%, 592, 286-1959:2-568	295 bits (755), Expect = 1e-77	Glycerol-3-phosphate O-acyltransferase	*Cryptococcus neoformans*
		XP_001874841, 641	51%, 581, 316-1959:2-557	291 bits (745), Expect = 2e-76	Glycerol-3-phosphate O-acyltransferase	*Laccaria bicolor*
		XP_817088, 700	52%, 557, 319-1965:142-664	271 bits (693), Expect = 2e-70	Glycerol-3-phosphate acyltransferase, putative	*Trypanosoma cruzi*
		EDP50502, 743	49%, 594, 331-1959:26-592	262 bits (670), Expect = 9e-68	Glycerol-3-phosphate acyltransferase	*Aspergillus fumigatus*
		XP_751693, 743	49%, 594, 331-1959:26-592	262 bits (670), Expect = 9e-68	Glycerol-3-phosphate acyltransferase	*Aspergillus fumigatus*
GTP-binding protein, 186	158727:159287	XP_711509, 214	92%, 167, 22-522:1-167	307 bits (786), Expect = 3e-82	Ran, Ras family GTP-binding protein	*Candida albicans*
		NP_001040274.1, 213	91%, 167, 22-522:1-167	306 bits (784), Expect = 5e-82	GTP-binding nuclear protein Ran	*Bombyx mori*
		XP_001900408, 215	93%, 164, 31-522:6-169	306 bits (784), Expect = 5e-82	GTP-binding nuclear protein RAN/TC4	*Brugia malayi*
		XP_001527056.1, 215	92%, 166, 25-522:3-168	306 bits (784), Expect = 5e-82	GTP-binding nuclear protein GSP1/Ran	*Lodderomyces elongisporus*
		CAE55862, 215	92%, 165, 28-522:5-169	306 bits (784), Expect = 5e-82	GTP-binding nuclear protein RAN1	*Chironomus tentans*

a The values can be represented as follows: %similarity, length of overlap(aa), length of query (bp), length of subject (bp)
b ORFS were TBLASTX against the NCBI non redundant nucleotide database http://blast.ncbi.nlm.nih.gov/

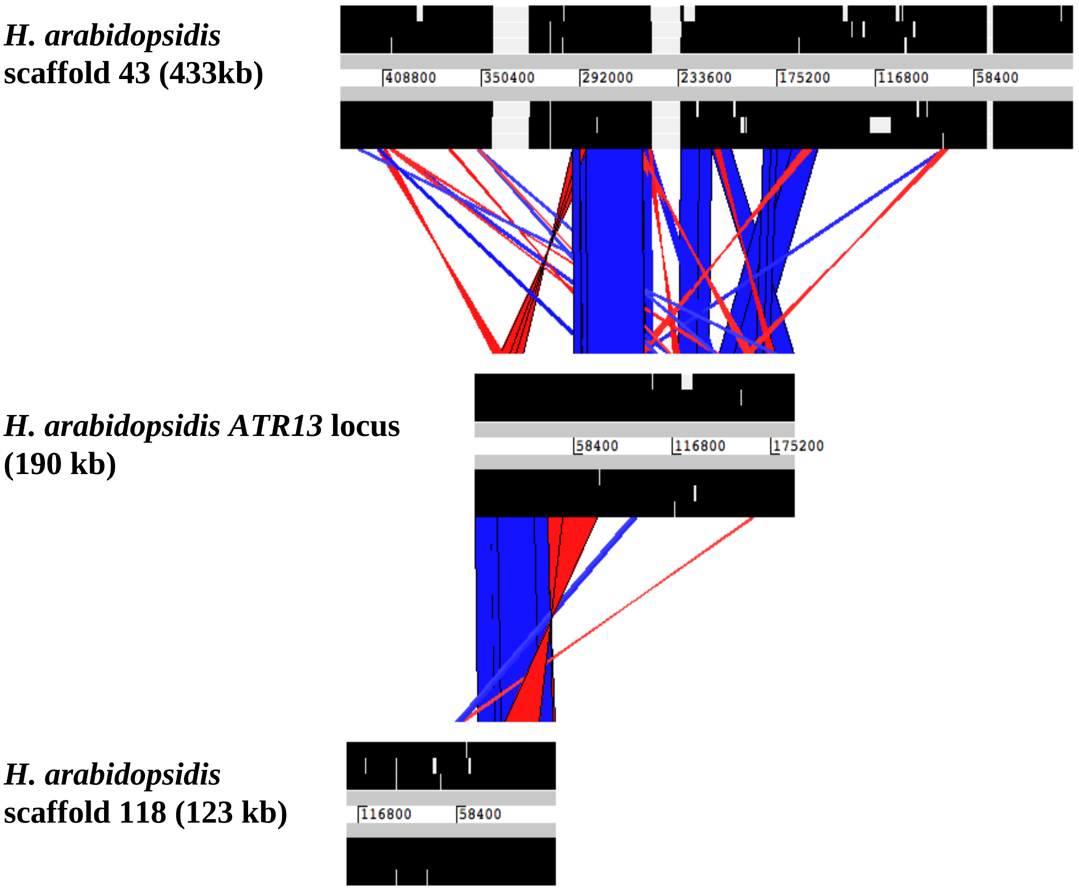


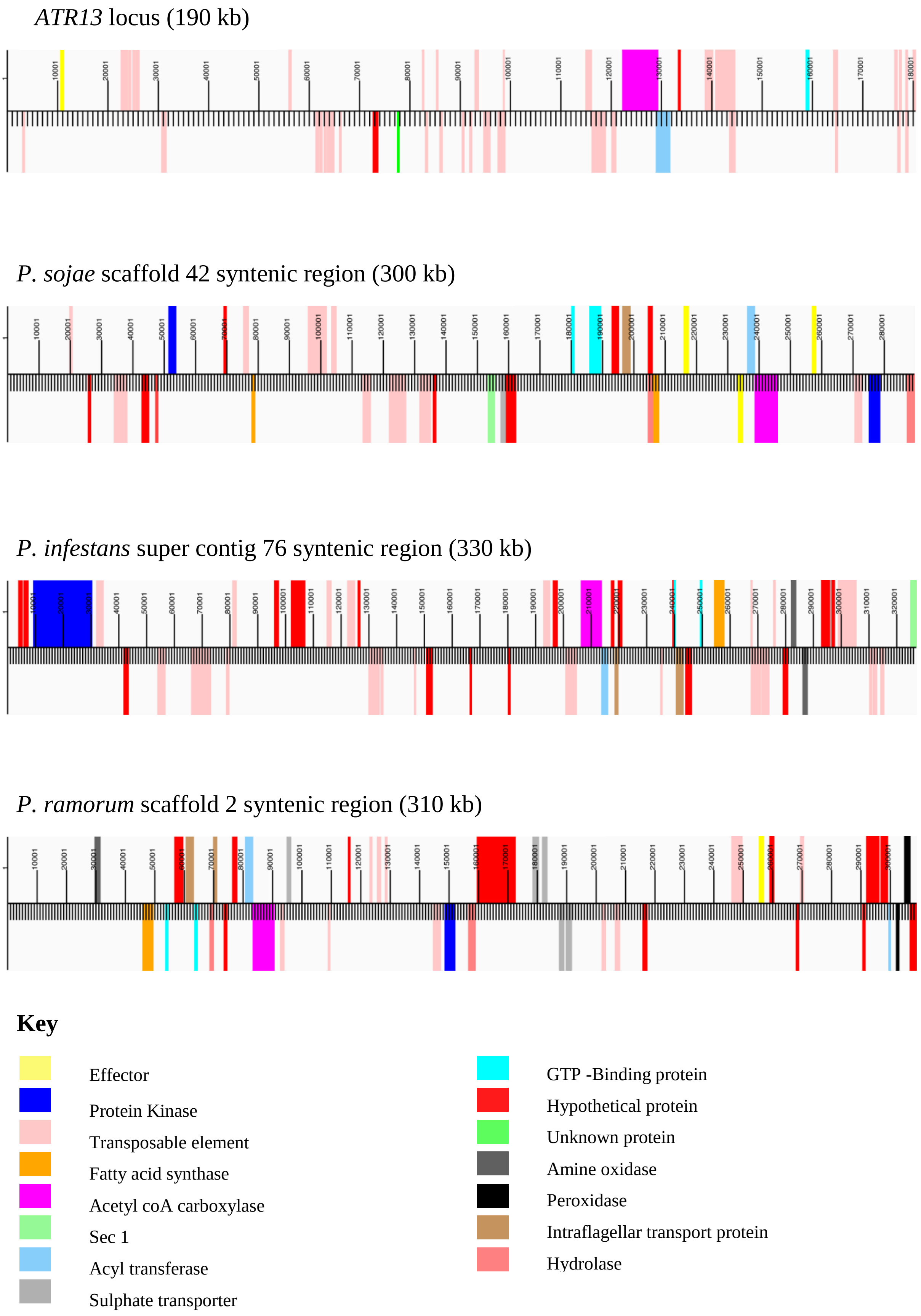
Figure. 2.4. Level of similarity between the sequences of the *ATR13* locus and scaffolds 43 and 118 of *H. arabidopsidis* version 6 using the Artemis comparison tool (ACT).·The blue lines represent BLAST hits between the sequences; the red lines show where the hit sequence has been flipped. Hits were considered significant if the BLAST score exceeded $^{e-5}$. The grey lines represent the nucleotide sequence, the black lines above it represent the 3 forward translational frames and the black lines below the grey line represent the 3 reverse translational frames and the white sections within it represent the open reading frames present and also un-sequenced regions of the scaffolds.

2.3.4 Synteny between the ATR13 locus and Phytophthora species

TBLASTX and BLASTP analysis of the *ATR13* locus with version 1 of all 3 genomes of *P. sojae*, *P. infestans* and *P. ramorum* have revealed synteny with scaffolds 42 of *P. sojae*, 76 of *P. infestans* and 2 of *P. ramorum.* Subsequent annotation of the syntenic regions of each *Phytophthora* scaffold has revealed that the syntenic region of *P. sojae* scaffold 42 was approximately 300 kb in length from nucleotide positions 1 to 300000 of the scaffold and contained 24 predicted genes and 22 predicted transposable elements. Analysis of the syntenic region for any possible RXLR effectors returned no candidates, however BLASTP analysis of the region did highlight the presence of 4 predicted Crinkler (CRN) proteins at positions 216017 to 217411, 257020 to 258090, 233389 to 233823 and 233824 to 234699. The genes were found to have similarity to classes 7, 8 and 9 of the CRN family respectively yet no orthologous CRN genes were found in the *ATR13* locus. The syntenic region of scaffold 76 of *P. infestans* spans a 330 kb region from bases 1 to 330000 and contains 35 genes and also 44 transposable elements. Investigation of the region for any effectors returned no candidates. The syntenic locale of scaffold 2 of *P. ramorum* encompasses a 310 kb region, from bases 1 to 309551 of the scaffold. The region contains 49 predicted genes and 13 transposable elements. Analysis of the region for effector candidates returned a single CRN8 –like predicted protein at base positions 255443 to 256879 of the syntenic region. No candidate RXLR effectors were identified in this region.

Of the predicted genes found in the *ATR13* region, only 3 genes, an acetyl CoA carboxylase, an acyl transferase and a GTP-binding protein, were found to be orthologous to the genes found in the syntenic region of all 3 *Phytophthora* scaffolds based on reciprocal best BLAST hits. Comparative analysis using a reciprocal BLASTP also revealed a low level of similarity between the *ATR13* locus and the 3 *Phytophthora* scaffolds in that synteny was limited to the positions of the 3 orthologous genes found in the locus (Table 2.6). No orthologous RXLR effectors were found between the syntenic regions of the 3 *Phytophthora* scaffolds and *ATR13* of the *ATR13* locus.

Analysis of the gene order and orientation between the Acetyl CoA Carboxylase, Acyl transferase and GTP-binding protein genes in the *ATR13* locus and their orthologue hits in syntenic regions of the 3 *Phytophthora* scaffolds reveal differences. The 3 *P. infestans* orthologue pairs were found to have the same gene orientation and gene order to that of the *ATR13* locus (Figure 2.5) however, the 3 orthologue pairs for both *P. sojae* and *P. ramorum* show a reversed gene orientation to that of the *ATR13* locus.



ATR13 locus (190 kb)

P. sojae scaffold 42 syntenic region (300 kb)

P. infestans super contig 76 syntenic region (330 kb)

P. ramorum scaffold 2 syntenic region (310 kb)

Key

Figure**.** 2.5**.** A diagram representing the genes found in the *ATR13* locus of *H. arabidopsidis* and its syntenic regions in *P. sojae*, *P. infestans* and *P. ramorum*. The different types of genes are characterised by the coloured bars along the length of the diagram and can be identified using the key. The yellow regions represent effector genes in general but in *H. arabidopsidis* ATR13 locus it refers to the ATR13 gene and *P. sojae* and *P. ramorum* the yellow lines represent CRN genes. The regions are not drawn to scale; each scaling unit below the horizontal line of the diagram represents 1 kb and above the horizontal line represents 10 kb.

Table. 2.6. Similarities of *ATR13* locus ORF products to potential orthologous genes in *P. sojae*, *P. infestans* and *P. ramorum*

ORF, size (aa)	Location on BAC P1202	Orthologue Transcript gene ID	Orthologue location	Degree of similarity	Organism
ATR13, 189	10615-11181	-	-	-	-
Acetyl coA Carboxylase, 2316	122293-129240	108918	Scaffold_42:238651-245670	96%, 1150, 3499-6948:3484-6933	P. sojae
		PITG_18706.1	Supercontig 76: 206693-214742	95%, 2322, 1-6945:27-6978	P. infestans
		81751	Scaffold_63:83434-90414	93%, 1069, 1-3207:25-3231	P. ramorum
Acyl Transferase, 895	128979-131663	135997	Scaffold_42:236186-238525	92%, 690, 190-2259:119-2188	P. sojae
		PITG_18707	Supercontig 76: 213816-216117	86%, 690, 190-2259:93-2061	P. infestans
		81750	Scaffold_63:80962-83231	94%, 635, 190-2094:91-1995	P. ramorum
GTP Binding protein, 187	158727-159287	108912	Scaffold_42:179771-181170	98%, 167, 22-522:49-549	P. sojae
		PITG_18718	Supercontig 76: 248941-250327	98%, 167, 22-522:21-519	P. infestans
		51857	Scaffold_63:53606-54326	98%, 25-522:4-501	P. ramorum

a The values can be represented as follows: %similarity, length of overlap(aa), length of query (bp), length of subject (bp)
b ORFS were TBLASTX against transcripts of *P. Sojae, P. Infestans, P. Ramorum* from the VBI
Note: *P = Phytophthora*

2.4 Discussion

2.4.1 Annotation of BAC P1202

Annotation of the BAC P1202 region confirmed the presence of *Ppat*s *3* and *8* as identified in the SSH cDNA library of *H. arabidopsidis* genes expressed during infection of the host plant *Arabidopsis thaliana* (Bittner-Eddy, Allen et al. 2003). BLAST analysis of *Ppat*s 3 and 8 revealed similarity to a H+ translocating inorganic pyrophosphatase and a serine/threonine protein kinase respectively. The investigation also identified an NAD- dependent epimerase, a Phosphatidylinositol-4-phosphate 5 kinase (PI4P5K) and a Glycoprotein in the region as well as 5 unknown proteins, 6 hypothetical proteins and 9 transposable elements.

The predicted functions of the identified genes suggest they do not play a direct role in the pathogenicity of *H. arabidopsidis*. *Ppat 3* has been shown to have similarity to a H+ translocating inorganic pyrophosphatase, an enzyme that act as H+ pumps using inorganic pyrophosphate as an energy source instead of ATP. This particular enzyme has been found in the genomes of numerous *Archea, Eubacteria* and *Eukarya* (Drozdowicz and Rea 2001) suggesting that that this particular enzyme is highly conserved and is thought to be required for the regulation of numerous essential cellular processes within *H. arabidopsidis* and is therefore unlikely to have any direct functions related to pathogenicity. This can also be said of the NAD-dependent epimerase that has been shown to use a sugar substrate to catalyse a diverse range of reactions and is conserved across numerous different species implying that it carries out essential “house keeping” duties.

In contrast, *Ppat 8* may have a more direct role in pathogenicity as it has particular similarity to the Ca^{2+} calmodulin-regulated class of serine/threonine kinase. A study focusing on kinase genes that were differentially expressed during zoosporogenesis in *P. infestans* highlighted the presence of a gene whose protein product resembled a Ca^{2+} and calmodulin-regulated protein kinase; the gene product was subsequently shown to be induced during sporangial cleavage leading to the formation of zoospores in *P. infestans* (Judelson and Roberts 2002). This suggests a possible role for *Ppat 8* in oospore formation in *H. arabidopsidis* that would be consistent with an increase in expression during infection of the Arabidopsis. Glycoproteins may also have a more direct role in pathogenicity as conidia from *H. arabidopsidis* have been shown to secrete glycoproteins along with β-1,3-glucans and polysaccharides to form an extracellular matrix which then mediate the attachment of spores, appressoria and germ tubes to plant surfaces as well as providing protection against environmental stresses (Carzaniga, Bowyer et al. 2001).

PI4P5K however, may have a more indirect link to pathogenicity. PI4P5K was found to have similarity to a G-protein coupled receptor phosphatidylinositol-4-phosphate 5 kinase (GPCR-PI4P5K) in *P. sojae*. The heterotrimeric G-protein pathway has been acknowledged as an essential regulator of development and physiology in plant-pathogenic fungi as it regulates numerous signal transduction cascades (Lengeler, Davidson et al. 2000). However, this particular class of GPCR contains a PIPK domain and is found uniquely in *Phytophthora* species. Some 12 GPCR-PIPKs have so far been annotated in *Phytophthora*, this high number suggests they have an important signalling role within the *Phytophthora* species but as yet there have been no studies to investigate the downstream signalling pathway of the protein (Meijer and Govers 2006; Tyler, Tripathy et al. 2006). However a study which silenced the PIGb1 Gα subunit of a G-protein in *P. infestans* resulted in a defect in sporangium formation (Latijnhouwers, de Wit et al. 2003). Hence, one possible function could be that GPCR-PIPKs play a role in the regulation of development of conidia in *H. arabidopsidis.*

The predicted function of some of the genes identified in the sequence of BAC P1202 may indicate involvement in both the formation of structures and the regulation of processes, which contribute towards the pathogenicity of *H. arabidopsidis*. This may go some way towards explaining why *Ppat*s 3 and 8 are up regulated during infection of the host. It has also been shown that all of these identified genes have been conserved in the syntenic regions of *P. sojae*, *P. ramorum* and *P. infestans*, suggesting a conserved role.

2.4.2 Co-linearity Between BAC P1202 and syntenic regions in *Phytophthora*

Comparative analysis between BAC P1202 and the genomes of *P. sojae*, *P. infestans* and *P. ramorum* revealed synteny with scaffold 16, Supercontig 7 and scaffold 2 respectively (Figure 2.2). The syntenic regions of the 3 *Phytophthora* scaffolds show a high level of conservation in terms of gene order and an orientation with each other despite the change in gene orientation seen in *P. ramorum*. This high level of co-linearity is in keeping with the work done by Tyler and associates (Tyler, Tripathy et al. 2006) on the genomes of *P. sojae* and *P. ramorum* which suggests that nearly all predicted genes between the 2 genomes form groups of homologous proteins. The work is further backed up by the findings of Jiang and associates (Jiang, Tyler et al. 2006), who showed that there was an overall high level of co-linearity across four regions of *P. sojae* and *P. ramorum* that was only interrupted in areas containing pathogenicity related elicitin genes that could be under diversifying selection at a higher rate than housekeeping genes. The high level of conserved genes across the *Phytophthora* species reflects their close evolutionary relationship.

The region of *H. arabidopsidis* defined by BAC P1202 also shows a high level of synteny with orthologous regions of the 3 *Phytophthora* genomes. The five genes for which functions could be predicted, including *Ppats* 3 and 8, and the 5 unknown genes identified in the region have been shown to have orthologues in the 3 *Phytophthora* species. However, in terms of gene order and orientation there are genes present in the 3 *Phytophthora* syntenic regions which are notably absent from the region spanned by BAC P1202; an extradiol ring-cleavage dioxygenase, a MORN repeat protein and a formin-like protein. The break in gene order caused by the loss of these genes from the region, and in fact from the *H. arabidopsidis* genome, could reflect the idea that the region is involved in cellular processes related to the pathogenic life cycle of *H. arabidopsidis*, placing it under more selective pressure thus causing it to evolve at a higher rate than genes not involved in those processes. Although *H. arabidopsidis* is an oomycete and closely related to *Phytophthora* it is evolutionarily more distant from *P. sojae*, *P. infestans* and *P. ramorum* than they are to each other and the loss of the genes could reflect the divergence of the *H. arabidopsidis* and *Phytophthora* genomes from one another.

2.4.3 Annotation of *ATR13* locus

Annotation of the *ATR13* locus confirmed the presence of the *ATR13* effector gene as identified in the work by Allen and associates (Allen, Bittner-Eddy et al. 2004). The annotation also revealed the presence of an Acetyl-coA carboxylase, an Acyl transferase and a GTP-binding protein as well as two hypothetical proteins, one unknown protein and 30 transposable elements in addition, two 61 kb inverted repeat regions were identified toward the 3’ end or “downstream” of *ATR13* (Figure. 2.4). The functions of the identified genes suggest they are involved in essential cellular processes such as fatty acid regulation for acetyl-coA carboxylase, acyl transfer across membranes for the acyl transferase and regulation of signal transduction cascades in the case of the GTP-binding protein. This suggests they would not have a specific role in pathogenicity. Analysis of the region for any potential effectors returned no candidates which suggests that unlike some avirulence genes found in the genomes of *P. sojae* and *P. infestans* which show clustering of avirulence genes (Whisson, Drenth et al. 1995; Gijzen, Forster et al. 1996; van der Lee, Robold et al. 2001), there are no other avirulence genes clustered near the *ATR13* gene region of *H. arabidopsidis*.

2.4.4 *ATR13* locus and syntenic regions of *Phytophthora*

Comparative analysis between the *ATR13* locus and the genomes of *P. sojae, P. infestans* and *P. ramorum* revealed synteny with scaffold 42, Supercontig 76 and scaffold 63, respectively. The levels of synteny between the *ATR13* locus and the syntenic regions of the 3 *Phytophthora* species are extremely low with synteny restricted to the positions of the three genes; an Acetyl-coA carboxylase, an Acyl transferase and a GTP-binding protein. Apart from these genes there appears to be little or no conservation of sequence between the genomes in this region. The lack of synteny and any kind of co-linearity between the syntenic regions and the *ATR13* locus appears consistent with the possibility that the region is under extreme diversifying selection, thought to be driven by the interaction of ATR13 with the matching RPP13 resistance protein in Arabidopsis (Allen, Bittner-Eddy et al. 2004).

Analysis of the syntenic *Phytophthora* regions for the presence of any RXLR motif containing effector sequences returned no candidates confirming that there are no orthologous effector genes to *ATR13* present in the syntenic regions or in the rest of the genomes. However, 3 CRN effector genes were identified in the syntenic region of *P. sojae* along with a single CRN gene in the syntenic region of *P. ramorum*, which suggests the syntenic regions are involved in pathogenicity.

2.4.5 The presence of transposable elements in BAC P1202 and the *ATR13* locus

The *ATR13* locus of *H. arabidopsidis* shows high levels of diversifying selection probably due to the co-evolutionary arms struggle occurring with the Arabidopsis RPP13 R protein. However the impact of Transposable Elements (TE) in both the *ATR13* locus and BAC P1202 must also be considered a contributory factor in the diversification of the *H. arabidopsidis* genomes from its *Phytophthora* relatives. The annotation of the BAC P1202 revealed the region contained 9 TEs, whereas syntenic regions in *P. sojae*, *P. ramorum* and *P. infestans* contained none. TEs are sequences of DNA found in all Phyla which are able to move around to different locations within a genome via a process called transposition whereby the ends of the TE are cleaved and then transferred to its target DNA and the 3' end is then joined to the target (Doak, Doerder et al. 1994).

The presence of TEs in BAC P1202 could explain the absence, in *H arabidopsidis*, of the extradiol ring-cleavage dioxygenase, MORN repeat protein and formin-like protein 20 from the region that were found in the *Phytophthora* syntenic regions. TEs have been found to mediate numerous types of chromosomal rearrangements such as deletions, inversions and duplications in many different classes of organism (Lonnig and Saedler 2002); hence their presence around and flanking the region of where the 3 genes are absent could suggest that the genes were deleted as a result of one of these rearrangements. The presence of TEs in the *ATR13* locus could also explain the discovery of the large inverted repeat regions flanking the sections containing the large clusters of TEs close to the *ATR13* gene (Figure 2.3) as inverted repeats are characteristic of the DNA transposon class of TE (Labrador and Corces 1997).

Chapter 3: The Secretome

3.1 Introduction

3.1.1 Methods of effector transport

As with most bacteria and fungi, oomycetes communicate and adapt to changes in their environment through the use of a set of extracellular proteins, which are exported from the cell and have been shown to be involved in a diverse range of important functions including cell signalling and communication (Kamoun 2006). This group of secreted proteins has thus been defined as the 'secretome' of a species. In terms of pathogenicity, the secretome plays a vital role in aiding the infection of the host via the targeting of a varied range of effector proteins towards two distinct host cell sites; the cytoplasm and the apoplastic space. These effectors manipulate and interfere with the physiological and biochemical functioning of the host cell not only to disrupt and suppress the host immune response, but in so doing, promote the pathogens' own ability to infect the host and parasitize it. The microbial pathogens however, differ in the method of transporting effector proteins into their respective host cell targets; phytopathogenic bacteria such as *Pseudomonas*, *Erwinia* and *Xanthomonas* species use a type III secretion system (TSS) to inject effector proteins into the cytoplasm of the host cell in plants (Alfano and Collmer 2004). The widespread use of the TSS in many bacterial species suggests this mechanism is conserved amongst many prokaryotes (Bhattacharjee, Hiller et al. 2006). In contrast, much less is known about the method of targeting and transporting eukaryotic effectors to their host targets. Studies of the protozoan malarial parasite *Plasmodium falciparum* showed that the translocation of effector proteins from *P. falciparum* to the host human erythrocyte is mediated by a host targeting RXLX (E/Q) motif occurring just after an N-terminus signal peptide in the effector sequence (Hiller, Bhattacharjee et al. 2004). The subsequent identification of effectors containing a similar RXLR DEER motif just after the signal peptide in *Phytophthora*s *infestans*, *ramorum* and *sojae* as well as in closely related oomycete *Hyaloperonospora arabidopsidis* and the presence of a signal peptide in the effectors of *Magnaporthe grisea* suggest a shared export pathway for hundreds of effectors in eukaryotic microbial pathogens (Hiller, Bhattacharjee et al. 2004; Rehmany, Gordon et al. 2005; Bhattacharjee, Hiller et al. 2006; Jiang, Tyler et al. 2006).

The common mode by which both phytopathogenic prokaryotes and eukaryotes invade their hosts i.e. through the secretion of effector proteins, has inadvertently allowed the host plant the opportunity to detect their presence. Plant resistance (R) proteins can recognise the presence of effectors, triggering the plant immune system to initiate a hypersensitive response (HR) resulting host cell death at the site of infection. Effectors eliciting a HR response have been termed Avirulence (Avr) proteins. Interestingly it has been shown that the genes that encode these AVR proteins are maintained in the genomes of many phytopathogenic bacteria, fungi and oomycetes despite the fact they are detected by their hosts (Kjemtrup, Nimchuk et al. 2000; Alfano and Collmer 2004; Gao, Knogge et al. 2004). Explanations for this ‘paradox’ in Avr effector activity are that these Avr effectors carry out a vital function for the pathogen that aides its virulence or that these Avr effectors perform functions essential to the normal growth and maintenance of the pathogen, so much so that the benefits of maintaining the effector genes outweigh the potential risks they bring (Fabritius and Judelson 2003).

The sequencing and release of the genomes of *P. infestans*, *P. sojae* and *P. ramorum* has enabled researchers to investigate and classify the different families of effectors present in the secretomes of these genomes as well as predict the numbers of effectors present in each type of effector family, based upon their sequences. Despite the huge efforts to isolate and identify the varying types of effector families within the oomycetes, knowledge of their specific function and contribution towards either the suppression of plant immunity or the promotion of pathogen growth is fairly limited. The classification of the different families of effectors and the current status on efforts to functionally characterise them to date are thus defined below.

The effector families can be loosely divided into 2 groups according to their target sites either in the cytoplasm of the host cell or in the apoplastic space around the host cell. The effectors can also be subdivided according to either their function where known in virulent isolates of the pathogen, or by the type and extent of host defence response Avr effectors elicit in their hosts. The emphasis of this study lies with the apoplastic class of effectors therefore only this class of effector will be discussed in this instance.

3.1.2 Apoplastic Effectors

During the process of infection of the host by the pathogen, several key events must take place to insure the pathogen is able to adhere to and colonize the host tissue. The basic mechanism of infection in *Phytophthora* and other related oomycetes is as follows; zoospores are released from the sporangia of the pathogen and then adhere to the surface of the leaf, become enclosed in a cyst and germinate, forming an appressorium that penetrates down through the host leaf surface between the cells. During the subsequent growth down between the cells, the appressorium forms nodule-like structures called haustoria that penetrate the host cells allowing the secretion of effectors to their specific targets. The pathogen then sporulates to form sporangia and repeats the cycle of infection (Kamoun, Huitema et al. 1999; Hardham 2001). During the encystment phase of the infection cycle, zoospores secrete an adhesive-like substance that enables it to adhere to the surface of the leaf. The Cellulose Binding Elicitor and Lectin-like (CBEL) glycoprotein was first identified and isolated in *Phytophthora parasitica* and was shown to be crucial to the zoospores ability to attach itself to leaf surfaces as studies by Gaulin and associates (Gaulin, Jauneau et al. 2002) showed that silencing of the CBEL gene in *P. parasitica* impaired the pathogen's ability to bind to cellophane membrane surfaces, but this impairment did not affect its overall ability to infect tobacco plants. However, aside from this function, CBEL has a pathogen-associated molecular pattern (PAMP) as it has been shown to trigger a HR response in *A. thaliana* (Khatib, Lafitte et al. 2004). GP42, is a glycoprotein found in the cell walls of most *Phytophthora* species and has been shown to trigger HR responses in the leaves of potato and parsley (Nurnberger, Nennstiel et al. 1994; Scheel, Hahlbrock et al. 1995). Further analysis of the protein revealed a domain comprising 13 amino acids (PEP-13) that was shown to be recognised by Toll-like receptor defence proteins of both potato and parsley, triggering the release of antimicrobial phytoalexins (Nurnberger, Nennstiel et al. 1994; Scheel, Hahlbrock et al. 1995). The PEP-13 motif however, has been shown to be involved in transglutaminase activity, important in numerous essential physiological functions (Liu, Cerione et al. 2002). Thus the PEP-13 motif is required for both the activation of this activity and plant recognition (Brunner, Rosahl et al.

2002). The fact that PEP-13 is essential for transglutaminase activity would explain its conservation despite its recognition by host defence proteins.

The penetration of the host cell phase of infection requires the breakdown of the host cell wall and this is achieved via the secretion of enzymes that degrade the barriers that prevent access to the main body of the cell. One such enzyme is Endopolygalacturonase PIPG1. The *pipg1* gene was identified and characterised in the genome of *P. infestans* and belongs to the endopolygalactonurase class of enzymes, which are cell wall degrading enzymes and have been found in the genomes of many phytopathogenic species of fungi and bacteria (Torto, Rauser et al. 2002). Another class of enzyme is the exo-1,3-β-glucanases, which are hydrolases that act on polysaccharides or glucans found in the cell wall of the plant. In fact, 3 genes encoding these enzymes (*Piexo1*, *Piexo2* and *Piexo3*) where identified in the genome of *P. infestans* (McLeod, Smart et al. 2003).

Although some effectors perform functions which are important for normal pathogen growth or improve the pathogen's ability to infect the host, such as those described above, the function of others is to suppress and evade the host defence response. Part of the host response includes the secretion of an array of enzymes such as glucan, serine and cysteine proteases, chitinases and other hydrolytic enzymes, which act to decompose the protein compounds of the pathogen. In the face of this arsenal, the oomycetes and indeed fungi and bacteria, have evolved genetically and also developed a collection of effectors to evade this enzymatic activity. Three distinct families of effectors have been identified in *Phytophthora* which specifically act to counteract the degradation caused by these enzymes; firstly, the Glucanase Inhibitor protein (GIP) family of effectors, which include GIP1 and GIP2 from *P. sojae* and PiGIP1 to 4 from *P. infestans*, that act to inhibit endo-β-1,3 glucanases from soybean and potato, thus stopping them from degrading the β-1,3 glucans within the cell walls of *P. sojae* and *P. infestans* (Rose, Ham et al. 2002; Bishop, Ripoll et al. 2005; Damasceno, Bishop et al. 2008). Secondly, the Serine protease inhibitor family, which contain domains similar to those of the Kazal family of serine protease inhibitors, previously believed only to exist in apicomplexan parasites and animals. The family include EPI10 and EPI1, identified in *P. infestans* and both have been shown to interact with and target the same P69B subtilisin-like serine protease from tomato. Further to this, analysis of the sequence databases of *P. infestans*, *P. sojae*, *P. ramorum*, *Phytophthora brassicae* and *Plasmopara halstedii* identified 35 predicted proteins with Kazal domains, suggesting that inhibition of serine proteases is a conserved strategy employed among many oomycete pathogens as well as many animal species (Tian, Huitema et al. 2004; Tian, Benedetti et al. 2005). Thirdly, there is the cysteine protease family of inhibitors, to which EPIC1 and EPIC2 from *P. infestans* belong. Studies show that these two inhibitors target a papain-like cysteine protease PIP1 from tomato. Interestingly, it has been shown that these inhibitors were degraded by the same P69B subtilisin-like serine protease from tomato described above and subsequently this degradation was halted by EPI1 indicating that EPI1 contributes to the virulence of the pathogen by protecting other pathogen protease inhibitors from degradation by defence-related proteases (Tian, Benedetti et al. 2005).

Unlike many of the proteins already described, some avirulence effectors have yet to be functionally defined and, therefore, can only be characterised based upon the type of defence response they generate. PEP-13, a motif present in the GP42 glycoprotein described earlier, is a pathogen surface-derived molecule and is an example of pathogen-associated molecular patterns (PAMPs). GP42 binds to plant pattern recognition receptors, thereby triggering the activation of immune response genes and the subsequent production of antimicrobial compounds (Brunner, Rosahl et al. 2002). Another example of this is Necrosis-Inducing *Phytophthora* Protein 1 (NPP1), originally purified from *P. parasitica*, which has been shown to induce hypersensitive death-like lesions in parsley (Fellbrich, Romanski et al. 2002).

NPP1 belongs to a family of effector proteins called the Necrosis and Ethylene inducing Protein 1 (NEP1) like effector family, and have been shown to induce a necrotic response in their hosts. NEP1 was originally isolated in *Fusarium oxysporum* and a great many NEP1 orthologues and NEP-like proteins (NLPs) have since been identified in both phytopathogenic and saprophytic bacteria, oomycetes, and fungi (Pemberton and Salmond 2004; Bae, Bowers et al. 2005; Pemberton, Whitehead et al. 2005). In the oomycetes, NLPs have been found to be widespread in *Pythium* and *Phytophthora* species, for example, PiNPP1 has been identified in *P. infestans* and has subsequently been shown to induce necrosis in *Nicotiana benthamiana* and the host plant tomato (Kanneganti, Huitema et al. 2006). *P. sojae* Necrosis Inducing Protein (PsojNIP) is expressed during the necrotic phase of *P. sojae* infection of soybean. PsojNIP is thought to act as a possible toxin thereby facilitating the colonisation of the host dying tissue (Qutob, Kamoun et al. 2002). For Pythium, a novel protein elicitor (PaNie) from *Pythium aphanidermatum* was identified and shown to induce defence responses from carrot cell cultures, and in intact plants of Arabidopsis and tobacco (Veit, Worle et al. 2001). To date, very little is known about the function of these proteins or the reason they are conserved in such a diverse range of organisms, though the research carried out by Qutob and associates, (Qutob, Kamoun et al. 2002) and Pemberton and associates (Pemberton, Whitehead et al. 2005) led to the hypothesis that NLPs act as toxins to degrade the host tissue during the necrotrophic phase of pathogen infection, though this has yet to be proven. However, nearly all members of the family classified as NLPs, have similar sequences and contain a conserved set of residues, which include two cysteine residues and a central GHRHDWE motif, which would suggest a common function.

3.1.3 Cysteine-rich effector family

Of all the apoplastic effector families, one of the most easily recognised is the small cysteine-rich class of effectors. These effectors are, unsurprisingly, identifiable by the number of cysteines present relative to the sequence length, which is usually less than 150 amino acids (aa) and also by the number of aa's between each cysteine. For example, PcF, an effector first discovered in *Phytophthora cactorum*, can be identified by its small size of 52 aa's and by a 6-cysteine aa domain with a 4-hydroxyproline at residue 49. PcF has been shown to elicit a HR in both tomato and strawberry (Orsomando, Lorenzi et al. 2003). Two PcF-like proteins, Secreted Cysteine Rich (SCR) 74 and SCR91 have also been detected in *P. infestans* and have been shown to induce a HR in tomato. SCR74 in particular is also under intense diversifying selection, related to the co-evolutionary conflict occurring between the pathogen and its host (Bos, Armstrong et al. 2003; Liu, Bos et al. 2005). *Ppats 12, 14*, *23* and *24* identified through suppression subtractive hybridisation (Bittner-Eddy, Allen et al. 2003) have been found to be cysteine rich, however, little is known of their exact role in virulence and they share no common sequence similarity to any cysteine-rich effectors in related oomycete species. Many eukaryotic *Avr* genes, such as *Cladosporium fulvum Avr2*, *Avr4*, and *Avr9*, encode small (<150 amino acids) secreted proteins with a large number of cysteine residues, which can induce defence responses from the host (van't Slot and Knogge 2002).

One of the most well studied families of effectors is the Elicitins (ELI) and Elicitin-like (ELL) effectors. These effectors are pumped into the apoplastic space between the plant cells from the haustoria and are thought to promote infection of the plant through interaction with host cell receptors on the cell surface. Extensive studies have shown that the primary function of ELIs is to bind sterols which in *Phytophthora* is important as they cannot synthesize their own sterol (Mikes, Milat et al. 1997; Mikes, Milat et al. 1998). Many of the ELIs, however, have been shown to elicit a HR in potato and tobacco species, which results in programmed cell death at the site of infection (Kamoun, vanWest et al. 1997; Kamoun, van West et al. 1998). The ELI domain can be characterised by a highly conserved 96 amino acid domain, containing the six cysteine residue pattern C_1-23-C_2-23-C_3-4-C_4-14-C_5-23-C_6 that form 3 disulphide bonds (Fefeu, Bouaziz et al. 1997). ELLs, although sharing the six cysteine spacing pattern, are more diverse in the size of the domain and spacing between the cysteine residues, particularly at the C- terminus. To date, the ELI and ELL gene products have only been found in other *Phytophthora* and *Pythium* species. Recently, extensive data mining of *Phytophthora* EST and genome databases for novel ELI and ELL candidates revealed the presence of 128 novel ELI and ELL sequences from *P. sojae, P. ramorum, P. brassicae* and *P. infestans.* Extensive analyses on the evolutionary relationships between ELI and ELLs of the different *Phytophthora* species showed that they could be grouped into 17 distinct clades based upon their sequence similarity and cysteine spacing patterns (Jiang, Tyler et al. 2006).

The sequencing of the *H. arabidopsidis* genome (Baxter et al., Submitted) has enabled the identification of new ELI and ELL candidates in the genome, and also further analyses on the evolutionary relationships between the ELIs and ELLs in *H. arabidopsidis* and other *Phytophthora* species. The release of the *H. arabidopsidis* has also provided an opportunity to investigate and classify those families of effectors that are more distinct, such as the Kazal serine protease family as well as the small cysteine-rich effector family, present in the genome. Thus the focus of this study is to identify and classify any candidates that could belong to one of these families of apoplastic effector using bioinformatics tools. To do this, I defined the secretome of *H. arabidopsidis* by identifying all open reading frames that encoded proteins predicted to contain an N-terminal signal peptide. Methods of data mining to identify sequences with this characteristic signal peptide and thus obtain a signal peptide positive dataset, have already been defined in research carried out to identify candidate effectors in *P. sojae* and *P. ramorum* (Jiang, Dawe et al. 2005). This method was used to obtain a signal peptide positive dataset for *H. arabidopsidis* and from that dataset, candidate effector sequences from the small cysteine rich class of effector and the kazal serine protease class of effector were identified based upon their characteristic domains.

3.2 Methods

3.2.1 *H. arabidopsidis* assembly and *Phytophthora* sequences

Version 3 of the *H. arabidopsidis* genome assembly is available at http://vmd.vt.edu. Known ELIs and ELLs from *P. sojae*, *P. brassicae*, *P. ramorum* and *P. infestans* were obtained through GenBank at http://www.ncbi.nlm.nih.gov.

3.2.2 Signal peptide positive dataset of secreted proteins

Identification of all open reading frames (ORFs) in the *H. arabidopsidis* genome and subsequent translation to its amino acid sequence was performed by EMBOSS 3.0 application getorf (Rice, Longden et al. 2000). Each ORF in the dataset was then trimmed to obtain a dataset of all possible sequences with a methionine start codon. The dataset was screened to extract all sequences with a signal peptide probability 0.6 or higher using SignalP 3.0 (Bendtsen, Nielsen et al. 2004) and the Bioperl module signalP (Stajich, Block et al. 2002). The identified ORFs were screened for possible transmembrane domains using SOSUI (Hirokawa, Boon-Chieng et al. 1998). This produced the final dataset for identification of small cysteine-rich effectors, which include candidate ELI and ELL amino acid sequences as well as kazal serine protease inhibitor and PcF candidates. All above processes were implemented using a custom perl script (Appendix E).

3.2.3 Identification of small cysteine rich effectors and kazal serine protease inhibitor candidates

Amino acid Sequences from the signal peptide positive dataset were screened against known ELI and ELL sequences from *P. sojae*, *P. ramorum* and *P. infestans* using BLASTX from the standalone Blast version 2.2.13 (Altschul, Madden et al. 1998) to identify possible elicitin domains. ELI and ELL sequences from *P. sojae* and *P. ramorum* was used to train standalone Hidden Markov Model Software (HMMER) version 2.3.2 (R. Durbin 1998) to retrieve amino acid sequences from *H. arabidopsidis* which share the same key features. Prediction of Glycosyl-Phosphatidylinositol (GPI) was performed using big-PI plant predictor (Eisenhaber, Wildpaner et al. 2003). Kazal-like serine protease inhibitor candidate sequences were identified using the custom Perl scripts and BLASTX of the signal peptide positive dataset against the NCBI non-redundant nucleotide database http://www.ncbi.nlm.nih.gov. Kazal-like serine protease inhibitors from P. infestans and P. sojae were used to construct a standalone Hidden Markov Model to identify kazal-like domains from the signal peptide positive dataset.

3.2.4 Phylogenetic analysis of candidate ELI and ELLs and Kazal-like serine protease inhibitors

Sequence alignments were performed with standalone Clustal W version 1.83 (Thompson, Higgins et al. 1994). Phylogenetic tree construction were performed by Phylogenetic Analysis Using Parsimony (PAUP) version 4.0 beta version (Swofford 2003) using a Neighbour-Joining analysis. Confidence in the phylogenetic groupings was estimated using 1000 bootstrap replicates. The bootstrapped tree generated was visualised using TreeView X (Page 1996) and splitstree version 4 (Huson and Bryant 2006).

3.2.5 Analysis of whether ELI and ELL sequences are under diversifying selection

Analysis of whether ELI and ELL sequences are under diversifying selection was done using DNAsp (Librado and Rozas 2009). This program estimates *Ka* (the number of non-synonymous substitutions per non-synonymous site), and *Ks* (the number of synonymous substitutions per synonymous site) for any pair of ELI and ELL DNA sequences, it also computes several measures of the extent of DNA polymorphism in protein coding regions, non-coding regions, or in regions with both protein coding and non-coding regions (i.e. regions with both exons and introns). The ratio of *Ka* to *Ks* is then calculated. Ka/Ks <1 indicates purifying selection Ka/Ks >1 indicates the DNA sequences are undergoing diversifying selection (Appendix D).

3.3 Results

3.3.1 Identification of candidate ELI and ELL sequences

The identification process revealed 14 ELI and ELL amino acid sequences in *H. arabidopsidis* (Table 3.1) (Appendix B). A separate analysis performed by Rays Jiang (Pers. Comm.) using HMMER on the *H. arabidopsidis* version 3 of the genome assembly revealed 18 candidates. Comparison of the two separate datasets showed them to be identical with the exception of four extra candidates produced by Rays Jiang's own dataset.

Further analysis of the extra candidates showed 2 to be identical to one another and were localised to the same contig in areas close to a gap in the assembly and the flanking sequences of both candidates were also identical. This indicates that the candidates are artefacts of the assembly and have been discounted as potential candidates. The third candidate was found to be an extended version of one that had already been identified in the identification process and had no predicted signal peptide. This implies that the extended version did not have the correct methionine start codon and was, therefore, discounted as a candidate. The fourth candidate, subsequently named HpELL11B, had an amino acid sequence similarity to *P. sojae* elicitin SOJ2D but the homologous region began at amino acid position 117 of SOJ2D onwards. HpELL11B, although similar to known elicitin domains, lacks the first cysteine residue of the characteristic six-cysteine residue spacing pattern found in all known elicitins. There was also no predicted signal peptide for this candidate at the N-terminus. This indicates that the methionine start

Table. 3.1. Details of the 15 candidate ELI and ELL sequences identified, their sequence length, predicted GPI anchor cleavage sites, cysteine spacing patterns and to which clade they belong.

Sequence Name	Sequence Length (amino acid)	Clade	Predicted GPI Cleavage Site	Cysteine Spacing Pattern
HpELI4	175	ELI4	-	C-23-C-23-C-4-C-14-C-22-C
HpELL1A	293	ELL1	268	C-16-C-22-C-4-C-14-C-18-C
HpELL1B	261	ELL1	-	C-16-C-22-C-4-C-14-C-18-C
HpELL1C	180	ELL1	157	C-16-C-22-C-4-C-14-C-18-C
HpELL4	168	ELL4	-	C-20-C-22-C-4-C-15-C-21-C
HpELL6	174	ELL6	-	C-17-C-24-C-4-C-14-C-21-C
HpELL8	190	ELL8	-	C-19-C-21-C-4-C-14-C-21-C
HpELL9	115	ELL9	-	C-18-C-23-C-4-C-11-C-21-C
HpELL11A	172	ELL11	148	C-20-C-25-C-4-C-12-C-21-C
HpELL11B	212	ELL11	-	C-x-C-23-C-4-C-14-C-23-C
HpELL11C	180	ELL11	-	C-24-C-23-C-4-C-14-C-23-C
HpELL13A	401	ELL13	376	C-20-C-15-C-4-C-13-C-17-C
HpELL13B	151	ELL13	-	C-20-C-14-C-4-C-13-C-18-C
HpELL13C	166	ELL13	148	C-19-C-12-C-4-C-13-C-17-C
HpELL13D	199	ELL13	173	C-20-C-22-C-4-C-14-C-17-C

codon of the sequence is not correct. Analysis of contig 135, in which the sequence lies, indicates a possible alternative methionine start codon at position 66128 of frame 6 of the contig, upstream of the original start codon at position 66055 of frame 4 of the contig. SignalP predicted the presences of a potential signal peptide following this alternative methionine start codon. There are also three cysteine residues present downstream of the alternative start codon that could be the missing cysteine residue from the elicitin domain. The existence of an alternative start site could be due to three possibilities; that there has been an assembly error; that the sequence contains a small intron which would cause there to be a frame shift in the sequence; or that there has been a real frame shift mutation in the sequence. If the latter is the case then that would signify that the sequence is no longer conserved and can, therefore, be discounted as a candidate. Unfortunately there are no EST matches to the sequence and, hence, it is impossible to state which of the possibilities is correct. Therefore, the candidate was included in further analyses but has a lower level of certainty.

A TBLASTN analysis of the 15 remaining candidates against the *H. arabidopsidis* EST database revealed a high degree of similarity to ESTs for 12 of the 15 candidates (Table 3.2). Hits were deemed significant if their similarity was over 90 percent. The lack of an EST for the remaining five candidates does not necessarily discredit them as candidates, reasons for the lack of hits could simply be that the appropriate EST was not sequenced from the library. In any case the lack of an EST simply means there is no confirmation that these predicted genes are expressed.

The DNA sequences of the 15 candidate ELI and ELLs were used to calculate the Ka/Ks ratio for each pair of sequences (Appendix D). Of the 78 ratios that could be calculated 40 were found to have ratios greater than one whereas 38 were found to have ratios less than one. Of those 38 ratios between the sequences that were not shown to be under diversifying selection, 5 were less than 0.05 below the threshold. In the instances where the sequences are just below the threshold, the reasons could be due to the pair of sequences being from ELL families more closely related to each other in the phylogenetic tree than to the other sequences but does not mean that they are not diverging from one another. The results would indicate that the sequences are undergoing diversifying selection.

Table. 3.2. Identification of ESTs[a] for predicted *H. arabidopsidis* ELI and ELL genes using TBLASTN

Gene Name	**EST hit**	**Bit score, E-value**	**Degree of similarity[b]**
HpELI4	CL3157Contig1	350 bits (898), Expect = 6e-98	99%, 175, 1-175:352-876
HpELL1A	CL3020Contig1	288 bits (737), Expect = 6e-79	100%, 152, 1-152:451-906
HpELL1B	Hp_ENSC_25f19	109 bits (273), Expect = 4e-25	98%, 55, 233-288:2-196
HpELL1C	CL3866Contig1	350 bits (898), Expect = 6e-98	100%, 180, 1-180:222-761
HpELL4	-	-	-
HpELL6	-	-	-
HpELL8	CL2728Contig1	393 bits (1010), Expect = e-111	100%, 190, 1-190:446-1016
HpELL9	CL2960Contig1	225 bits (574), Expect = 1e-60	100%, 115, 1-115:405-61
HpELL11A	CL171Contig	340 bits (873), Expect = 5e-95	100%, 172, 1-172:610-95
HpELL11B	-	-	-
HpELL11C	CL21Contig2	355 bits (912), Expect = 2e-99	100%, 180, 1-180:45-584
HpELL13A	CL2482Contig1	785 bits (2026), Expect = 0.0	98%, 409, 14-419:1394-177
HpELL13B	CL4824Contig1	293 bits (749), Expect = 9e-81	99%, 151, 1-151:574-122
HpELL13C	CL363Contig1	319 bits (817), Expect = 1e-88	100%, 166, 1-166:76-573
HpELL13D	CL214Contig1	389 bits (998), Expect = e-109	100%, 199, 1-199:681-85

a ORFS were TBLASTN against Hp Unigenes from the VBI http://vmd.vbi.vt.edu to obtain EST hits.

b The values can be represented as follows: %similarity, length of overlap (aa), length of query (aa), length of subject (bp)

3.3.2 Phylogenetic reconstruction of ELI and ELL candidates

To determine the relationship between the elicitin domains in other *Phytophthora* species and those candidates found in *H. arabidopsidis*, 128 ELI and ELL amino acid sequences obtained from *P. sojae, P. ramorum, P. brassicae* and *P. infestans* and the 15 candidate ELI and ELL sequences from *H. arabidopsidis* were used to construct a bootstrapped phylogenetic tree with values greater than 50 (Figure 3.2). The 15 candidates fell into the same 17 clades (bootstrap values of over 80) as identified by Jiang, Tyler et al. (2006).

Of the 15 *H. arabidopsidis* candidates used to construct the tree (Table 3.2), 1 candidate was found phylogenetically to belong to clade ELI-4 and subsequently named HpELI4. Interestingly, although there is a high sequence similarity between HpELI4 and the ELI-4 sequences of the four *Phytophthora* species, the cysteine spacing pattern of HpELI4 diverges from the highly conserved C_1-23-C_2-23-C_3-4-C_4-14-C_5-23-C_6 pattern found in all the elicitins of the other four species. HpELI4 shows a variation in its elicitin domain as it has 97 amino acids instead of 98 and therefore variation between the C_5 and C_6 cysteine residues in only having 22 amino acids instead of 23. This variation is reflected in the splitstree representation of the phylogenetic tree (Figure 3.1) as HpELI4 diverges from the main branch that represents ELI-4. The remaining 14 candidates were found to be more diverse in their sequence domains and were all classified as ELLs with 4 of the

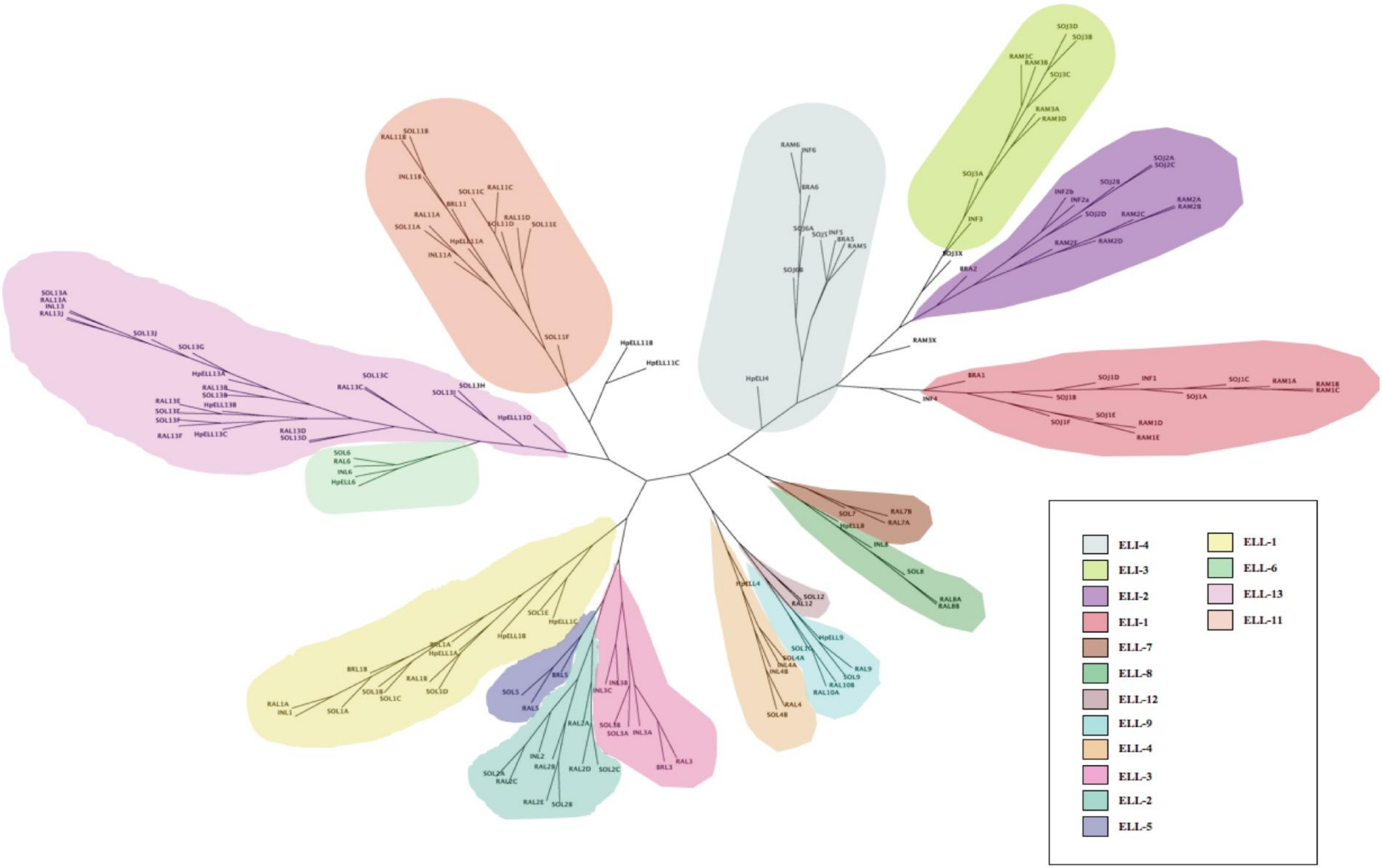


Figure. 3.1 Splitstree of ELI and ELL amino acid sequences from *H. parasitica*, *P sojae*, *P. ramorum*, *P. brassicae* and *P. infestans*. The amino acid sequences from each of the species were used to construct and unrooted Splitstree using a Neighbor-Joining analysis. The shaded areas represent the different ELI and ELL groups. The key shows the shaded area that represents the individual ELI or ELL clade. The lines do not represent any physical or evolutionary distances between the sequences.

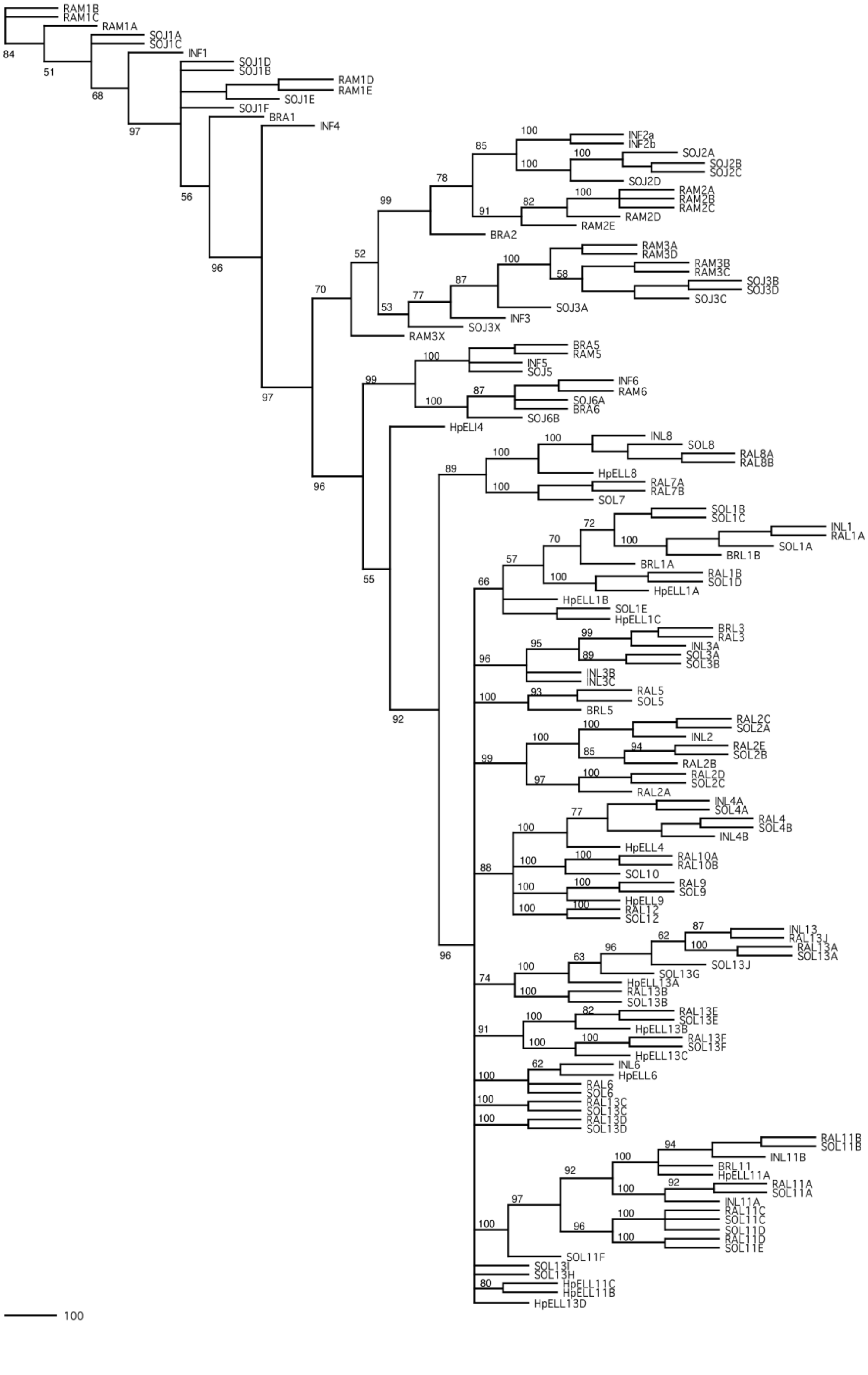


Figure. 3.2 – Phylogram of ELI and ELL amino acid sequences from *H. parasitica*, *P. sojae*, *P. ramorum*, *P. brassicae* and *P. infestans*. The amino acid sequences from each species were used to create an unrooted phylogram using a Neighbor-Joining analysis. The percentage support for each node on the phylogram was estimated using 1000 bootstrapped replicates and are represented by the numbers at each branch point. The lines represent the differences, with the scales showing a difference of 100 steps between the sequences. The lengths of lines do not represent the evolutionary distances between the sequences.

candidates belonging to the most divergent of the ELL clades, ELL-13. Candidates HpELL11B and HpELL11C were left as outliers in the splitstree representation of the clades (Figure 3.1) because although phylogenetically they show a higher sequence similarity to clade ELL-11, the cysteine spacing patterns in their domains are much more closely associated with that of ELI-4.

6 of the candidates from clades ELL-1, ELL-11 and ELL-13 (Table 3.2) have predicted hydrophobic regions at their C-terminus which is consistent with them being predicted to contain a Glycosylphosphatidylinositol (GPI) anchor site. GPI anchors are glycolipids that are attached to the C-terminus of the effector protein during post-translational modification. First the protein is directed in to the Endoplasmic reticulum (ER) and the hydrophobic region at the C-terminus is cleaved off and a GPI added. The modified protein then passes through the *H. arabidopsidis* haustoria into the apoplastic space, the GPI then anchors the effector protein to the plasma membrane.

3.3.3 Identification of Kazal serine protease inhibitor candidates

Five candidate Kazal serine protease inhibitors were identified. Further investigation of the candidates revealed that two were false positives as BLASTP analysis of the candidate amino acid sequence against the NCBI non redundant protein database returned only a 50 percent hit to the defined kazal domain, hence they were discounted as potential candidates**.** A TBLASTN analysis of the three candidates against version 1 of the *P. sojae*, *P. ramorum* and *P. infestans* assemblies returned significant similarity scores to orthologous kazal serine protease inhibitors in the 3 *Phytophthora* species (Table 3.3) (Appendix C)**.** Hits were deemed significant if their similarity was over 90 percent. The *H. arabidopsidis* EST database contained sequences that showed high levels of similarity to the candidate genes (Table 3.3 gives actual figures), confirming their structure and that they are expressed sequences. To determine the association and similarity between the *Phytophthora* species and the candidates found in *H. arabidopsidis*, 34 predicted kazal-like serine protease

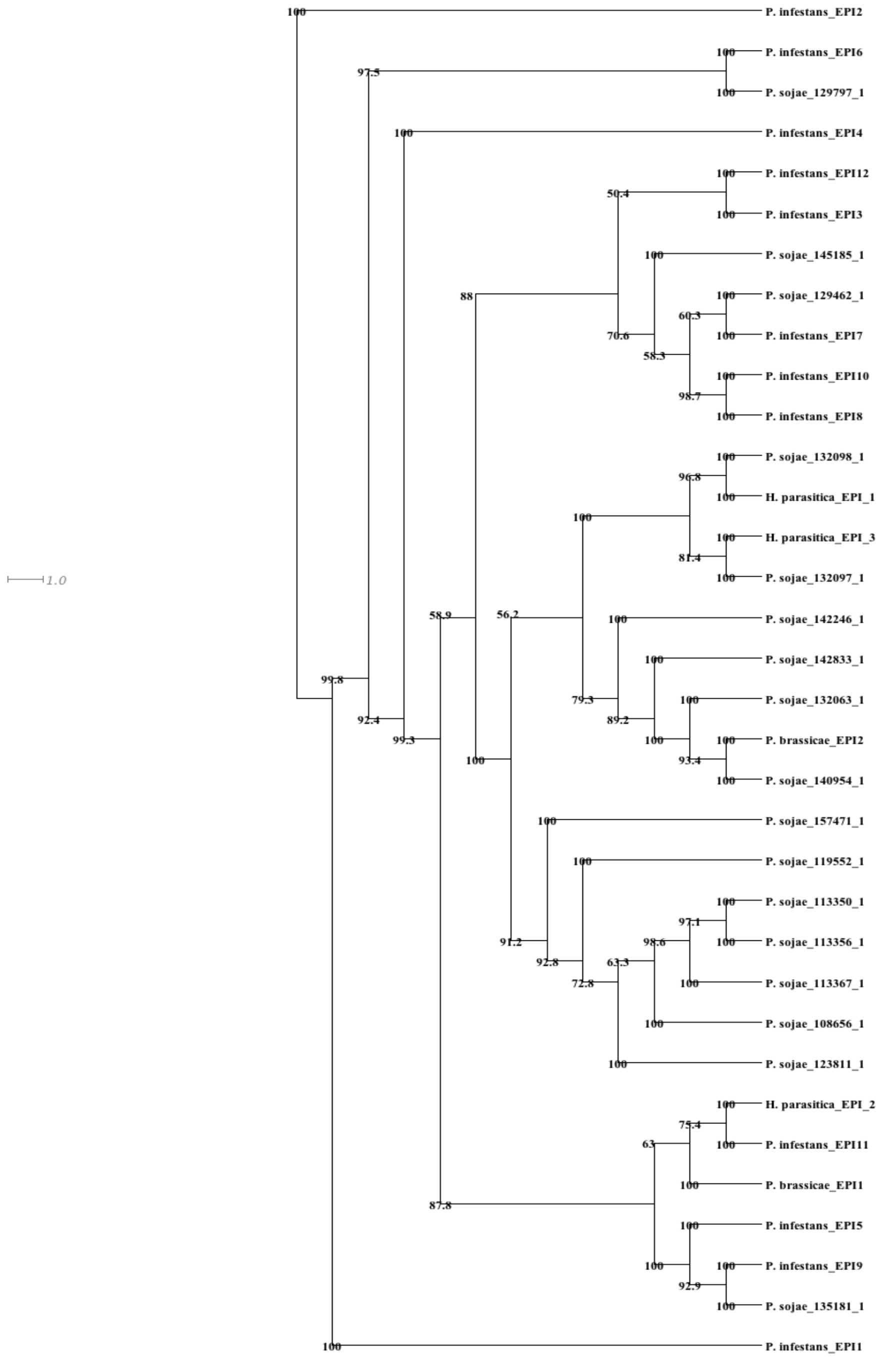


Figure. 3.3. A phylogram of Kazal_like serine protease inhibitor amino acid sequences from *H. arabidopsidis, P. sojae*, *P. infestans* and *P. brassicae*. The amino acid sequences from each species was used to create an unrooted phylogram using a Neighbour-joining analysis. The percentage support for each node on the phylogram was estimated using 1000 bootstrapped replicates and are represented by the numbers at each branch point. The lines represent the differences with the scale showing a difference of 100 steps between the sequences. The length of the lines do not represent the evolutionary distances between the sequences. The assignment of the names of the *H. arabidopsidis* sequences reflect the order in which they were found.

Table 3.3 Identification of ESTs[a] and Phytophthora orthologues for candidate *H. arabidopsidis* kazal-like serine protease inhibitors.

Candidate Gene	***H. arabidopsidis*** **EST**	Degree of similarity[c]	Orthologue Transcript ID	Orthologue location	Degree of similarity	Organism
HpEPI_1	CL104Contig1	100%, 110, 1-110:141-470	132098 (PsojEPI_190048)	Scaffold_19:242253-242636	68%, 107, 7-110:25-345	*P. sojae*
			76995	Scaffold_21:431281-431631	69%, 101, 12-109:10-309	*P. ramorum*
			PITG_09840.1	Supercontig 17: 347369-347713	57%, 107, 6-109:4-324	*P. infestans*
HpEPI_2	CL1067Contig1	100%, 283, 1-283:41-889	143590 (PbraEPI1)	Scaffold_144:100275-101338	55%, 299, 1-275:50-925	*P. sojae*
			79145	Scaffold_37:300928-303234	56%, 286, 9-275:16-837	*P. ramorum*
			PITG_07096.1	Supercontig 10: 2242586-2244039	53%, 266, 28-278:571-1317	*P. infestans*
HpEPI_3	CL2680Contig1	100%, 117, 1-117:42-392	132097 (psojEPI1)	Scaffold_19:241289-241654	70%, 112, 7-117:31-363	*P. sojae*
			76994	Scaffold_21:430501-430854	67%, 118, 1-117:1-351	*P. ramorum*
			PITG_09845.1	Supercontig 17: 356047-356412	73%, 98, 21-117:67-357	*P. infestans*

a ORFS were TBLASTN against Hp Unigenes from the VBI http://vmd.vbi.vt.edu/ to obtain EST hits.

b ORFS were TBLASTN against Transcripts of *P. sojae, P. infestans* and *P. ramorum* from the VBI Microbial Database http://vmd.vbi.vt.edu/, orthologues were assigned using reciprocal best BLAST hits.

c The values can be represented as follows: %similarity, length of overlap (aa), length of query (aa), length of subject (bp) inhibitor amino acid sequences obtained from *P. sojae, P. brassicae* and *P. infestans* and the three predicted sequences from *H. arabidopsidis* were used to construct a bootstrapped phylogenetic tree with values greater than 50 (Figure 3.3). The candidates show a clear similarity to the predicted kazal-like serine protease inhibitors of *P. sojae* and *P. infestans* giving added weight to the validity of these candidates. The phylogenetic tree also highlights the difference in the number of candidates found between *H. arabidopsidis* and the *Phytophthora* species; 12 were found for *P. infestans*, 17 *P. sojae*, 2 in *P. brassicae* compared to only 3 in *H. arabidopsidis.*

3.3.4 Identification of *Ppat* 24 and *Ppat* 14 like sequences

Analysis of the signal peptide positive dataset to identify *Ppat* like sequences returned no candidates for *Ppats* 12 and 23, it would appear that these two *Ppats* are entirely unique within *H. arabidopsidis* and TBLASTN analysis revealed no potentially orthologous sequences were found in the genomes of *P. sojae*, *P. ramorum* and *P. infestans.* However analysis of the signal peptide positive data set to identify *Ppat* 24 like sequences returned 4 candidates in total, one of which was *Ppat* 24 itself. ESTs were identified for 3 of the 4 candidates, indicating they are expressed. BLASTP analysis of the candidates against the NCBI non-redundant protein sequence database returned no significant hits to other genes that would indicate any previously determined function associated with this novel group of cysteine rich proteins. TBLASTN analysis of the *Ppat* 24 like candidates against the genomes of *P. sojae*, *P. ramorum* and *P. infestans* returned no potentially orthologous sequences sharing the distinctive C-5-C-14-C-8-CC-3-C-4-C-5-C-17-C-12-CC-3-C cysteine spacing pattern found in the *Ppat* 24 like sequences. This would indicate that this potential class of cysteine rich effector, in the currently sequenced oomycete genomes, is present only in *H. arabidopsidis*.

Similar analyses of the signal peptide positive dataset for *Ppat* 14 like candidates returned two sequences, *Ppat* 14 and one other candidate. Using the *Ppat*14 sequence to scan the EST database returned no hits to any *H. arabidopsidis* EST clusters that could indicate expression. BLASTP analysis of the sequences against the NCBI non-redundant protein database, as with the analysis of *Ppat* 24 candidates, returned no significant hits to any gene that would indicate any possible function associated with the sequences. And as with the *Ppat* 24 candidates, there no orthologues to *Ppat* 14, sharing the C-6-C-10-C-12-CC-3-C-10-C-6-C-11-C-12-CC3-C gene structure, could be identified in the genomes of *P. sojae*, *P. ramorum* and *P. infestans*. This suggests that this class of cysteine rich effector is only present in *H. arabidopsidis.*

3.4 Discussion

The analysis of the *H. arabidopsidis* genome and subsequent extraction of a signal peptide positive dataset from it has enabled the identification of 15 candidate ELI and ELL sequences based upon their cysteine rich spacing pattern. For 12 of these 15 candidates a matching EST was identified. Interestingly only one candidate, HpELI4, could be classed as an elicitin, based on the highly conserved 96 amino acid domain C_1-23-C_2-23-C_3-4-C_4-14-C_5-23-$C_{6.}$

Elicitins have been identified in *Phytophthora* and *Pythium* species. In fact, of the 128 sequences used to construct the phylogenetic tree, 18 were *P. sojae* elicitins, 22 were *P. ramorum* elicitins, 10 were *P. infestans* elicitins and 5 were *P. brassicae* elicitins. These numbers are in sharp contrast to the single elicitin, HpELI4, found in *H. arabidopsidis*. The lack of elicitins found in *H. arabidopsidis* implies they are no longer being conserved within the genome unlike in the *Phytophthora* species. Analysis was performed on ELIs in *P. sojae*, *P. ramorum*, *P. brassicae* and *P. infestans* to determine if they were under selective pressure (Jiang, Tyler et al. 2006). This was achieved by comparison of the rate of non-synonymous nucleotide substitutions to that of synonymous substitutions, revealing that these ELIs are, in contrast, undergoing purifying selection. This is in sharp contrast to the similar analysis carried out on *H. arabidopsidis* ELI and ELL sequences, which indicated the sequences were under diversifying selective pressure. The vast differences seen in the levels of conservation of ELI genes between the *Phytophthora* species and *H. arabidopsidis* could be explained by their differing behaviour in *planta*. In recent years, elicitins have been shown to elicit a HR in *Nicotiana* species; for example *P.*

infestans ELI-1 elicitin INF1 has been proven to induce HR in *Nicotiana benthamiana* (Kamoun, vanWest et al. 1997; Kamoun, van West et al. 1998). Elicitins SOJ3 and SOJ6 from *P. sojae* have also been shown to elicit a host defence response (Qutob, Huitema et al. 2003). The main functions of elicitins in *Phytophthora* are to act as sterol carriers, as sterols are required by many Oomycete species but they are unable to synthesize on their own. This need to obtain sterol for the pathogens own structural maintenance could outweigh the problem of being detected by the host defence proteins; this would explain the conservation of elicitins in the *Phytophthora* species. However, for *H. arabidopsidis* the obligate biotrophic lifestyle may make evasion of detection paramount. Therefore, the surprising discovery that it contains only 1 ELI and 14 ELL candidates may be logical, but this leaves some question as to how *H. arabidopsidis* obtains sterols. Currently it has only 1 gene known to function as a sterol carrier. It can be speculated that perhaps the ELL sequences are involved in sterol uptake but there is some ambiguity as to what the function of these divergent ELL sequences is as essential amino acid residues shown to be involved in sterol binding in *cryptogein* are not conserved within the ELL domains. A general involvement in lipid binding can be assigned to ELLs of *Phytophthora* as characterisation studies of ELLs found in *Phytophthora capsici* showed Phospholipid activity (Nespoulous, Gaudemer et al. 1999) though the exact nature of this activity has yet to be defined.

A second explanation for the conservation of ELIs in *Phytophthora* but not in *H. arabidopsidis* is that (although not their primary function) the benefit of conserving HR inducing elicitins in hemi-biotrophic *Phytophthora* species lies in the ability of *Phytophthora* pathogens to change from a biotrophic mode to a necrotrophic mode between 12 to 14 hrs after infection of the host as part of its life-cycle (Moy, Qutob et al. 2004). Once the host detects the elicitin, cell death ensues and the pathogen is then able to synthesize necrosis-inducing enzymes to break down the necrotic tissue (Qutob, Kamoun et al. 2002). However as *H. arabidopsidis* is an obligate biotrophic oomycete and therefore requires that not only does the host survive, but that it does not detect the pathogens' presence; an HR would impair its ability to obtain nutrients from the host, therefore, *H. arabidopsidis* would not benefit from an elicitin-induced HR. This could perhaps explain why ELIs are not conserved within *H. arabidopsidis*. This maybe further supported in that the remaining 14 *H. arabidopsidis* candidates were instead found to have more divergent domains and were classed as elicitin-like. The diversity in the spacing patterns of the different clades of ELL sequences, which has in turn led to altered structure of the protein, is in keeping with the theory that *H. arabidopsidis* is involved in an evolutionary "arms race" with its Arabidopsis plant host; this diversity may contribute towards the pathogen's ability to avoid detection (Birch, Rehmany et al. 2006) (Slusarenko and Schlaich 2003). Interestingly, it is important to note that none of the members of the ELL clades in *Phytophthora* have been found to elicit a HR response (Qutob, Huitema et al. 2003), this could represent a successful evasion technique employed by *Phytophthora* to avoid plant defences. The predominance of ELLs in *H. arabidopsidis* over ELIs could indicate that *H. arabidopsidis* could be using this same technique to avoid detection; however the exact role of ELLs is still unknown.

This study identified the presence of three candidates for Kazal-like serine protease inhibitors, which had strong similarity to potential orthologues from *P. sojae*, *P. infestans* and *P. ramorum*. The candidates were also found to each have matching

ESTs from the *H. arabidopsidis* EST database. Kazal-like serine protease inhibitors have been shown to protect many secreted proteins in *P. infestans* from degradation by the P69B protease secreted in tomato. Current research suggests that kazal-like domains are conserved particularly within Oomycete species, as some 56 kazal-like domains were discovered in 35 predicted proteins from the genomes of *P. sojae*, *P. ramorum*, *P. infestans*, *P. brassicae* and *Pl. halstedii* (Tian, Huitema et al. 2004). The presence of Kazal-like serine protease inhibitors in the genomes of both plant and mammalian parasites suggests that a wide range of hosts use proteases to defend against parasite invasion. It can be argued that inhibition of host plant proteases is even more important for obligate biotrophic organisms such as *H. arabidopsidis* as their *modus operandi* is primarily to evade detection. Despite this reasoning, the low numbers of kazal-like serine protease inhibitors found in *H. arabidopsidis* is in sharp contrast to the numbers found in other oomycete species (12 in *P. infestans* and 18 in *P. sojae*) but the numbers are more akin to *P. brassicae* which currently has only two candidates (Kamoun 2006). This suggests that the extent to which *H. arabidopsidis* uses protease inhibitors as a means of counter defence and aiding pathogen virulence is much less than those of most of the Phytophthora species. The lack of Kazal-like serine protease inhibitors in *H. arabidopsidis* is surprising when considering *H. arabidopsidis's* infection strategy. However, the lack of conservation of this type of cysteine rich effector could be because it has been surpassed by another class of protease inhibitor such as the cysteine protease inhibitors in aiding pathogen virulence, or simply that the pathogens' virulence strategy is not dependent upon inhibition of host plant proteases within the apoplastic space.

The final analysis of the signal peptide positive dataset centred on identification of candidates similar to *Ppat* 24 and *Ppat* 14, which have been shown to be involved in pathogenicity (Bittner-Eddy, Allen et al. 2003). The analysis returned three candidates similar to Ppat 24 based on amino acid sequence similarity and EST analysis identified an EST cluster for three of the four candidates, indicating they are expressed and are likely to be secreted. A similar analysis of the signal peptide positive dataset for *Ppat* 14 like candidates identified only one other candidate similar to Ppat 14. EST analysis of the sequences returned no strong hits to any *H. arabidopsidis* EST clusters that would confirm expression of the predicted ORF. TBLASTN analysis of both Ppat 24 and Ppat 14 against the genomes of *P. sojae*, *P. infestans* and *P. ramorum* identified no orthologues, indicating that these sequences are unique to *H. arabidopsidis*. This outcome is in keeping with the results of the analysis of BAC P1202 (see Chapter. 2.), which contain Ppats 3 and 8; both the genes were found to be unique to *H. arabidopsidis* and have functions that are essential for the maintenance of the pathogen’s own cellular processes, rather than directly contributing towards the virulence of the pathogen. Thus it is tempting to suppose that both Ppat 24 and 14 and their homologous sequences are involved in similar functions. However, a BLASTP analysis of the NCBI database using the amino acid sequences of Ppat 24 and 14 and their homologous sequences found in the signal peptide positive dataset identified no homologous sequences that could be used to indicate potential function for *Ppat24* and *Ppat14.*

The process of discovery and classification of candidates for these three types of cysteine rich effectors has both revealed and reinforced the idea that the *H. arabidopsidis* genome is under intense dynamic, diversifying selective pressure. This can be seen in the diverse range of cysteine spacing patterns of *H. arabidopsidis* ELL candidates with only a single candidate showing the rigid 96 amino acid domain C_1-23-C_2-23-C_3-4-C_4-14-C_5-23-C_6 characteristic of elicitins. The lack of kazal-like serine protease inhibitors within the *H. arabidopsidis* genome suggests that perhaps this not a major form of virulence when compared with that of its closely related relatives *P. sojae* and *P. infestans*. The extreme differences shown in the diversity and numbers of effectors seen in *H. arabidopsidis* paint a picture of a highly adaptive pathogen, ever changing to evade detection.

Chapter 4: Modelling transcriptional networks from pathogen induced and developmental microarray time course experiments.

4.1 Introduction

Previous chapters have focussed on the annotation of *H. arabidopsidis* regions containing genes known to be expressed during infection in an attempt to identify the presence of novel *Avr* genes, and identify the repertoire of secreted proteins secreted during infection of the host. A natural extension to the identification of genes governing infection of the host is to determine the transcriptional networks that govern the regulation of expression of these genes during infection. Recent advances in microarray technology have enabled the measurement of the expression levels of the majority of genes within a genome over a number of time points in a single experiment (Zareparsi, Hero et al. 2004). This in turn has highlighted those genes that show varying expression levels over time under particular conditions. It is reasonable to suppose that a proportion of these genes play an important role in regulating gene expression under these conditions. Such genes could then be used to reverse engineer transcriptional networks based on their function. Unfortunately, elucidation of regulatory pathways requires a high level of functional annotation of the *H. arabidopsidis* genome in order to identify reasonable pathways. Because the *H. arabidopsidis* genome has only provisionally been sequenced it is still relatively un-annotated with mostly genes involved in pathogenicity having been annotated (Allen, Bittner-Eddy et al. 2004; Rehmany, Gordon et al. 2005). The lack of high quality annotations in *H. arabidopsidis* makes it impossible to use as a model organism from which to identify transcriptional regulatory networks. The Complete

Arabidopsis Transcriptome MicroArray (CATMA) project (http://www.catma.org/) was formed in 2000. The aim was to use the newly completed *A. thaliana* genome sequence to develop a complete and specific microarray for *A. thaliana* by producing a specific gene sequence tag (GST) for every known or predicted gene found in the genome sequence, currently 30,343 genes (http://www.tigr.org) (Crowe, Serizet et al. 2003; Sclep, Allemeersch et al. 2007). The CATMA GSTs are specific only to their target gene. The GST design was based on both the TIGR annotations and the predictions of protein coding genes obtained from Eugene v1.0 software (Schiex, Moisan et al. 2001). By combining different information (transcripts, splicing sites, translation initiation sites, coding potential and protein similarities), the Eugene prediction software provided an alternative Arabidopsis genome annotation that in turn have vastly improved the functional annotation of Arabidopsis (Aubourg, Martin-Magniette et al. 2007). Furthermore, because the full sequence of each GST is known, they can be used for microarray experiments. Hence these GSTs were used to perform a microarray time course experiment of Arabidopsis gene expression in response to inoculation with two isolates of *H. arabidopsidis*. Thus the focus of this chapter is on Arabidopsis and elucidating the transcriptional networks that are altered by the effector complement of *H. arabidopsidis*.

A generalised description of Arabidopsis response pathways to pathogen infection is as follows. It is generally accepted that there are two types of plant immune system, in the first type; transmembrane pattern recognition receptors (PRR) detect slow evolving pathogenic elicitors/PAMPs (pathogen-associated molecular patterns) leading to PAMP triggered immunity (PTI). The second type of defence is by R receptor proteins. There are two basic classes of R protein. One is membrane spanning and extracellular with distinct leucine rich repeat regions (LRR), such as RPP13 (Bittner-Eddy, Can et al. 1999) or RFO1 (Bae, Kim et al. 2006). The second class is a cytoplasmic group that may or may not be membrane associated, e.g. RPP5 in Arabidopsis (Parker, Coleman et al. 1997). The R receptor proteins are thought to act as a second line of defence against specific pathogenic effectors released by the pathogen after successfully avoiding PTI. R receptor proteins indirectly or directly recognise effectors through monitoring the integrity of the sites of effector targets largely within the plant cell, any interference will result in effector recognition (Jones and Dangl 2006). Detection of the pathogen by theses two types of defence systems induces downstream signalling. This signalling causes the activation of calcium channels and the subsequent increase in cytoplasmic calcium levels triggers the activation of NADPH oxidases and peroxidases (PEXs) (Torres, Jones et al. 2005), resulting in hydrogen peroxide (H2O2) production and oxidative burst (ROS) (Rao and Davis 2001). Subsequently this initiates a MAP-kinase cascade to active three separate downstream defensive pathways, mediated by the plant hormones salicylic acid (SA), ethylene (ET), jasmonic acid (JA) and abscisic acid (ABA) (Glazebrook 2001). These defensive signal transduction pathways then initiate the expression of defence genes such as Pathogenesis-Related Gene1 (*PR1*) and a hypersensitive response at the site of infection. Control of the expression of these defence genes is maintained through a set of transcription factors, which include WRKY, AtMYB and ERF classes. However, exactly how these transcription factors act together in a transcriptional network to regulate the expression of these defence genes is still largely unknown. Thus microarray time series data can be used to identify genes differentially regulated after infection with *H. arabidopsidis* and used to impute networks describing the transcriptional regulation of expression of these genes. However generating valid network models is a major challenge for Systems Biology.

The problem of how to reverse engineer these networks from expressed data has resulted in numerous approaches being taken such as Boolean Networks whereby a gene's state can be described by a Boolean flag as being active with a 1 or inactive with a 0. Thus in Boolean networks a gene's state can be predicted based upon whether other genes are also active or inactive (Akutsu T 1999). The use of Dynamic Bayesian Networks (DBN) to model microarray data was first explored by Friedman (Friedman, Linial et al. 2000) by using *Saccharomyces cerevisiae* cell cycle measurements at a single time point (Spellman, Sherlock et al. 1998). This resulted in inference of causal relationships between genes thought to initiate cell cycle and its control. However this approach only looked at a single time point and assumed that all variables that contribute towards explaining the expressed data are present on the microarray. This overlooks the possibility that some genes could have been missed, which would significantly contribute towards explaining the expression levels seen. This approach also uses discretised expression data where expression levels are seen as independent separate values rather than as part of a continuum. Although this approach has many problems it does illustrate the usefulness of using a Bayesian modelling approach to generate networks from gene expression data without the use of any prior information. A general evaluation of the Bayesian method of generating regulatory networks was carried out by Husmeier (Husmeier 2003) using data simulated from a realistic biological network. This was then used to infer regulatory networks using the Bayesian learning method. The study concluded that small local regulatory networks could to a certain extent be recovered using this approach. However this is dependent upon the quality of the prior information given to the model as inevitably accurate priors enable more accurate prediction to be made by the model. Husmeier (Husmeier 2003) emphasised the importance of proceeding with caution as such networks will almost certainly return highly spurious regulatory relationship nodes which will only increase with an increase in the number of genes to be modelled. Thus the real networks one wishes to find are obscured by spurious relationships, hence the user must compromise between the number of true regulatory relationships one wishes to find and the number of spurious nodes one is willing to accept.

In an attempt to rectify the problems highlighted previously by Friedman (Friedman, Linial et al. 2000), a new model has been developed by Beal (Beal, Falciani et al. 2005). The approach employed by Beal et al uses a class of DBN known as Linear Dynamical Systems (LDS) otherwise known as Kalman Filters (Kalman 1960) to model the data. The Beal model has extended upon the principles of the original Kalman filter model so that it can accommodate continuous gene expression data using an “output to input” feed forward loop and also model unknown factors that contribute towards explaining the expressed data. The Beal model also uses new sampling techniques to try and improve the accuracy of generated networks. The ability to model unknown factors is crucial as mRNA levels are a complex mix of a variety of events including the rate of transcription and mRNA degradation which would undoubtedly have an impact when modelling the data. The principles behind this model are explained in more depth later in this chapter. The Beal model has been used previously to successfully model genes associated with T-cell activation and the resulting models reflect many of the processes supported in the biological literature

(Rangel, Angus et al. 2004). For example the gene *Fyb* was shown to be involved in many regulatory relationships in the recovered model. The genes that were shown in the model to be directly regulating *Fyb* were also shown to have biological functions related to the inflammation response. This fits well with the literature which states that *Fyb* acts as an adaptor molecule for T-cell signalling (Rangel, Angus et al. 2004). This work was then expanded by Beal to include a new method of estimating the optimum number of hidden states that best explains the expressed data. The method uses a Variational Bayesian Expectation Maximisation Algorithm (VB), which acts to integrate out all hidden variables that do not contribute towards explaining the expressed data (Beal, Falciani et al. 2005). This Variational Bayesian Learning method is discussed in more detail later in this chapter. The models generated using this new method of hidden variable estimation compared favourably with the bootstrapping method previously used by Rangel (Rangel, Angus et al. 2004). At least 60% of models generated using the VB method returned the same regulatory relationships between *Fyb* and its *IL-2* target gene as first shown in the study performed by Rangel (Rangel, Angus et al. 2004). The method also estimated a different optimum number of hidden states *k* for the dataset (*k* =14), which revealed a new sub network representing the regulatory relationships between 3 members of the Jun protein family. This relationship fits well with a hypothesis proposed by Weitzman (Weitzman, Fiette et al. 2000) that Jun proteins interact with one another to form dimers to form the AP-1 transcription factor. Experiments performed to over express these genes led to the hypothesis that these genes play an important role in regulating cell proliferation and apoptosis. The studies performed by Rangel (Rangel, Angus et al. 2004) and Beal (Beal, Falciani et al. 2005) highlight the usefulness of using the Beal model to generate regulatory networks from microarray expression data and the VB algorithm to estimate the optimum number of hidden states.

The aims of this study are to firstly identify potential signalling networks up-regulated during plant defence responses to infection by *H. arabidopsidis*. Secondly, and most importantly, it was to critically assess the viability of a set of methods (see introduction to modelling methodology section) to generate a feasible transcriptional model from which hypotheses of transcriptional regulation of downstream targets could be built.

4.2 Introduction to modelling methodology

This section describes the principal components of the modelling methodology and how these components have been modified and applied in an attempt to infer transcriptional networks from gene expression profiling data.

4.2.1 Timecourse algorithm

The first step in the process of elucidating possible transcriptional networks in this particular dataset is to determine which of the genes are showing a significant change in gene expression over the course of the experiment. This is seen as an indicator that the gene is involved in biological processes motivating the experiment, particularly in time course experiments where gene expression levels are measured at a series of time points after applying a particular treatment. In this experiment the treatment is the spraying of three batches of seven-day-old Arabidopsis seedlings with *H. arabidopsidis* isolates EMCO5, MAKS9 and a water control. One of the challenges of differentiating between genes that vary in expression over time and those which are invariant, is obtaining data from sufficient time points to enable the use of the appropriate statistical techniques. The approach taken to distinguish between significantly and non-significantly changing genes is to use Hotellings T^2 statistic (Hotelling 1931) in the program Timecourse developed by Tai and Speed (Tai and Speed 2006). The program will perform Hotellings T^2 statistic on the expression levels of each biological replicate at each time point simultaneously and determine how different they are from a null hypothesis of zero representing no change over time, whilst taking into account the amount of variation seen between the biological replicates. The greater the variation between the biological replicates the less significant any overall variation in gene expression will be. The statistic will score and rank those notably changing genes in order of the magnitude of change.

4.2.2 Bayesian Hierarchical Clustering algorithm

As part of the strategy to infer regulatory networks from genes whose expression profiles are deemed to be significantly changing over time, it is important to be able to choose which genes to model as the modelling software has limitations on the number of genes that can be modelled. It is therefore imperative that the selection of genes is based upon the functions they carry out. This in turn is dependent upon the quality and accuracy of the Gene Ontology (GO) annotations used to describe the genes. Unfortunately the process of annotation for Arabidopsis is still very much in its infancy with many genes having little or no information on function. This is a hindrance in the process of gene selection as these genes may play an important role within the regulatory network but may not be selected because their function is still largely unknown. In order to better aid the selection of significant genes a Bayesian Hierarchical clustering algorithm was used to cluster the microarray time course data. Cluster analyses of microarray time course data have been used in the past to identify groups of genes that are functionally related because they respond similarly to the same treatment i.e. inoculation of the host by a pathogen. The theory is that because their responses are similar it can be hypothesised that at the transcriptional level they are likely to be controlled by the same transcription factors and regulatory pathways (Heard, Holmes et al. 2005). Therefore by clustering the genes based upon their expression profiles, putative functions can be assigned to genes with unknown function if they fall into clusters with well-annotated genes with similar profiles. This not only acts to provide candidates for functional characterisation studies but it allows the modelling of uncharacterised genes based on the assumption that genes with similar expression profiles will have the same function.

As mentioned before this approach centres upon the use of a Hierarchical clustering algorithm developed by Heard (Heard, Holmes et al. 2005). Traditionally clustering methods use a "bottom up" or agglomerative approach whereby each data point is assigned to its own cluster and then the program iteratively merges the closest clusters together until the data belongs to a single cluster. The closest clusters are usually chosen based upon the distance measure i.e. the Euclidean distance or distance between the nearest data points. The problem with using these approaches is that there is no accurate way of determining the best number of clusters to use for the data or to know the distance at which two clusters should be merged. There is also the problem that traditional algorithms require a single data point in a cluster, which is no good for modelling multivariate time course microarray expression data. Heard and associates (Heard, Holmes et al. 2005) have developed a Bayesian model based agglomerative approach to cluster time course microarray expression data. The program begins with assigning each gene to its own cluster. The method then uses a Bayesian nonlinear regression model for the time series to describe each gene in each separate cluster. This model is defined as follows:

$$y_g = X\beta_g + \sigma_g \varepsilon_g$$

Where y_g represents a concatenation of the gene expression levels at each of the time points and X the sum of the design matrix, β_g represents a concatenation of hidden variables affecting gene expression at the different time points, σ_g is the error variance and ε_g represents a concatenation of the normally distributed standard error at each time point.

The next process is to cluster the genes based upon how closely their models resemble one another and find the optimum number of clusters for the dataset. This is achieved by following five steps:

First the marginal posterior probability of each cluster is calculated. This means that for each cluster containing a gene calculate the posterior probability of that cluster, then calculate the posterior probability that the optimum number of clusters is the sum of the clusters. For example, if every gene has its own cluster and there are 36, 000 genes then what is the posterior marginal probability that the optimum number of clusters is 36,000?

For the second step the program then takes every possible pair of clusters and calculates the multiplicative increase in the marginal posterior that would be gained by merging the two clusters. This way a "map" of cluster closeness can be generated.

In the third step take a cluster and identify the cluster that is closest to it and merge them if the merged value significantly increases the marginal posterior probability.

For the fourth step recalculate the marginal posterior probability that this is the optimum number of clusters.

For the fifth step calculate a new "closeness map" to identify a new nearest cluster.

This is an iterative process so steps 1-5 are repeated until there is only a single cluster remaining containing all the genes. Since each value of the marginal posterior probability is calculated at each iteration and stored, the optimum number of clusters can be determined by finding the cluster number that gives the maximum marginal posterior probability. Thus the outcome is an optimum number of clusters containing genes with a similar model. To ensure that the genes are correctly placed in a cluster the marginal likelihood is estimated for the model representing each cluster. This means that for each model, remove those parameters that do not contribute towards explaining the final expression values of the genes and which over complicate the estimation process. The marginal likelihoods for each cluster are multiplied together to give the likelihood function for the entire dataset. This contains all parameters used to explain the expression values for the entire dataset. The expectation maximisation algorithm is then used to maximise the parameters that explain the final expression values. Thus the final output is a set of clusters containing genes whose profiles are similar to one another.

4.2.3 Go Stat

The genes in each cluster were analysed using the Gene Ontology (GO) statistic to find overrepresented GO terms within each cluster (Beissbarth and Speed 2004; Beissbarth 2006). The program first determines the GO terms for all the genes analysed in each group. The program then counts the number of appearances of each GO term for the genes in the group and then for each GO term in a reference group (in this case The Arabidopsis Information Resource (TAIR) database (Rhee, Beavis et al. 2003)). A p-value is calculated representing the probability that the number of counts of the GO term occurred through random distribution of the term in the test group and reference group. The genes with the most specific GO terms will have the lowest p-values. The resulting output is a list of genes ranked in order of the lowest p-values and therefore the most specific GO terms.

4.2.4 The principles of the Kalman filter

The Beal model is a variation of the Linear Dynamical System or Kalman filter, which in turn is a subclass of Dynamic Bayesian Networks (DBN) used to model times series data. DBNs use probabilities given a sequence of observed variables to calculate the relationships between them. For the purposes of explanation, in this example the observed variables can represent a sequence of expression measurements for a single gene at a set of given time points on a microarray. At each time point there are factors that can affect the measured expression levels of the gene on the microarray such as poor RNA extraction or low mRNA levels. These factors cannot be quantified directly and are therefore "hidden" from the user and can be referred to as the hidden state. At each time point the hidden state variables impact upon the observed expression values, therefore when modelling regulatory relationships between genes, the hidden variables must be taken into account. The Kalman filter captures this process of change in the hidden state from time point to time point, which in turn impacts upon the observed expression values at each time point (Kalman 1960). The basic Kalman filter model is as follows:

$$x_t = Ax_{t-1} + w_t, \qquad \text{w. ~Gaussian (0, } Q\text{)}$$

$$y_t = Cx_t + w_t, \qquad \text{v. ~Gaussian (0, } R\text{)}$$

Where **x** represents a *k*-vector of hidden state variables that cannot be observed directly but impact upon **y**, a *p*-vector of observed variables that can be measured. *A* represents a (*k* x *k*) transition matrix, which captures the process of change in the state of hidden variables over time and *C* is a (*p* x *k*) observation matrix that captures the change in observed variables over time. w and v are variables that represent the state and observed variable noise respectively and *Q* and *R* represent the covariance matrices associated with them. The noise represents imperfections in the data (in this case a microarray) caused by a random set of variables such as temperature, vibrations or even dust specks on the laser. These noise variables are thought to occur randomly irrespective of any time index hence why **w** and **v** are followed by a "." to emphasise their independence from any time step. It is important to take this noise into account when estimating the hidden state and observed variables as they invariably have an impact. The state noise is also considered Gaussian; that is to say normally distributed. Because the model captures the process of change in the hidden state (and its subsequent impact on the observed values) over a single time step it is defined as being a 1st order Markov model. This is because it has a memory of 1; therefore the probabilities of the possible values of the next hidden state depend on the values of the previous state.

In practical terms the model answers the following question:

"*Given a set of observed variables and parameters what can be said of the hidden state at time point **t**?* " (Roweis and Ghahramani 1999).

Thus the model offers the user information as to what caused the change in the observed set of variables across the time points.

In the context of this chapter, determining the regulatory relationships between genes over time is the main aim of this study, therefore it is important that the effect of both the hidden and observed gene expression levels at each time point is not only captured but that its impact is also incorporated into the values at the next time point. The Beal model incorporates the principles of the Kalman filter whilst extending it to include matrix ***B***, which captures the influence of the observed variables from a previous time point on the current hidden state and matrix ***D***, which captures the influence of the observed variables from a previous time point on the current observed variables. By including these matrices Beal is able to model the influence of previous observed measurements back on to the current hidden state and observed measurements. The Beal model is thus defined below:

$$x_t = Ax_{t-1} + By_{t-1} + w_t, \qquad \text{w. \sim Gaussian } (0, Q)$$

$$y_t = Cx_t + Dy_{t-1} + w_t, \qquad \text{v. \sim Gaussian } (0, R)$$

Where x represents a set of hidden state variables, y represents a set of observed variables. Matrix *A* represents a transition matrix capturing the change in the hidden state variables over time *t*. Matrix *B* represents the effect of observed values from a previous time point on the current hidden state. *C* symbolises an observation matrix capturing the change in observed variables over time and *D*, a matrix containing the effect of the previous observed values on the current observed values. w and v symbolise the state and observed variable noise respectively. By incorporating the influence of observed variables from a previous time point into the values of the current time point, the model acts as a feedback loop whereby the outputted observed and hidden variables at one time point act as input values for the next. Thus the gene expression levels and hidden state variables at time *t*-1 are used as input values to help determine the hidden state and explain the gene expression values at *t*. Whilst the Beal model is still a 1st order Markov model the inclusion of matrices *B* and *D* has allowed the capture of gene relationships which are higher than 1st order.

Until now the example provided centred upon modelling the expression values of a single gene over a given set of time points, however the model is designed to estimate the influences of a set of genes on one another over a given set of time points. In this case all matrices (*A*, *B*, *C* and *D*) would contain values for a set of genes at time *t*. The model can then be used to characterise both direct gene-gene regulatory relationships and those that occur through the hidden state. For example, to observe the direct effect of gene *a* at *t*-1 on gene *b* at *t*, one must look at matrix element $[D]_{ba}$. Thus to capture all the effects of the hidden and observed states of one gene on another over a single time step, the matrices must be combined. A function of the model to do this is shown below:

$$y_t = (CB + D)y_{t-1} + r_t$$

Where y_t represents the observed expression level at time *t* and *CB* + *D* represents the influences of the hidden state on gene expression values, the effect of gene levels from a previous time point on the current hidden state and the effect of gene levels from a previous time point on the current gene expression levels incorporated into a single matrix. Also, **r** represents contributions from both hidden state and observed variable noise from all previous time points. The *CBD* matrix generated encapsulates all direct and indirect gene-gene regulatory relationships over a single time step. The

Standard score or Z score values from this matrix (CBDZ) will be used to determine which genes are showing “significant” regulatory relationships. This is defined as those genes whose relationship scores differ from a default normal distribution of zero. Values at or close to zero indicate the genes involved do not have any regulatory influence on one another. For genes to be considered as having regulatory influences on one another, either directly or indirectly, their CBDZ values must be at least 1.69 standard deviations away from the mean of zero, which equates to around a 90% confidence that the values are significantly different.

4.2.5 Parameter learning

The use and extension of the Kalman filter in Beal’s model to estimate the hidden factors present in a dataset given some observations and parameters, has enabled the inference of gene- gene regulatory relationships over time. There is however, a second problem that must be solved to accurately determine these regulatory relationships:

“*Given only an observed sequence (or perhaps several sequences) of outputs find the parameters which maximise the likelihood of the observed data*” (Roweis and Ghahramani 1999).

In other words, given the observed values find the path or parameters that is most likely to have caused the data. The parameters can be defined as *A, C Q, R* that were defined previously. To solve this problem in the past Beal has used the *expectation-maximisation* (*EM*) algorithm (Baum and Petrie 1966). The *EM* algorithm uses the Kalman filter to estimate the number of hidden states using the current model parameters (Stoffer 1982). The estimated hidden states are then used to estimate and generate new model parameters that are then used again to estimate new hidden states and so forth. Thus the *EM* algorithm is able to iteratively sum up all possible ways in which the observed sequences could have been generated by the various parameters and calculate their log likelihood.

The problems using the maximum likelihood methods in parameter estimation is that the likelihood will usually be higher for more complex explanations for the observed data as it incorporates a greater number of estimated hidden variables (Beal and Ghahramani 2002). This can lead to two of the biggest problems that occur when trying to infer regulatory relationships, the first being that of under fitting; this is where a model with too few hidden states is chosen and as a result not all of the possible indirect gene-gene regulatory influences are captured. The model will try and infer direct gene-gene regulatory relationships where indirect gene-gene relationships would explain the data more accurately and the result is a loss of information about the data. The second problem is that of over fitting, which applies more to the *EM* algorithm, is where a model with too many hidden states is chosen to describe the data and as a result the model will try and infer incorrect regulatory relationships from random noise variables that occur in the data. As a result of these issues Beal uses a *Variational Bayesian EM* algorithm to overcome the problem of over fitting as it treats all parameters of the model as unknown quantities and integrates out all those parameters which are not thought to meaningfully contribute towards explaining the data. By doing this, the algorithm will penalise overly complicated models and promote models with simpler explanations. This type of integration is referred to as adhering to the principle of *Ockham's razor*, which is to always favour the simplest explanation for a result.

Thus with this algorithm the optimum number of hidden states for the model can be estimated whilst avoiding the problem of over fitting. An important factor in assisting the estimation of new model parameters and extending the principles of *Ockham's razo*r is incorporating prior information or knowledge to remove complexity. This not only allows the user to incorporate existing prior biological knowledge into the model but it also helps to distinguish between hidden variables contributing towards meaningful regulatory relationships and those generated from random noise variables. This process is called Automatic Relevance Determination (ARD) whereby parameters such as hidden variables shown to be irrelevant by a prior are no longer included when estimating new model parameters to find the optimum number of hidden states. Thus using the *Variational Bayesian EM* algorithm the model whose parameters give the greatest log likelihood for the observed data is then used to calculate the gene- gene regulatory relationships.

4.2.6 Developmental senescence

Senescence is a highly regulated process of resource remobilisation that represents the final stage of the leaf growth and development cycle. During senescence leaf cells undergo both structural and metabolic changes. One of the earliest cell structure changes is the breakdown of organelles such as the chloroplast. On the metabolic level there is an increase in catabolism of cellular components such as lipids, proteins and nucleic acids, into nutrients that can be reallocated to new growth areas of the plant such as the developing seeds. The process however, results in the gradual degradation of plant leaf tissue and organs leading to the death of the leaf (Lim, Kim et al. 2007).

Senescence is controlled by the developmental age of the plant leaf and is, therefore, regulated by numerous external and internal factors. External factors affecting the leaf developmental age include stresses caused by extreme changes in temperature, drought conditions, nutrient depletion, shading and infection by a pathogen. Internal factors comprise several different hormones from plant hormone signalling pathways that are thought to regulate the developmental age of the leaf and ultimately the onset of senescence. Therefore, it is likely that many of the processes involved will overlap significantly with those in disease resistance (Chen, Provart et al. 2002; Lim, Kim et al. 2007; Saibo, Lourenco et al. 2009).

4.2.7 Key hormones affecting the regulation of the onset of senescence

The identification of a number of senescence-enhanced genes in the past decade has greatly increased our knowledge of the process of senescence, however the signalling pathways that initiate subsequent transcription factor activity are still undefined. Plant hormonal signalling pathways are thought to have a role at all stages of senescence from early initiation towards the final stages of cell death as different hormones affect different stages. The plant hormone cytokinin has a negative effect on senescence as it has been found to significantly prolong the life span of plant organs and cytokinin levels drop during leaf senescence. Transgenic modification of cytokinin biosynthesis during the senescence phase causes a delay in senescence in plant organs including the leaves (Kim, Ryu et al. 2006). These results are consistent with the findings that genes involved in cytokinin synthesis are down regulated during senescence, whereas genes involved in cytokinin degradation are up regulated. Ethylene, unlike cytokinin, has been shown to be have a major role in leaf senescence, fruit ripening and flowering senescence (Abeles, Dunn et al. 1988) as a positive regulator of leaf senescence (Van der Graaff, Schwacke et al. 2006). Studies show an up regulation in the genes coding for 1-aminocyclopropane -2 carboxylate (ACC) synthase, ACC oxidase and nitrilase followed by an increase in ethylene levels in senescing leaf tissue (Mishina, Lamb et al. 2007). The Arabidopsis mutants *etr1* and *eir1*, deficient in ethylene detection and sensitivity respectively, show a considerable delay in leaf senescence (Oh, Park et al. 1997). However, it must be noted that over-expression of ethylene does not induce an earlier onset of senescence, indicating that senescence onset is not triggered by ethylene levels alone but may also rely on other age dependent factors (Jing, Sturre et al. 2002). Jasmonic acid (JA) was first verified to have a key role in the promotion of senescence in oat leaves (*Avena sativa*) (Hisamatsu, Goto et al. 2006). Its senescence promoting effect was found to be stronger than that of Abscisic acid (ABA) (Ueda and Kato 1980). Other experiments have shown that exogenous application of JA caused early senescence in attached and detached leaves in wild-type Arabidopsis, but failed to induce the same effect in JA-insensitive mutant plants, suggesting that the JA-signalling pathway is essential for JA to promote leaf senescence (He, Fukushige et al. 2002).

Many signalling pathways are used for both plant stress response and for senescence. For instance, the plant hormone ABA is known to have roles in plant reactions to stress induced by environmental factors such as drought, extreme temperature change and osmotic pressure. Although ABA’s exact role in leaf senescence has yet to be directly verified, evidence suggests there is an increase in ABA levels during senescence and accordingly there is also an increase in the expression of genes that code for enzymes involved in ABA synthesis during senescence (Buchanan-Wollaston, Harrison et al. 2007). It is also thought that increases in levels of ABA trigger the expression of senescence-associated genes (SAG) (Weaver, Gan et al. 1998); (Skriver and Mundy 1990). Salicylic acid (SA) has also been identified as being an important component in many plant responses to environmental stresses such as pathogen attack, ozone exposure and ultra-violet radiation. Many biotic and abiotic stresses can initiate the onset of senescence as numerous genes that are commonly induced in response to stress are also expressed during senescence (Gupta, Willits et al. 2000; Rao and Davis 2001; Rao, Lee et al. 2002). This indicates that there are likely to be common pathways leading to gene expression during senescence and in response to stress; for example it has recently been shown that the SA signalling pathway is involved in the control of gene expression during senescence such as the senescence enhanced gene SAG12. Transgenic Arabidopsis plants deficient in the SA pathway were used to test senescence enhanced gene expression which result in altered gene expression of many of the genes including SAG12 which was vastly reduced or undetectable (Morris, Mackerness et al. 2000).

4.2.8 Regulatory Genes involved in senescence

Although key classes of plant signalling hormone have been identified as having important roles in the promotion and onset of senescence, still little is known about the actual regulatory mechanisms governing senescence. The identification of a large number of SAG genes using microarray analysis has revealed some to encode regulatory components involved in signal transduction. Closer analysis of gene function revealed them to be important regulators in the overall process of senescence. The key regulatory factors can be separated into several families based on function and expression pattern and are further discussed below. Of the regulatory factors identified in Arabidopsis, some 96 were identified as transcription factors that were up-regulated during senescence. The largest families included in this group are the AtMYB, NAC, C2H2-type zinc finger, and WRKY and AP2/EREBP transcription factor proteins. To date, only three have been characterised for function; *WRKY6, WRKY53* and *AtNAP*. Both *WRKY53* and *WRKY6* have been shown to play a positive role in inducing senescence as over expression of both of these genes has lead to the early onset of leaf senescence (Miao, Laun et al. 2004). Further to this, both *WRKY53* and *WRKY6* have been implicated in regulating plant stress responses to pathogen infection (Robatzek and Somssich 2002). The *AtNAP* gene is part of the *NAC* family of genes that are involved in plant shoot meristem growth and development and defence responses. *AtNAP* encodes a NAC transcription factor and is a positive regulator of senescence as transgenic mutant knockouts of *At*NAP show a significant delay in senescence (Guo and Gan 2006).

Examination of the hormones and genes involved in senescence has indicated a close link between the pathways that regulate plant stress response and those that regulate senescence as the majority mediate signalling pathways involved in both processes. In fact, it can be speculated that senescence itself is a plant stress response to the environmental and age dependent factors that affect the plant. Despite the work currently being undertaken to characterise the function of senescence regulators, still little is known of the exact pathway that mediates the process. Thus the aims of this work were redirected to using and evaluating the modelling methodology described earlier as a way of identifying established networks governing senescence and also predicting new networks.

4.3 Aim

To use systems modelling approaches to infer Arabidopsis gene regulatory networks altered during stress:

1. Previous analyses have shown that different isolates of *H. arabidopsidis* contain varying complements of effector alleles. Therefore, the concept of this analysis was to compare the network profiles generated by different isolates to give an insight into the effects resulting from varying effector complements.
2. Developmental senescence shows many of the gene expression stress response profiles that occur during biotic and abiotic stress. Therefore, I carried out network inference on a senescence time series.

4.4 Methods

4.4.1 Arabidopsis Pathogen Experiment

All laboratory based experimental procedures relating to the Arabidopsis pathogen experiment were performed by Sharon Hall, Peter Bittner-Eddy, Mary Coates and Kate Fisher.

4.4.1.1 Plant Growth and Inoculation

7 day-old wild type Arabidopsis seedlings (Col-0) were inoculated with 25ml of a water solution containing 100 spores of *H. arabidopsidis* isolates *Maks9*, *Emco5* and a water control. Subsequently after inoculation, four cotyledons were harvested (each treated as a biological replicate) for each treatment at 0, 8, 16, 24, 32, 40, 48, and 56 hours. This resulted in 8 sample time points with four biological replicates at each time point.

4.4.2 Microarray analysis

4.4.2.1 RNA preparation and labelling

For each experimental treatment total RNA was isolated from individual cotyledons using TRIzol reagent (Invitrogen) and amplified using the MessageAmp II aRNA Amplification kit (Ambion) in accordance with the kit protocol with a single round of amplification. Cy3 and Cy5 labelled cDNA probes were prepared by reverse transcribing 5µg of aRNA with Cy3- or Cy5- dCTP (GE Healthcare) and a modified dNTP mix (10mM each dATP, dGTP and dTTP; 2mM dCTP) using random primers (Invitrogen) and SuperScript II reverse transcriptase (Invitrogen). Labelled probes were purified using QiaQuick PCR Purification columns (Qiagen), freeze-dried, resuspended in hybridization buffer (25% formamide, 5xSSC, 0.1% SDS, 0.5µg/µl yeast tRNA [Invitrogen]).

4.4.2.2 Microarray experiments

The microarray experiments were carried out using the CATMA (version 3) microarray (Allemeersh *et al.* 2005; http://www.catma.org). A complex loop design was applied to extract the maximum information from the two colour hybridisation experiments. Following the layout of the experimental design (Appendix F), combinations of labelled samples were hybridized to slides overnight at 42^{o}C. Following hybridization, slides were washed and scanned using an Affymetrix 428 array scanner at 532nm (Cy3) and 635nm (Cy5). Scanned data were quantified using Imagene 7.5.0 software (BioDiscovery, Inc.).

4.4.3 Senescence Experiment

Vicky Buchanan–Wollaston and associates performed all laboratory based experimental procedures relating to the Arabidopsis senescence long day experiment. Further details of this experiment will be described in their forthcoming paper in preparation.

4.4.3.1 Plant Growth

Arabidopsis seeds (Col-0) were stratified at 4^{o}C in the dark for 48 hours and then sown onto Arabidopsis compost mix (Levingtons F2 compost:sand:vermiculite 6:1:1). Plants were grown at 20^{o}C under a 16h/8h light/dark cycle at 70% relative humidity and 250µmol m^{-2} s^{-1} light intensity. Leaf 7 was tagged with cotton 18 days after sowing (DAS). Sampling of leaf 7 started at 20 DAS and continued every other day until full senescence was reached (40 DAS). Leaves were harvested twice on each sampling day, 7h and 14h into the light period. This resulted in 22 sample time points. Ten leaves were collected at each time point.

4.4.4 Microarray analysis

4.4.4.1 RNA preparation and labelling

Total RNA was isolated from four individual leaves for each time point (biological replicates) using TRIzol reagent (Invitrogen) and amplified using the MessageAmp II aRNA Amplification kit (Ambion) in accordance with the kit protocol with a single round of amplification. Cy3 and Cy5 labelled cDNA probes were prepared by reverse transcribing 5µg of aRNA with Cy3- or Cy5- dCTP (GE Healthcare) and a modified dNTP mix (10mM each dATP, dGTP and dTTP; 2mM dCTP) using random primers (Invitrogen) and SuperScript II reverse transcriptase (Invitrogen). Labelled probes were purified using QiaQuick PCR Purification columns (Qiagen), freeze-dried, resuspended in hybridization buffer (25% formamide, 5xSSC, 0.1% SDS, 0.5µg/µl yeast tRNA [Invitrogen]).

4.4.4.2 Microarray experiments

The microarray experiments were carried out using the CATMA (version 3) microarray (Allemeersh *et al.* 2005; http://www.catma.org). A complex loop design was applied to extract the maximum information from the two colour hybridisation experiments (A Mead, in preparation). Following the layout of the experimental design (Appendix G), combinations of labelled samples were hybridized to slides overnight at 42^{o}C. Following hybridization, slides were washed and scanned using an Affymetrix 428 array scanner at 532nm (Cy3) and 635nm (Cy5). Scanned data were quantified using Imagene 7.5.0 software (BioDiscovery, Inc.).

4.4.5 Software

4.4.5.1 Image processing

Image processing was carried out using ImaGene® software version 7.5.0 BioDiscovery Inc. http://www.biodiscovery.com/index/imagene to yield the raw signal intensities.

4.4.5.2 Normalisation of raw signal intensities

Raw data was normalized and analyzed in GeneSpring® GX software version 7.0 (Silicon Genetics, Redwood City, CA). The raw array data was normalised using GeneSpring by normalisation per chip to the 50th percentile, per spot divided by the control channel and per gene normalised to the median. GeneSpring was also used to filter out genes that were flagged absent. The resulting output was a set of normalised signal intensities from which the gene expression over time could be calculated and the gene ranked from highest change to lowest change in expression over time.

4.4.5.3 Ranking, clustering and modelling differentially expressed genes

Calculation of change of gene expression over time and subsequent ranking was performed using BioConductor (Gentleman, Rossini, Dudoit and Hornik 2003), "The Bioconductor FAQ", http://www.bioconductor.org/docs/faq/. The package was called Timecourse (Tai and Speed 2006). Clustering of gene expression profiles was carried out using SplineCluster, (Heard, Holmes et al. 2005) http://stats.ma.ic.ac.uk/n/naheard/public_html/software/splinecluster/index.html. Over represented GO terms within each cluster were identified using GO stat, (Beissbarth and Speed 2004). Genes were modelled using the VBSSM V3.3.5 tool kit (Beal, Falciani et al. 2005), using MATLAB version R2007a, http://www.mathworks.com.

4.5 Results

4.5.1 Identification of gene networks involved in responses to pathogen invasion

The $\log_2$ gene expression ratio values following infection of Arabidopsis by *H. arabidopsidis* isolates *Emco5, Maks9* and a water control were analysed using the Bioconductor package Timecourse (Tai and Speed 2006). The aims were firstly to differentiate between genes showing a significant change in expression over time under the 3 separate treatments and those that do not change and rank them according to the intensity of change. The second aim was to identify genes showing the biggest difference in expression between the treatments..

Timecourse analysis of the differences between *Maks9* gene expression compared to that of *Emco5* returned Arabidopsis gene *At2g42530* as the top ranked most differentially expressed gene over time based on Hotellings T^2 statistic. The annotations revealed that *At2g42530* codes for the cold-regulated protein *cor15b.* However, inspection of the line graphs (Figure 4.1) produced for each ranked gene as part of the *Timecourse* package show a disparity between the biological replicates of the gene. There is no correlation between the biological replicates of the gene under either pathogen treatment. The relative change in expression over time of *At2g42530* is also low with the gene showing around a two-fold change in expression with most changes seemingly occurring as part of the circadian rhythm. This relative lack of change is reflected in the low Hotellings T^2 statistical score of 803.2.

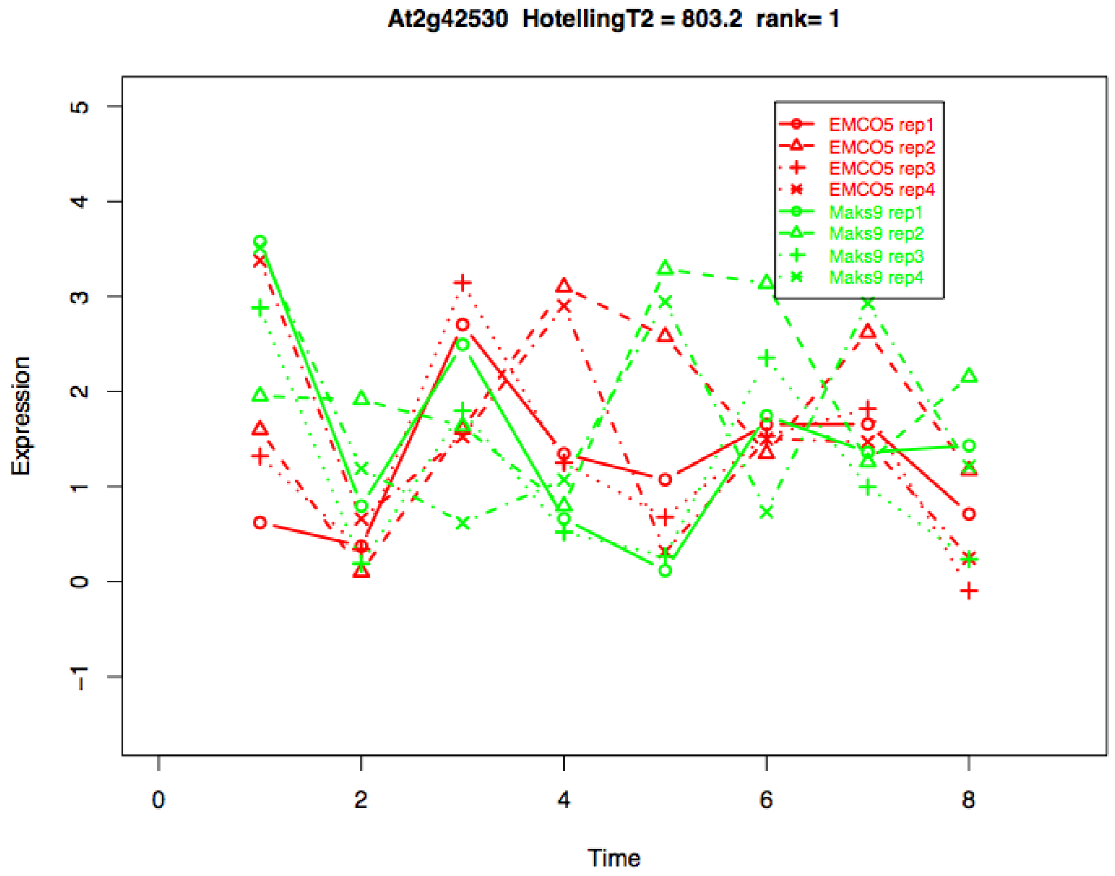


Figure. 4.1. A line graph representing the relative expression level of gene *At2g42530* during pathogen infection. This gene shows the greatest relative change in expression between *H. arabidopsidis* isolates *Maks9* and *Emco5* over the eight time points. Each green line represents one of the four *Maks9* biological replicates and each red line represents one of the four *Emco5* biological replicates. The horizontal axis represents the eight time points at which Arabidopsis cotyledon samples were taken. The time period between each sample is six hours, so the time course occurred over a period of 48 hours. The vertical axis represents the fold change in expression between the sample points over time.

The trend of uncorrelated biological replicates with low Hotellings T^2 scores is replicated in the next four top ranked genes (Appendix H). Timecourse analysis of the differences between *Maks9* gene expressions over time compared to that of the water control returned *At5g11420*, an un-annotated gene as the most differentially expressed gene over time under the two conditions. Examination of the line graphs (Figure 4.2) revealed a similar trend to that seen with the top ranked genes in the analysis of *Maks9* compared to *Emco5*, non-correlation of the biological replicates along with a low Hotellings T^2 score.

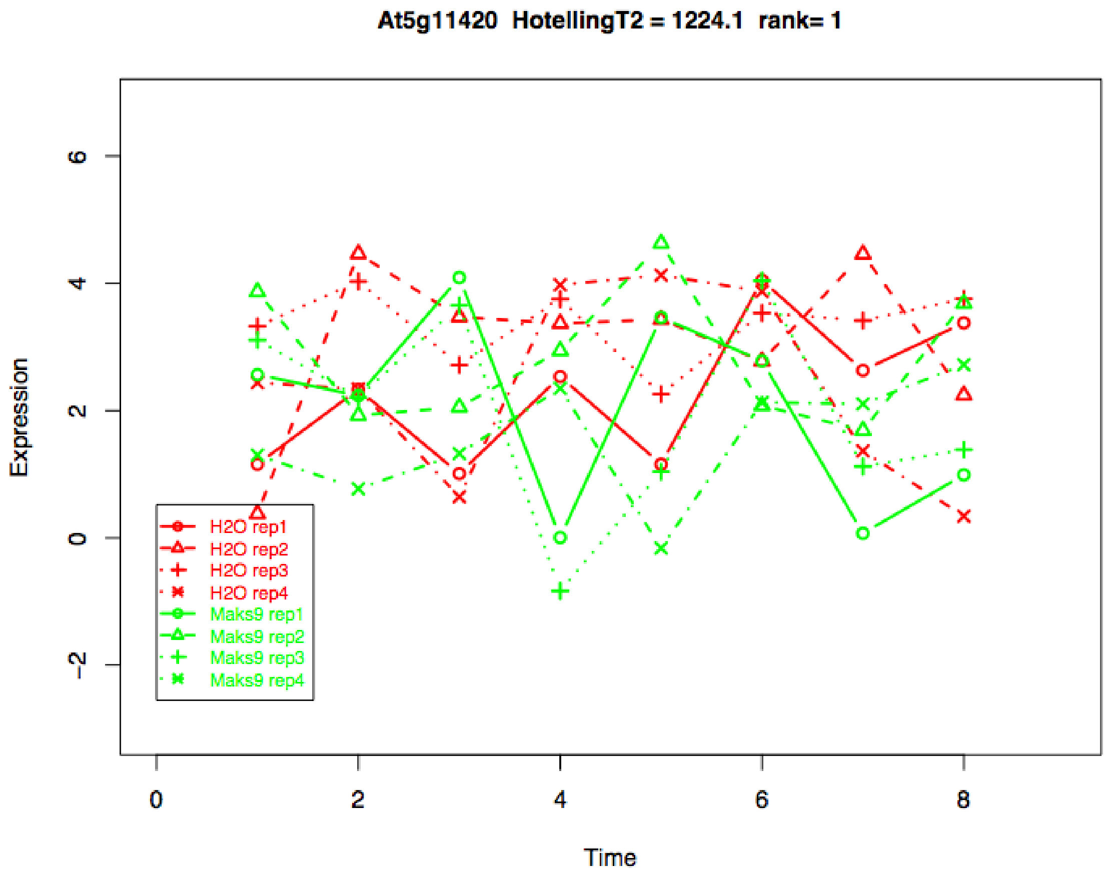


Figure. 4.2. A line graph showing relative expression levels of gene *At5g11420*, which shows the greatest relative change in expression between *H. arabidopsidis* isolate *Maks9* and *$H_2 0$* over the eight time points. Each green line represents one of the four *Maks9* biological replicates and each red line represents one of the four *$H_2 0$* biological replicates. The horizontal axis represents the eight time points at which Arabidopsis cotyledon samples were taken. The time period between each sample is six hours, so the time course occurred over a period of 48 hours. The vertical axis represents the fold change in expression between the sample points over time.

Again this trend is seen in the analysis of the differences in expression over time between *Emco5* and the water control (Figure 4.3). With the top ranked gene *At2g01990* showing similar non-correlation of biological replicates and low Hotellings T^2 scores, which are also seen in other highly ranked genes.

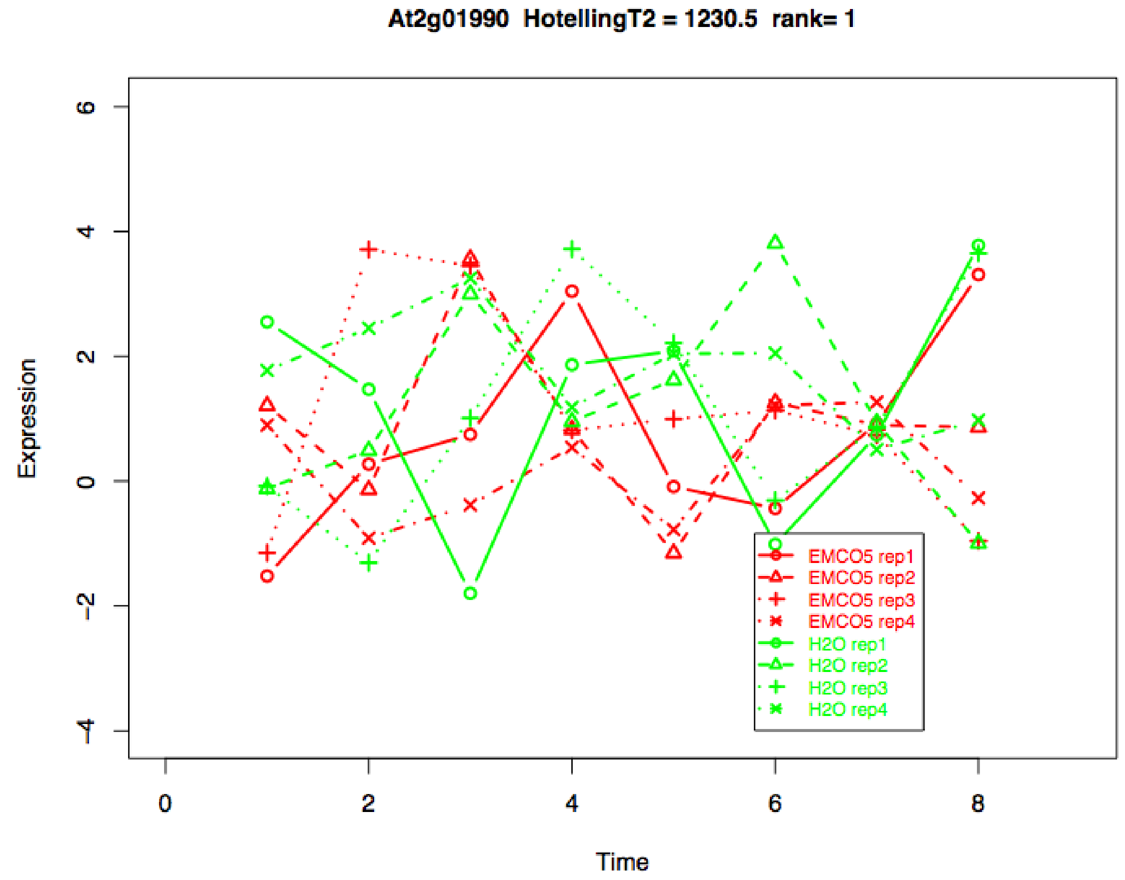


Figure. 4.3. A line graph showing the change in gene expression over time for Arabidopsis gene *At2g01990* following infection with *H. Arabidopsis* isolate *Emco5*. Each green line represents one of the four *H_20* biological replicates and each red line represents one of the four *Emco5* biological replicates. The horizontal axis represents the eight time points at which Arabidopsis cotyledon samples were taken. The time period between each sample is six hours, so the time course occurred over a period of 48 hours. The vertical axis represents the fold change in expression between the sample points over time.

The lack of correlation between the biological replicates in the different samples coupled with the lack of change in expression over time indicate the experiment was unsuccessful in eliciting a change in expression in the host over time in response to the treatments. It was, therefore, impossible to model putative transcriptional networks based on the results of the Timecourse analysis. The lack of correlation between the replicates in the samples would have resulted in networks being generated that reflected the noise associated with the experiment rather than biological significance.

4.5.2 Identification of networks regulating the onset of senescence

Analysis of the Arabidopsis long day developmental time course expression data was carried out using the methodology outlined earlier. The results of the analysis for each section of the methodology are described below.

4.5.2.1 Timecourse

Normalised (see methods section) Arabidopsis RNA $\log_2$ intensity values for the 22 time points for each biological replicate were used to run the Timecourse program (Tai and Speed 2006). The Timecourse algorithm uses Hotellings T^2 statistic, a variation of Students T^2, to calculate the degree of differential expression. Rather than calculate the means of the sample at each time point separately, Hotellings T^2 calculates the means of the sample at all 22 time points simultaneously and determines the level of change in gene expression over time from a null hypothesis of 0. Timecourse analysis was performed using the Arabidopsis developmental series microarray data and revealed that of the 31106 genes on the array 21090 genes had a Hotellings T^2 statistic of under 500 and therefore a low differential expression level. However, 4665 genes had a Hotellings T^2 statistic of over 1000 (Figure 4.4). Thus based upon these scores it would seem that the vast majority of genes are showing virtually no change in expression over time in response to the onset of senescence.

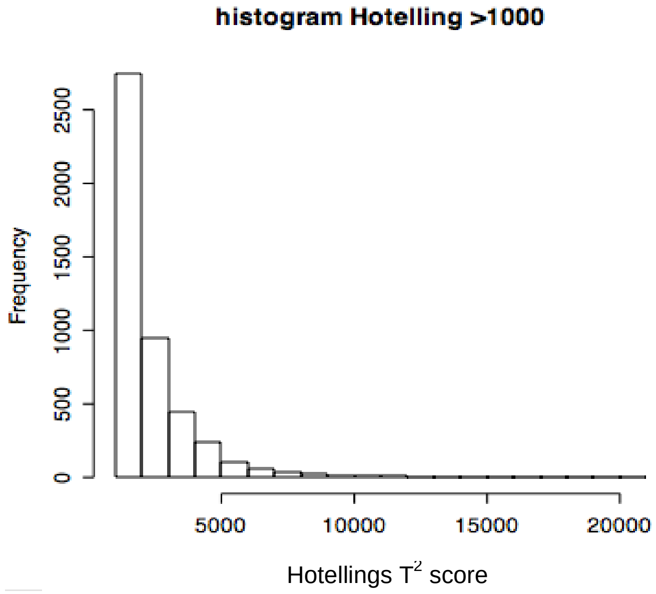


Figure. 4.4. A histogram of the proportion of Arabidopsis genes that have a Hotellings T^2 statistic of over 1000 during developmental senescence.

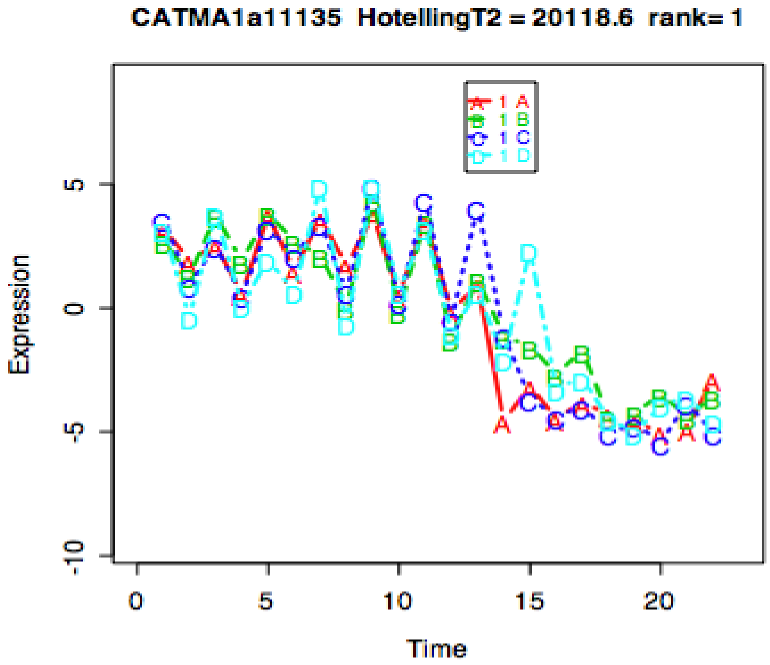


Figure. 4.5. A line graph representing the change in the gene expression profile *At1g12090* coding for a lipid transport protein (LTP), which shows the greatest relative change in Arabidopsis expression over the 22 time points. Each coloured line represents one of the four biological replicates. The horizontal axis represents the 22 time points at which Arabidopsis leaf samples were taken. The vertical axis represents the fold change in expression between the sample points over time.

Analysis of the Timecourse output revealed the most differentially expressed gene was CATMA1a11135 (Lipid Transfer Protein (LTP) *At1g12090*) (Figure. 4.5) over the 22 time points with a Hotellings T^2 statistic of 20118.6. The profile of the gene shows a steady decline in expression from sample point 12 onwards. Of the 4665 genes whose Hotellings T^2 statistic was greater than 1000, the intention was to cluster the genes based upon their expression profiles. The assumption is that genes with similar expression profiles are more likely to be regulated by the same transcriptional elements and can therefore be implicated in related processes. However, closer inspection of genes ranked between 2000 - 4000 showed a number of genes not known to play any type of role in plant stress response or senescence. One explanation for the high rank of these genes could be due to cross-hybridisation of RNA onto the microarray probes if the probes are not specific. To minimise the inclusion of genes with non-specific binding of RNA, an arbitrary cut-off of 2000 was chosen as no genes inspected in this top 2000 appeared to be there as a result of cross-hybridisation as no genes known to show no changes in expression during senescence were found. Thus only the top 2000 genes were chosen for clustering.

4.5.2.2 Agglomerative Hierarchical Clustering

The top 2000 most differentially expressed genes were clustered using the agglomerative hierarchical clustering program SplineCluster (Heard, Holmes et al. 2005). The program clusters the genes based upon their expression profiles and determines the optimum number of clusters that represent the differing expression profiles exhibited by the genes (see Chapter 2). The program clustered the genes into 82 separate clusters and also generated a file containing line graphs of all the expression profiles of the genes in each of the clusters over the 22 time points. The program also produced a dendrogram that shows the how each cluster was iteratively merged during the clustering process.

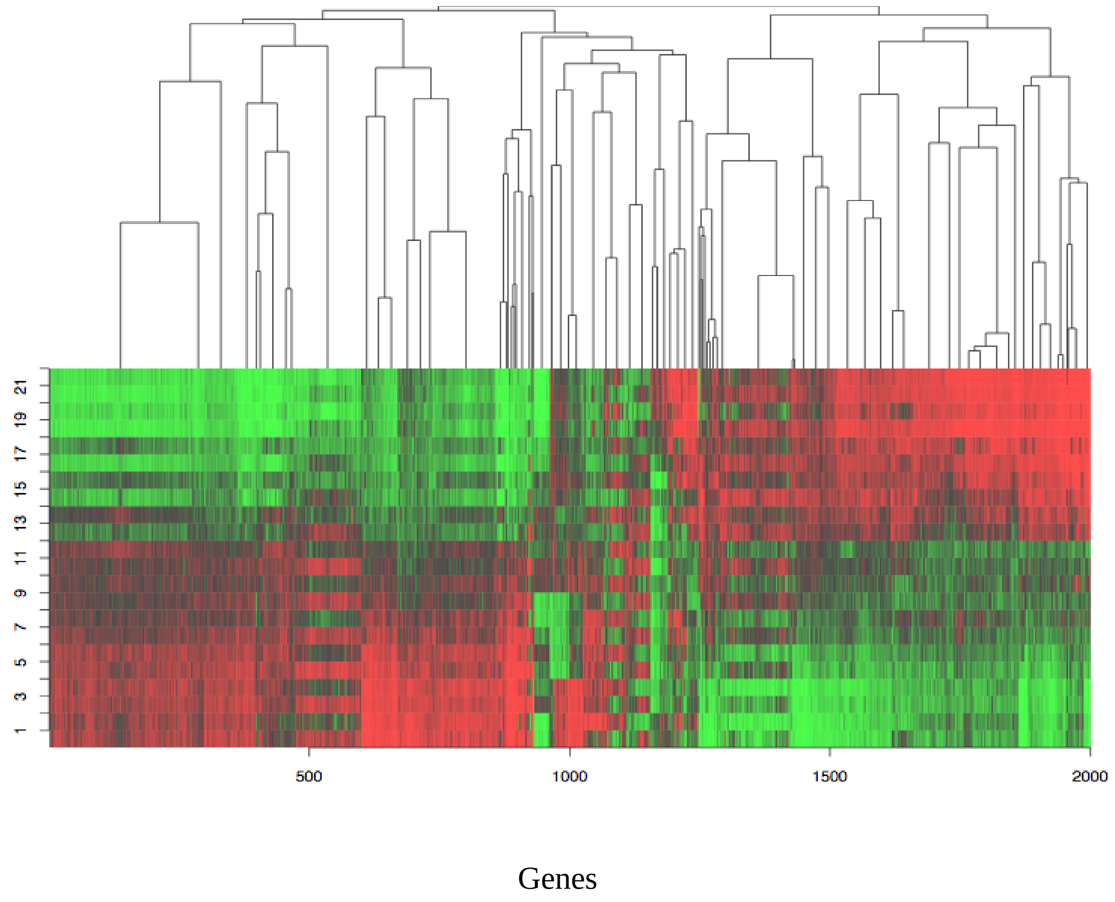


Figure. 4.6. A dendrogram of the top 2000 differentially expressed Arabidopsis genes, from the leaf developmental time course experiment, over 22 time points. Green represents down regulation and red represents up regulation of gene expression. The horizontal axis represents each of the 2000 genes clustered. The vertical axis represents the 22 time points.

Figure 4.6 represents a dendrogram of the top 2000 most differentially expressed genes. It shows the genes clustered hierarchically based on similarity of expression pattern over the course of the experiment. The laddering effect seen in some of the clusters on the dendrogram represent a rise and fall in gene expression on a daily cycle consistent with them being under the regulation of the circadian clock. Initial analyses show that the clusters can loosely be divided into 3 groups; those which show a decrease in expression, those which show an increase in expression and those which show no real change in expression over time. This is demonstrated by the dendrogram, which shows a clear breakpoint at sample point 12 where the genes begin to show either an increase, decrease or no change in expression.

Closer inspection of the clusters also reflected this clear change in expression at sample point 12. Figure 4.7 shows the expression profiles of cluster 10 of the 82 separate clusters generated by the SplineCluster program. The expression profile shows sharp increases and decreases in expression from pm to am, representing circadian clock control of gene expression. However the overall trend over the 22 time points is a two-fold decrease in expression from sample point 12 onwards. This is in keeping with the idea of a gradual deterioration of gene product function following the onset of senescence. The ontologies of the genes in this cluster show them to be chlorophyll A-B binding proteins, protochlorophyllide reductases and photosystem I reaction centre subunits, which are all involved in photosynthesis and known to be genes that are down-regulated early in senescence (Hortensteiner 2009) (Appendix J, Appendix M).

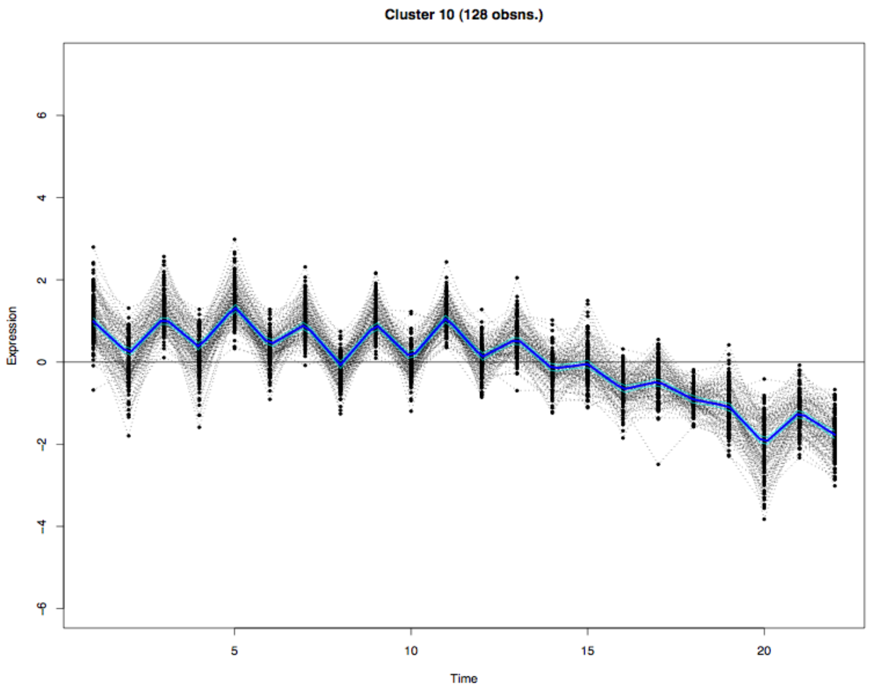


Figure. 4.7. A line graph of gene profiles from cluster 10 of the 82 clusters generated for the top 2000 most differentially expressed Arabidopsis genes using SplineCluster. The horizontal axis represents each of the 22 time points and the vertical axis represents the fold change in expression of the $\log_2$ signal intensities.

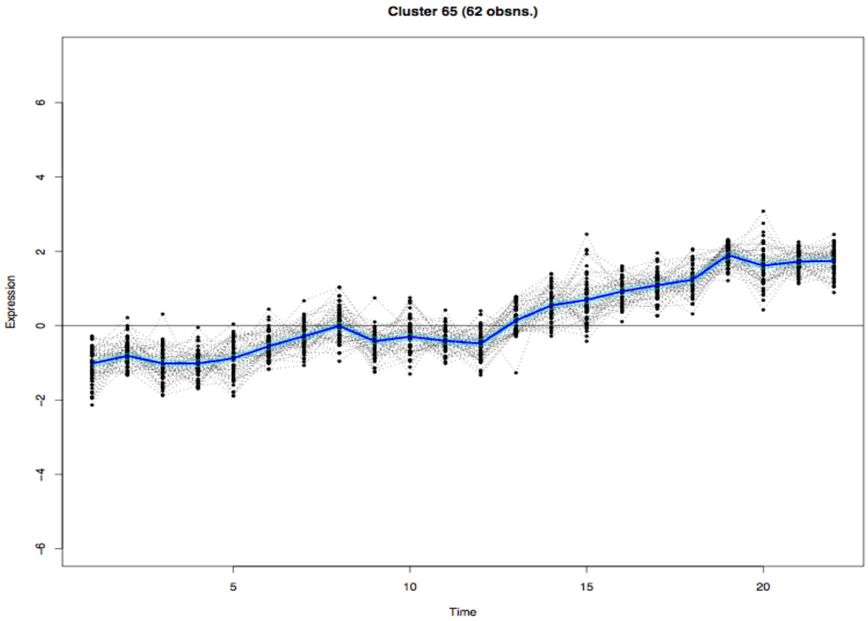


Figure 4.8. A line graph of gene profiles from cluster 65 of the 82 clusters generated for the top 2000 most differentially expressed Arabidopsis genes using SplineCluster. The horizontal axis represents each of the 22 time points and the vertical axis represents the fold change in expression of the $\log_2$ signal intensities.

In contrast, Figure 4.8 represents cluster 65 and shows a two-fold increase in expression over the 22 time points. However, as with cluster 10, there is a clear change in expression levels in cluster 65 from sample point 12 onwards, except this time there is a clear increase in expression suggesting that the genes in this cluster are positively regulated during the onset of senescence (Appendix I). These genes also lack a diurnal circadian clock directed expression. This theory is supported by the presence in the cluster of genes encoding AP2 domain containing transcription factors, WRKY transcription factors, NAC family proteins and genes coding for ethylene that are known to play important regulatory roles in the pathways governing senescence as they overlap with those controlling plant response to stress (Appendix K).

Analysis of the genes within the clusters highlight the success of SplineCluster as a tool for clustering genes based on expression profile as the clusters contain genes with similar known functions. For example, a number of clusters appear to be senescence enhanced, in particular 41, 42, 43, 54, 63 and 64 (see Appendix. F) which show a minimum of a two fold increase in expression over the 22 time points and further inspection of the genes in the clusters shows them to contain transcriptions factors such as WRKY, AtMYB75 and AP2 binding domain containing transcription factors known to be up-regulated during senescence. In contrast clusters 1-6 (see Appendix. G) show a minimum of a twofold decrease in expression over the 22 time points and contain a number of photosynthesis related genes such as glyceraldehyde-3-phosphate dehydrogenase B (GAPB), and the chloroplast precursor phosphoglycerate kinase. The clusters also contain the transcription factor AtMYB29, which has been implicated in regulation of glucosinolates. This is a class of secondary metabolites that act as natural pesticides in response to plant physical injury caused by herbivores (Hirai, 2007). The down regulation of expression of these genes directly supports the idea that the plant is undergoing senescence and is recycling plant material back into the developing seeds.

The clustering of the genes was done to aid the process of gene selection for modelling of potentially novel regulatory networks by categorising them on expression profile. The second phase of this process was to then identify the functions of the genes in the clusters and whether these functions are over represented within the cluster. Thus based on this the genes to be modelled can be decided using biological intuition.

4.5.2.3 Gene selection for modelling

The genes in each cluster were analysed using the GO Stat statistic to find overrepresented GO terms within each cluster. The GO annotations can be divided into 3 categories:

- Component- where the gene product can be found.
- Molecular function- the activity the gene product performs.
- Biological process– the series of activities to which the gene product belongs.

The GO Stat statistic returned the over-represented GO terms for the 3 categories of annotation for each cluster. Unfortunately the annotations in the TAIR 6 database were relatively poor and in some cases inaccurate due to the fact that many of the genes were annotated automatically and had not been verified by an annotator. For around 15% of the genes there were no annotations at all making it impossible to find overrepresented genes in those clusters. It became clear that as the effectiveness of the GO Stat method depends upon the accuracy of the GO annotations this approach could not be used to select genes for modelling.

To overcome these limitations it was decided that the building of smaller networks based upon known regulatory gene-gene relationships was a more logical way to proceed. The use of known regulatory relationships, in the first instance, was an effective way to test the accuracy and sensitivity of the modelling software. Once robust networks were established new ones could be added, so increasing the complexity of the network. The profiles of 38 genes were selected from the clusters on the basis that they were senescence-enhanced genes and included ethylene responsive elements, WRKY, DREB and AP2 domain transcription factors (see Appendix. K). The list also contained the profiles of 10 genes known to be involved in the flavonoid biosynthetic pathway. The expression values of the genes were then modelled using Variational Bayesian State Space modelling (VBSSM) software (Beal, Falciani et al. 2005) to determine whether it would retrieve the known flavonoid pathway genes in a network and whether the predicted network was accurate.

4.5.2.4 Modelling of flavonoid biosynthetic pathway

Thirty-eight genes (Appendix K) were selected from the 82 clusters of genes generated by the agglomerative hierarchical clustering program SplineCluster. The genes were modelled using the VBSSM program, the principles of which are described in the methodology section. As part of the process of determining the optimum number of hidden variables *K* that can be used to best describe the model, the program increments the number of hidden variables from 1 to 20. For each increment of the number of hidden variables, the program builds 10 models each beginning at a random gene. The program then calculates *F*, which represents the median value for the lower bound on the marginal likelihood or probability that the number of variables is optimum. The set of 10 models whose number of hidden variables gives the most positive value for the marginal likelihood are the ones that will be used. Of the 10 models generated with the optimum number of hidden variables, only predicted regulatory relationships that appear in at least 8 of the 10 models were considered as being strong candidate regulatory relationships.

Figure 4.9 shows the median value for the lower bound on the marginal likelihood for the number of hidden variables between 1 to 20 for the models of the 38 genes. The graph shows that the optimum number of hidden variables that best describe the network models generated for the 38 genes is 7. Hence the 10 randomly initialized models generated for this number of hidden variables, were inspected. Only the predicted regulatory relationships that satisfied the following criteria were considered as significant candidate gene-gene relationships: those with CBDZ scores (see methodology section) of at least 1.69 standard deviations (sds) away (positively or negatively) from a standard normal distribution of 0 (equating to a 90.10% confidence value that the predicted relationship is significantly different from the mean) and predicted regulatory relationships that appeared in at least 8 of the 10 models. This was done to increase the confidence in the robustness of the predictions generated. Both criteria act to ensure that only significant gene interactions are reported in the final output.

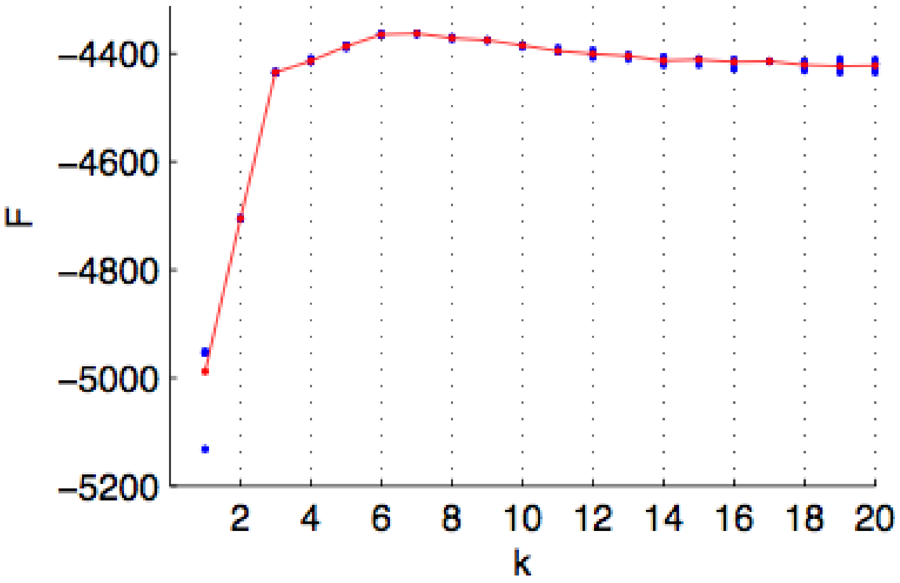


Figure 4.9. A line graph representing the median value for the lower bound on the marginal likelihood for the number of hidden variables between 1 to 20 for the models of 38 selected genes. The vertical axis represents the value for the marginal likelihood (F) and the horizontal axis represents the number of hidden variables (K) used to generate the model.

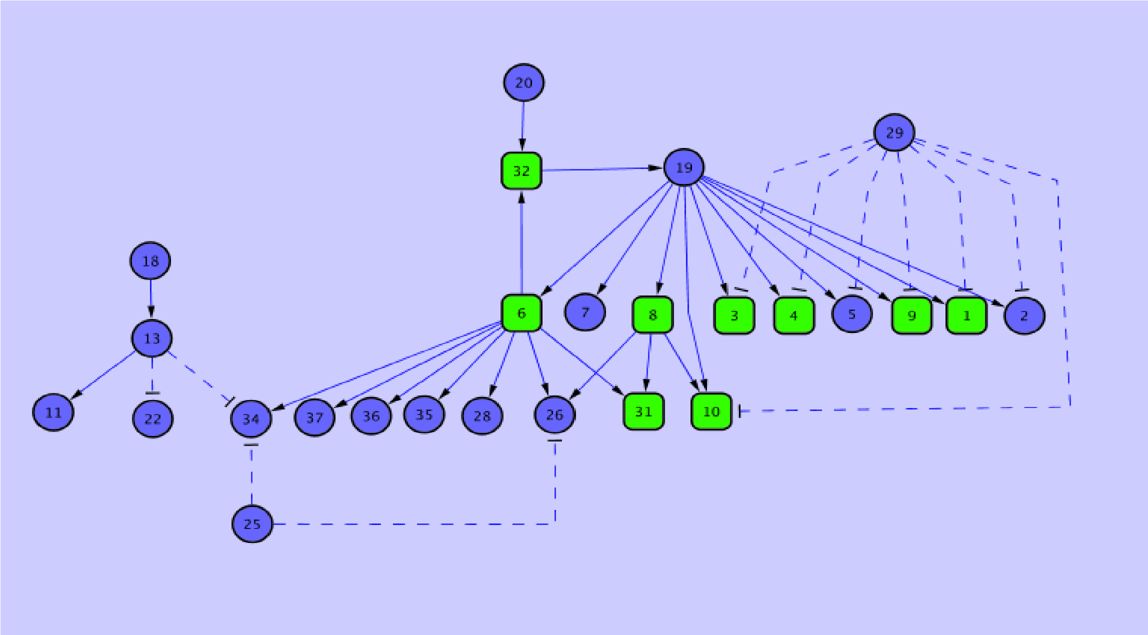


Figure 4.10. The network generated from 38 senescence altered genes. The green nodes represent the transcription factors and blue nodes represent all other genes. The number on each node corresponds to the position of the gene in the list. The blue lines represent positive regulatory relationships and the dotted lines represent negative regulatory relationships.

Figure 4.10 represents a visualisation of the network generated by the VBSSM program and visualised by Cytoscape. The regulatory relationships predicted by the model (see Appendix. L) show 6 main predicted hubs; *At3g23250* (AtMYB15 transcription factor) (6), *At3g48520* (cytochrome P450 family protein) (19), *At4g21830* (methionine sulfoxide reductase) (20), *At4g22470* (lipid transfer protein) (22), *At5g28237* (a tryptophan synthase) (29) and *At1g66390* (AtMYB90 transcription factor that codes for anthocyanin pigment 2 protein PAP2) (32). One of the most interesting set of positive regulatory relationships appears to be between the *At3g23250* AtMYB15 transcription factor (6), *At3g48520* the cytochrome P450 family protein (19) and *At1g66390* AtMYB90 (32). The model suggests these genes are directly or indirectly involved in a positive feedback loop. Interestingly, *At3g48520* the cytochrome P450 family protein is shown to be positively regulating a number of stress associated transcription factors such as dehydration stress related AP2 domain containing transcription factors (*At1g19210* (1), *At1g74930* (3), *At4g34410* (9)), defence stress associated WRKY transcription factors (*At1g80840* (4), *At4g23810* (8)), and *At2g44840* an ethylene-responsive binding protein (5). Further to this the model shows *At4g23810* WRKY53 transcription factor (8) positively regulating *At1g56650* the anthocyanin regulator AtMYB75 (31) and *At5g21960* a DREB binding element (10) often induced during stress response to dehydration. This is all in keeping with the current theory that stress response, anthocyanin biosynthesis and senescence signalling pathways are interlinked as the WRKY family of transcription factors have been shown to be involved in regulating plant defence responses to pathogen invasion (Eulgem and Somssich 2007).

Analysis of the network generated revealed the model had retrieved the flavonoid pathway associated genes. *At5g05270* a chalcone-flavanone isomerase family protein (26), *At5g13930* a chalcone synthase (28), *At1g56650* the AtMYB family transcription factor AtMYB75 (31), *At4g22880* a leucoanthocyanidin dioxygenase (34), *At5g07990* a flavonoid 3'-hydroxylase (F3'H) protein (35), *At5g17220* a glutathione S-transferase protein (36) and *At5g42800* a dihydroflavonol 4-reductase (DFR) protein (37) were all retrieved by the model. In the flavonoid biosynthetic pathway, the first step is the synthesis of naringenin chalcone from chalcone synthase (CHS). Chalcone is then isomerised to a flavanone by chalcone flavanone isomerase (CHI). The pathway then diverges into several different side pathways, each resulting in a different class of flavonoids. For example, Flavanone 3-hydroxylase (F3H) catalyzes the hydroxylation of flavanones to dihydroflavonols. Anthocyanins are synthesised via the reduction of dihydroflavonols to leucoanthocyanins by the enzyme dihydroflavonol reductase (DFR), which in turn are converted to anthocyanidins by anthocyanidin synthase (ANS). The fact that the model has been able to retrieve all of these genes increases confidence in the ability of the program to return genes known to be active during the process of senescence. Interestingly all the flavonoid pathway associated genes are shown to be positively regulated by the *At3g23250* AtMYB15 (6) transcription factor, which in turn is shown to positively regulate *At1g66390* the known senescence regulating transcription factor AtMYB90 in the positive feedback loop alluded to earlier. Further to this, all of the regulatory relationships predicted in the model were at least 1.69 sds away from the standard distribution mean of 0 (Appendix L). This suggests the predicted regulatory relationships between the genes and *At3g23250* AtMYB15 (6) are strong, significant ones. Also, the model shows *At3g23250* AtMYB15 (6) is strongly regulating *At1g56650* AtMYB75 (31), which along with *At1g66390* AtMYB90 (32) is thought to regulate flavonoid biosynthesis (Borevitz, Xia et al. 2000). It is interesting to note that the model does not show *At1g56650* AtMYB75 (31) as a major hub of flavonoid biosynthesis despite the experimental evidence that suggest both *At1g56650* AtMYB75 (31) and *At1g66390* AtMYB90 (32) positively regulate anthocyanin production in Arabidopsis (Borevitz, Xia et al. 2000). However, the model does show *At1g56650* AtMYB75 (31) to be positively regulated by *At3g23250* AtMYB15 (6). Overall this part of the model suggests that flavonoid biosynthesis is in part, regulated by a combination of transcription factors *At3g23250* AtMYB15 (6), *At1g66390* AtMYB90 (32), *At1g56650* AtMYB75 (31) and *At3g48520* a cytochrome P450 gene (19).

4.6 Discussion

4.6.1 Modelling transcriptional networks in pathogen infected Arabidopsis

The results suggest that the experiment was unsuccessful as there was little change in Arabidopsis gene expression over time despite treatment with two isolates of *H. arabidopsidis.* This is reflected in the low Hotellings T^2 statistical scores generated from the datasets (data not shown). A second observation that would validate this theory is the lack of correlation between the biological replicates. This suggests the replicates are not showing a similar gene expression response to stress induced via pathogen infection as would be expected. As a result of this it was impossible to select genes showing a significant change in expression over time directly or indirectly in response to treatment with *H. arabidopsidis*. Most of the change in expression seems to be due to “noise” associated with performing the experiment or as part of the daily life cycle of the plant.

The failure of the experiment appears to be due to a fundamental error in its design in that it does not take the behaviour of *H. arabidopsidis* into account. Upon infection, related *Phytophthora* species such as *P. ramorum* and *P. sojae* will parasitize their respective hosts for around 12 to 14 hours after infection at which point their behaviour changes to become necrotrophic. As a result the Phytophthoras then kill their host plants and feed off the dead tissue, thus they have been defined as hemi-biotrophs (Moy, Qutob et al. 2004). *H. arabidopsidis* is an obligate biotroph and, therefore, its survival depends upon remaining undiscovered after parasitisation of the host plant. The two isolates of *H. arabidopsidis*, *Maks9* and *Emco5,* have been shown not to elicit a host defence response in the Arabidopsis Col-0 wild type (Rose, Bittner-Eddy et al. 2004). This would suggest that the intensity of expression from genes involved in plant defence response, if any, would be very low. Thus any changes in gene expression levels as part of plant response to this would be so low to be indistinguishable from random noise.

It is clear that there is an issue with the sensitivity of the microarray as background noise levels are distorting the intensity values. An alternative to using a microarray time course experiment would be to use single cell sampling instead as described by Tomos (Tomos and Sharrock 2000). This is a method whereby a micro-capillary tip containing an RNAse inhibitor solution is inserted into an epidermal cell. The turgor pressure exerted by the water within the cell will force the cell sap outward into the microcapillary tube. The RNA within the cell sap is converted to complementary DNA (cDNA) using reverse transcriptase. Once this is complete the cDNA is then amplified using standard PCR. This method can be combined with Quantitative PCR (Q-PCR) as a way of quantifying the amount of mRNA in the sample. Q-PCR differs from RT-PCR whereby during the PCR step a fluorescent reporter probe coupled with fluorescence inhibitor molecule is added to its DNA target. During the polymerisation stage the *Taq* polymerase degrades the fluorescence inhibitor therefore enabling the reporter probe to fluoresce. The relative levels of cDNA can then be quantified by comparing specific cDNA levels to those produced by ordinary “house keeping” genes. The use of RT-PCR and Q-PCR at successive time points enables the detection and subsequent quantification of very low copy mRNA. This could enable the measurement of very subtle changes in gene transcription levels like those seen during *H. arabidopsidis* infection of Arabidopsis that wouldn't be picked up by a microarray time course experiment, but would be limited to only a few target genes. Karrer and associates (Karrer, Lincoln et al. 1995) have used the method described above to measure the levels of mRNA from Rubisco, cyclophillin and actin genes from tomato cells. However, this method also has problems, first described by Karrer (Karrer, Lincoln et al. 1995) in that the efficiency of amplification of cDNA is lower in samples with lower mRNA templates, this has been termed the "Monte Carlo" effect. This is where the lower the abundance of any template the less likely true mRNA levels will be reflected by the amplified cDNA library. However this should only apply to more complex rare mRNAs and thus in theory this method should be useful in identifying genes showing a significantly altered gene expression level under the different conditions.

A second alternative to microarray time course would be to use a high-throughput sequencing technique known as RNA-seq to quantify the changing expression levels of each transcript under the different conditions. RNA- seq works by first converting the RNA to cDNA using RT-PCR and then sheared into a set of fragments. Adaptors are then ligated to the 5' and 3' ends of each fragment and sequenced using one of a set of high-throughput sequencing technologies. After sequencing, the resulting reads can then be aligned to the appropriate Arabidopsis reference genome using a read aligner such as MAQ (Li, Ruan et al. 2008). The result is a genome wide map of the transcription structure of each gene and transcription levels for each gene as transcription levels can determined by simply measuring the number of reads that overlap the position. The read numbers can be normalised using a method called Reads Per Kilo base of exon Model (RPKM) (Mortazavi, Williams et al. 2008). This is a measure of read density taking into account the concentration of RNA in the original sample. This value can be calculated by taking the number of mappable reads that hit to an exon, dividing it by the total number of reads in the experiment and multiplying by the total combined length of all the exons in base pairs. This allows the direct comparison of transcript levels between samples. One of the main advantages of this technique is that it has low background signal compared to microarrays as the lack of cross hybridisation issues allows reads to be mapped relatively unambiguously to unique locations on the genome. Secondly, the required amount of RNA is low because there are no cloning steps. Lastly, because there is no lower limit for quantification of expression as it is defined by the number of reads that map to the position, genes with low expression can still be detected making it more sensitive than using a DNA Microarray. However, this approach does have its problems in that larger RNAs must be fragmented into smaller pieces to be able to be sequenced as all sequencing technologies are restricted in the length of read. The fragmentation of the RNAs in constructing the cDNA library can lead to biases in the outcome. Also there are bioinformatics issues in developing programs to remove reads with low quality base calls from the final output and issues with the sensitivity of the programs to map the reads to unique positions their respective reference genomes. However, these are all related to the novelty of the technology and will improve over time. This approach has already been used to map and quantify the mouse transcriptome (Mortazavi, Williams et al. 2008). More importantly this has been used in the Arabidopsis wild type transcriptome as part of an attempt to produce highly integrated map of the genomic distributions of small RNAs, methylcytosines and transcripts in Arabidopsis (Lister, O'Malley et al. 2008). The fact that quantification and mapping of the transcriptome in Arabidopsis wild type has already been achieved using next generation sequencing technologies adds credibility to the idea of using it as an alternative to Microarrays in identifying differentially expressed genes.

4.6.2 Modelling transcriptional networks in Arabidopsis senescence

The results of the model generated indicate that the Beal modelling software (Beal, Falciani et al. 2005) is capable of predicting a model in which the entire set of anthocyanin biosynthesis pathway related genes were included. Further to this the anthocyanin biosynthesis pathway associated genes were predicted to be downstream of the AtMYB90 transcription factor known to regulate anthocyanin production (Borevitz, Xia et al. 2000). Interestingly the model also predicted the presence of AtMYB15 as a positive regulator of anthocyanin biosynthesis. Hitherto AtMYB15 had only been shown to be a negative regulator of C-repeat/DRE-Binding Factor (CBF) genes that confer increased plant tolerance to cold stress (Agarwal, Hao et al. 2006) and as a positive regulator of the Shikimate wounding response pathway (Chen, Zhang et al. 2006). The presence of AtMYB15 in the model suggests a complex web of overlapping transcriptional regulation of both senescence and plant responses to various stresses including cold and wounding stress due to shared signalling pathways (Lim, Kim et al. 2007). It is not, therefore, inconceivable that AtMYB15 could play a role in transcriptional regulation of anthocyanin biosynthesis. This highlights one of the main benefits of using the Beal modelling software to model transcriptional regulation of signalling pathways: the results offer a hypothesis that can be tested based on the principle that the gene is known to be significantly differentially expressed during senescence.

Approaches that could be taken to test the extent to which AtMYB15 is involved in regulation of anthocyanin production are many. One approach would be to over express the AtMYB15 gene using an inducible promoter such as has been used by Chen and associates during the over expression of AtMYB15 (Chen, Zhang et al. 2006) and then determine the effect on anthocyanin gene expression via Q-PCR. A second approach would be to silence AtMYB15 via small RNAs. These RNAs are comprised of several families including small interfering RNAs (siRNAs) (Elbashir, Lendeckel et al. 2001) and microRNAs (miRNAs) (Lee and Ambros 2001). Small RNAs form part of a protein mechanism known as the RNA silencing induced complex (RISC) (Dugas and Bartel 2004), which use small RNAs to recognise motifs in specific nucleotide targets. The binding of these RNAs, specifically siRNAs, acts to guide RISC to the mRNA target site and cleave the mRNA, which is then degraded. Through the degradation of the mRNA gene expression can effectively by silenced. This process is known as RNA interference (RNAi) and can be used to silence AtMYB15 and then determine the effect it has on the expression of downstream anthocyanin associated genes again via Q-PCR. However, one of the downsides to using RNAi is that some endogenous miRNAs can bind less specifically to other target mRNA sites known as off targets, causing reduced expression of these genes. This makes quantifying the effect of silencing a specific gene more difficult. One solution to this is to design artificial miRNAs (amiRNAs) to be more specific to the gene of interest and insert them into miRNA precursors. Also, because the amiRNA is of known sequence it is then easier to predict any potential off targets that may come about through less specific binding (Alvarez, Pekker et al. 2006; Schwab, Ossowski et al. 2006). A third approach would be to perform a microarray time course experiment on an AtMYB15 knockout mutant. A previous experiment to identify the effect of AtMYB15 on CBF genes conferring increased tolerance to cold stress (Agarwal, Hao et al. 2006) used an Arabidopsis AtMYB15 mutant, the seeds of which (SALK_151976) are available from the Nottingham Arabidopsis Stock Centre (NASC), http://arabidopsis.info, 2009. A microarray time course experiment could be carried out using mRNA isolated from this mutant and the gene expression profiles modelled to determine the predicted effects of the AtMYB15 knockout on anthocyanin biosynthesis.

In order to further investigate whether AtMYB15 transcription factor regulates expression of AtMYB90 directly, yeast one hybrid could be used. This is a technique that determines whether there are any direct protein-DNA interactions occurring. The technique uses a single fusion protein in which the activator domain Gal4 is linked directly to the DNA binding domain from AtMYB15. This can be tested to determine if the fusion protein causes specific activation of AtMYB90 gene promoter target sequence. This could provide experimental evidence that AtMYB15 is a direct regulator of AtMYB90. An alternative to this method would be to use ChIP-Seq (Chromatin Immunoprecipitation – Sequencing). In this technique the ChIP process enhances specific cross-linked DNA-protein complexes using an antibody against the protein of interest, in this instance AtMYB15. After this the protein is removed leaving only the sequence it bound to. The DNA sequence is then PCR amplified and sequenced using next generation sequencing technology. The sequenced reads can then be mapped to the Arabidopsis reference genome, enabling identification of regions that are overrepresented in the number of mapped reads, which may correspond to transcription factor binding sites upstream of AtMYB90. This identification can be done using software such as PeakSeq (Rozowsky, Euskirchen et al. 2009). If this technique was successful and the model prediction correct, it would elucidate the target genes of AtMYB15 regulation including, if true AtMYB90, and the location of their transcription factor binding sites.

The identification of a candidate transcriptional regulator of anthocyanin biosynthesis is an exciting hypothesis to test, it serves to highlight the ability of the model to not only return a reasonable model but that it can infer AtMYB15 as being potentially involved in regulation of the anthocyanin biosynthesis pathway. In addition to this it serves to add weight to the idea that this set of methods can be used to generate transcriptional networks in Arabidopsis as until now the methods had only been employed in generating networks from *Saccharomyces cerevisiae* and human T-cells (Rangel, Angus et al. 2004; Beal, Falciani et al. 2005). However, the methods do have issues that need to be resolved to be able to make the most of the information available in the dataset. Firstly, the ranking of the genes based on the level of differential expression over time worked but using a 2000 gene threshold to avoid the inclusion of genes up-regulated due to cross hybridisation is rather arbitrary and could serve to introduce errors. The main problem is that many genes showing a more subtle but significant change in expression may be missed based on the ranking and cut off system. The obvious solution to this is to use no cut off and simply take all genes to be clustered. However this only acts to delay the main problem with the methodology, deciding which genes to model due to the severe limitations on the maximum number of genes that can be modelled at any one time, which is around 150. In the experiment a known pathway was used to test the accuracy of the modelling software and subsequently can be used as a baseline from which to build outwards. Using prior biological knowledge as a method of choosing genes for modelling would seem to be the most logical way to proceed as the models generated can be tested and validated much more easily. However, it must also be stated that the anthocyanin biosynthesis pathway and its regulation is arguably one of the most well known pathways associated with Arabidopsis stress response (Winkel-Shirley 2001). Therefore, it will be much more difficult to elucidate novel regulatory networks for pathways where literature based prior biological knowledge is much more limited. This remains the biggest challenge facing systems biologists today as greater prior knowledge will always lead to a much more accurate model. An approach to try and overcome this problem would be to start with a set of known gene regulatory relationships validated by the literature and then repeatedly model these genes with other genes shown to be differentially regulated under the same conditions. Alongside this a regular review of the literature must be conducted to identify important prior knowledge which would improve the accuracy of the model. There are currently several databases such as TAIR, Aracyc and Virtual Plant containing information on up to date regulatory relationships (Mueller, Zhang et al. 2003; Katari, Nowicki et al. 2008; Swarbreck, Wilks et al. 2008) as well as literature search tools available such as ONdex (Kohler, Baumbach et al. 2006) to carry out this sort of review. Thus incorporation of priors must be a major goal of bioinformatics in the future. The process can then be repeated again in an iterative fashion. This method would involve the production of hundreds of models for all the different possible genes, which would be computationally intensive but theoretically eventually there will be a consensus between the differing models as to which genes are thought to be regulating one another. These genes can then be tested experimentally for predicted effects.

The strength of this method lies in its ability to offer predicted novel regulatory pathways that can be tested. Also, the modelling software has the capability to adapt the model generated in the light of any subsequent prior information fed to it allowing the user to iteratively generate an increasingly accurate model. The method is not without its flaws but offers a powerful set of tools to be used in conjunction with conventional expression studies such as those outlined earlier.

Chapter 5: Modelling transcriptional networks using Wild Type and MYB90 mutant microarray time course experiments.

5.1 Introduction

Anthocyanin biosynthesis occurs as part of senescence, a process that represents the final stage of leaf development (Lim, Kim et al. 2007). The previous chapter focussed on establishing a methodology as a way of identifying both established networks and predicting new ones. The process returned a model that predicted many of the genes associated with regulating anthocyanin biosynthesis to be downstream of the transcription factor AtMYB90, which has been shown to be involved in the regulation of the anthocyanin biosynthesis pathway (Borevitz, Xia et al. 2000; Winkel-Shirley 2001) adding validity to the modelling methodology used. The modelling process also highlighted the presence of AtMYB15, which was predicted to be an upstream regulator of AtMYB90. AtMYB15 had until now only been shown experimentally to be involved in plant response to wounding through the shikimate wounding pathway (Chen, Zhang et al. 2006) and as a negative regulator of CBF genes conferring increased plant tolerance to cold stress (Agarwal, Hao et al. 2006). This prediction is supported by the knowledge that the pathway provides phenylalanine required for anthocyanin biosynthesis as part of the by-products of the shikimate pathway (Deikman and Hammer 1995) and, therefore, in this respect, AtMYB15 may well act as a regulator of anthocyanin biosynthesis. This prediction is also interesting as it acts to support previous research carried out by Nicola Warner (Warner 2008) which suggests that although MYB90 has a role in controlling anthocyanin biosynthesis during senescence it is not *essential* for senescence associated anthocyanin biosynthesis.

The previous research centred upon investigating the effects of the loss of MYB90 on senescence by comparing the phenotype of the MYB90 knock out, IM28 with the wild type (WT) Col-0. A time course microarray experiment was carried out to identify differentially expressed genes between WT and IM28 (see methods). Gene expression studies on the data showed a significant decrease in expression of key genes *At5g42800* and *At5g13930* that encode dihydroflavonol reductase (DFR) and chalcone synthase (CHS) respectively in the mutant IM28, suggesting that AtMYB90 is regulating these genes during senescence. In light of this decreased gene expression it was expected that anthocyanin levels would fall in the IM28 mutant, however to the contrary anthocyanin levels increased in the IM28 mutant during senescence. These results suggest that the absence of MYB90 may have affected the expression of the key genes normally associated with anthocyanin biosynthesis but anthocyanin levels increase despite this. The continued rise in anthocyanin levels in IM28 suggests that there may be an alternative signalling pathway controlling the biosynthesis of the anthocyanin. Thus the results imply that AtMYB90 is not essential for anthocyanin biosynthesis during senescence. The theory that an alternative signalling pathway is controlling anthocyanin and anthocyanin component production is supported in a study by Buchanan-Wollaston *et al.* (2005) which identified a possible alternative flavonoid signalling pathway expressed during dark induced senescence. The lack of increased expression however, shown by these alternative genes in the IM28 mutant would suggest that these alternative genes are not inducing anthocyanin biosynthesis in the absence of AtMYB90.

The increased levels of anthocyanins despite the loss of AtMYB90 indicate that there is an alternative gene compensating for the absence of AtMYB90 expression and also an alternative pathway for anthocyanin biosynthesis, which does not depend on AtMYB90. Research carried out in other plant species suggests that this alternative gene is most likely to be a MYB transcription factor-encoding gene because this family of transcription factors has been shown to be required for anthocyanin biosynthesis in other species (Goff, Cone et al. 1992; Davies and Schwinn 2003; Gonzalez, Zhao et al. 2008). The Warner study (Warner 2008) identified six differentially expressed MYB genes between the WT and the IM28 mutant; *At1g66370* (AtMYB113), *At1g18570* (AtMYB51), *At5g07690* (AtMYB29) and *At3g28910* (AtMYB30).

The presence of WT and IM28 microarray time course datasets represent an opportunity to model these differentially expressed genes with the original 38 genes used to model the anthocyanin biosynthesis pathway in Chapter 4. The generated models will hopefully predict a candidate MYB transcription factor as regulating the expression of anthocyanin components in the absence of AtMYB90 that can then be tested experimentally. The second aim would be to determine whether the generated model would support the previous predictions made by the modelling methodology using the long day microarray dataset that suggest AtMYB15 is an upstream regulator of AtMYB90.

5.2 Methods

5.2.1 Arabidopsis anthocyanin biosynthesis Microarray Experiment

Vicky Buchanan –Wollaston and associates performed all laboratory based experimental procedures relating to the Arabidopsis anthocyanin biosynthesis experiment. Further details of this experiment will be described in their forthcoming paper in preparation.

5.2.1.1 Plant Growth

Arabidopsis seeds (wild type Col-0 and IM28 mutant) were stratified at 4^{o}C in the dark for 48 hours and then sown onto Arabidopsis compost mix (Levingtons F2 compost:sand:vermiculite 6:1:1). Plants were grown at 20^{o}C under a 16h/8h light/dark cycle at 70% relative humidity and 250µmol m^{-2} s^{-1} light intensity. Leaf 7 was tagged with cotton 18 days after sowing (DAS). Sampling of leaf 7 started at 30 DAS and continued every other day until full senescence was reached (40 DAS). The seventh rosette leaf was harvested at days 30, 34, 35, 36, 37, 38, 39 and 40 after sowing. This resulted in 2 biological replicates and 4 technical replicates at each time point for each treatment.

5.2.2 Microarray analysis

5.2.2.1 RNA preparation and labelling

For each experimental treatment (WT and IM28), total RNA was isolated from four individual leaves for each time point (Biological replicates A, B, C and D) using

TRIzol reagent (Invitrogen) and amplified using the MessageAmp II aRNA Amplification kit (Ambion) in accordance with the kit protocol with a single round of amplification. Cy3 and Cy5 labelled cDNA probes were prepared by reverse transcribing 5µg of aRNA with Cy3- or Cy5- dCTP (GE Healthcare) and a modified dNTP mix (10mM each dATP, dGTP and dTTP; 2mM dCTP) using random primers (Invitrogen) and SuperScript II reverse transcriptase (Invitrogen). Labelled probes were purified using QiaQuick PCR Purification columns (Qiagen), freeze-dried, resuspended in hybridization buffer (25% formamide, 5xSSC, 0.1% SDS, 0.5µg/µl yeast tRNA [Invitrogen]).

5.2.2.2 Microarray experiments

The microarray experiments were carried out using the CATMA (version 3) microarray (Allemeersh *et al.* 2005; http://www.catma.org). A complex loop design was applied to extract the maximum information from the two colour hybridisation experiments (A Mead et al, in preparation). Following the layout of the experimental design, combinations of labelled samples were hybridized to slides overnight at 42^{o}C. Following hybridization, slides were washed and scanned using an Affymetrix 428 array scanner at 532nm (Cy3) and 635nm (Cy5). Scanned data were quantified using Imagene 7.5.0 software (BioDiscovery, Inc.).

5.2.3 Software

5.2.3.1 Image processing

Image processing was carried out using ImaGene® software version 7.5.0 BioDiscovery Inc. http://www.biodiscovery.com/index/imagene to yield the raw signal intensities.

5.2.3.2 Normalisation and identification of differentially expressed genes

Raw data was normalized and analyzed with R version 2.9.0 using the R package MAANOVA version 1.13.1 http://research.jax.org/faculty/churchill/. MicroArray ANalysis Of VAriance or MAANOVA is a collection of functions for statistical analysis of gene expression data from two-colour cDNA microarray experiments. The program accepts a set of raw signal intensities from the set of microarray slides obtained after processing the microarray images using ImaGene (see above) and the microarray experimental design file. The data is then analysed to identify regions of variation in the data that may have come from reasons other than expected sources of variation. This undesirable type of variation may have come from preparation of the slides such as smearing from slide handling, spatially biased dye binding or slide printing problems. The data was then logarithmically transformed to normalize the data and remove any anomalies. The data was then fit to an ANOVA model of expected sources of variation such as dye, array, time, biological replicate and treatment. The data was then analysed for each of the terms using F-tests to see how variation they provide to the model. Thus genes showing the greatest variation over time and between treatments were extracted from the data. From this final data set, genes for modelling were chosen.

5.2.3.3 Modelling of differentially expressed genes

Genes were modelled using the VBSSM V3.3.5 tool kit (Beal, Falciani et al. 2005), using MATLAB version R2007a , http://www.mathworks.com.

5.3 Results

The results of the microarray experiment are as follows. Of the 31106 genes used in the microarray, 1262 were significantly differentially expressed over both time and between the wild type (WT) and the *At1g66390* (AtMYB90) knockout (IM28) using the microarray analysis of variance package MAANOVA. The list also contained six MYBs besides AtMYB90 showing significant differences in expression over time and between cell lines; *At5g59780* (AtMYB59), *At1g22640* (AtMYB3), *At3g50060* (AtMYB77), *At5g44190* (GLK2), *At1g71030* (ATMYBL2) and *At5g08520* (MYB-like transcription factor). As the aim of this analysis was to identify potential candidates in an alternative pathway regulating anthocyanin biosynthesis in the absence of AtMYB90, these genes were added to the original list of 38 genes used to model the anthocyanin biosynthesis pathway in Chapter 4. Further to this four genes identified by Warner (2008) as being potential candidates regulating an alternative pathway; *At1g66370* (AtMYB113), *At1g18570* (AtMYB51), *At5g07690* (AtMYB29) and *At3g28910* (AtMYB30) were also added to the list, thus in total 48 genes were submitted to the Variational Bayesian State Space Modelling (VBSSM) software for modelling (Table 5.1).

Table 5.1. Known or predicted functions of the 48 genes submitted for VBSSM modelling. The functions of the genes were obtained from the TAIR7 database.

Number	*At* Number	Gene Function
1	*At1g19210*	AP2 domain-containing transcription factor, putative, encodes a member of the DREB subfamily A-5 of ERF/AP2 transcription factor family.
2	*At1g72520*	Lipoxygenase, putative, iron ion binding / lipoxygenase/ metal ion binding / oxidoreductase, acting on single donors with incorporation of molecular oxygen.
3	*At1g74930*	ORA47; DNA binding / transcription factor, encodes a member of the DREB subfamily A-5 of ERF/AP2 transcription factor family.
4	*At1g80840*	WRKY40 (WRKY DNA-binding protein 40); transcription factor, Pathogen-induced transcription factor. Binds W-box sequences in vitro. Forms protein complexes with itself and with WRKY40 and WRKY60.
5	*At2g44840*	ATERF13/EREBP (ETHYLENE-RESPONSIVE ELEMENT BINDING FACTOR 13); DNA binding / transcription factor, encodes a member of the ERF (ethylene response factor) subfamily B-3 of ERF/AP2 transcription factor family.)
6	*At3g23250*	AtMYB15/AtY19/MYB15 (AtMYB domain protein 15); DNA binding / transcription factor, Member of the R2R3 factor gene family.
7	*At4g23800*	High mobility group (HMG1/2) family protein,
8	*At4g23810*	WRKY53 (WRKY DNA-binding protein 53); DNA binding / protein binding / transcription activator/ transcription factor.
9	*At4g34410*	AP2 domain-containing transcription factor, putative, encodes a member of the ERF (ethylene response factor) subfamily B-3 of ERF/AP2 transcription factor family.
10	*At5g21960*	AP2 domain-containing transcription factor, putative, encodes a member of the DREB subfamily A-5 of ERF/AP2 transcription factor family.
11	*At1g34020*	Transporter-related, similar to transporter-related [Arabidopsis thaliana] (TAIR:AT4G09810.1)
12	*At1g43160*	RAP2.6 (related to AP2 6); DNA binding / transcription factor, encodes a member of the ERF (ethylene response factor) subfamily B-4 of ERF/AP2 transcription factor family (RAP2.6).
13	*At1g49900*	Zinc finger (C2H2 type) family protein.

14	*At2g02990*	RNS1 (RIBONUCLEASE 1); endoribonuclease, member of the ribonuclease T2 family, responds to inorganic phosphate starvation, and inhibits production of anthocyanin. Also involved in wound-induced signalling independent of jasmonic acid.
15	*At2g38240*	Oxidoreductase, 2OG-Fe(II) oxygenase family protein, similar to oxidoreductase, 2OG-Fe(II) oxygenase family protein [Arabidopsis thaliana] (TAIR:AT5G05600.1).
16	*At2g38380*	Peroxidase 22 (PER22) (P22) (PRXEA) / basic peroxidase E, Identical to Peroxidase 22 precursor (PER22) [Arabidopsis Thaliana]
17	*At2g43870*	Polygalacturonase, putative / pectinase, putative, , similar to polygalacturonase, putative / pectinase, putative [Arabidopsis thaliana] (TAIR:AT3G59850.1).
18	*At3g11480*	BSMT1; S-adenosylmethionine-dependent methyltransferase, The gene encodes a SABATH methyltransferase that methylates both salicylic acid and benzoic acid. It is highly expressed in flowers, induced by biotic and abiotic stress and thought to be involved in direct defense mechanism
19	*At3g48520*	CYP94B3 (cytochrome P450, family 94, subfamily B, polypeptide 3); oxygen binding, member of CYP94B, similar to CYP94B1 (cytochrome P450, family 94, subfamily B, polypeptide 1), oxygen binding [Arabidopsis thaliana] (TAIR:AT5G63450.1).
20	*At4g21830*	Methionine sulfoxide reductase domain-containing protein / SeIR domain-containing protein.
21	*At4g21850*	Methionine sulfoxide reductase domain-containing protein / SeIR domain-containing protein.
22	*At4g22470*	Protease inhibitor/seed storage/lipid transfer protein (LTP) family protein.
23	*At4g35160*	O-methyltransferase family 2 protein.
24	*At4g36950*	MAPKKK21; ATP binding / protein kinase, member of MEKK subfamily.
25	*At5g02490*	Heat shock cognate 70 kDa protein 2 (HSC70-2) (HSP70-2)
26	*At5g05270*	Chalcone-flavanone isomerase family protein.
27	*At5g13220*	JAS1/JAZ10/TIFY9 (JASMONATE-ZIM-DOMAIN PROTEIN 10).
28	*At5g13930*	ATCHS/CHS/TT4 (CHALCONE SYNTHASE); naringenin-chalcone synthase, Encodes chalcone synthase (CHS), a key enzyme involved in the biosynthesis of flavonoids. Required for the accumulation of purple anthocyanins in leaves and stems.

29	*At5g28237*	Tryptophan synthase, beta subunit.
30	*At5g28238*	Tryptophan synthase, beta subunit.
31	*At1g56650*	PAP1 (PRODUCTION OF ANTHOCYANIN PIGMENT 1); DNA binding / transcription factor, Encodes a putative AtMYB domain containing transcription factor involved in anthocyanin metabolism and radical scavenging. Essential for the sucrose-mediated expression of the dihydroflavonol reductase gene, Identical to Transcription factor AtMYB75 (AtMYB75) [Arabidopsis Thaliana].
32	*At1g66390*	PAP2 (PRODUCTION OF ANTHOCYANIN PIGMENT 2); DNA binding / transcription factor, production of anthocyanin pigment 2 protein (PAP2), Identical to Transcription factor AtMYB90 (AtMYB90) [Arabidopsis Thaliana] (GB:Q9ZTC3); similar to PAP1 (PRODUCTION OF ANTHOCYANIN PIGMENT 1), DNA binding / transcription factor [Arabidopsis thaliana]
33	*At4g22880*	LDOX encodes leucoanthocyanidin dioxygenase, which is involved in proanthocyanin biosynthesis. Mutant analysis suggests that this gene is also involved in vacuole formation.
34	*At5g07990*	TT7 (TRANSPARENT TESTA 7); flavonoid 3'-monooxygenase/ oxygen binding, Required for flavonoid 3' hydroxylase activity. Identical to Flavonoid 3'-monooxygenase (CYP75B1) [Arabidopsis Thaliana]
35	*At5g17220*	ATGSTF12 (GLUTATHIONE S-TRANSFERASE 26); glutathione transferase, Encodes glutathione transferase belonging to the phi class of GSTs.
36	*At5g42800*	DFR (DIHYDROFLAVONOL 4-REDUCTASE); dihydrokaempferol 4-reductase, dihydroflavonol reductase. Catalyzes the conversion of dihydroquercetin to leucocyanidin in the biosynthesis of anthocyanins.
37	*At5g54060*	UF3GT (UDP-GLUCOSE:FLAVONOID 3-O-GLUCOSYLTRANSFERASE); transferase, transferring glycosyl groups.
38	*At5g56840*	DNA-binding family protein, similar to AtMYB family transcription factor [Arabidopsis thaliana] (TAIR:AT3G16350.1).
39	*At5g59780*	MYB59 (AtMYB domain protein 59); DNA binding / transcription factor.
40	*At1g22640*	MYB3 (AtMYB domain protein 3); DNA binding / transcription factor, AtMYB-type transcription factor (AtMYB3) that represses phenylpropanoid biosynthesis gene expression.
41	*At3g50060*	MYB77; DNA binding / transcription factor, Member of the R2R3 factor gene family. similar to ATMYB44/ATMYBR1/MYBR1 (AtMYB DOMAIN

		PROTEIN 44), DNA binding / transcription factor [Arabidopsis thaliana] (TAIR:AT5G67300.1).
42	*At5g44190*	GLK2 (GOLDEN2-LIKE 2); DNA binding / transcription factor, Encodes a protein containing a GARP DNA-binding domain which interacts with the Pro-rich regions of GBF1.
43	*At1g71030*	ATMYBL2 (Arabidopsis AtMYB-like 2); DNA binding / transcription factor, Encodes a putative AtMYB family transcription factor.
44	*At5g08520*	Myb family transcription factor, similar to AtMYB family transcription factor [Arabidopsis thaliana] (TAIR:AT5G23650.1).
45	*At1g66370*	MYB113 (AtMYB domain protein 113); DNA binding / transcription factor, Encodes a putative transcription factor (AtMYB113)., similar to PAP2 (PRODUCTION OF ANTHOCYANIN PIGMENT 2), DNA binding / transcription factor [Arabidopsis thaliana] (TAIR:AT1G66390.1).
46	*At1g18570*	MYB51 (AtMYB DOMAIN PROTEIN 51); DNA binding / transcription factor, putative transcription factor: R2R3-MYB transcription family, similar to AtMYB122 (AtMYB domain protein 122), DNA binding / transcription factor [Arabidopsis thaliana] (TAIR:AT1G74080.1).
47	*At5g07690*	MYB29 (AtMYB domain protein 29); DNA binding / transcription factor, Encodes a putative transcription factor (AtMYB29 Homeodomain-like (InterPro:IPR009057); contains InterPro domain AtMYB, DNA-binding (InterPro:IPR014778).
48	*At3g28910*	MYB30 (AtMYB domain protein 30); DNA binding / transcription factor.

Figures 5.1 and 5.2 represent a visualisation of the networks generated by the VBSSM program, using the network visualisation application Cytoscape, for WT and IM28 models respectively. The optimum number of hidden states *K* for the WT model is 3 and for the IM28 model *K* is 1. The influences on expression between genes predicted by the model of the WT cell line are in Table 5.2 whilst those of the IM28 are in Table 5.3. All predicted regulatory gene-gene relationships have a

standard deviation of 1.69 away from a mean of 0, which equates to a 90.10 % confidence in the prediction.

The regulatory gene-gene relationships predicted by the WT model (Table 5.2, Figure 5.1) show 11 predicted hubs; *At1g72520* (lipoxygenase) (2), *At1g74930* (ORA47 DNA binding transcription factor) (3), *At3g23250* (AtMYB15) (6), *At4g23810* (AtWRKY53) (8), *At1g43160* (RAP2.6 ethylene response transcription factor) (12), *At2g02990* (RNS1 - Ribonuclease 1) (14), *At2g38240* (oxidoreductase) (15), *At2g43870* (polygalactunorase) (17), *At5g44190* (GLK2 - GOLDEN2-LIKE 2) (42), *At1g71030* (AtMYBL2 - Arabidopsis AtMYB-like 2) (43), and *At5g07690* (AtMYB29 - AtMYB domain protein 29) (47). The results showed AtWRKY53 as a major hub strongly regulating expression of stress related AP2 domain containing transcription factors *At1g19210*, *At4g34410* (1, 9) as well as anthocyanin biosynthesis associated genes *At4g22880* (LDOX - leucoanthocyanidin dioxygenase) (33) and *At5g13930* (CHS) (Chalcone Synthase) (28).

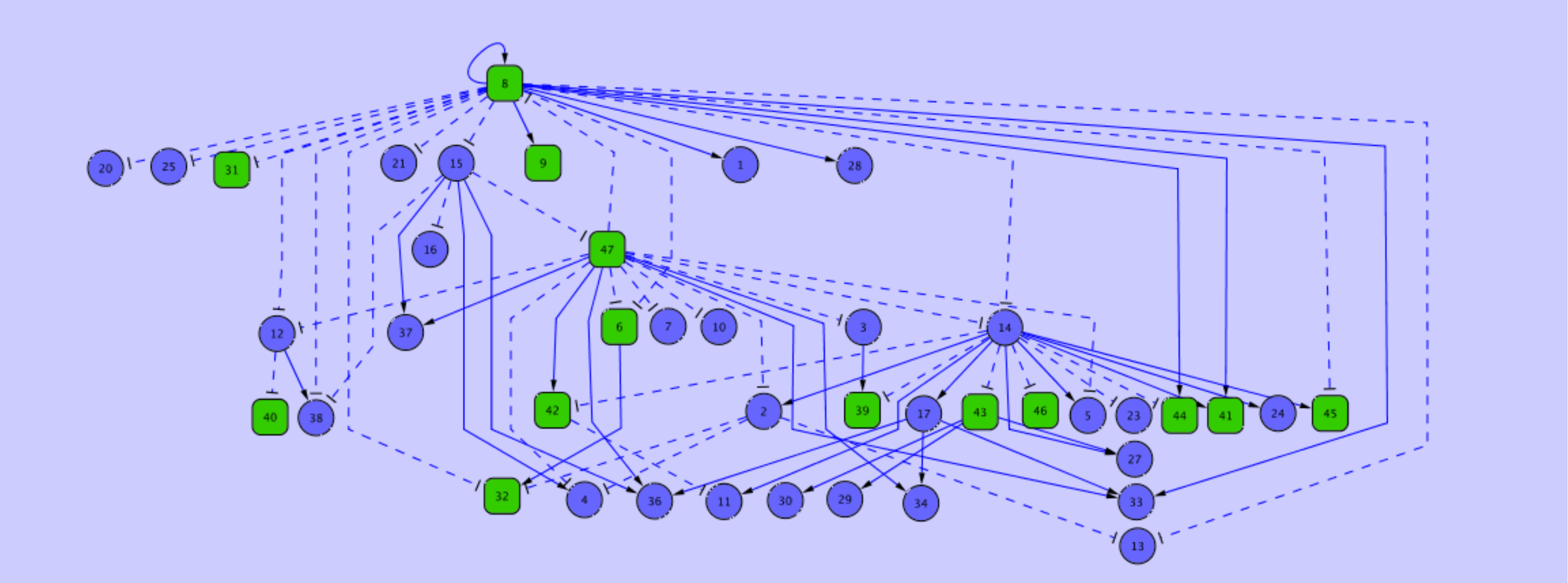


Figure. 5.1. The network generated from the transcription profiles of 48 genes alerted in expression during a senescence time series using wild type Col-0 plants. The green nodes represent the transcription factors and blue nodes represent all other gene classes. The number on each node corresponds to the position of the gene in Table 5.1. The blue lines represent positive interactions and the dotted lines represent negative interactions.

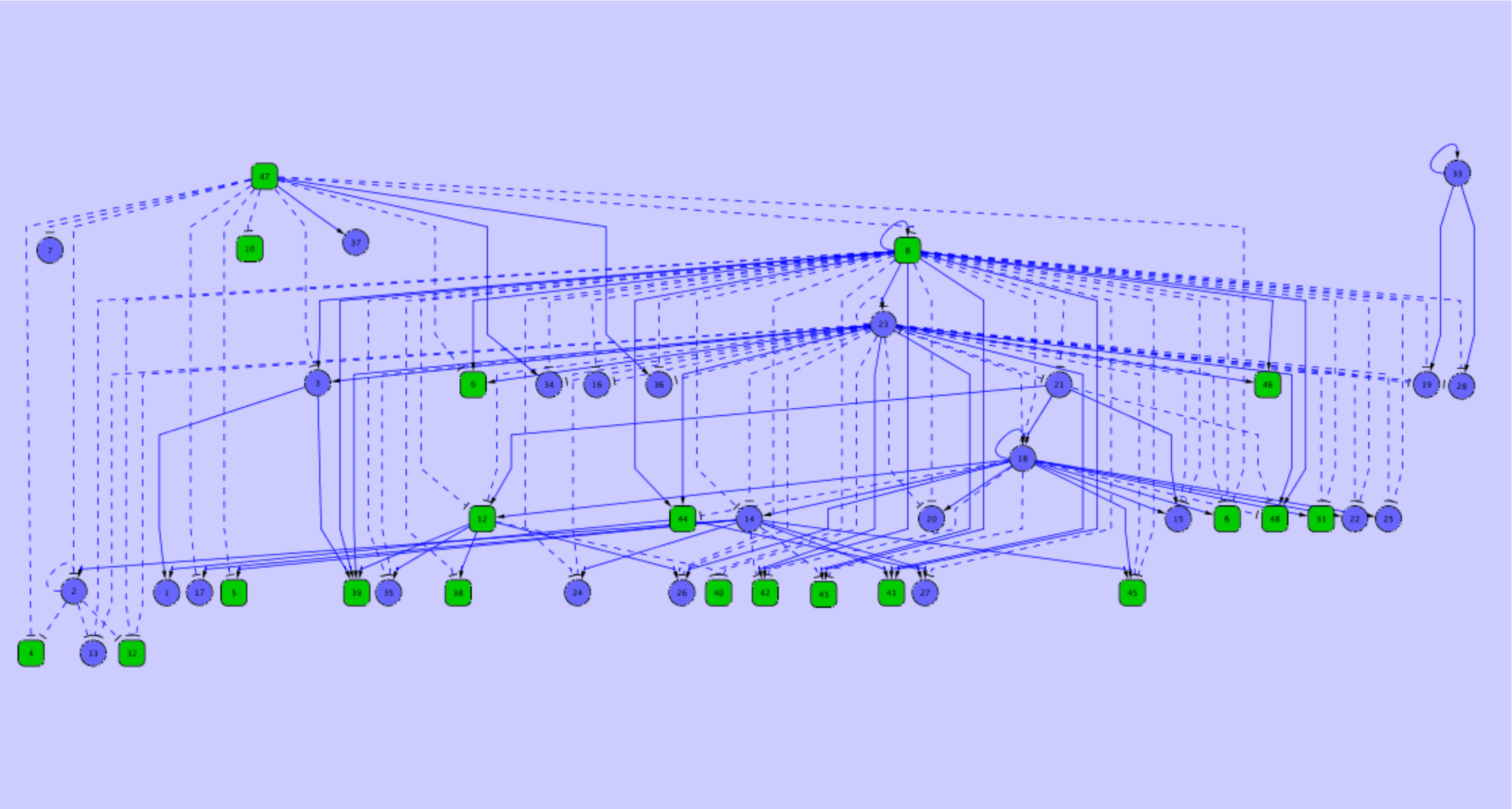


Figure. 5.2. The network generated from the transcription profiles of 48 genes alerted in expression during a senescence time series using IM28 plants that contain an insertional knockout mutation in AtMYB*90*. The green nodes represent the transcription factors and blue nodes represent all other gene classes. The number on each node corresponds to the position of the gene in Table 5.2. The blue lines represent positive interactions and the dotted lines represent negative interactions.

Table 5.2 Predicted regulatory relationships between the 48 genes chosen for modelling for the WT dataset. The interaction strength is determined from the CBDZ score generated by the model, which is calculated in part from the standard deviation. Regulatory gene-gene relationships are considered significant if their CBDZ scores are at least +/- 1.69 deviations from a standard normal distribution of 0, representing no relationship between the genes.

Source Gene Number	**At Number**	**Target Gene Number**	**At Number**	**Standard Deviation (sds)**
2	*At1g72520*	4	*At1g80840*	-1.83611
2	*At1g72520*	13	*At1g49900*	-1.75495
2	*At1g72520*	32	*At1g66390*	-2.51134
3	*At1g74930*	39	*At5g59780*	1.75631
6	*At3g23250*	32	*At1g66390*	1.70336
8	*At4g23810*	1	*At1g19210*	2.19773
8	*At4g23810*	6	*At3g23250*	-2.50662
8	*At4g23810*	8	*At4g23810*	2.18456
8	*At4g23810*	9	*At4g34410*	2.35501
8	*At4g23810*	12	*At1g43160*	-2.52232
8	*At4g23810*	13	*At1g49900*	-2.51218
8	*At4g23810*	14	*At2g02990*	-1.76747
8	*At4g23810*	15	*At2g38240*	-2.67667
8	*At4g23810*	20	*At4g21830*	-2.56464
8	*At4g23810*	21	*At4g21850*	-2.97293
8	*At4g23810*	25	*At5g02490*	-1.76725
8	*At4g23810*	28	*At5g13930*	2.04729
8	*At4g23810*	31	*At1g56650*	-2.27264
8	*At4g23810*	32	*At1g66390*	-2.67454
8	*At4g23810*	33	*At4g22880*	1.84409
8	*At4g23810*	38	*At5g56840*	-2.89362
8	*At4g23810*	41	*At3g50060*	3.17627
8	*At4g23810*	44	*At5g08520*	2.48659
8	*At4g23810*	45	*At1g66370*	-2.23916
12	*At1g43160*	8	*At4g23810*	-2.06719
12	*At1g43160*	38	*At5g56840*	2.60191
12	*At1g43160*	40	*At1g22640*	-1.88974
14	*At2g02990*	2	*At1g72520*	2.03399
14	*At2g02990*	5	*At2g44840*	2.03154
14	*At2g02990*	11	*At1g34020*	1.70544
14	*At2g02990*	17	*At2g43870*	3.60485
14	*At2g02990*	23	*At4g35160*	-2.01626
14	*At2g02990*	24	*At4g36950*	2.1606
14	*At2g02990*	27	*At5g13220*	2.67894
14	*At2g02990*	39	*At5g59780*	-2.62684

14	*At2g02990*	41	*At3g50060*	3.57515
14	*At2g02990*	42	*At5g44190*	-3.11879
14	*At2g02990*	43	*At1g71030*	-3.50631
14	*At2g02990*	44	*At5g08520*	-2.202
14	*At2g02990*	45	*At1g66370*	2.66108
14	*At2g02990*	46	*At1g18570*	-2.13173
15	*At2g38240*	4	*At1g80840*	1.71628
15	*At2g38240*	16	*At2g38380*	-3.37679
15	*At2g38240*	36	*At5g42800*	1.73355
15	*At2g38240*	37	*At5g54060*	2.51969
15	*At2g38240*	38	*At5g56840*	-3.39347
15	*At2g38240*	47	*At5g07690*	-2.00319
17	*At2g43870*	33	*At4g22880*	1.78185
17	*At2g43870*	34	*At5g07990*	2.01896
17	*At2g43870*	36	*At5g42800*	1.71967
42	*At5g44190*	11	*At1g34020*	-2.22675
43	*At1g71030*	27	*At5g13220*	1.72989
43	*At1g71030*	29	*At5g28237*	1.9067
43	*At1g71030*	30	*At5g28238*	1.9067
47	*At5g07690*	2	*At1g72520*	-2.71106
47	*At5g07690*	3	*At1g74930*	-2.0558
47	*At5g07690*	4	*At1g80840*	-2.45986
47	*At5g07690*	5	*At2g44840*	-2.12229
47	*At5g07690*	6	*At3g23250*	-2.38052
47	*At5g07690*	7	*At4g23800*	-2.46979
47	*At5g07690*	8	*At4g23810*	-1.69055
47	*At5g07690*	10	*At5g21960*	-1.82167
47	*At5g07690*	12	*At1g43160*	-1.87064
47	*At5g07690*	14	*At2g02990*	-1.85104
47	*At5g07690*	33	*At4g22880*	1.98511
47	*At5g07690*	34	*At5g07990*	1.9511
47	*At5g07690*	36	*At5g42800*	2.07126
47	*At5g07690*	37	*At5g54060*	2.06753
47	*At5g07690*	42	*At5g44190*	1.79505

Table 5.3 Predicted regulatory relationships between the 48 genes chosen for modelling for the IM28 dataset. The relationship strength is determined from the CBDZ score generated by the model, which is calculated in part from the standard deviation. Regulatory gene-gene relationships are considered significant if their CBDZ scores are at least +/- 1.69 deviations from a standard normal distribution of 0, representing no relationship between the genes.

Source Gene Number	*At Number*	Target Gene Number	*At Number*	Standard Deviation (sds)
2	*At1g72520*	2	*At1g72520*	-2.01359
2	*At1g72520*	4	*At1g80840*	-2.3826
2	*At1g72520*	13	*At1g49900*	-2.19574
2	*At1g72520*	32	*At1g66390*	-2.90232
3	*At1g74930*	1	*At1g19210*	1.97568
3	*At1g74930*	39	*At5g59780*	2.0222
8	*At4g23810*	3	*At1g74930*	2.40946
8	*At4g23810*	6	*At3g23250*	-3.36532
8	*At4g23810*	8	*At4g23810*	2.32032
8	*At4g23810*	9	*At4g34410*	2.47818
8	*At4g23810*	12	*At1g43160*	-4.46437
8	*At4g23810*	13	*At1g49900*	-2.2376
8	*At4g23810*	14	*At2g02990*	-3.90836
8	*At4g23810*	15	*At2g38240*	-4.29053
8	*At4g23810*	16	*At2g38380*	-2.53284
8	*At4g23810*	18	*At3g11480*	-4.11316
8	*At4g23810*	19	*At3g48520*	-2.28726
8	*At4g23810*	20	*At4g21830*	-4.35895
8	*At4g23810*	21	*At4g21850*	-3.38331
8	*At4g23810*	22	*At4g22470*	-4.06228
8	*At4g23810*	23	*At4g35160*	-3.7997
8	*At4g23810*	24	*At4g36950*	-2.56587
8	*At4g23810*	25	*At5g02490*	-3.30979
8	*At4g23810*	26	*At5g05270*	-2.95651
8	*At4g23810*	27	*At5g13220*	-1.77738
8	*At4g23810*	28	*At5g13930*	-2.23846
8	*At4g23810*	31	*At1g56650*	-3.41065
8	*At4g23810*	32	*At1g66390*	-1.88994
8	*At4g23810*	34	*At5g07990*	-2.53159
8	*At4g23810*	35	*At5g17220*	-2.04787
8	*At4g23810*	36	*At5g42800*	-2.09349
8	*At4g23810*	38	*At5g56840*	-1.80919
8	*At4g23810*	39	*At5g59780*	1.89943
8	*At4g23810*	40	*At1g22640*	-2.15302
8	*At4g23810*	41	*At3g50060*	2.3585
8	*At4g23810*	42	*At5g44190*	3.1875

8	*At4g23810*	43	*At1g71030*	3.02189
8	*At4g23810*	44	*At5g08520*	3.68474
8	*At4g23810*	45	*At1g66370*	-2.91846
8	*At4g23810*	46	*At1g18570*	1.89848
8	*At4g23810*	48	*At3g28910*	3.79605
12	*At1g43160*	8	*At4g23810*	-2.05558
12	*At1g43160*	26	*At5g05270*	1.90647
12	*At1g43160*	35	*At5g17220*	2.32373
12	*At1g43160*	38	*At5g56840*	3.0594
12	*At1g43160*	39	*At5g59780*	1.75919
12	*At1g43160*	40	*At1g22640*	-2.79255
14	*At2g02990*	2	*At1g72520*	2.0258
14	*At2g02990*	5	*At2g44840*	2.19044
14	*At2g02990*	17	*At2g43870*	2.69397
14	*At2g02990*	24	*At4g36950*	2.36444
14	*At2g02990*	27	*At5g13220*	1.95934
14	*At2g02990*	41	*At3g50060*	2.81512
14	*At2g02990*	42	*At5g44190*	-2.45959
14	*At2g02990*	43	*At1g71030*	-2.65123
14	*At2g02990*	45	*At1g66370*	2.11189
18	*At3g11480*	6	*At3g23250*	1.81493
18	*At3g11480*	12	*At1g43160*	1.99403
18	*At3g11480*	14	*At2g02990*	1.91656
18	*At3g11480*	15	*At2g38240*	1.96886
18	*At3g11480*	18	*At3g11480*	1.94511
18	*At3g11480*	20	*At4g21830*	1.97746
18	*At3g11480*	21	*At4g21850*	1.80911
18	*At3g11480*	22	*At4g22470*	1.94031
18	*At3g11480*	23	*At4g35160*	1.90236
18	*At3g11480*	25	*At5g02490*	1.79475
18	*At3g11480*	26	*At5g05270*	1.72781
18	*At3g11480*	31	*At1g56650*	1.82092
18	*At3g11480*	42	*At5g44190*	-1.77831
18	*At3g11480*	43	*At1g71030*	-1.74145
18	*At3g11480*	44	*At5g08520*	-1.89392
18	*At3g11480*	45	*At1g66370*	1.71457
18	*At3g11480*	48	*At3g28910*	-1.90912
21	*At4g21850*	12	*At1g43160*	1.75016
21	*At4g21850*	15	*At2g38240*	1.79111
21	*At4g21850*	18	*At3g11480*	1.80587
21	*At4g21850*	48	*At3g28910*	-1.77017
23	*At4g35160*	3	*At1g74930*	2.1335
23	*At4g35160*	6	*At3g23250*	-2.77636
23	*At4g35160*	8	*At4g23810*	2.05836
23	*At4g35160*	9	*At4g34410*	2.17899
23	*At4g35160*	12	*At1g43160*	-3.35098
23	*At4g35160*	13	*At1g49900*	-1.99128
23	*At4g35160*	14	*At2g02990*	-3.08385
23	*At4g35160*	15	*At2g38240*	-3.26647
23	*At4g35160*	16	*At2g38380*	-2.23445

23	*At4g35160*	18	*At3g11480*	-3.19126
23	*At4g35160*	19	*At3g48520*	-2.0594
23	*At4g35160*	20	*At4g21830*	-3.30331
23	*At4g35160*	21	*At4g21850*	-2.77815
23	*At4g35160*	22	*At4g22470*	-3.16856
23	*At4g35160*	23	*At4g35160*	-3.03857
23	*At4g35160*	24	*At4g36950*	-2.23952
23	*At4g35160*	25	*At5g02490*	-2.74308
23	*At4g35160*	26	*At5g05270*	-2.53011
23	*At4g35160*	28	*At5g13220*	-2.03522
23	*At4g35160*	31	*At1g56650*	-2.79781
23	*At4g35160*	32	*At1g66390*	-1.71332
23	*At4g35160*	34	*At5g07990*	-2.24494
23	*At4g35160*	35	*At5g17220*	-1.86166
23	*At4g35160*	36	*At5g42800*	-1.90149
23	*At4g35160*	39	*At5g59780*	1.74446
23	*At4g35160*	40	*At1g22640*	-1.95059
23	*At4g35160*	41	*At3g50060*	2.08491
23	*At4g35160*	42	*At5g44190*	2.66693
23	*At4g35160*	43	*At1g71030*	2.566
23	*At4g35160*	44	*At5g08520*	2.9619
23	*At4g35160*	45	*At1g66370*	-2.48927
23	*At4g35160*	46	*At1g18570*	1.7459
23	*At4g35160*	48	*At3g28910*	3.02548
33	*At4g22880*	19	*At3g48520*	2.47979
33	*At4g22880*	28	*At5g13220*	2.21836
33	*At4g22880*	33	*At4g22880*	2.21169
44	*At5g08520*	1	*At1g19210*	1.78499
44	*At5g08520*	27	*At5g13220*	1.78734
47	*At5g08520*	2	*At1g72520*	-2.34815
47	*At5g07690*	3	*At1g74930*	-2.32566
47	*At5g07690*	4	*At1g80840*	-2.2169
47	*At5g07690*	5	*At2g44840*	-2.09381
47	*At5g07690*	6	*At3g23250*	-2.2175
47	*At5g07690*	7	*At4g23800*	-2.58813
47	*At5g07690*	8	*At4g23810*	-1.91016
47	*At5g07690*	9	*At4g34410*	-1.96152
47	*At5g07690*	10	*At5g21960*	-2.17513
47	*At5g07690*	17	*At2g43870*	-1.75538
47	*At5g07690*	34	*At5g07990*	1.77964
47	*At5g07690*	36	*At5g42800*	1.94542
47	*At5g07690*	37	*At5g54060*	1.73041

These regulatory gene-gene relationships are in keeping with current research suggesting AtWRKY53 is a key regulator of leaf senescence (Ay, Irmler et al. 2009) through regulation of expression of downstream senescence associated transcription factors that directly regulate anthocyanin biosynthesis and other senescence-associated processes. Further to this the WT model shows *At4g23810* (AtWRKY53) (8) interacting with the anthocyanin biosynthesis regulatory transcription factor *At1g56650* (AtMYB75) (31) as was predicted using the microarray results of the Arabidopsis senescence experiment (Chapter 4). However, in direct contrast to the senescence experiment, where the predicted relationship between *At4g23810* (AtWRKY53) and *At1g56650* (AtMYB75) was positive, in this experiment the predicted relationship between them was negative (Table 5.2) suggesting that *At4g23810* (WRK53) is inhibiting *At1g56650* (AtMYB75) expression. It should also be noted that in both microarray experiments *At1g56650* (AtMYB75) was not predicted to be a major regulator of anthocyanin biosynthesis.

One of the most interesting aspects of the WT model (Table 5.2, Figure 5.1) is of the predicted role of *At5g07690* (AtMYB29) (47). The model predicts *At5g07690* (AtMYB29) (47) to be a major hub of anthocyanin biosynthesis having direct positive regulation of anthocyanin biosynthesis genes *At4g22880* (LDOX) (33), *At4g22880* (TT7/F3’H - Transparent Testa 7/flavonoid-3-monooxygenase) (34), *At5g07990* (DFR - Dihydroflavonol 4-reductase) (36) and *At5g54060* (UF3GT - UDP-Glucose:Flavonoid 3-O-Glucosyltransferase) (37). It can also be argued that AtMYB29 is indirectly regulating the expression of other AtMYB transcription factors *At3g50060* (AtMYB77) (41), *At1g71030* (AtMYBL2) (43), *At5g08520* (MYB-like transcription factor) (44), *At1g66370* (AtMYB113) (45), *At1g18570*

(AtMYB51) (46) and other senescence-associated genes through *At2g02990* (RNS1). *At5g07690* (AtMYB29) is also predicted to regulate two AtMYB transcription factors *At5g44190* (GLK2) (42) and *At3g23250* (AtMYB15) (6), which in turn is predicted to regulate *At1g66390* (AtMYB90) (32). The positive relationship between *At3g23250* (AtMYB15) and *At1g66390* (AtMYB90) is consistent with predictions made by the model generated using the Arabidopsis long day microarray data. However in sharp contrast to the Arabidopsis long day model neither *At3g23250* (AtMYB15) nor *At1g66390* (AtMYB90) are predicted as either direct or indirect regulators of anthocyanin biosynthesis, instead *At5g07690* (AtMYB29) is predicted to be upstream of both AtMYBs and is predicted to directly regulate anthocyanin biosynthesis genes. Thus *At5g07690* (AtMYB29) could be a major regulator of an alternative pathway.

The regulatory gene- gene relationships predicted by the IM28 mutant (Table 5.3, Figure 5.2) show 11 predicted hubs; *At1g72520* (lipoxygenase) (2), *At1g74930* (ORA47 DNA binding transcription factor) (3), *At3g23250* (AtMYB15) (6), *At4g23810* (AtWRKY53) (8), *At1g43160* (RAP2.6 ethylene response transcription factor) (12), *At2g02990* (RNS1) (14), *At2g38240* (oxidoreductase) (15), *At2g43870* (polygalacturonase) (17), *At5g44190* (GLK2) (42), *At1g71030* (AtMYBL2) (43), and *At5g07690* (AtMYB29) (47). The IM28 model shows that much like the WT model AtWRKY53 is shown to be a major regulatory hub, directly regulating the expression of 34 of the 48 genes including four anthocyanin biosynthesis genes; *At5g05270* (CHI) (26), *At5g13930* (CHS) (28), *At5g07990* (TT7) (34) and *At5g42800* (DFR) (36). The model shows that the regulatory relationship between *At4g23810* (*At*WRK53) (8) and the anthocyanin genes are negative. In addition to this the IM28 model shows *At5g07690* (AtMYB29) (47) as not only as a major regulatory hub but that it positively regulates the same anthocyanin biosynthesis genes. This suggests that *At5g07690* (AtMYB29) (47) and *At4g23810* (AtWRKY53) both act to co-regulate expression of anthocyanin biosynthesis genes.

The WT and IM28 models predict the same regulatory hubs and the majority of predicted regulatory gene- gene relationships are the same, in fact many of the predictions made by both models reflect the ongoing process of senescence. For instance in both models *At4g23810* (AtWRKY53) is shown to be regulating AP2 domain transcription factors *At1g19210*, *At4g34410* (1, 9), *At1g43160* (RAP2.6 DNA binding transcription factor) (12), *At2g02990* (RNS1) (14), *At5g13220* (JAZ10) (27) and *At3g11480* (BSMT1 - S-adenosylmethionine-dependent methyltransferase) (18).

Although there is a high degree of similarity between the models in terms of the regulatory gene- gene relationships predicted by them there are some notable differences. In the WT model *At5g07690* (AtMYB29) is directly regulating *At3g23250* (AtMYB15), which in turn is shown as directly regulating *At1g66390* (AtMYB90) (32). In sharp contrast to this, the IM28 model shows *At4g23810* (AtWRKY53) (8) as directly regulating *At1g66390* (AtMYB90) (32) although the regulatory relationship in this instance is a negative one. Also, the IM28 model predicts *At4g22880* (LDOX) as positively regulating *At5g13930* (CHS) (28) with no regulator up stream of it whereas in the WT model *At4g22880* (LDOX) is predicted to be regulated by *At5g07690* (AtMYB29).

5.4 Discussion

The modelling of the forty eight anthocyanin biosynthesis related pathway genes for WT and IM28 knockout yielded some interesting results. Both models show *At*4g23810 (AtWRKY53) as a major regulator of senescence associated processes. The prediction is consistent with the current literature in which AtWRKY53 has been shown to be an important regulatory transcription factor during the early stages of senescence. Expression studies have shown that RNAi silencing of the AtWRKY53 leads to a delayed onset of senescence in comparison to the WT (Col-0), whereas over-expression of AtWRKY53 leads to an early onset of senescence (Miao, Laun et al. 2004). Furthermore both models show positive regulatory relationships between AtWRKY53 and AP2/ethylene responsive binding proteins (EREBP), and negative relationships with JAZ10. These regulatory relationships also concur with the current literature which show AP2/EREBP is regulated by numerous biotic and abiotic stresses such as cold, drought, pathogen infection, wounding or treatment with ethylene, Salicylic Acid (SA) or Jasmonic acid (JA) as part of their response pathways (Singh, Foley et al. 2002). The negative regulatory relationship between AtWRKY53 and JAZ10 is in keeping with current scientific theory that JAZ10 acts as a repressor of the methyl jasmonate (MeJA) signalling pathway (Chung and Howe 2009). A repression of JAZ10 expression by AtWRKY53 leading to increased MeJA signalling pathway activity is consistent with the known role of AtWRKY53 as a positive regulator of senescence and ties in with the idea that SA and JA stress response pathways and senescence associated signalling pathways overlap (Love, Milner et al. 2008). These results suggest that the VBSSM modelling method is effective at identifying regulatory networks that have been previously been described in other studies.

One of the most interesting results shown in both models is that of AtMYB29. Both WT and IM28 models pin point AtMYB29 as a positive regulator of anthocyanin biosynthesis. Both models predict AtMYB29 as directly positively regulating anthocyanin biosynthesis genes. Further to this neither model show AtMYB90 nor AtMYB75 as being a direct regulators of anthocyanin production. The results showing that AtMYB90 is not a major regulatory hub in the WT model suggests a partial redundancy of AtMYB90. Furthermore, the IM28 mutant model predicts regulatory relationships between anthocyanin-associated genes despite the silencing of AtMYB90. The model provides an explanation for this in that AtMYB29 is shown to be involved in positive regulatory relationships with the anthocyanin associated genes. The prediction that AtMYB90 is not essential for anthocyanin biosynthesis is consistent with the findings made by Nichola Warner (Warner 2008). In fact the lack of change between the WT model and the AtMYB90 knockout IM28 model would suggest that the loss of AtMYB90 function impacts very little on the signalling pathways. However one of the differences highlighted between the WT and IM28 models with respect to AtMYB90 is the role of AtMYB15. In the previous chapter AtMYB15 has been predicted to be a positive regulator of AtMYB90 using the Arabidopsis long day microarray dataset. This prediction is confirmed in the current WT model. As anticipated this linkage is not present in the IM28 model as AtMYB90 is mutated in this Arabidopsis line.

This new prediction for the role of AtMYB29 is fascinating, as hitherto it had only been shown to be involved in the production of glucosinolates as part of the MeJA signalling pathway in response to wounding (Gigolashvili, Engqvist et al. 2008). The WT and IM28 models suggest that both AtWRKY53 and AtMYB29 act to co-regulate anthocyanin production. This adds further evidence to the concept that during senescence the SA and JA stress response pathways are used to regulate the expression of genes associated with senescence. AtMYB29 could also be the regulator of an alternative anthocyanin biosynthesis pathway in the absence of AtMYB90 or indeed as the predictions suggest, the essential regulator of anthocyanin production upstream of AtMYB90. How AtMYB29 expression could regulate anthocyanin biosynthesis is still yet to be determined, whether AtMYB29 directly influences anthocyanin associated genes or whether it simply acts as mediator of an upstream signal from another source through the MeJA pathway, perhaps in response to a rise in JA levels as JA have been shown to stimulate anthocyanin accumulation in many plant systems (Loreti, Povero et al. 2008). One possibility is that with the onset of senescence, chlorophyll and other photosynthesis associated apparatus are broken down and recycled into the developing seeds (Lim, Kim et al. 2007). This process temporarily makes the leaves more vulnerable to UV associated oxidative stress leading to the damaging of leaf cells. This damage could stimulate the MeJA stress response pathway leading to stimulation of AtMYB29; this in turn could lead to a stimulation of anthocyanin-associated genes.

There are several approaches that could be taken to determine the extent to which AtMYB*29* is involved in the regulation of anthocyanin biosynthesis. One approach would be to over express the AtMYB*29* gene using an inducible promoter such as has been used by Gigolashvili (Gigolashvili, Engqvist et al. 2008) and determine the effect on anthocyanin gene expression via Q-PCR. A second approach would be to silence AtMYB*29* using RNAi (Dugas and Bartel 2004) using specifically designed artificial miRNAs (Alvarez, Pekker et al. 2006; Schwab, Ossowski et al. 2006). The effect the silencing has on the expression of downstream anthocyanin associated genes can be determined again via Q-PCR. A third approach would be to perform a microarray time course experiment on an AtMYB*29* knockout mutant. A previous experiment to characterise AtMYB28 and AtMYB29 as regulators of aliphatic glucosinolate production under non-stress and also in response to stressful environmental conditions, used an Arabidopsis AtMYB*29* mutant, the seeds of which (SALK_ N55242) are available from The Nottingham Arabidopsis Stock Centre (NASC), http://arabidopsis.info, 2009. The RNA could be extracted from the leaves of plants harvested at several regular time points and hybridised to a microarray and the expression values extracted. This time course could be used to model and determine the effects of the loss of AtMYB29 on anthocyanin production. Furthermore, an Arabidopsis line could be constructed that is mutant in both AtWRKY53 and AtMYB29 to determine whether both of these genes co-regulate anthocyanin associated gene expression and whether the loss of expression of these genes causes a permanent decrease in anthocyanin biosynthesis gene expression through Q-PCR.

Although AtMYB29 is predicted to be directly interacting with the promoter regions of several anthocyanin genes; *At4g22880* (LDOX), *At4g22880* (TT7/F3'H -

Transparent Testa 7/flavonoid-3-monooxygenase), *At5g07990* (DFR - Dihydroflavonol 4-reductase) and *At5g54060* (UF3GT - UDP-Glucose: Flavonoid 3-O-Glucosyltransferase), there is no current literature, which supports this prediction. In order to further investigate whether AtMYB29 transcription factor regulates expression of anthocyanin biosynthesis directly, yeast one hybrid could be used. This is a technique that determines whether there are any direct protein-DNA interactions occurring. The technique uses a single fusion protein in which the activator domain of Gal4 is linked directly to the DNA binding domain from AtMYB29. This can be tested to determine if the fusion protein causes specific activation of an anthocyanin biosynthesis gene promoter target sequence, which is inserted in the promoter region of the reporter gene construct. If the AtMYB29 transcription factor successfully binds to the promoter regions then this will trigger the transcription of the reporter gene. This technique could provide experimental evidence that AtMYB29 is a direct regulator of anthocyanin biosynthesis. An alternative to this method would be to use ChIP-Seq (Chromatin Immunoprecipitation – Sequencing). The process combines ChIP with next generation DNA sequencing to identify binding sites of DNA-associated proteins (Jothi, Cuddapah et al. 2008). In this technique the ChIP process enriches specific cross-linked DNA-protein complexes and is then detected using an antibody against the protein of interest, in this instance AtMYB29. After this the protein is removed leaving only the sequence it bound to. The DNA sequence is then ligated to two oligonucleotide adapters, PCR amplified and sequenced using next generation sequencing technology. The sequenced reads can then be mapped to the reference genome, in this instance Arabidopsis. This enables identification of regions that are overrepresented in the number of mapped reads, which may correspond to transcription factor binding sites upstream of known anthocyanin gene locations.

This identification can be done using software such as PeakSeq (Rozowsky, Euskirchen et al. 2009). If this technique was successful and the model prediction correct, it would elucidate the target genes of AtMYB29 regulation and the location of their transcription factor binding sites. It would also provide experimental evidence of a direct regulatory relationship between AtMYB29 and anthocyanin associated genes. This evidence could then be used as prior knowledge for any future modelling of these genes thus improving the accuracy of the model and hopefully elucidating new predicted interactions.

Chapter 6: General Discussion

6.1 Annotation of regions of the *H. arabidopsidis* genome potentially involved in pathogenicity

Annotation of the BAC P1202 region confirmed the presence of *Ppat*s *3* and *8* as identified in the SSH cDNA library of *H. arabidopsidis* genes expressed during infection of the host plant *Arabidopsis thaliana* (Bittner-Eddy, Allen et al. 2003). The investigation also identified an NAD-dependent epimerase, a phosphatidylinositol-4-phosphate 5 kinase (PI4P5K) and a glycoprotein in the region as well as 5 unknown proteins, 6 hypothetical proteins and 9 transposable elements. The predicted function of some of the genes identified in the sequence of BAC P1202 may indicate involvement in both the formation of structures and the regulation of processes, which contribute towards the pathogenicity of *H. arabidopsidis*. Furthermore all of these identified genes have been conserved in the syntenic regions of *P. sojae*, *P. ramorum* and *P. infestans*, also suggesting a conserved role for the genes in this region in the general maintenance of the pathogen pathogenicity structures.

Annotation of the *ATR13* locus confirmed the presence of the *ATR13* effector gene (Allen, Bittner-Eddy et al. 2004). The annotation also revealed the presence of an Acetyl-coA carboxylase, an Acyl transferase and a GTP-binding protein as well as two hypothetical proteins, one unknown protein and 30 transposable elements and two 61 kb inverted repeat regions were identified toward the 3 prime end of *ATR13*. The functions of the identified genes suggest they are involved in essential cellular processes. This suggests they would not have a specific role in pathogenicity.

Analysis of the region for any potential effectors returned no candidates, which is in contrast to avirulence loci analysed in the genomes of *P. sojae* and *P. infestans* which show clustering of avirulence genes (Whisson, Drenth et al. 1995; Gijzen, Forster et al. 1996; van der Lee, Robold et al. 2001).

6.2 Analysis of the apoplastic effector family within the *H. arabidopsidis* secretome

The analysis of the *H. arabidopsidis* secretome has enabled the identification of 15 candidate ELI and ELL sequences based upon their cysteine rich spacing patterns. However, of the 15 candidates only one candidate, HpELI4, could be classed as an elicitin based on the highly conserved 96 amino acid domain C_1-23-C_2-23-C_3-4-C_4-14-C_5-23-$C_{6.}$ Comparison of the rate of non-synonymous nucleotide substitutions to that of synonymous substitutions carried out on *H. arabidopsidis* ELI and ELL sequences indicated the sequences were under diversifying selective pressure which could explain the proliferation of ELL sequences and the conservation of only one ELI in the *H. arabidopsidis* genome. However this leaves some question as to how *H. arabidopsidis* obtains sterols, as the main functions of elicitins identified in *Phytophthora* are to act as sterol carriers. Currently *H. arabidopsidis* has only one gene with a high likelihood to function as a sterol carrier. One possibility is that the ELL sequences are involved in sterol uptake but essential amino acid residues involved in sterol binding are not conserved in the ELL sequences. Despite this, a general involvement in lipid binding can be assigned to ELLs of *Phytophthora* as characterisation studies of ELLs found in *Phytophthora capsici* showed phospholipid activity (Nespoulous, Gaudemer et al. 1999).

Nespoulous et al. (1999) also identified the presence of three candidates for Kazal-like serine protease inhibitors in *P. capsici* and orthologues from *P. sojae*, *P. infestans* and *P. ramorum.* Current research suggests that kazal-like domains are conserved particularly within Oomycete species particularly in the genomes of *P. sojae*, *P. ramorum*, *P. infestans*, *P. brassicae* and *Pl. halstedii* (Tian, Huitema et al. 2004). However, the low numbers of kazal-like serine protease inhibitors found in *H. arabidopsidis* is in sharp contrast to the numbers found in other oomycete species (12 in *P. infestans* and 18 in *P. sojae*) but the numbers are more akin to *P. brassicae* which currently has only two candidates (Kamoun 2006). This suggests that the extent to which *H. arabidopsidis* uses protease inhibitors as a means of counter defence is much less than those of most of the Phytophthora species.

Searches for genes similar to *Ppat* 24 and *Ppat* 14, shown to be involved in *H. arabidopsidis* pathogenicity (Bittner-Eddy, Allen et al. 2003), returned three candidates similar to *Ppat* 24 and one other candidate similar to *Ppa*t 14. BLAST analysis of both *Ppat* 24 and *Ppat* 14 against the genomes of *P. sojae*, *P. infestans* and *P. ramorum* identified no orthologues, indicating that the sequences are unique to *H. arabidopsidis.* However, no function could be identified for *Ppat* 24 and *Ppat* 14.

6.3 Modelling transcriptional networks from pathogen induced and developmental microarray time course experiments

The aims of this chapter were to identify potential signalling networks up regulated during plant defence responses to infection by *H. arabidopsidis* and developmental senescence using a new model that has been developed by Beal et al (Beal, Falciani et al. 2005) to reverse engineer transcriptional networks. Unfortunately this first aim failed due to technical problems with the experiment (see pathogen results section). Unlike related *P. ramorum* and *P. sojae* species that have been defined as hemi-biotrophs (Moy, Qutob et al. 2004) *H. arabidopsidis* is an obligate biotroph and, therefore, its survival depends upon remaining undiscovered after parasitisation of the host plant. It extends haustoria into mesophyll cells that surround the growing hypha and delivers effectors into the contacted cells. Therefore, in any leaf or cotyledon the number of infected host cells relative to uninfected is very small. Therefore early time points in infection, which are the most informative, are likely to be masked by mRNA isolated form unaffected host cells. An alternative mRNA sampling technique would be to use single cell sampling (Tomos and Sharrock 2000) and isolate mRNA from haustoria associated cells, then a high-throughput sequencing technique known as RNA-seq to quantify the changing expression levels of each transcript under the different conditions. Another possibility would be to express individual effectors *in planta*, under the control of constitutive or inducible promoters, and monitor their effect on transcription in infected and uninfected tissues.

Network inference was also carried out on a dataset created to identify transcriptional profiles altered during leaf developmental senescence. The results of the model generated indicate that the Beal modelling software (Beal, Falciani et al. 2005) was capable of predicting a model in which the entire set of anthocyanin biosynthesis pathway related genes were included. Furthermore the anthocyanin biosynthesis pathway associated genes were predicted to be downstream of the AtMYB90 transcription factor known to regulate anthocyanin production (Borevitz, Xia et al. 2000). The model also predicted the presence of AtMYB15 as a positive regulator of anthocyanin biosynthesis. The presence of AtMYB15 in the model suggests a complex web of overlapping transcriptional regulation of both senescence and plant responses to various stresses as it had been previously implicated in transcriptional responses to cold and wounding stress (Lim, Kim et al. 2007). Approaches that could be taken to test the extent to which AtMYB15 is involved in regulation of anthocyanin production would be to over express the AtMYB15 gene using an inducible promoter such as that of Dexamethasone has been used by Chen et al during the over expression of AtMYB15 (Chen, Zhang et al. 2006). A second approach would be to silence AtMYB15 via small RNAs. The process known as RNA interference (RNAi) can be used to silence AtMYB15 and then the effect it has on the expression of downstream anthocyanin associated genes can be determined via Q-PCR. A third approach would be to perform a microarray time course experiment on an AtMYB15 knockout mutant previously used to identify the effect of AtMYB15 on CBF genes conferring increased tolerance to cold stress (Agarwal, Hao et al. 2006). The gene expression profiles from the microarray could be used to determine the predicted effects of the AtMYB15 knockout on anthocyanin biosynthesis.

6.4 Modelling transcriptional networks using Wild Type and MYB90 mutant microarray time course experiments.

This study investigated the effects of the absence of MYB90 on senescence by comparing the phenotype of the MYB90 knock out, IM28, to wild type (WT) Col-0 using a time course microarray to identify differentially expressed genes. The results suggested that the absence of MYB90 might have affected the expression of the key genes normally associated with anthocyanin biosynthesis but anthocyanin levels increase despite this. The continued rise in anthocyanin levels in IM28 suggested that there may be an alternative signalling pathway controlling the biosynthesis of the anthocyanin. The presence of time course datasets represented an opportunity to model the differentially expressed genes with the original 38 genes used to model the anthocyanin biosynthesis pathway and determine whether the generated model would support the previous predictions made by the modelling methodology suggested AtMYB15 is an upstream regulator of AtMYB90.

In the network inference models generated for WT and IM28 models AtMYB29 was identified as a positive regulator of anthocyanin biosynthesis. Further to this neither model show AtMYB90 nor AtMYB75 as being direct regulators of anthocyanin production. This suggests a partial redundancy with AtMYB90. In addition, the IM28 mutant model shows continued regulatory relationships between anthocyanin associated genes and AtMYB29 despite the silencing of AtMYB90. This indicates that AtMYB90 is not essential for the regulation of anthocyanin biosynthesis. There are several approaches that could be taken to determine the extent to which AtMYB29 is involved in the regulation of anthocyanin biosynthesis. One approach would be to over express the AtMYB*29* gene using an inducible promoter such as has been used by Gigolashvili (Gigolashvili, Engqvist et al. 2008) to show AtMYB29 is involved in the production glucosinolates as part of the MeJA signalling pathway in response to wounding. This can be used to determine the effect on anthocyanin gene expression via Q-PCR. A second approach would be to silence AtMYB*29* using RNAi (Dugas and Bartel 2004) and the effect on the expression of downstream anthocyanin associated genes assessed. A third approach would be to perform a microarray time course experiment on an AtMYB*29* knockout mutant. In order to further investigate whether AtMYB29 transcription factor regulates expression of anthocyanin biosynthesis directly yeast one hybrid could be used. This is a technique that determines whether there are any direct protein-DNA interactions occurring. The technique uses a single fusion protein in which the activator domain from Gal4 is linked directly to the DNA binding domain from AtMYB29. If the fusion protein causes specific activation of an anthocyanin biosynthesis gene promoter target sequence, this will trigger the transcription of the reporter gene. This technique could provide experimental evidence that AtMYB29 is a direct regulator of anthocyanin biosynthesis. All the evidence provided by the experimental techniques discussed could then be used as prior knowledge for any future modelling of these genes thus improving the accuracy of the model and hopefully elucidating new predicted interactions.

Chapter 7: Bibliography

Abeles, F. B., L. J. Dunn, et al. (1988). "Induction of 33-Kd and 60-Kd Peroxidases During Ethylene-Induced Senescence of Cucumber Cotyledons." Plant Physiology **87**(3): 609-615.

Agarwal, M., Y. J. Hao, et al. (2006). "A R2R3 type MYB transcription factor is involved in the cold regulation of CBF genes and in acquired freezing tolerance." Journal of Biological Chemistry **281**(49): 37636-37645.

Akutsu T, M. S., Kuhara S (1999). "Identification of genetic networks from a small number of gene expression patterns under the Boolean network model." Pac Symp Biocomput: 17-28.

Albert, R. (2007). "Network inference, analysis, and modeling in systems biology." Plant Cell **19**(11): 3327-3338.

Alfano, J. R. and A. Collmer (2004). "Type III secretion system effector proteins: Double agents in bacterial disease and plant defense." Annual Review of Phytopathology **42**: 385-414.

Allen, R. L., P. D. Bittner-Eddy, et al. (2004). "Host-parasite coevolutionary conflict between Arabidopsis and downy mildew." Science **306**(5703): 1957-1960.

Altschul, S. (1999). "Hot papers - Bioinformatics - Gapped BLAST and PSI-BLAST: a new generation of protein database search programs by S.F. Altschul, T.L. Madden, A.A. Schaffer, J.H. Zhang, Z. Zhang, W. Miller, D.J. Lipman - Comments." Scientist **13**(8): 15-15.

Altschul, S., T. Madden, et al. (1998). "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs." Faseb Journal **12**(8): A1326-A1326.

Alvarez, J. P., I. Pekker, et al. (2006). "Endogenous and synthetic microRNAs stimulate simultaneous, efficient, and localized regulation of multiple targets in diverse species." Plant Cell **18**(5): 1134-1151.

Armstrong, M. R., S. C. Whisson, et al. (2005). "An ancestral oomycete locus contains late blight avirulence gene Avr3a, encoding a protein that is recognized in the host cytoplasm." Proceedings of the National Academy of Sciences of the United States of America **102**(21): 7766-7771.

Aubourg, S., M. L. Martin-Magniette, et al. (2007). "Analysis of CATMA transcriptome data identifies hundreds of novel functional genes and improves gene models in the Arabidopsis genome." Bmc Genomics **8**.

Ay, N., K. Irmler, et al. (2009). "Epigenetic programming via histone methylation at WRKY53 controls leaf senescence in Arabidopsis thaliana." Plant Journal **58**(2): 333-346.

Bae, H., M. S. Kim, et al. (2006). "Necrosis- and ethylene-inducing peptide from Fusarium oxysporum induces a complex cascade of transcripts associated with signal transduction and cell death in arabidopsis." Plant Physiology **141**(3): 1056-1067.

Bae, H. H., J. H. Bowers, et al. (2005). "NEP1 orthologs encoding necrosis and ethylene inducing proteins exist as a multigene family in Phytophthora megakarya, causal agent of black pod disease on cacao." Mycological Research **109**: 1373-1385.

Bateman, A., L. Coin, et al. (2004). "The Pfam protein families database." Nucleic Acids Research **32**: D138-D141.

Baum, L. E. and T. Petrie (1966). "STATISTICAL INFERENCE FOR PROBABILISTIC FUNCTIONS OF FINITE STATE MARKOV CHAINS." Annals of Mathematical Statistics **37**(6): 1554-&.

Beal, M. J., F. Falciani, et al. (2005). "A Bayesian approach to reconstructing genetic regulatory networks with hidden factors." Bioinformatics **21**(3): 349-356.

Beal, M. J. and Z. Ghahramani (2002). The variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures. 7th Valencia International Meeting on Bayesian Statistics, Valencia, SPAIN.

Beissbarth, T. (2006). Interpreting experimental results using gene ontologies. DNA Microarrays, Part B: Databases and Statistics. **411:** 340-352.

Beissbarth, T. and T. P. Speed (2004). "GOstat: find statistically overrepresented Gene Ontologies within a group of genes." Bioinformatics **20**(9): 1464-1465.

Bendtsen, J. D., H. Nielsen, et al. (2004). "Improved prediction of signal peptides: SignalP 3.0." Journal of Molecular Biology **340**(4): 783-795.

Besemer, J. and M. Borodovsky (2005). "GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses." Nucleic Acids Research **33**: W451-W454.

Bhattacharjee, S., N. L. Hiller, et al. (2006). "The malarial host-targeting signal is conserved in the Irish potato famine pathogen." Plos Pathogens **2**(5): 453-465.

Birch, P. R. J., A. P. Rehmany, et al. (2006). "Trafficking arms: oomycete effectors enter host plant cells." Trends in Microbiology **14**(1): 8-11.

Bishop, J. G., D. R. Ripoll, et al. (2005). "Selection on glycine beta-1,3-endoglucanase genes differentially inhibited by a phytophthora glucanase inhibitor protein." Genetics **169**(2): 1009-1019.

Bittner-Eddy, P., C. Can, et al. (1999). "Genetic and physical mapping of the RPP13 locus, in arabidopsis, responsible for specific recognition of several Peronospora parasitica (downy mildew) isolates." Molecular Plant-Microbe Interactions **12**(9): 792-802.

Bittner-Eddy, P. D., R. L. Allen, et al. (2003). "Use of suppression subtractive hybridization to identify downy mildew genes expressed during infection of Arabidopsis thaliana." Molecular Plant Pathology **4**(6): 501-507.

Boeckmann, B., A. Bairoch, et al. (2003). "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003." Nucleic Acids Research **31**(1): 365-370.

Borevitz, J. O., Y. J. Xia, et al. (2000). "Activation tagging identifies a conserved MYB regulator of phenylpropanoid biosynthesis." Plant Cell **12**(12): 2383-2393.

Bos, J. I. B., M. Armstrong, et al. (2003). "Intraspecific comparative genomics to identify avirulence genes from Phytophthora." New Phytologist **159**(1): 63-72.

Botella, M. A., J. E. Parker, et al. (1998). "Three genes of the arabidopsis RPP1 complex resistance locus recognize distinct Peronospora parasitica avirulence determinants." Plant Cell **10**(11): 1847-1860.

Brady, S. M., D. A. Orlando, et al. (2007). "A high-resolution root spatiotemporal map reveals dominant expression patterns." Science **318**(5851): 801-806.

Breeze, E., E. Harrison, et al. (2008). "Transcriptional regulation of plant senescence: from functional genomics to systems biology." Plant Biology **10**: 99-109.

Bruggeman, F. J. and H. V. Westerhoff (2007). "The nature of systems biology." Trends in Microbiology **15**(1): 45-50.

Brunner, F., S. Rosahl, et al. (2002). "Pep-13, a plant defense-inducing pathogen-associated pattern from Phytophthora transglutaminases." Embo Journal **21**(24): 6681-6688.

Buchanan-Wollaston, V., E. Harrison, et al. (2007). "Elucidating signaling pathways that control Arabidopsis leaf senescence." Comparative Biochemistry and Physiology a-Molecular & Integrative Physiology **146**(4): S53-S54.

Buchanan-Wollaston, V., T. Page, et al. (2005). "Comparative transcriptome analysis reveals significant differences in gene expression and signalling pathways between developmental and dark/starvation-induced senescence in Arabidopsis." Plant Journal **42**(4): 567-585.

Buzzell, R. I. and T. R. Anderson (1992). "Inheritance and Race Reaction of a New Soybean Rps1 Allele." Plant Disease **76**(6): 600-601.

Carver, T. J., K. M. Rutherford, et al. (2005). "ACT: the Artemis comparison tool." Bioinformatics **21**(16): 3422-3423.

Carzaniga, R., P. Bowyer, et al. (2001). "Production of extracellular matrices during development of infection structures by the downy mildew Peronospora parasitica." New Phytologist **149**(1): 83-93.

Chang, W. C., C. W. Li, et al. (2005). "Quantitative inference of dynamic regulatory pathways via microarray data." Bmc Bioinformatics **6**.

Chen, W. Q., N. J. Provart, et al. (2002). "Expression profile matrix of Arabidopsis transcription factor genes suggests their putative functions in response to environmental stresses." Plant Cell **14**(3): 559-574.

Chen, Y. H., X. B. Zhang, et al. (2006). "Overexpression of the wounding-responsive gene AtMYB15 activates the shikimate pathway in Arabidopsis." Journal of Integrative Plant Biology **48**(9): 1084-1095.

Chung, H. S. and G. A. Howe (2009). "A Critical Role for the TIFY Motif in Repression of Jasmonate Signaling by a Stabilized Splice Variant of the JASMONATE ZIM-Domain Protein JAZ10 in Arabidopsis." Plant Cell **21**(1): 131-145.

Crowe, M. L., C. Serizet, et al. (2003). "CATMA: a complete Arabidopsis GST database." Nucleic Acids Research **31**(1): 156-158.

Damasceno, C. M. B., J. G. Bishop, et al. (2008). "Structure of the glucanase inhibitor protein (GIP) family from Phytophthora species suggests coevolution with plant endo-beta-1,3-glucanases." Molecular Plant-Microbe Interactions **21**(6): 820-830.

Davies, K. M. and K. E. Schwinn (2003). "Transcriptional regulation of secondary metabolism." Functional Plant Biology **30**(9): 913-925.

Deikman, J. and P. E. Hammer (1995). "INDUCTION OF ANTHOCYANIN ACCUMULATION BY CYTOKININS IN ARABIDOPSIS-THALIANA." Plant Physiology **108**(1): 47-57.

Delcher, A. L., D. Harmon, et al. (1999). "Improved microbial gene identification with GLIMMER." Nucleic Acids Research **27**(23): 4636-4641.

Devoto, A. and J. G. Turner (2005). "Jasmonate-regulated Arabidopsis stress signalling network." Physiologia Plantarum **123**(2): 161-172.

Doak, T. G., F. P. Doerder, et al. (1994). "A Proposed Superfamily of Transposase Genes - Transposon-Like Elements in Ciliated Protozoa and a Common D35e Motif." Proceedings of the National Academy of Sciences of the United States of America **91**(3): 942-946.

Doherty, H. M., R. R. Selvendran, et al. (1988). "THE WOUND RESPONSE OF TOMATO PLANTS CAN BE INHIBITED BY ASPIRIN AND RELATED HYDROXYBENZOIC ACIDS." Physiological and Molecular Plant Pathology **33**(3): 377-384.

Drozdowicz, Y. M. and P. A. Rea (2001). "Vacuolar H+ pyrophosphatases: from the evolutionary backwaters into the mainstream." Trends in Plant Science **6**(5): 206-211.

Dugas, D. V. and B. Bartel (2004). "MicroRNA regulation of gene expression in plants." Current Opinion in Plant Biology **7**(5): 512-520.

Eisenhaber, B., M. Wildpaner, et al. (2003). "Glycosylphosphatidylinositol lipid anchoring of plant proteins. Sensitive prediction from sequence- and genome-wide studies for arabidopsis and rice." Plant Physiology **133**(4): 1691-1701.

Elbashir, S. M., W. Lendeckel, et al. (2001). "RNA interference is mediated by 21- and 22-nucleotide RNAs." Genes & Development **15**(2): 188-200.

Ellis, J. G., P. N. Dodds, et al. (2007). "Flax rust resistance gene specificity is based on direct resistance-avirulence protein interactions." Annual Review of Phytopathology **45**: 289-306.

Eulgem, T. and I. E. Somssich (2007). "Networks of WRKY transcription factors in defense signaling." Current Opinion in Plant Biology **10**(4): 366-371.

Fabritius, A. L. and H. S. Judelson (2003). "A mating-induced protein of Phytophthora infestans is a member of a family of elicitors with divergent structures and stage-specific patterns of expression." Molecular Plant-Microbe Interactions **16**(10): 926-935.

Falquet, L., M. Pagni, et al. (2002). "The PROSITE database, its status in 2002." Nucleic Acids Research **30**(1): 235-238.

Fefeu, S., S. Bouaziz, et al. (1997). "Three-dimensional solution structure of beta cryptogein, a beta elicitin secreted by a phytopathogenic fungus Phytophthora cryptogea." Protein Science **6**(11): 2279-2284.

Fellbrich, G., A. Romanski, et al. (2002). "NPP1, a Phytophthora-associated trigger of plant defense in parsley and Arabidopsis." Plant Journal **32**(3): 375-390.

Feys, B. J. and J. E. Parker (2000). "Interplay of signaling pathways in plant disease resistance." Trends in Genetics **16**(10): 449-455.

Friedman, N., M. Linial, et al. (2000). Using Bayesian networks to analyze expression data. 4th Annual International Conference on Computational Biology (RECOMB 2000), Tokyo, Japan.

Fung, R. W. M., C. Y. Wang, et al. (2006). "Characterization of alternative oxidase (AOX) gene expression in response to methyl salicylate and methyl jasmonate pre-treatment and low temperature in tomatoes." Journal of Plant Physiology **163**(10): 1049-1060.

Gao, L. L., W. Knogge, et al. (2004). "Expression patterns of defense-related genes in different types of arbuscular mycorrhizal development in wild-type and mycorrhiza-defective mutant tomato." Molecular Plant-Microbe Interactions **17**(10): 1103-1113.

Gaulin, E., A. Jauneau, et al. (2002). "The CBEL glycoprotein of Phytophthora parasitica var. nicotianae is involved in cell wall deposition and adhesion to cellulosic substrates." Journal of Cell Science **115**(23): 4565-4575.

Gebhardt, C. and J. P. T. Valkonen (2001). "Organization of genes controlling disease resistance in the potato genome." Annual Review of Phytopathology **39**: 79-102.

Ghanem, M. E., A. Albacete, et al. (2008). "Hormonal changes during salinity-induced leaf senescence in tomato (Solanum lycopersicum L.)." Journal of Experimental Botany **59**(11): 3039-3050.

Gigolashvili, T., M. Engqvist, et al. (2008). "HAG2/MYB76 and HAG3/MYB29 exert a specific and coordinated control on the regulation of aliphatic glucosinolate biosynthesis in Arabidopsis thaliana." New Phytologist **177**(3): 627-642.

Gijzen, M., H. Forster, et al. (1996). "Cosegregation of Avr4 and Avr6 in Phytophthora sojae." Canadian Journal of Botany-Revue Canadienne De Botanique **74**(5): 800-802.

Glazebrook, J. (2001). "Genes controlling expression of defense responses in Arabidopsis - 2001 status." Current Opinion in Plant Biology **4**(4): 301-308.

Goff, S. A., K. C. Cone, et al. (1992). "FUNCTIONAL-ANALYSIS OF THE TRANSCRIPTIONAL ACTIVATOR ENCODED BY THE MAIZE-B GENE - EVIDENCE FOR A DIRECT FUNCTIONAL INTERACTION BETWEEN 2 CLASSES OF REGULATORY PROTEINS." Genes & Development **6**(5): 864-875.

Gonzalez, A., M. Zhao, et al. (2008). "Regulation of the anthocyanin biosynthetic pathway by the TTG1/bHLH/Myb transcriptional complex in Arabidopsis seedlings." Plant Journal **53**(5): 814-827.

Grbic, V. and A. B. Bleecker (1995). "ETHYLENE REGULATES THE TIMING OF LEAF SENESCENCE IN ARABIDOPSIS." Plant Journal **8**(4): 595-602.

Guo, Y. F. and S. S. Gan (2006). "AtNAP, a NAC family transcription factor, has an important role in leaf senescence." Plant Journal **46**(4): 601-612.

Gupta, V., M. G. Willits, et al. (2000). "Arabidopsis thaliana EDS4 contributes to salicylic acid (SA)-dependent expression of defense responses: Evidence for inhibition of jasmonic acid signaling by SA." Molecular Plant-Microbe Interactions **13**(5): 503-511.

Hardham, A. R. (2001). The cell biology behind Phytophthora pathogenicity. 2nd Australasian Soilborne Diseases Symposium, Lorne, Australia.

He, Y. H., H. Fukushige, et al. (2002). "Evidence supporting a role of jasmonic acid in Arabidopsis leaf senescence." Plant Physiology **128**(3): 876-884.

He, Y. H. and S. S. Gan (2002). "A gene encoding an acyl hydrolase is involved in leaf senescence in Arabidopsis." Plant Cell **14**(4): 805-815.

Heil, M. and J. C. Silva Bueno (2007). "Within-plant signaling by volatiles leads to induction and priming of an indirect plant defense in nature." Proceedings of the National Academy of Sciences of the United States of America **104**(13): 5467-5472.

Heard, N. A., C. C. Holmes, et al. (2005). "Bayesian coclustering of Anopheles gene expression time series: Study of immune defense response to multiple experimental challenges." Proceedings of the National Academy of Sciences of the United States of America **102**(47): 16939-16944.

Hiller, N. L., S. Bhattacharjee, et al. (2004). "A host-targeting signal in virulence proteins reveals a secretome in malarial infection." Science **306**(5703): 1934-1937.

Hirokawa, T., S. Boon-Chieng, et al. (1998). "SOSUI: classification and secondary structure prediction system for membrane proteins." Bioinformatics **14**(4): 378-379.

Hisamatsu, Y., N. Goto, et al. (2006). "Senescence-promoting effect of arabidopside A." Zeitschrift Fur Naturforschung C-a Journal of Biosciences **61**(5-6): 363-366.

Holub, E. B. and J. L. Beynon (1997). Symbiology of mouse-ear cress (Arabidopsis thaliana) and oomycetes. Advances in Botanical Research Incorporating Advances in Plant Pathology, Vol 24. **24:** 227-273.

Hortensteiner, S. (2009). "Stay-green regulates chlorophyll and chlorophyll-binding protein degradation during senescence." Trends in Plant Science **14**(3): 155-162.

Hotelling, H. (1931). "The Generalization of Student's Ratio." The Annals of Mathematical Statistics **2**(3): 360-378.

Huitema, E., J. I. B. Bos, et al. (2004). "Linking sequence to phenotype in Phytophthora-plant interactions." Trends in Microbiology **12**(4): 193-200.

Husmeier, D. (2003). "Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks." Bioinformatics **19**(17): 2271-2282.

Huson, D. H. and D. Bryant (2006). "Application of phylogenetic networks in evolutionary studies." Molecular Biology and Evolution **23**(2): 254-267.

Jia, Y., S. A. McAdams, et al. (2000). "Direct interaction of resistance gene and avirulence gene products confers rice blast resistance." Embo Journal **19**(15): 4004-4014.

Jiang, R. H. Y., A. L. Dawe, et al. (2005). "Elicitin genes in Phytophthora infestans are clustered and interspersed with various transposon-like elements." Molecular Genetics and Genomics **273**(1): 20-32.

Jiang, R. H. Y., B. M. Tyler, et al. (2006). "Comparative analysis of Phytophthora genes encoding secreted proteins reveals conserved synteny and lineage-specific gene duplications and deletions." Molecular Plant-Microbe Interactions **19**(12): 1311-1321.

Jiang, R. H. Y., B. M. Tyler, et al. (2006). "Ancient origin of elicitin gene clusters in Phytophthora genomes." Molecular Biology and Evolution **23**(2): 338-351.

Jing, H. C., J. H. M. Schippers, et al. (2005). "Ethylene-induced leaf senescence depends on age-related changes and OLD genes in Arabidopsis." Journal of Experimental Botany **56**(421): 2915-2923.

Jing, H. C., M. J. G. Sturre, et al. (2002). "Arabidopsis onset of leaf death mutants identify a regulatory pathway controlling leaf senescence." Plant Journal **32**(1): 51-63.

Jones, J. D. G. and J. L. Dangl (2006). "The plant immune system." Nature **444**(7117): 323-329.

Jones, M. L. and W. R. Woodson (1997). "Pollination-induced ethylene in carnation - Role of stylar ethylene in corolla senescence." Plant Physiology **115**(1): 205-212.

Jothi, R., S. Cuddapah, et al. (2008). "Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data." Nucleic Acids Research **36**(16): 5221-5231.

Judelson, H. S. and S. Roberts (2002). "Novel protein kinase induced during sporangial cleavage in the oomycete Phytophthora infestans." Eukaryotic Cell **1**(5): 687-695.

Kachroo, P., K. Yoshioka, et al. (2000). "Resistance to turnip crinkle virus in Arabidopsis is regulated by two host genes and is salicylic acid dependent but NPR1, ethylene, and jasmonate independent." Plant Cell **12**(5): 677-690.

Kalman, R. E. (1960). "A new approach to linear filtering and prediction problems." Journal of Basic Engineering(82): 95-108.

Kamoun, S. (2006). "A catalogue of the effector secretome of plant pathogenic oomycetes." Annual Review of Phytopathology **44**: 41-60.

Kamoun, S., P. Hraber, et al. (1999). "Initial assessment of gene diversity for the oomycete pathogen Phytophthora infestans based on expressed sequences." Fungal Genetics and Biology **28**(2): 94-106.

Kamoun, S., E. Huitema, et al. (1999). "Resistance to oomycetes: a general role for the hypersensitive response?" Trends in Plant Science **4**(5): 196-200.

Kamoun, S. and C. D. Smart (2005). "Late blight of potato and tomato in the genomics era." Plant Disease **89**(7): 692-699.

Kamoun, S., P. van West, et al. (1998). "Resistance of Nicotiana benthamiana to Phytophthora infestans is mediated by the recognition of the elicitor protein INF1." Plant Cell **10**(9): 1413-1425.

Kamoun, S., P. vanWest, et al. (1997). "A gene encoding a protein elicitor of Phytophthora infestans is down-regulated during infection of potato." Molecular Plant-Microbe Interactions **10**(1): 13-20.

Kanneganti, T. D., E. Huitema, et al. (2006). "Synergistic interactions of the plant cell death pathways induced by Phytophthora infestans Nep1-like protein PiNPP1.1 and INF1 elicitin." Molecular Plant-Microbe Interactions **19**(8): 854-863.

Karrer, E. E., J. E. Lincoln, et al. (1995). "IN-SITU ISOLATION OF MESSENGER-RNA FROM INDIVIDUAL PLANT-CELLS - CREATION OF CELL-SPECIFIC CDNA LIBRARIES." Proceedings of the National Academy of Sciences of the United States of America **92**(9): 3814-3818.

Katari, M. S., S. D. Nowicki, et al. (2008). "VirtualPlant: A software platform to support systems biology research in the post-genomic era." Plant Biology (Rockville) **2008**: 45-46.

Khatib, M., C. Lafitte, et al. (2004). "The CBEL elicitor of Phytophthora parasitica var. nicotianae activates defence in Arabidopsis thaliana via three different signalling pathways." New Phytologist **162**(2): 501-510.

Kim, H. J., H. Ryu, et al. (2006). "Cytokinin-mediated control of leaf longevity by AHK3 through phosphorylation of ARR2 in Arabidopsis." Proceedings of the National Academy of Sciences of the United States of America **103**(3): 814-819.

Kjemtrup, S., Z. Nimchuk, et al. (2000). "Effector proteins of phytopathogenic bacteria: bifunctional signals in virulence and host recognition." Current Opinion in Microbiology **3**(1): 73-78.

Klee, E. W. and L. B. M. Ellis (2005). "Evaluating eukaryotic secreted protein prediction." Bmc Bioinformatics **6**.

Kloek, A. P., M. L. Verbsky, et al. (2001). "Resistance to Pseudomonas syringae conferred by an Arabidopsis thaliana coronatine-insensitive (coi1) mutation occurs through two distinct mechanisms." Plant Journal **26**(5): 509-522.

Kohler, J., J. Baumbach, et al. (2006). "Graph-based analysis and visualization of experimental results with ONDEX." Bioinformatics **22**(11): 1383-1390.

Labrador, M. and V. G. Corces (1997). "Transposable element-host interactions: Regulation of insertion and excision." Annual Review of Genetics **31**: 381-404.

Latijnhouwers, M., P. de Wit, et al. (2003). "Oomycetes and fungi: similar weaponry to attack plants." Trends in Microbiology **11**(10): 462-469.

Lee, R. C. and V. Ambros (2001). "An extensive class of small RNAs in Caenorhabditis elegans." Science **294**(5543): 862-864.

Lengeler, K. B., R. C. Davidson, et al. (2000). "Signal transduction cascades regulating fungal development and virulence." Microbiology and Molecular Biology Reviews **64**(4): 746-+.

Li, H., J. Ruan, et al. (2008). "Mapping short DNA sequencing reads and calling variants using mapping quality scores." Genome Research **18**(11): 1851-1858.

Librado, P. and J. Rozas (2009). "DnaSP v5: a software for comprehensive analysis of DNA polymorphism data." Bioinformatics **25**(11): 1451-1452.

Lim, P. O., H. J. Kim, et al. (2007). "Leaf senescence." Annual Review of Plant Biology **58**: 115-136.

Lister, R., R. C. O'Malley, et al. (2008). "Highly integrated single-base resolution maps of the epigenome in Arabidopsis." Cell **133**(3): 523-536.

Liu, S. P., R. A. Cerione, et al. (2002). "Structural basis for the guanine nucleotide-binding activity of tissue transglutaminase and its regulation of transamidation activity." Proceedings of the National Academy of Sciences of the United States of America **99**(5): 2743-2747.

Liu, Z. Y., J. I. B. Bos, et al. (2005). "Patterns of diversifying selection in the phytotoxin-like scr74 gene family of Phytophthora infestans (vol 22, pg 659, 2005)." Molecular Biology and Evolution **22**(4): 1159-1159.

Lonnig, W. E. and H. Saedler (2002). "Chromosome rearrangements and transposable elements." Annual Review of Genetics **36**: 389-410.

Loreti, E., G. Povero, et al. (2008). "Gibberellins, jasmonate and abscisic acid modulate the sucrose-induced expression of anthocyanin biosynthetic genes in Arabidopsis." New Phytologist **179**(4): 1004-1016.

Love, A. J., J. J. Milner, et al. (2008). "Timing is everything: regulatory overlap in plant cell death." Trends in Plant Science **13**(11): 589-595.

Mathews, A., B. J. Carroll, et al. (1990). "THE GENETIC INTERACTION BETWEEN NON-NODULATION AND SUPERNODULATION IN SOYBEAN - AN EXAMPLE OF DEVELOPMENTAL EPISTASIS." Theoretical and Applied Genetics **79**(1): 125-130.

McDowell, J. M., M. Dhandaydham, et al. (1998). "Intragenic recombination and diversifying selection contribute to the evolution of downy mildew resistance at the RPP8 locus of arabidopsis." Plant Cell **10**(11): 1861-1874.

McLeod, A., C. D. Smart, et al. (2003). "Characterization of 1,3-beta-glucanase and 1,3;1,4-beta-glucanase genes from Phytophthora infestans." Fungal Genetics and Biology **38**(2): 250-263.

Meijer, H. J. G. and F. Govers (2006). "Genomewide analysis of phospholipid signaling genes in Phytophthora spp.: Novelties and a missing link." Molecular Plant-Microbe Interactions **19**(12): 1337-1347.

Meinke, D. W., J. M. Cherry, et al. (1998). "Arabidopsis thaliana: A model plant for genome analysis." Science **282**(5389): 662-+.

Miao, Y., T. Laun, et al. (2004). "Targets of the WRKY53 transcription factor and its role during leaf senescence in Arabidopsis." Plant Molecular Biology **55**(6): 853-867.

Mikes, V., M. L. Milat, et al. (1998). "Elicitins, proteinaceous elicitors of plant defense, are a new class of sterol carrier proteins." Biochemical and Biophysical Research Communications **245**(1): 133-139.

Mikes, V., M. L. Milat, et al. (1997). "The fungal elicitor cryptogein is a sterol carrier protein." Febs Letters **416**(2): 190-192.

Mishina, T. E., C. Lamb, et al. (2007). "Expression of a nitric oxide degrading enzyme induces a senescence programme in Arabidopsis." Plant Cell and Environment **30**(1): 39-52.

Moreau, P., P. Thoquet, et al. (1998). "Genetic mapping of Ph-2, a single locus controlling partial resistance to Phytophthora infestans in tomato." Molecular Plant-Microbe Interactions **11**(4): 259-269.

Morgan, W. and S. Kamoun (2007). "RXLR effectors of plant pathogenic oomycetes." Current Opinion in Microbiology **10**(4): 332-338.

Morris, K., S. A. H. Mackerness, et al. (2000). "Salicylic acid has a role in regulating gene expression during leaf senescence." Plant Journal **23**(5): 677-685.

Mortazavi, A., B. A. Williams, et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nature Methods **5**(7): 621-628.

Moy, P., D. Qutob, et al. (2004). "Patterns of gene expression upon infection of soybean plants by Phytophthora sojae." Molecular Plant-Microbe Interactions **17**(10): 1051-1062.

Mueller, L. A., P. F. Zhang, et al. (2003). "AraCyc: A biochemical pathway database for Arabidopsis." Plant Physiology **132**(2): 453-460.

Mur, L. A. J., Y. M. Bi, et al. (1997). "Compromising early salicylic acid accumulation delays the hypersensitive response and increases viral dispersal during lesion establishment in TMV-infected tobacco." Plant Journal **12**(5): 1113-1126.

Mur, L. A. J., I. R. Brown, et al. (2000). "A loss of resistance to avirulent bacterial pathogens in tobacco is associated with the attenuation of a salicylic acid-potentiated oxidative burst." Plant Journal **23**(5): 609-621.

Nespoulous, C., O. Gaudemer, et al. (1999). "Characterization of elicitin-like phospholipases isolated from Phytophthora capsici culture filtrate." Febs Letters **452**(3): 400-406.

Nurnberger, T., D. Nennstiel, et al. (1994). "HIGH-AFFINITY BINDING OF A FUNGAL OLIGOPEPTIDE ELICITOR TO PARSLEY PLASMA-MEMBRANES TRIGGERS MULTIPLE DEFENSE RESPONSES." Cell **78**(3): 449-460.

Oh, S. A., J. H. Park, et al. (1997). "Identification of three genetic loci controlling leaf senescence in Arabidopsis thaliana." Plant Journal **12**(3): 527-535.

Orsomando, G., M. Lorenzi, et al. (2003). "PcF protein from Phytophthora cactorum and its recombinant homologue elicit phenylalanine ammonia lyase activation in tomato." Cellular and Molecular Life Sciences **60**(7): 1470-1476.

Page, R. D. M. (1996). "TreeView: An application to display phylogenetic trees on personal computers." Computer Applications in the Biosciences **12**(4): 357-358.

Parker, J. E., M. J. Coleman, et al. (1997). "The Arabidopsis downy mildew resistance gene RPP5 shares similarity to the toll and interleukin-1 receptors with N and L6." Plant Cell **9**(6): 879-894.

Pemberton, C. L. and G. P. C. Salmond (2004). "The Nep1-like proteins - a growing family of microbial elicitors of plant necrosis." Molecular Plant Pathology **5**(4): 353-359.

Pemberton, C. L., N. A. Whitehead, et al. (2005). "Novel quorum-sensing-control led genes in Erwinia carotovora subsp carotovora: Identification of a fungal elicitor homologue in a soft-rotting bacterium." Molecular Plant-Microbe Interactions **18**(4): 343-353.

Penninckx, I., K. Eggermont, et al. (1996). "Pathogen-induced systemic activation of a plant defensin gene in Arabidopsis follows a salicylic acid-independent pathway." Plant Cell **8**(12): 2309-2323.

Qutob, D., P. T. Hraber, et al. (2000). "Comparative analysis of expressed sequences in Phytophthora sojae." Plant Physiology **123**(1): 243-253.

Qutob, D., E. Huitema, et al. (2003). "Variation in structure and activity among elicitins from Phytophthora sojae." Molecular Plant Pathology **4**(2): 119-124.

Qutob, D., S. Kamoun, et al. (2002). "Expression of a Phytophthora sojae necrosis-inducing protein occurs during transition from biotrophy to necrotrophy." Plant Journal **32**(3): 361-373.

R. Durbin, S. E., A. Krogh and G. Mitchison (1998). " Biological sequence analysis: probabilistic models of proteins and nucleic acids." Cambridge University Press.

Rangel, C., J. Angus, et al. (2004). "Modeling T-cell activation using gene expression profiling and state-space models." Bioinformatics **20**(9): 1361-1372.

Rao, M. V. and K. R. Davis (2001). "The physiology of ozone induced cell death." Planta **213**(5): 682-690.

Rao, M. V., H. Lee, et al. (2002). "Ozone-induced ethylene production is dependent on salicylic acid, and both salicylic acid and ethylene act in concert to regulate ozone-induced cell death." Plant Journal **32**(4): 447-456.

Rehmany, A. P., A. Gordon, et al. (2005). "Differential recognition of highly divergent downy mildew avirulence gene alleles by RPP1 resistance genes from two Arabidopsis lines." Plant Cell **17**(6): 1839-1850.

Rhee, S. Y., W. Beavis, et al. (2003). "The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community." Nucleic Acids Research **31**(1): 224-228.

Rice, P., I. Longden, et al. (2000). "EMBOSS: The European molecular biology open software suite." Trends in Genetics **16**(6): 276-277.

Rizzo, D. M., M. Garbelotto, et al. (2005). "Phytophthora ramorum: Integrative research and management of an emerging pathogen in California and Oregon forests." Annual Review of Phytopathology **43**: 309-335.

Robatzek, S. and I. E. Somssich (2002). "Targets of AtWRKY6 regulation during plant senescence and pathogen defense." Genes & Development **16**(9): 1139-1149.

Rose, J. K. C., K. S. Ham, et al. (2002). "Molecular cloning and characterization of glucanase inhibitor proteins: Coevolution of a counterdefense mechanism by plant pathogens." Plant Cell **14**(6): 1329-1345.

Rose, L. E., P. D. Bittner-Eddy, et al. (2004). "The maintenance of extreme amino acid diversity at the disease resistance gene, RPP13, in Arabidopsis thaliana." Genetics **166**(3): 1517-1527.

Roweis, S. and Z. Ghahramani (1999). "A unifying review of linear gaussian models." Neural Computation **11**(2): 305-345.

Rozowsky, J., G. Euskirchen, et al. (2009). "PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls." Nature Biotechnology **27**(1): 66-75.

Rutherford, K., J. Parkhill, et al. (2000). "Artemis: sequence visualization and annotation." Bioinformatics **16**(10): 944-945.

Saibo, N. J. M., T. Lourenco, et al. (2009). "Transcription factors and regulation of photosynthetic and related metabolism under environmental stresses." Annals of Botany **103**(4): 609-623.

Scheel, D., K. Hahlbrock, et al. (1995). "Peptide Elicitor Recognition and Signal-Transduction in Plant Defense." Journal of Cellular Biochemistry: 472-472.

Schenk, P. M., K. Kazan, et al. (2000). "Coordinated plant defense responses in Arabidopsis revealed by microarray analysis." Proceedings of the National Academy of Sciences of the United States of America **97**(21): 11655-11660.

Schiex, T., A. Moisan, et al. (2001). "Eugène, an eukaryotic gene finder that combines several sources of evidence." Computational Sciences **2066**: 111-125.

Schwab, R., S. Ossowski, et al. (2006). "Highly specific gene silencing by artificial microRNAs in Arabidopsis." Plant Cell **18**(5): 1121-1133.

Sclep, G., J. Allemeersch, et al. (2007). "CATMA, a comprehensive genome-scale resource for silencing and transcript profiling of Arabidopsis genes." BMC Bioinformatics **8**.

Shan, W. X., M. Cao, et al. (2004). "The Avr1b locus of Phytophthora sojae encodes an elicitor and a regulator required for avirulence on soybean plants carrying resistance gene Rps1b." Molecular Plant-Microbe Interactions **17**(4): 394-403.

Singh, K. B., R. C. Foley, et al. (2002). "Transcription factors in plant defense and stress responses." Current Opinion in Plant Biology **5**(5): 430-436.

Skriver, K. and J. Mundy (1990). "Gene-Expression in Response to Abscisic-Acid and Osmotic-Stress." Plant Cell **2**(6): 503-512.

Slusarenko, A. J. and N. L. Schlaich (2003). "Downy mildew of Arabidopsis thaliana caused by Hyaloperonospora parasitica (formerly Peronospora parasitica)." Molecular Plant Pathology **4**(3): 159-170.

Smalle, J. and D. VanderStraeten (1997). "Ethylene and vegetative development." Physiologia Plantarum **100**(3): 593-605.

Spellman, P. T., G. Sherlock, et al. (1998). "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization." Molecular Biology of the Cell **9**(12): 3273-3297.

Stajich, J. E., D. Block, et al. (2002). "The bioperl toolkit: Perl modules for the life sciences." Genome Research **12**(10): 1611-1618.

Stoffer, S. a. (1982). "An approach to time series smoothing and forecasting using the EM algorithm." Journal of Time Series Analysis **3**(4): 253 - 264.

Suzuki, A., T. Kikuchi, et al. (1997). "Changes of ethylene evolution, ACC content, ethylene forming enzyme activity and respiration in fruits of highbush blueberry." Journal of the Japanese Society for Horticultural Science **66**(1): 23-27.

Swarbreck, D., C. Wilks, et al. (2008). "The Arabidopsis Information Resource (TAIR): gene structure and function annotation." Nucleic Acids Research **36**: D1009-D1014.

Swofford, D. L. (2003). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Sunderland, Massachusetts., Sinauer Associates.

Tai, Y. C. and T. P. Speed (2006). "A multivariate empirical Bayes statistic for replicated microarray time course data." Annals of Statistics **34**(5): 2387-2412.

Thomma, B., K. Eggermont, et al. (1998). "Separate jasmonate-dependent and salicylate-dependent defense-response pathways in Arabidopsis are essential for resistance to distinct microbial pathogens." Proceedings of the National Academy of Sciences of the United States of America **95**(25): 15107-15111.

Thompson, J. D., D. G. Higgins, et al. (1994). "Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice." Nucleic Acids Research **22**(22): 4673-4680.

Tian, M. Y., B. Benedetti, et al. (2005). "A second kazal-like protease inhibitor from Phytophthora infestans inhibits and interacts with the apoplastic pathogenesis-related protease P69B of tomato." Plant Physiology **138**(3): 1785-1793.

Tian, M. Y., E. Huitema, et al. (2004). "A Kazal-like extracellular serine protease inhibitor from Phytophthora infestans targets the tomato pathogenesis-related protease P69B." Journal of Biological Chemistry **279**(25): 26370-26377.

Tomos, A. D. and R. A. Sharrock (2000). Cell sampling and analysis (SiCSA): metabolites measured at single cell resolution. Annual Meeting of the Society-for-Experimental-Biology, Exeter, England.

Torres, M. A., J. D. G. Jones, et al. (2005). "Pathogen-induced, NADPH oxidase-derived reactive oxygen intermediates suppress spread of cell death in Arabidopsis thaliana." Nature Genetics **37**(10): 1130-1134.

Torto, T. A., L. Rauser, et al. (2002). "The pipg 1 gene of the oomycete Phytophthora infestans encodes a fungal-like endopolygalacturonase." Current Genetics **40**(6): 385-390.

Tyler, B. M. (2002). "Molecular basis of recognition between Phytophthora pathogens and their hosts." Annual Review of Phytopathology **40**: 137-167.

Tyler, B. M. (2007). "Phytophthora sojae: root rot pathogen of soybean and model oomycete." Molecular Plant Pathology **8**(1): 1-8.

Tyler, B. M., S. Tripathy, et al. (2006). "Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis." Science **313**(5791): 1261-1266.

Ueda, J. and J. Kato (1980). "Isolation and Identification of a Senescence-Promoting Substance from Wormwood (Artemisia-Absinthium L)." Plant Physiology **66**(2): 246-249.

Van der Graaff, E., R. Schwacke, et al. (2006). "Transcription analysis of arabidopsis membrane transporters and hormone pathways during developmental and induced leaf senescence." Plant Physiology **141**(2): 776-792.

van der Lee, T., A. Robold, et al. (2001). "Mapping of avirulence genes in Phytophthora infestans with amplified fragment length polymorphism markers selected by bulked segregant analysis." Genetics **157**(3): 949-956.

van't Slot, K. A. E. and W. Knogge (2002). "A dual role for microbial pathogen-derived effector proteins in plant disease and resistance." Critical Reviews in Plant Sciences **21**(3): 229-271.

Veit, S., J. M. Worle, et al. (2001). "A novel protein elicitor (PaNie) from Pythium aphanidermatum induces multiple defense responses in carrot, Arabidopsis, and tobacco." Plant Physiology **127**(3): 832-841.

Wang, Z., G. G. Cao, et al. (2008). "Identification and characterization of COI1-dependent transcription factor genes involved in JA-mediated response to wounding in Arabidopsis plants." Plant Cell Reports **27**(1): 125-135.

Warner, N. (2008). Unravelling the roles of two senescent enhanced MYB transcription factors in the regulation of anthocyanin biosynthesis in Arabidopsis thaliana. Warwick HRI. Warwick, University of Warwick. **PhD:** 200.

Weaver, L. M., S. S. Gan, et al. (1998). "A comparison of the expression patterns of several senescence-associated genes in response to stress and hormone treatment." Plant Molecular Biology **37**(3): 455-469.

Weitzman, J. B., L. Fiette, et al. (2000). "JunD protects cells from p53-dependent senescence and apoptosis." Molecular Cell **6**(5): 1109-1119.

Weng, C., K. Yu., et al. (2001). "Mapping Genes Conferring Resistance to Phytophthora Root Rot of Soybean, Rps1a and Rps7." The Journal of Heredity **92**(5): 442-446.

Whisson, S. C., P. C. Boevink, et al. (2007). "A translocation signal for delivery of oomycete effector proteins into host plant cells." Nature **450**: 115-+.

Whisson, S. C., A. Drenth, et al. (1995). "Phytophthora sojae avirulence genes, RAPD, and RFLP markers used to construct a detailed genetic linkage map." Molecular Plant-Microbe Interactions **8**(6): 988-995.

Winkel-Shirley, B. (2001). "Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology." Plant Physiology **126**(2): 485-493.

Zareparsi, S., A. Hero, et al. (2004). "Seeing the unseen: Microarray-based gene expression profiling in vision." Investigative Ophthalmology & Visual Science **45**(8): 2457-2462.

Appendices

Appendix. A The gettingRXLRs.pl perl code used to identify gene products containing the RXLR motif synonymous with avirulence proteins.

```
#opens the file of orfs and sends to the subroutine fix sequence
open(FILEHANDLE, "/Users/hrseaw/Perl_scripts/results/$seq_file" )or die ("cannot
open \n\n");

my @temp = <FILEHANDLE>;
chomp @temp;
my @orfs = perl_Modules->fix_sequence(@temp);

open (OUTFILE,">>/Users/hrseaw/Perl_scripts/results/5_heptads_Hp.fasta");
open (OUTFILE3,">>/Users/hrseaw/Perl_scripts/results/4_heptads_Hp.fasta");
open (OUTFILE4,">>/Users/hrseaw/Perl_scripts/results/3_heptads_Hp.fasta");

#get sequences beginning with a methionine start followed by 27 amino acids then
look for the RXLR motif between 28-40 amino acids after the methionine start
followed by ILorV repeats

foreach my $pos (@orfs){
  if ($pos =~
/M.{28,40}R.LR.+([ILV].{6}[ILV].{6}[ILV].{6}[ILV].{6}[ILV].{6})/g){
            print OUTFILE "$1\n";
            print OUTFILE "$pos\n";
            }
            elsif ($pos =~
/M.{28,40}R.LR.+([ILV].{6}[ILV].{6}[ILV].{6}[ILV].{6})/g){
            print OUTFILE3 "$1\n";
            print OUTFILE3 "$pos\n";
            }
            elsif ($pos =~ /M.{28,40}R.LR.+([ILV].{6}[ILV].{6}[ILV].{6})/g){
            print OUTFILE4 "$1\n";
            print OUTFILE4 "$pos\n";
            }
  }

# get any sequence containg the RXLR motif

 my @RXLR = RXLRs(@orfs);



#print RXLR containing repeats
```

```
  open (OUTFILE,">>/Users/hrseaw/Perl_scripts/results/RXLR-MR_Hp.fasta");

   foreach my $gotone (@RXLR){

            print OUTFILE "$gotone\n";
            }

 close OUTFILE;
 close OUTFILE3;
 close OUTFILE4;

exit;

##########################################

#RXLR subroutine

sub RXLRs{

my @RXLR;
my @temp;


foreach my $seq2  (@_){

            if ($seq2=~ /r.lr/g){
                        push (@RXLR,$seq2);
            }
}
return @RXLR;
}
```

Appendix. B. Candidate ELI and ELL amino acid sequences.

Gene Name	Sequence (aa)
HpELI4	MNTHFAIAAIALAVATSVNGQGDCSPEVTKAAYTSMSSLLKRAELMSCGDRSHYNFMTAEQPANHEQELAMCGVAECHTLIAEVKELNPPDCVISIPGRFPINIKAMADAFEGKCKSPNPARSAIEEPTESAMLTAAAPSPDVSEETDVIQQDNDFTEKNTTAYTPGKVLDPFTF
HpELL1A	MKVVASSLIAAALTVANVHADVCDSASVTVLLTSGEVAACTSSSGYSPSSLRSPAIAQLEIMCSAKACQTMLSIVTTMFPEECTINTFALHSGLLAPVSTYCGGSSNSSSLTVTPTDTDLLASTSTTASTETEGATSESSLTDAMMSSALFDNADQKTTGSLDGSVSSTDYADQMTTGSLDGSVSLTNLNESMAWDDYNGSMAWDNYNDSTASAIFDDDMMMDVSGSGSEFTIETMPPSTSSDSLSPDDEGSVSGDPDAVDETPDRASGSVTVGPSSAMLLGSAVVATAALFL
HpELL1B	MKLIFRFSFCYSSTTRLLPSLHQVLVAMKLAVIISALALVANANDNDAPDAYDAHDDHDDVKPVNAVDCNVAALTPLITDPTTIKCANESGYEMTALTAPTQEQSAKMCVNSACQSVLKQVEAIAPTECKLMNFHLHYDLLDPLDRACDDGKPTIGYKDAAPPATGTTPATGTAPATGTTSTTSYTPATGTTPVAGTVTTMNPTTPTTMTPAGGANGATSTEQETTGSTGPEGTTPVVAPGAQSSETPSPTSSDTINTTSGSVSGHSGADLTTILAGAVLTAFVTAFF
HpELL1C	MNIALGSALLLVVAAFSSIAAAPCNTVALSKLFVTGNVTLCRADPGYDPTSMALPTDAQITAVCNSDACKKSISAIKEVAPEECTVGPIRMYADVLNPLSERCGLSSGSRPDAGSVAGNTSYPANTNTTGSAANISQPNARSNSSSPSPNAAASNEGADALTIPMFAAIAVLVAMVTLLL
HpELL4	MQALLALFVTIDLFLSTVVVKAEPCTASEISSIVKPIASNPDYASCQSESNYTLSAFPSPSAAQLRDFCSSSACQGILSATLKSNLLPDCEVVVGSQAFNLIEVAAVLAATCGPAVHELDSLTEGLVDKPDSENPVQRASDRVASLLGHSAPIEKVGIVAALLSLFRE
HpELL6	MIQLSALVLLTLIGSGIGLTNAKPCSTKELSVFNEVSGQVNKCVQDSKLNFQIPPRSSLLMSQQSALCKSEACKDMIGAMDDLDIPNCEAVFDKKNMTLQRSLDMFVSSCDTTTPSPSPIKRRKSLESSSSEGSDVGSKKRRDINPATAAPFGTAHQLVVLLVVGILSLGLVLP
HpELL8	MHSSLFLSFLLTSYGAAADDSCPPATIAKLGELYTNPHLHSCQKIISDTPPANGSTSEQVKALCTSDECSTLINDVLNFKHTDCNMSLVGVELDVRELVNTFEKACQNNKDDSRDSKKYERPPTTKNDTMNHSGQGELNFVKDGNVTQVAENTSEKDDYKLKPPMNNSTAKEFFPMPNTTYKAVVPESAH
HpELL9	MKLVTFVVAAAAVLGSLIAAEQCENTTLSFALIPLEAVSKQCTDDSGYSPFPFKAMPSQAETRAMCTSEACNKLLETAAMPDCTLLVGDSAYNMKESFRLMRAACLVVNVNELAS
HpELL11A	MTGRKFVPTIVTTLLVTSVVSAQECPPDVSESFVATVDNSSYFSTCAEGTTFNVTSVFDVFNFTANDLLQFCNSSSCLEPIHELMGSLDCNISYMGTPRNLSSEVSDLHDVCHEVLDAAEGSGQAAKKPMDLSDHADGSSHSDPTTSDASSSVVLSAVFSVASAAIVAVFLA
HpELL11B	MPQVAKTRVSFRARFYKSTALKSIKEVFEISSTIWTCADGTGYDFYTVMVLPKKHHMPGICASPACRGFVDALELVDPPNCDLHIPWSNTTFNLLDLVYKVKNECDVPPQNSTILV
HpELL11C	MFKYFAITTTILALVAPARGANADAGEDDKCGMIETFRMGSNIGQVFQKSGNIALCASATDFNMLTASVLPDAATMAKMCKDSACRSTIGSLEEFDLPDCTLHVPLVSANLNFARLLKQFQTTCTDSSTSDSASNSTSDSGSASTSTSDSGSASTSTSDSTSNSDEDDDSDEDEDEDEDK
HpELL13A	MSLSNNFSRAQVFVAHFIMVRVLAGFLLPLGLTFSVSEAAECTDAETATAKSVWETTASTSACAPYVTQSEPIYINAPCSATDCISLVEIMVVDLPNCTFNGINNKNDVQRALAACNAVYFKRAELMSAGSLSSDPLNGGSLNVGSLNPDSLSTGSLNAGSLSTDSLSLNAGSLSTGSPSPDSGSLGSGSLTTMATDASGSSTPDVSNTNSDVIFASSASSLDLTTAMTGSSSSASPSGQTTASEGSTVSCSMAEVKKTWNLFVYTATSNECIEASTTGGYGVEILALCGSDCARNVETLADKLPNCYYDNESVNKKEDVRSQLTSGCNGASKFVSVAIVADSSMMFVSSSGSMVVSESDASGSTSFNSGTRPDDETLESSNFATGSSRTNESSAAPSFDAQLHLWTLFLIGIVVAYAS
HpELL13B	MTLLRTFAASFVVLTALQCTAVSASNCTEDEQSTIDSVYATLANGTACSDLMADSGVSSLDYCMQNDCISELSTAVEELPACTGEDGAERKTGLQSIVDYCADVTEVTDQSASGSGPGNDPVVSAASRDVLATSAVITQLFMMIYFVAALS

HpELL13C	MNISALVLAVAIATSTFVVAEDCTVDDLTAISNVYSTASDESCPEMTKMTGGTDYCTFADCLTFMTDMVEKLPDCSTGGINVKESVRAALTVCETGTADISKVFANSSSNATSGSKDAATSPSAGSISSDKASGDLTASDASSSSLSSFALTISSGTFAVILFAGL
HpELL13D	MQAFKPVVGFVAIVAAAALTASAQADSGSLVASPELAAIAECTSTQLDEGQVVLTSNQRAEQCETALNLTAGTMLQVTTASATEMCDTASCRAALQELYNTLPNCRYDLWGLQYSAMKLLEYCGITPVNASESNSTDSTSGSVGWRNTSGSASFAPVGVPTDDPATPRPAPDSSAGAAPITVVSAAVATTVGLVAAYLA

Appendix. C. Candidate kazal-like serine protease inhibitor amino acid sequences.

Gene Name	Sequence (aa)
HpEPI_1	MRQNAKITATMVLVVLCQVTAATTDPEKMLRVVSQVHANRNVYVSAEVTRGRDCVDTGETKNEPVCASNGVRYLNEDMFKYHKCVIQATWDKTIELVDMKMCAEALQEDN
HpEPI_2	MKLSVCLLLAVAAAAIAPIHGQEVDPRCAIRCAFTGDRVCGSNNVTYPNLCLLTLANCANPGEDITVASEGECVPIESEPVVAQLPSKTTGGKSGVTISLPADCADACPMIFAPVCGSDSITYGNGCLLGIAHCESKGTITQTSEGQCPDPSSSGPGDNFSNCPDLCVQTYEPVCGSDGVTHNNICMLRAVACYDPSITLAYEGACETIEDSSQSNETMTKPGKDMASCPDVCLAVFAPVCGSNDVTYGNECELGIVSCNNPGLQLTKVSDGACSNEQPHPNC
HpEPI_3	MQLFFVSIMIVMVVMTSVVNALEREKLLEVLSPVHGRTVDFQYDDLLENQQEHLCYDYELTREENPVCASNGKKYSNPSTFEFHKCLIAAMDQLDIQIVDMKICRDAELEDLDHHDK

Appendix. D. A table showing the synonymous and non-synonymous sites for each pair of elicitin and elicitin –like DNA sequences. Ks (the number of synonymous substitutions per synonymous site) and Ka (the number of non-synonymous substitutions per non-synonymous site) are used to calculate the Ka/Ks ratio. Ka/Ks ratios are used to determine whether the sequences are under diversifying selective pressure, Ka/Ks <1 indicates purifying selection and Ka/Ks >1 indicates diversifying selection.

Seq-1[a]	**Seq-2**[b]	**SilentDif**[c]	**SilentPos**[d]	**Ks**	**SynDif**[e]	**SynPos**[f]	**Ka**	**Pairwise Ka/ks**[g]
HpELL11B	HpELL9	42.5	70.58	1.2178	119.5	178.42	1.6765	1.376662835
HpELL11B	HpELL8	45.58	65.33	1.9974	117.42	183.67	1.4349	0.718383899
HpELL11B	HpELL4	43.33	72.25	1.2059	119.67	176.75	1.7476	1.44920806
HpELL11B	HpELL1C	47.17	70	1.7151	126.83	179	2.172	1.266398461
HpELL11B	HpELL13C	52.08	70.5	3.1512	120.92	178.5	1.7514	0.555788271
HpELL11B	HpELL13A	47.75	66.67	2.3258	128.25	182.33	2.0836	0.895863789
HpELL11B	HpELL13B	47.75	70.25	1.7756	124.5	179.5	1.6031	0.902849741
HpELL11B	HpELL13D	53.92	72.42	3.6911	118.25	178.75	2.7569	0.746904717
HpELL11B	HpELL1B	42.42	70.17	1.23	129.08	176.58	1.8396	1.495609756
HpELL11B	HpELL11C	42.42	69.5	1.2605	122.58	178.83	1.2615	1.000793336
HpELL11B	HpELL6	46.25	66.92	1.9089	109.58	179.5	1.9748	1.0345225
HpELL11B	HpELL1A	49.58	71.92	1.8875	126.75	182.08	1.1839	0.627231788
HpELL11B	HpELI4	42.83	66.25	1.4857	105.42	177.08	2.0511	1.380561352
HpELL11B	HpELL11A	57.83	69.67	n.a*.	128.17	182.75	1.4946	n.a.
HpELL9	HpELL8	39.92	63.08	1.3919	1.3919	124.08	1.6547	1.188806667
HpELL9	HpELL4	46.42	70	1.6164	1.6164	118.58	1.6111	0.996721109
HpELL9	HpELL1C	42.5	67.75	1.3578	1.3578	131.5	2.5666	1.890263662
HpELL9	HpELL13C	47.17	68.25	1.908	1.908	118.83	1.5692	0.822431866
HpELL9	HpELL13A	48.33	64.42	n.a.	129.67	184.58	2.0692	n.a.
HpELL9	HpELL13B	52.67	68	n.a.	114.33	181	1.385	n.a.
HpELL9	HpELL13D	55.5	70.17	n.a.	118.5	178.83	1.6124	n.a.
HpELL9	HpELL1B	51.67	67.92	n.a.	109.33	181.08	1.2262	n.a.
HpELL9	HpELL11C	49.92	67.25	3.4298	126.08	181.75	1.9423	0.566301242
HpELL9	HpELL6	45.58	64.67	2.1083	126.42	184.33	1.8436	0.874448608
HpELL9	HpELL1A	54.67	69.67	n.a.	126.33	179.33	2.1011	n.a.
HpELL9	HpELI4	41.33	64	1.4806	123.67	185	1.6643	1.124071322
HpELL9	HpELL11A	57.08	67.42	n.a.	136.92	181.58	n.a.	n.a.
HpELL8	HpELL4	42.58	64.75	1.5709	123.42	184.25	1.677	1.0675409
HpELL8	HpELL1C	45.17	62.5	2.484	115.83	186.5	1.3207	0.53168277
HpELL8	HpELL13C	39.75	63	1.3804	129.25	186	1.9581	1.418501884
HpELL8	HpELL13A	44.67	59.17	n.a.	133.33	189.83	2.0675	n.a.
HpELL8	HpELL13B	45	62.75	2.3457	121	186.25	1.5087	0.643176877
HpELL8	HpELL13D	44	64.92	1.7554	131	184.08	2.2297	1.270194827
HpELL8	HpELL1B	43.33	62.67	1.9131	127.67	186.33	1.836	0.959698918
HpELL8	HpELL11C	49.25	62	n.a.	119.75	187	1.4423	n.a.
HpELL8	HpELL6	39.75	59.42	1.6693	129.25	189.58	1.7978	1.076978374
HpELL8	HpELL1A	50.75	64.42	n.a.	135.25	184.58	2.8284	n.a.
HpELL8	HpELI4	37.5	58.75	1.4282	125.5	190.25	1.5874	1.111468982
HpELL8	HpELL11A	49.92	62.17	n.a.	133.08	186.83	2.243	n.a.
HpELL4	HpELL1C	45.75	69.42	1.5824	123.25	179.58	1.8495	1.168794237
HpELL4	HpELL13C	54.75	69.92	n.a.	119.25	179.08	1.641	n.a.
HpELL4	HpELL13A	47.08	66.08	2.2465	130.92	182.92	2.3141	1.030091253

HpELL4	HpELL13B	51.67	69.67	3.3713	118.33	179.33	1.589	0.471331534
HpELL4	HpELL13D	57.17	71.83	n.a.	128.83	177.17	2.6196	n.a.
HpELL4	HpELL1B	44	69.58	1.3892	115	179.42	1.4463	1.041102793
HpELL4	HpELL11C	57.25	68.92	n.a.	122.75	180.08	1.7963	n.a.
HpELL4	HpELL6	46.58	66.33	2.0657	129.42	182.67	2.1705	1.050733408
HpELL4	HpELL1A	54.83	71.33	n.a.	133.17	177.67	5.5329	n.a.
HpELL4	HpELI4	42.42	65.67	1.4813	130.58	183.33	2.2423	1.513737933
HpELL4	HpELL11A	50.83	69.08	2.9765	137.17	179.92	n.a.	n.a.
HpELL1C	HpELL13C	44.33	67.67	1.551	112.67	181.33	1.3221	0.852417795
HpELL1C	HpELL13A	37.25	63.83	1.129	126.75	185.17	1.8287	1.619751993
HpELL1C	HpELL13B	48.83	67.42	2.5317	116.17	181.58	1.4379	0.567958289
HpELL1C	HpELL13D	55.25	69.58	n.a.	127.75	179.42	2.2375	n.a.
HpELL1C	HpELL1B	48.08	67.33	2.2797	114.92	181.67	1.3907	0.610036408
HpELL1C	HpELL11C	45.25	66.67	1.7654	125.75	182.33	1.8902	1.070692194
HpELL1C	HpELL6	41.33	64.08	1.4745	124.67	184.92	1.7188	1.165683282
HpELL1C	HpELL1A	42.67	69.08	1.3007	125.33	179.92	1.982	1.52379488
HpELL1C	HpELI4	44.92	63.42	2.1668	130.08	185.58	2.0453	0.943926528
HpELL1C	HpELL11A	51.92	66.83	n.a.	123.08	182.17	1.7336	n.a.
HpELL13C	HpELL13A	45.75	64.33	2.2201	128.25	184.67	1.9527	0.879554975
HpELL13C	HpELL13B	46.42	67.92	1.8164	107.58	181.08	1.1782	0.648645673
HpELL13C	HpELL13D	51.5	70.08	2.926	120.5	178.92	1.7121	0.585133288
HpELL13C	HpELL1B	47.83	67.83	2.1127	127.17	181.17	2.0606	0.975339613
HpELL13C	HpELL11C	47.25	67.17	2.085	130.75	181.83	2.3911	1.146810552
HpELL13C	HpELL6	47.17	64.58	2.7305	126.83	184.42	1.8667	0.683647684
HpELL13C	HpELL1A	47.58	69.58	1.8209	121.42	179.42	1.7444	0.957987808
HpELL13C	HpELI4	46.25	63.92	2.51	125.75	185.08	1.7725	0.706175299
HpELL13C	HpELL11A	47.58	67.33	2.1386	128.42	181.67	2.1421	1.001636585
HpELL13A	HpELL13B	46.08	64.08	2.3924	129.92	184.92	2.0706	0.865490721
HpELL13A	HpELL13D	46.5	66.25	2.0599	123.5	182.75	1.7348	0.842176805
HpELL13A	HpELL1B	48.25	64	n.a.	126.75	185	1.8358	n.a.
HpELL13A	HpELL11C	44.5	63.33	2.0716	134.5	185.67	2.5336	1.223016026
HpELL13A	HpELL6	41.67	60.75	1.8444	139.33	188.25	3.2495	1.761819562
HpELL13A	HpELL1A	51.5	65.75	n.a.	117.5	183.25	1.4479	n.a.
HpELL13A	HpELI4	42.67	60.08	2.2007	131.33	188.92	1.9622	0.891625392
HpELL13A	HpELL11A	47.92	63.5	n.a.	130.08	185.5	2.0501	n.a.
HpELL13B	HpELL13D	47.58	69.83	1.7937	114.42	179.17	1.4302	0.797346267
HpELL13B	HpELL1B	32.75	67.58	0.7791	123.25	181.42	1.772	2.274419202
HpELL13B	HpELL11C	44.67	66.92	1.6554	125.33	182.08	1.8737	1.131871451
HpELL13B	HpELL6	43.42	64.33	1.7256	129.58	184.67	2.0572	1.192165044
HpELL13B	HpELL1A	50.83	69.33	2.8478	131.17	179.67	2.7203	0.955228598
HpELL13B	HpELI4	38.92	63.67	1.2656	123.08	185.33	1.6253	1.284213021
HpELL13B	HpELL11A	49.33	67.08	2.9545	122.67	181.92	1.72	0.582162803
HpELL13D	HpELL1B	56	69.75	n.a.	121	179.25	1.7273	n.a.
HpELL13D	HpELL11C	44.42	69.08	1.46	130.58	179.92	2.5753	1.76390411
HpELL13D	HpELL6	47.92	66.5	2.4281	122.08	182.5	1.6688	0.687286356
HpELL13D	HpELL1A	50.42	71.5	2.1122	122.58	177.5	1.902	0.900482909
HpELL13D	HpELI4	45.58	65.83	1.925	131.42	183.17	2.3534	1.222545455
HpELL13D	HpELL11A	42.33	69.25	1.2659	115.67	179.75	1.4639	1.156410459
HpELL1B	HpELL11C	47.42	66.83	2.1886	120.58	182.17	1.6065	0.734030887
HpELL1B	HpELL6	46.83	64.25	2.6789	123.17	184.75	1.6479	0.615140543
HpELL1B	HpELL1A	48.5	69.25	2.0365	122.5	179.75	1.795	0.881414191
HpELL1B	HpELI4	37.33	63.58	1.1455	127.67	185.42	1.8763	1.637974684
HpELL1B	HpELL11A	51.08	67	n.a.	129.58	184.67	2.9754	n.a.
HpELL11C	HpELL6	45.42	63.58	2.2834	131.17	179.67	1.3341	0.584260314
HpELL11C	HpELL1A	38.33	68.58	1.0256	123.08	185.33	1.3083	1.275643526
HpELL11C	HpELI4	41.5	62.92	1.5869	122.67	181.92	2.4875	1.567521583
HpELL11C	HpELL11A	49.42	66.33	3.7542	121	179.25	2.7685	0.737440733
HpELL6	HpELL1A	46.75	66	2.1678	130.58	179.92	1.7121	0.789786881
HpELL6	HpELI4	40.17	60.33	1.6397	122.08	182.5	1.9389	1.182472403
HpELL6	HpELL11A	51.92	63.75	n.a.	122.58	177.5	2.0649	n.a.
HpELL1A	HpELI4	42.25	65.33	1.4867	131.42	183.17	2.8701	1.930517253
HpELL1A	HpELL11A	52.83	68.75	n.a.	115.67	179.75	2.1185	n.a.
HpELI4	HpELL11A	58.25	63.08	n.a.	120.58	182.17	1.9264	n.a.

***n.a.** , not applicable. When the proportion of differences is equal or higher than 0.75, the Jukes and Cantor correction cannot be computed.

[a,b]**Seq-1** and **Seq-2**, the two sequences compared.

[e]**SynDif**, the total number of synonymous differences.

[f]**SynPos**, the total number of synonymous sites.

[c]**SilentDif**, the total number of silent differences.

[d]**SilentPos**, the total number of silent sites.

[g]**Pairwise Ka/ks ratio of non-synonymous and synonymous substitutions**

Appendix. E. Scripts showing how the signal peptide positive dataset was obtained

```
#! /usr/bin/perl

use lib "/Users/hrseaw/perl_scripts/modules";
use perl_Modules;

print "Please enter the name of the sequence file: \n" ;
        my $seq_file = <STDIN> ;
         chomp $seq_file ;

# calls and executes the emboss application getorf which retrieves all orfs with a
methionine start and writes to a compulsory file
qx( getorf -options -sequence $seq_file -outseq
/Bioinf/home/hughesl/perl/output/orfs_$seq_file.fasta);

# opens the file of orfs and sends to the subroutine fix sequence
open(FILEHANDLE, "/Bioinf/home/hughesl/perl/output/orfs_$seq_file.fasta" )or die
("cannot open eeeee \n\n");

my @temp = <FILEHANDLE>;
chomp @temp;
my @orfs = perl_Modules->fix_sequence(@temp);

my @trimmed = trim_sequence(@orfs);



my $count;
my $count2;
my $j = 1;

my $D =1;
my $s;
$k= 0;
# splits the seqs into files of no more than 350 seqs each as sigp has limitations on its
entries ie no seq can be more than 6000aa and wont accept more than about 2000
seqs and the total num of aa not more than 200000, i think etc see sigp server for
more info
open (OUTFILE,">>/Bioinf/home/hughesl/perl/output/forSigp_$seq_file$j.fasta");

foreach my $gotone (@trimmed){
        $count .= $gotone;
        $count2 = length ($count);
        my $eek = length ($gotone);


        if ($eek >= 5500){
                print "$gotone\n";
                print "oi thats aint right!!!!!!\n";
                exit;
        }
        $k++;

         if ($count2 >= 100000 || $k == 1000){
                close OUTFILE;
                $j++;
                open
(OUTFILE,">>/Bioinf/home/hughesl/perl/output/forSigp_$seq_file$j.fasta");
                $k =0;
                $count= '';
         }
         if ($eek > 90){

                        $gotone =~ s/>\d*|>/>$D/;
                        print OUTFILE "$gotone\n";
                        $D++;
                }
                $s = $j;
}



# executes the signalp program and outputs to a file



for (my $i= 1; $i<= $s; $i++){

qx( signalp -t euk /Bioinf/home/hughesl/perl/output/forSigp_$seq_file$i.fasta
>/Bioinf/home/hughesl/perl/output/allorf_sigpeps_$seq_file$i.fasta );

}
exit;

########################################################
```

Perl script 2

```
# A perl script which calls the bioperl module signalp.pm, the script accepts a signalp output file as input and must be stated on the command line along with the script name.

#The signalp module is a parser which extracts the significant results from a signalp file by printing all the features of signalp out put as long as the score is over 0.6,



use warnings;

use strict;

use Bio::Tools::Signalp;





my $file = $ARGV[0];

open (OUT,">>$ARGV[0]_parsed.txt");



my $parser = new Bio::Tools::Signalp(-file => $file);



while (my $feat = $parser->next_result) {



        my $name = $feat->seq_id;
        my $start = $feat->start;
        my $end = $feat->end;
        my $score = $feat->score();
        my $tag = $feat->source_tag();


        my ($peptideProb) = $feat->get_tag_values('peptideProb');

        my ($anchorProb) = $feat->get_tag_values('anchorProb');
        my ($evalue) = $feat->get_tag_values('evalue');
        my ($percent_id) = $feat->get_tag_values('percent_id');
        my ($hid) = $feat->get_tag_values('hid');
        my ($SignalpProediction) = $feat->get_tag_values('SignalpProediction');


        if ($score >= 0.6){



        print OUT " name: $name  $start-$end  score:$score  percent_id:$percent_id
evalue:$evalue  peptideProb:$peptideProb  anchorProb:$anchorProb
SignalpProediction:$SignalpProediction\n";





        }

}

close OUT;

exit;
```

Appendix F. The table shows the loop design of the *H. arabidopsidis* inoculated Arabidopsis time course microarray experiment. The aim of the design is to allow a comparison of expression levels between the Cy3 and Cy5 array slides at all time points, biological replicates and the three treatments of the experiment.

Dye	Slide Number	Slide Name	Sample	BioRep	Time	Inoculation
Cy5	109	13541394	1	A	0h	EM
Cy3	53	13541407	1	A	0h	EM
Cy5	30	13541521	1	A	0h	EM
Cy3	31	13541522	1	A	0h	EM
Cy5	108	13541399	2	B	0h	EM
Cy3	42	13541539	2	B	0h	EM
Cy3	177	13541647	2	B	0h	EM
Cy5	99	13586933	2	B	0h	EM
Cy5	57	13541281	3	C	0h	EM
Cy5	76	13541297	3	C	0h	EM
Cy3	79	13541515	3	C	0h	EM
Cy3	174	13587107	3	C	0h	EM
Cy5	27	13541125	4	D	0h	EM
Cy5	62	13541413	4	D	0h	EM
Cy3	184	13541656	4	D	0h	EM
Cy3	135	13586944	4	D	0h	EM
Cy5	8	13537200	33	A	0h	H2O
Cy3	151	13537209	33	A	0h	H2O
Cy5	54	13541408	33	A	0h	H2O
Cy3	67	13541418	33	A	0h	H2O
Cy3	15	13541458	34	B	0h	H2O
Cy5	31	13541522	34	B	0h	H2O
Cy3	22	13541639	34	B	0h	H2O
Cy5	117	13587011	34	B	0h	H2O
Cy5	155	13537213	35	C	0h	H2O
Cy3	55	13541410	35	C	0h	H2O
Cy3	62	13541413	35	C	0h	H2O
Cy5	7	13541662	35	C	0h	H2O
Cy3	162	13530678	36	D	0h	H2O
Cy5	187	13537194	36	D	0h	H2O
Cy3	188	13537195	36	D	0h	H2O
Cy5	94	13541308	36	D	0h	H2O
Cy3	167	13530684	65	A	0h	MA
Cy3	108	13541399	65	A	0h	MA
Cy5	110	13541400	65	A	0h	MA
Cy5	36	13541530	65	A	0h	MA
Cy3	190	13537196	66	B	0h	MA
Cy5	151	13537209	66	B	0h	MA
Cy3	51	13541405	66	B	0h	MA
Cy5	39	13541535	66	B	0h	MA
Cy3	94	13541308	67	C	0h	MA
Cy3	106	13541311	67	C	0h	MA
Cy5	13	13541456	67	C	0h	MA
Cy5	48	13541545	67	C	0h	MA
Cy5	137	13530671	68	D	0h	MA

Cy3	68	13541419	68	D	0h	MA
Cy3	134	13586923	68	D	0h	MA
Cy5	174	13587107	68	D	0h	MA
Cy5	162	13530678	5	A	8h	EM
Cy3	192	13537199	5	A	8h	EM
Cy5	53	13541407	5	A	8h	EM
Cy3	116	13586626	5	A	8h	EM
Cy5	157	13541495	6	B	8h	EM
Cy5	42	13541539	6	B	8h	EM
Cy3	125	13586503	6	B	8h	EM
Cy3	91	13586927	6	B	8h	EM
Cy3	152	13537210	7	C	8h	EM
Cy5	51	13541405	7	C	8h	EM
Cy5	79	13541515	7	C	8h	EM
Cy3	124	13586499	7	C	8h	EM
Cy3	77	13541298	8	D	8h	EM
Cy3	63	13541414	8	D	8h	EM
Cy5	135	13586944	8	D	8h	EM
Cy5	121	13586947	8	D	8h	EM
Cy5	67	13541418	37	A	8h	H2O
Cy3	157	13541495	37	A	8h	H2O
Cy3	34	13541527	37	A	8h	H2O
Cy5	134	13586923	37	A	8h	H2O
Cy5	139	13530673	38	B	8h	H2O
Cy3	153	13537211	38	B	8h	H2O
Cy3	50	13541404	38	B	8h	H2O
Cy5	15	13541458	38	B	8h	H2O
Cy3	86	13541303	39	C	8h	H2O
Cy5	55	13541410	39	C	8h	H2O
Cy3	64	13541422	39	C	8h	H2O
Cy5	177	13541647	39	C	8h	H2O
Cy5	188	13537195	40	D	8h	H2O
Cy3	144	13537214	40	D	8h	H2O
Cy3	32	13541525	40	D	8h	H2O
Cy5	124	13586499	40	D	8h	H2O
Cy3	139	13530673	69	A	8h	MA
Cy3	140	13530674	69	A	8h	MA
Cy5	167	13530684	69	A	8h	MA
Cy5	184	13541656	69	A	8h	MA
Cy5	190	13537196	70	B	8h	MA
Cy5	192	13537199	70	B	8h	MA
Cy3	72	13541520	70	B	8h	MA
Cy3	118	13586627	70	B	8h	MA
Cy5	106	13541311	71	C	8h	MA
Cy3	18	13541453	71	C	8h	MA
Cy5	22	13541639	71	C	8h	MA
Cy3	121	13586947	71	C	8h	MA
Cy3	165	13530682	72	D	8h	MA
Cy3	74	13541295	72	D	8h	MA
Cy5	68	13541419	72	D	8h	MA
Cy5	64	13541422	72	D	8h	MA
Cy5	165	13530682	9	A	16h	EM
Cy3	25	13541119	9	A	16h	EM
Cy3	171	13541652	9	A	16h	EM
Cy5	116	13586626	9	A	16h	EM
Cy3	93	13541306	10	B	16h	EM
Cy3	132	13541505	10	B	16h	EM
Cy5	182	13541654	10	B	16h	EM
Cy5	125	13586503	10	B	16h	EM

Cy5	152	13537210	11	C	16h	EM
Cy5	50	13541404	11	C	16h	EM
Cy3	96	13586930	11	C	16h	EM
Cy3	122	13586948	11	C	16h	EM
Cy3	150	13537208	12	D	16h	EM
Cy3	3	13541113	12	D	16h	EM
Cy5	77	13541298	12	D	16h	EM
Cy5	38	13541534	12	D	16h	EM
Cy3	65	13537202	41	A	16h	H2O
Cy5	63	13541414	41	A	16h	H2O
Cy5	34	13541527	41	A	16h	H2O
Cy3	43	13541540	41	A	16h	H2O
Cy5	153	13537211	42	B	16h	H2O
Cy3	145	13537215	42	B	16h	H2O
Cy5	25	13541119	42	B	16h	H2O
Cy3	166	13587115	42	B	16h	H2O
Cy5	86	13541303	43	C	16h	H2O
Cy3	159	13541499	43	C	16h	H2O
Cy3	38	13541534	43	C	16h	H2O
Cy5	118	13586627	43	C	16h	H2O
Cy5	144	13537214	44	D	16h	H2O
Cy5	146	13537216	44	D	16h	H2O
Cy3	90	13586619	44	D	16h	H2O
Cy3	100	13586620	44	D	16h	H2O
Cy5	140	13530674	73	A	16h	MA
Cy3	26	13541124	73	A	16h	MA
Cy5	32	13541525	73	A	16h	MA
Cy3	182	13541654	73	A	16h	MA
Cy3	75	13541296	74	B	16h	MA
Cy5	72	13541520	74	B	16h	MA
Cy5	43	13541540	74	B	16h	MA
Cy3	21	13541637	74	B	16h	MA
Cy3	146	13537216	75	C	16h	MA
Cy5	18	13541453	75	C	16h	MA
Cy5	91	13586927	75	C	16h	MA
Cy3	114	13587010	75	C	16h	MA
Cy3	107	13537307	76	D	16h	MA
Cy5	74	13541295	76	D	16h	MA
Cy3	170	13541501	76	D	16h	MA
Cy5	122	13586948	76	D	16h	MA
Cy3	37	13541532	13	A	24h	EM
Cy3	23	13541640	13	A	24h	EM
Cy5	171	13541652	13	A	24h	EM
Cy5	90	13586619	13	A	24h	EM
Cy3	138	13530672	14	B	24h	EM
Cy5	93	13541306	14	B	24h	EM
Cy3	56	13541537	14	B	24h	EM
Cy5	180	13541651	14	B	24h	EM
Cy3	186	13537193	15	C	24h	EM
Cy3	20	13541636	15	C	24h	EM
Cy5	21	13541637	15	C	24h	EM
Cy5	96	13586930	15	C	24h	EM
Cy5	150	13537208	16	D	24h	EM
Cy3	87	13541304	16	D	24h	EM
Cy5	101	13541310	16	D	24h	EM
Cy3	120	13586946	16	D	24h	EM
Cy5	65	13537202	45	A	24h	H2O
Cy5	107	13537307	45	A	24h	H2O
Cy3	176	13541646	45	A	24h	H2O
Cy3	180	13541651	45	A	24h	H2O

Cy5	145	13537215	46	B	24h	H2O
Cy3	17	13541452	46	B	24h	H2O
Cy5	46	13541543	46	B	24h	H2O
Cy3	189	13587111	46	B	24h	H2O
Cy3	142	13537203	47	C	24h	H2O
Cy3	60	13541284	47	C	24h	H2O
Cy5	159	13541499	47	C	24h	H2O
Cy5	132	13541505	47	C	24h	H2O
Cy3	97	13541309	48	D	24h	H2O
Cy3	158	13541498	48	D	24h	H2O
Cy5	20	13541636	48	D	24h	H2O
Cy5	100	13586620	48	D	24h	H2O
Cy5	3	13541113	77	A	24h	MA
Cy5	26	13541124	77	A	24h	MA
Cy3	46	13541543	77	A	24h	MA
Cy3	179	13541650	77	A	24h	MA
Cy5	75	13541296	78	B	24h	MA
Cy3	45	13541542	78	B	24h	MA
Cy5	23	13541640	78	B	24h	MA
Cy3	115	13586624	78	B	24h	MA
Cy3	101	13541310	79	C	24h	MA
Cy3	126	13541487	79	C	24h	MA
Cy5	114	13587010	79	C	24h	MA
Cy5	60	13541284	80	D	24h	MA
Cy5	170	13541501	80	D	24h	MA
Cy3	80	13541516	80	D	24h	MA
Cy3	47	13541544	80	D	24h	MA
Cy3	4	13541114	17	A	32h	EM
Cy5	80	13541516	17	A	32h	EM
Cy3	35	13541528	17	A	32h	EM
Cy5	37	13541532	17	A	32h	EM
Cy3	88	13541305	18	B	32h	EM
Cy3	61	13541412	18	B	32h	EM
Cy5	56	13541537	18	B	32h	EM
Cy5	84	13586925	18	B	32h	EM
Cy5	186	13537193	19	C	32h	EM
Cy5	17	13541452	19	C	32h	EM
Cy3	19	13587008	19	C	32h	EM
Cy3	6	13541661	19	C	32h	EM
Cy3	149	13537207	20	D	32h	EM
Cy5	73	13541293	20	D	32h	EM
Cy5	87	13541304	20	D	32h	EM
Cy3	2	13541660	20	D	32h	EM
Cy3	154	13537212	49	A	32h	H2O
Cy3	85	13541302	49	A	32h	H2O
Cy5	176	13541646	49	A	32h	H2O
Cy5	120	13586946	49	A	32h	H2O
Cy3	136	13530669	50	B	32h	H2O
Cy5	4	13541114	50	B	32h	H2O
Cy3	41	13541538	50	B	32h	H2O
Cy5	189	13587111	50	B	32h	H2O
Cy5	142	13537203	51	C	32h	H2O
Cy3	73	13541293	51	C	32h	H2O
Cy3	131	13541504	51	C	32h	H2O
Cy5	45	13541542	51	C	32h	H2O
Cy3	163	13530679	52	D	32h	H2O
Cy5	156	13541490	52	D	32h	H2O
Cy5	158	13541498	52	D	32h	H2O
Cy3	185	13541657	52	D	32h	H2O
Cy5	97	13541309	81	A	32h	MA

Cy3	24	13541641	81	A	32h	MA
Cy5	179	13541650	81	A	32h	MA
Cy3	84	13586925	81	A	32h	MA
Cy3	168	13530685	82	B	32h	MA
Cy3	191	13537198	82	B	32h	MA
Cy5	154	13537212	82	B	32h	MA
Cy5	115	13586624	82	B	32h	MA
Cy5	138	13530672	83	C	32h	MA
Cy3	113	13541416	83	C	32h	MA
Cy5	126	13541487	83	C	32h	MA
Cy3	156	13541490	83	C	32h	MA
Cy3	82	13541292	84	D	32h	MA
Cy5	47	13541544	84	D	32h	MA
Cy5	19	13587008	84	D	32h	MA
Cy3	89	13586926	84	D	32h	MA
Cy3	103	13541396	21	A	40h	EM
Cy5	35	13541528	21	A	40h	EM
Cy5	185	13541657	21	A	40h	EM
Cy3	181	13587109	21	A	40h	EM
Cy5	88	13541305	22	B	40h	EM
Cy5	49	13541403	22	B	40h	EM
Cy3	133	13541508	22	B	40h	EM
Cy3	178	13541649	22	B	40h	EM
Cy5	168	13530685	23	C	40h	EM
Cy3	111	13541401	23	C	40h	EM
Cy5	6	13541661	23	C	40h	EM
Cy3	92	13586928	23	C	40h	EM
Cy5	10	13541134	24	D	40h	EM
Cy3	12	13541455	24	D	40h	EM
Cy3	44	13541541	24	D	40h	EM
Cy5	2	13541660	24	D	40h	EM
Cy3	5	13541115	53	A	40h	H2O
Cy5	85	13541302	53	A	40h	H2O
Cy3	49	13541403	53	A	40h	H2O
Cy5	89	13586926	53	A	40h	H2O
Cy5	136	13530669	54	B	40h	H2O
Cy3	59	13541283	54	B	40h	H2O
Cy5	66	13541417	54	B	40h	H2O
Cy3	123	13586949	54	B	40h	H2O
Cy3	104	13541397	55	C	40h	H2O
Cy5	61	13541412	55	C	40h	H2O
Cy5	131	13541504	55	C	40h	H2O
Cy3	98	13586931	55	C	40h	H2O
Cy5	163	13530679	56	D	40h	H2O
Cy3	102	13541315	56	D	40h	H2O
Cy5	111	13541401	56	D	40h	H2O
Cy3	16	13541459	56	D	40h	H2O
Cy5	149	13537207	85	A	40h	MA
Cy3	66	13541417	85	A	40h	MA
Cy3	130	13541502	85	A	40h	MA
Cy5	24	13541641	85	A	40h	MA
Cy5	191	13537198	86	B	40h	MA
Cy3	29	13541127	86	B	40h	MA
Cy3	119	13586629	86	B	40h	MA
Cy5	181	13587109	86	B	40h	MA
Cy3	10	13541134	87	C	40h	MA
Cy3	52	13541406	87	C	40h	MA
Cy5	113	13541416	87	C	40h	MA
Cy5	41	13541538	87	C	40h	MA
Cy5	82	13541292	88	D	40h	MA

Cy5	104	13541397	88	D	40h	MA
Cy3	33	13541526	88	D	40h	MA
Cy3	95	13586929	88	D	40h	MA
Cy3	11	13537201	25	A	48h	EM
Cy5	103	13541396	25	A	48h	EM
Cy3	69	13541420	25	A	48h	EM
Cy5	33	13541526	25	A	48h	EM
Cy3	9	13541122	26	B	48h	EM
Cy5	78	13541301	26	B	48h	EM
Cy5	133	13541508	26	B	48h	EM
Cy3	175	13541645	26	B	48h	EM
Cy3	161	13530677	27	C	48h	EM
Cy5	59	13541283	27	C	48h	EM
Cy3	105	13541398	27	C	48h	EM
Cy5	92	13586928	27	C	48h	EM
Cy3	169	13530686	28	D	48h	EM
Cy5	44	13541541	28	D	48h	EM
Cy5	173	13541643	28	D	48h	EM
Cy3	1	13541658	28	D	48h	EM
Cy3	160	13530676	57	A	48h	H2O
Cy5	5	13541115	57	A	48h	H2O
Cy3	81	13541291	57	A	48h	H2O
Cy5	12	13541455	57	A	48h	H2O
Cy3	28	13541126	58	B	48h	H2O
Cy3	83	13541294	58	B	48h	H2O
Cy5	69	13541420	58	B	48h	H2O
Cy5	123	13586949	58	B	48h	H2O
Cy3	129	13541500	59	C	48h	H2O
Cy3	173	13541643	59	C	48h	H2O
Cy5	119	13586629	59	C	48h	H2O
Cy5	98	13586931	59	C	48h	H2O
Cy3	164	13530680	60	D	48h	H2O
Cy5	70	13541421	60	D	48h	H2O
Cy5	16	13541459	60	D	48h	H2O
Cy3	40	13541536	60	D	48h	H2O
Cy3	78	13541301	89	A	48h	MA
Cy5	102	13541315	89	A	48h	MA
Cy3	127	13541488	89	A	48h	MA
Cy5	130	13541502	89	A	48h	MA
Cy3	148	13537205	90	B	48h	MA
Cy5	29	13541127	90	B	48h	MA
Cy5	81	13541291	90	B	48h	MA
Cy3	128	13541489	90	B	48h	MA
Cy5	52	13541406	91	C	48h	MA
Cy3	70	13541421	91	C	48h	MA
Cy5	178	13541649	91	C	48h	MA
Cy3	147	13586504	91	C	48h	MA
Cy3	141	13530675	92	D	48h	MA
Cy5	161	13530677	92	D	48h	MA
Cy3	143	13537204	92	D	48h	MA
Cy5	95	13586929	92	D	48h	MA
Cy5	164	13530680	29	A	56h	EM
Cy5	11	13537201	29	A	56h	EM
Cy3	58	13541282	29	A	56h	EM
Cy3	109	13541394	29	A	56h	EM
Cy5	9	13541122	30	B	56h	EM
Cy5	71	13541519	30	B	56h	EM
Cy3	48	13541545	30	B	56h	EM
Cy3	99	13586933	30	B	56h	EM
Cy5	148	13537205	31	C	56h	EM

Cy3	76	13541297	31	C	56h	EM
Cy5	105	13541398	31	C	56h	EM
Cy3	172	13541642	31	C	56h	EM
Cy5	169	13530686	32	D	56h	EM
Cy3	27	13541125	32	D	56h	EM
Cy3	54	13541408	32	D	56h	EM
Cy5	183	13541655	32	D	56h	EM
Cy5	160	13530676	61	A	56h	H2O
Cy3	8	13537200	61	A	56h	H2O
Cy5	143	13537204	61	A	56h	H2O
Cy3	71	13541519	61	A	56h	H2O
Cy3	57	13541281	62	B	56h	H2O
Cy5	83	13541294	62	B	56h	H2O
Cy5	112	13541395	62	B	56h	H2O
Cy3	117	13587011	62	B	56h	H2O
Cy3	155	13537213	63	C	56h	H2O
Cy3	14	13541457	63	C	56h	H2O
Cy5	129	13541500	63	C	56h	H2O
Cy5	175	13541645	63	C	56h	H2O
Cy3	187	13537194	64	D	56h	H2O
Cy3	110	13541400	64	D	56h	H2O
Cy5	40	13541536	64	D	56h	H2O
Cy5	172	13541642	64	D	56h	H2O
Cy3	112	13541395	93	A	56h	MA
Cy5	127	13541488	93	A	56h	MA
Cy3	36	13541530	93	A	56h	MA
Cy5	1	13541658	93	A	56h	MA
Cy5	58	13541282	94	B	56h	MA
Cy5	128	13541489	94	B	56h	MA
Cy3	39	13541535	94	B	56h	MA
Cy3	7	13541662	94	B	56h	MA
Cy5	28	13541126	95	C	56h	MA
Cy3	13	13541456	95	C	56h	MA
Cy3	183	13541655	95	C	56h	MA
Cy5	147	13586504	95	C	56h	MA
Cy3	137	13530671	96	D	56h	MA
Cy5	141	13530675	96	D	56h	MA
Cy5	14	13541457	96	D	56h	MA
Cy3	30	13541521	96	D	56h	MA

Appendix G. The table shows the loop design of the Arabidopsis long day developmental time course microarray experiment. The aim of the design is to allow a comparison of expression levels between the Cy3 and Cy5 array slides at all time points of the experiment and between all biological replicates of the experiment.

Array	**Dye**	**Sample**	**Time Point**	**Day**	**ToD***	**BioRep**
25	Cy3	1	01 am	01	am	A
36	Cy3	1	01 am	01	am	A
85	Cy5	1	01 am	01	am	A
123	Cy5	1	01 am	01	am	A
29	Cy3	2	01 am	01	am	B
41	Cy5	2	01 am	01	am	B
53	Cy5	2	01 am	01	am	B
66	Cy3	2	01 am	01	am	B
7	Cy5	3	01 am	01	am	C
79	Cy3	3	01 am	01	am	C
130	Cy3	3	01 am	01	am	C
137	Cy5	3	01 am	01	am	C
13	Cy3	4	01 am	01	am	D
23	Cy3	4	01 am	01	am	D
52	Cy5	4	01 am	01	am	D
77	Cy5	4	01 am	01	am	D
42	Cy5	5	01 pm	01	pm	A
50	Cy5	5	01 pm	01	pm	A
94	Cy3	5	01 pm	01	pm	A
152	Cy3	5	01 pm	01	pm	A
3	Cy5	6	01 pm	01	pm	B
113	Cy5	6	01 pm	01	pm	B
126	Cy3	6	01 pm	01	pm	B
157	Cy3	6	01 pm	01	pm	B
49	Cy5	7	01 pm	01	pm	C
51	Cy5	7	01 pm	01	pm	C
95	Cy3	7	01 pm	01	pm	C
135	Cy3	7	01 pm	01	pm	C
40	Cy5	8	01 pm	01	pm	D
97	Cy5	8	01 pm	01	pm	D
103	Cy3	8	01 pm	01	pm	D
149	Cy3	8	01 pm	01	pm	D
16	Cy3	9	02 am	02	am	A
36	Cy5	9	02 am	02	am	A
135	Cy5	9	02 am	02	am	A
141	Cy3	9	02 am	02	am	A
128	Cy3	10	02 am	02	am	B
130	Cy5	10	02 am	02	am	B
147	Cy3	10	02 am	02	am	B
149	Cy5	10	02 am	02	am	B
29	Cy5	11	02 am	02	am	C
56	Cy3	11	02 am	02	am	C
94	Cy5	11	02 am	02	am	C
171	Cy3	11	02 am	02	am	C
13	Cy5	12	02 am	02	am	D

15	Cy3	12	02 am	02	am	D
68	Cy3	12	02 am	02	am	D
126	Cy5	12	02 am	02	am	D
23	Cy5	13	02 pm	02	pm	A
47	Cy3	13	02 pm	02	pm	A
55	Cy3	13	02 pm	02	pm	A
152	Cy5	13	02 pm	02	pm	A
25	Cy5	14	02 pm	02	pm	B
26	Cy3	14	02 pm	02	pm	B
157	Cy5	14	02 pm	02	pm	B
161	Cy3	14	02 pm	02	pm	B
5	Cy3	15	02 pm	02	pm	C
11	Cy3	15	02 pm	02	pm	C
79	Cy5	15	02 pm	02	pm	C
95	Cy5	15	02 pm	02	pm	C
66	Cy5	16	02 pm	02	pm	D
103	Cy5	16	02 pm	02	pm	D
118	Cy3	16	02 pm	02	pm	D
138	Cy3	16	02 pm	02	pm	D
47	Cy5	17	03 am	03	am	A
56	Cy5	17	03 am	03	am	A
71	Cy3	17	03 am	03	am	A
136	Cy3	17	03 am	03	am	A
5	Cy5	18	03 am	03	am	B
115	Cy3	18	03 am	03	am	B
141	Cy5	18	03 am	03	am	B
155	Cy3	18	03 am	03	am	B
104	Cy3	19	03 am	03	am	C
138	Cy5	19	03 am	03	am	C
144	Cy3	19	03 am	03	am	C
147	Cy5	19	03 am	03	am	C
31	Cy3	20	03 am	03	am	D
68	Cy5	20	03 am	03	am	D
108	Cy3	20	03 am	03	am	D
161	Cy5	20	03 am	03	am	D
15	Cy5	21	03 pm	03	pm	A
55	Cy5	21	03 pm	03	pm	A
102	Cy3	21	03 pm	03	pm	A
107	Cy3	21	03 pm	03	pm	A
16	Cy5	22	03 pm	03	pm	B
26	Cy5	22	03 pm	03	pm	B
93	Cy3	22	03 pm	03	pm	B
168	Cy3	22	03 pm	03	pm	B
11	Cy5	23	03 pm	03	pm	C
46	Cy3	23	03 pm	03	pm	C
128	Cy5	23	03 pm	03	pm	C
151	Cy3	23	03 pm	03	pm	C
116	Cy3	24	03 pm	03	pm	D
118	Cy5	24	03 pm	03	pm	D
150	Cy3	24	03 pm	03	pm	D
171	Cy5	24	03 pm	03	pm	D
6	Cy3	25	04 am	04	am	A
12	Cy3	25	04 am	04	am	A
46	Cy5	25	04 am	04	am	A
115	Cy5	25	04 am	04	am	A
17	Cy3	26	04 am	04	am	B
104	Cy5	26	04 am	04	am	B
116	Cy5	26	04 am	04	am	B
145	Cy3	26	04 am	04	am	B
71	Cy5	27	04 am	04	am	C

91	Cy3	27	04 am	04	am	C
102	Cy5	27	04 am	04	am	C
172	Cy3	27	04 am	04	am	C
31	Cy5	28	04 am	04	am	D
44	Cy3	28	04 am	04	am	D
92	Cy3	28	04 am	04	am	D
93	Cy5	28	04 am	04	am	D
65	Cy3	29	04 pm	04	pm	A
67	Cy3	29	04 pm	04	pm	A
107	Cy5	29	04 pm	04	pm	A
108	Cy5	29	04 pm	04	pm	A
112	Cy3	30	04 pm	04	pm	B
114	Cy3	30	04 pm	04	pm	B
155	Cy5	30	04 pm	04	pm	B
168	Cy5	30	04 pm	04	pm	B
38	Cy3	31	04 pm	04	pm	C
144	Cy5	31	04 pm	04	pm	C
151	Cy5	31	04 pm	04	pm	C
170	Cy3	31	04 pm	04	pm	C
35	Cy3	32	04 pm	04	pm	D
64	Cy3	32	04 pm	04	pm	D
136	Cy5	32	04 pm	04	pm	D
150	Cy5	32	04 pm	04	pm	D
4	Cy3	33	05 am	05	am	A
12	Cy5	33	05 am	05	am	A
38	Cy5	33	05 am	05	am	A
43	Cy3	33	05 am	05	am	A
28	Cy3	34	05 am	05	am	B
35	Cy5	34	05 am	05	am	B
58	Cy3	34	05 am	05	am	B
145	Cy5	34	05 am	05	am	B
65	Cy5	35	05 am	05	am	C
91	Cy5	35	05 am	05	am	C
146	Cy3	35	05 am	05	am	C
154	Cy3	35	05 am	05	am	C
37	Cy3	36	05 am	05	am	D
48	Cy3	36	05 am	05	am	D
92	Cy5	36	05 am	05	am	D
114	Cy5	36	05 am	05	am	D
44	Cy5	37	05 pm	05	pm	A
67	Cy5	37	05 pm	05	pm	A
142	Cy3	37	05 pm	05	pm	A
175	Cy3	37	05 pm	05	pm	A
6	Cy5	38	05 pm	05	pm	B
112	Cy5	38	05 pm	05	pm	B
117	Cy3	38	05 pm	05	pm	B
131	Cy3	38	05 pm	05	pm	B
10	Cy3	39	05 pm	05	pm	C
17	Cy5	39	05 pm	05	pm	C
121	Cy3	39	05 pm	05	pm	C
170	Cy5	39	05 pm	05	pm	C
45	Cy3	40	05 pm	05	pm	D
64	Cy5	40	05 pm	05	pm	D
164	Cy3	40	05 pm	05	pm	D
172	Cy5	40	05 pm	05	pm	D
4	Cy5	41	06 am	06	am	A
10	Cy5	41	06 am	06	am	A
69	Cy3	41	06 am	06	am	A
96	Cy3	41	06 am	06	am	A
14	Cy3	42	06 am	06	am	B

22	Cy3	42	06 am	06	am	B
28	Cy5	42	06 am	06	am	B
45	Cy5	42	06 am	06	am	B
154	Cy5	43	06 am	06	am	C
160	Cy3	43	06 am	06	am	C
163	Cy3	43	06 am	06	am	C
175	Cy5	43	06 am	06	am	C
18	Cy3	44	06 am	06	am	D
48	Cy5	44	06 am	06	am	D
117	Cy5	44	06 am	06	am	D
167	Cy3	44	06 am	06	am	D
27	Cy3	45	06 pm	06	pm	A
37	Cy5	45	06 pm	06	pm	A
129	Cy3	45	06 pm	06	pm	A
142	Cy5	45	06 pm	06	pm	A
43	Cy5	46	06 pm	06	pm	B
62	Cy3	46	06 pm	06	pm	B
131	Cy5	46	06 pm	06	pm	B
140	Cy3	46	06 pm	06	pm	B
1	Cy3	47	06 pm	06	pm	C
58	Cy5	47	06 pm	06	pm	C
83	Cy3	47	06 pm	06	pm	C
121	Cy5	47	06 pm	06	pm	C
57	Cy3	48	06 pm	06	pm	D
134	Cy3	48	06 pm	06	pm	D
146	Cy5	48	06 pm	06	pm	D
164	Cy5	48	06 pm	06	pm	D
1	Cy5	49	07 am	07	am	A
30	Cy3	49	07 am	07	am	A
69	Cy5	49	07 am	07	am	A
70	Cy3	49	07 am	07	am	A
14	Cy5	50	07 am	07	am	B
125	Cy3	50	07 am	07	am	B
134	Cy5	50	07 am	07	am	B
158	Cy3	50	07 am	07	am	B
54	Cy3	51	07 am	07	am	C
63	Cy3	51	07 am	07	am	C
129	Cy5	51	07 am	07	am	C
160	Cy5	51	07 am	07	am	C
18	Cy5	52	07 am	07	am	D
62	Cy5	52	07 am	07	am	D
100	Cy3	52	07 am	07	am	D
166	Cy3	52	07 am	07	am	D
24	Cy3	53	07 pm	07	pm	A
27	Cy5	53	07 pm	07	pm	A
73	Cy3	53	07 pm	07	pm	A
167	Cy5	53	07 pm	07	pm	A
96	Cy5	54	07 pm	07	pm	B
140	Cy5	54	07 pm	07	pm	B
148	Cy3	54	07 pm	07	pm	B
165	Cy3	54	07 pm	07	pm	B
22	Cy5	55	07 pm	07	pm	C
83	Cy5	55	07 pm	07	pm	C
99	Cy3	55	07 pm	07	pm	C
156	Cy3	55	07 pm	07	pm	C
57	Cy5	56	07 pm	07	pm	D
153	Cy3	56	07 pm	07	pm	D
162	Cy3	56	07 pm	07	pm	D
163	Cy5	56	07 pm	07	pm	D
34	Cy3	57	08 am	08	am	A

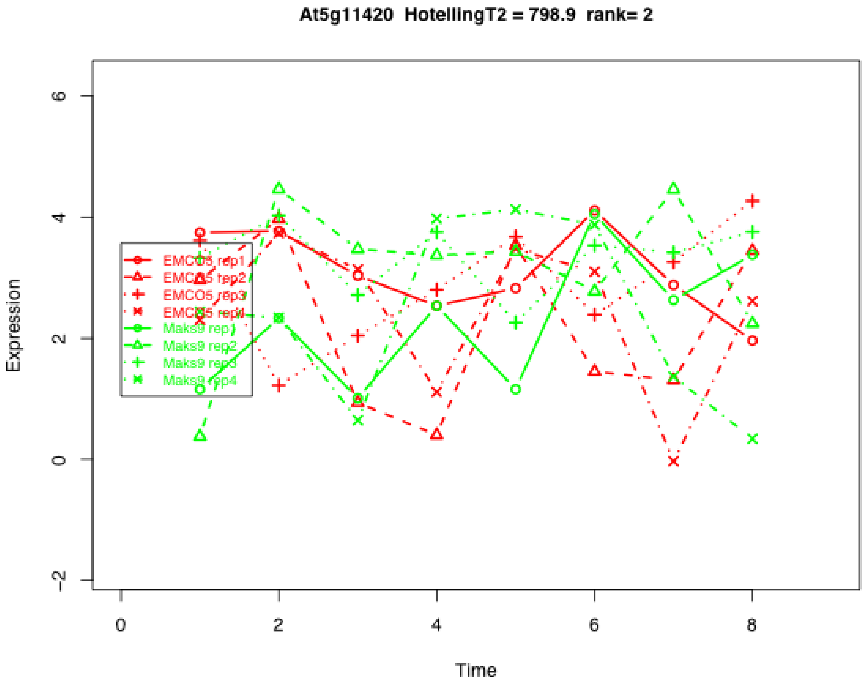
165	Cy5	57	08 am	08	am	A
166	Cy5	57	08 am	08	am	A
173	Cy3	57	08 am	08	am	A
70	Cy5	58	08 am	08	am	B
99	Cy5	58	08 am	08	am	B
133	Cy3	58	08 am	08	am	B
159	Cy3	58	08 am	08	am	B
20	Cy3	59	08 am	08	am	C
125	Cy5	59	08 am	08	am	C
153	Cy5	59	08 am	08	am	C
174	Cy3	59	08 am	08	am	C
33	Cy3	60	08 am	08	am	D
54	Cy5	60	08 am	08	am	D
73	Cy5	60	08 am	08	am	D
124	Cy3	60	08 am	08	am	D
60	Cy3	61	08 pm	08	pm	A
63	Cy5	61	08 pm	08	pm	A
109	Cy3	61	08 pm	08	pm	A
162	Cy5	61	08 pm	08	pm	A
24	Cy5	62	08 pm	08	pm	B
59	Cy3	62	08 pm	08	pm	B
61	Cy3	62	08 pm	08	pm	B
100	Cy5	62	08 pm	08	pm	B
19	Cy3	63	08 pm	08	pm	C
30	Cy5	63	08 pm	08	pm	C
39	Cy3	63	08 pm	08	pm	C
148	Cy5	63	08 pm	08	pm	C
74	Cy3	64	08 pm	08	pm	D
139	Cy3	64	08 pm	08	pm	D
156	Cy5	64	08 pm	08	pm	D
158	Cy5	64	08 pm	08	pm	D
76	Cy3	65	09 am	09	am	A
101	Cy3	65	09 am	09	am	A
139	Cy5	65	09 am	09	am	A
159	Cy5	65	09 am	09	am	A
34	Cy5	66	09 am	09	am	B
39	Cy5	66	09 am	09	am	B
98	Cy3	66	09 am	09	am	B
143	Cy3	66	09 am	09	am	B
60	Cy5	67	09 am	09	am	C
72	Cy3	67	09 am	09	am	C
169	Cy3	67	09 am	09	am	C
174	Cy5	67	09 am	09	am	C
8	Cy3	68	09 am	09	am	D
33	Cy5	68	09 am	09	am	D
61	Cy5	68	09 am	09	am	D
87	Cy3	68	09 am	09	am	D
59	Cy5	69	09 pm	09	pm	A
82	Cy3	69	09 pm	09	pm	A
120	Cy3	69	09 pm	09	pm	A
173	Cy5	69	09 pm	09	pm	A
106	Cy3	70	09 pm	09	pm	B
109	Cy5	70	09 pm	09	pm	B
119	Cy3	70	09 pm	09	pm	B
124	Cy5	70	09 pm	09	pm	B
19	Cy5	71	09 pm	09	pm	C
90	Cy3	71	09 pm	09	pm	C
105	Cy3	71	09 pm	09	pm	C
133	Cy5	71	09 pm	09	pm	C
20	Cy5	72	09 pm	09	pm	D

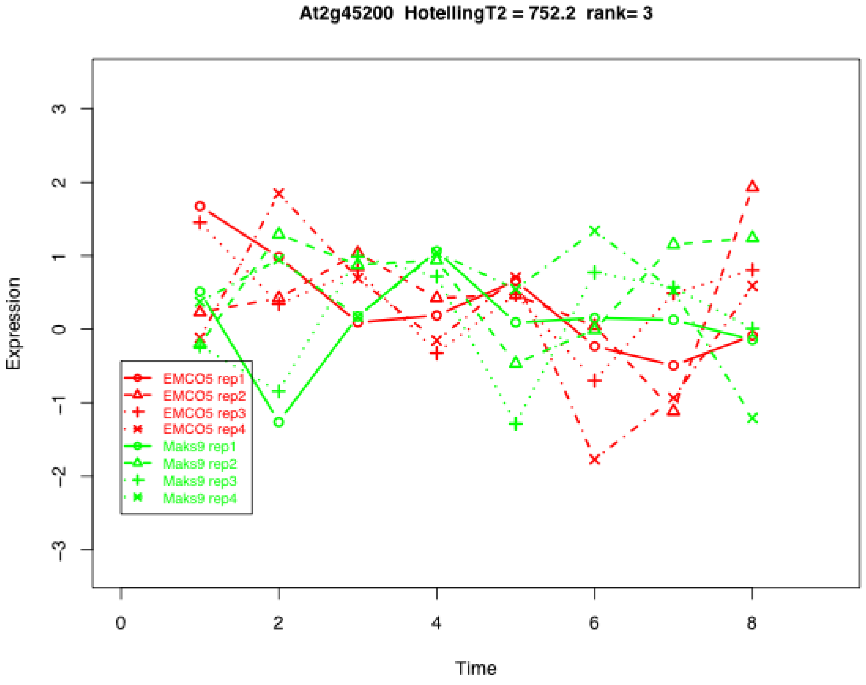
21	Cy3	72	09 pm	09	pm	D
74	Cy5	72	09 pm	09	pm	D
80	Cy3	72	09 pm	09	pm	D
8	Cy5	73	10 am	10	am	A
111	Cy3	73	10 am	10	am	A
120	Cy5	73	10 am	10	am	A
176	Cy3	73	10 am	10	am	A
21	Cy5	74	10 am	10	am	B
76	Cy5	74	10 am	10	am	B
110	Cy3	74	10 am	10	am	B
127	Cy3	74	10 am	10	am	B
9	Cy3	75	10 am	10	am	C
84	Cy3	75	10 am	10	am	C
106	Cy5	75	10 am	10	am	C
169	Cy5	75	10 am	10	am	C
78	Cy3	76	10 am	10	am	D
88	Cy3	76	10 am	10	am	D
105	Cy5	76	10 am	10	am	D
143	Cy5	76	10 am	10	am	D
2	Cy3	77	10 pm	10	pm	A
82	Cy5	77	10 pm	10	pm	A
98	Cy5	77	10 pm	10	pm	A
122	Cy3	77	10 pm	10	pm	A
75	Cy3	78	10 pm	10	pm	B
81	Cy3	78	10 pm	10	pm	B
90	Cy5	78	10 pm	10	pm	B
101	Cy5	78	10 pm	10	pm	B
72	Cy5	79	10 pm	10	pm	C
80	Cy5	79	10 pm	10	pm	C
86	Cy3	79	10 pm	10	pm	C
132	Cy3	79	10 pm	10	pm	C
32	Cy3	80	10 pm	10	pm	D
87	Cy5	80	10 pm	10	pm	D
89	Cy3	80	10 pm	10	pm	D
119	Cy5	80	10 pm	10	pm	D
3	Cy3	81	11 am	11	am	A
86	Cy5	81	11 am	11	am	A
110	Cy5	81	11 am	11	am	A
123	Cy3	81	11 am	11	am	A
9	Cy5	82	11 am	11	am	B
32	Cy5	82	11 am	11	am	B
49	Cy3	82	11 am	11	am	B
137	Cy3	82	11 am	11	am	B
2	Cy5	83	11 am	11	am	C
53	Cy3	83	11 am	11	am	C
97	Cy3	83	11 am	11	am	C
111	Cy5	83	11 am	11	am	C
50	Cy3	84	11 am	11	am	D
52	Cy3	84	11 am	11	am	D
75	Cy5	84	11 am	11	am	D
88	Cy5	84	11 am	11	am	D
41	Cy3	85	11 pm	11	pm	A
42	Cy3	85	11 pm	11	pm	A
78	Cy5	85	11 pm	11	pm	A
122	Cy5	85	11 pm	11	pm	A
77	Cy3	86	11 pm	11	pm	B
81	Cy5	86	11 pm	11	pm	B
113	Cy3	86	11 pm	11	pm	B
127	Cy5	86	11 pm	11	pm	B
51	Cy3	87	11 pm	11	pm	C

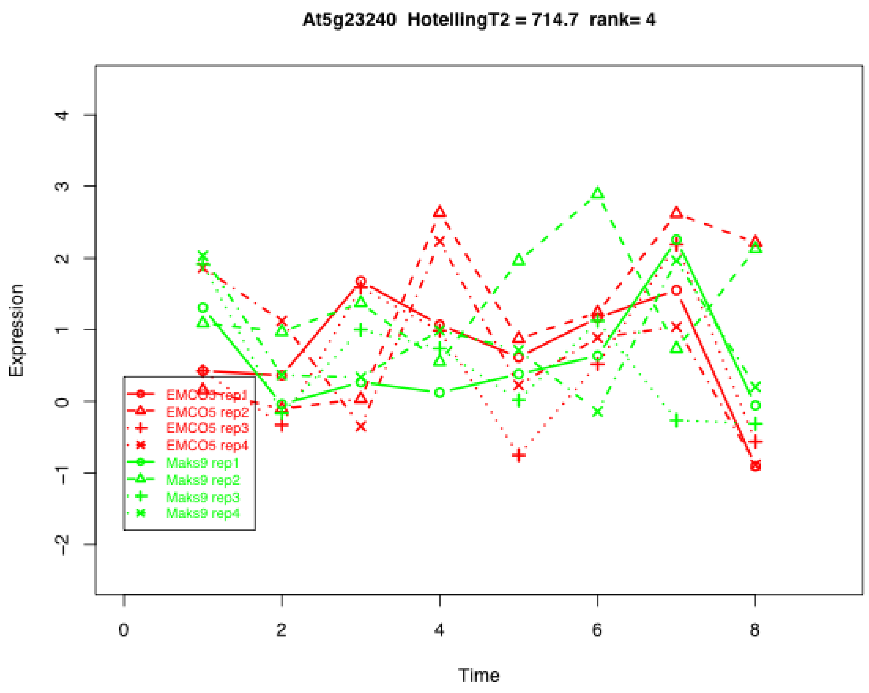
84	Cy5	87	11 pm	11	pm	C
85	Cy3	87	11 pm	11	pm	C
132	Cy5	87	11 pm	11	pm	C
7	Cy3	88	11 pm	11	pm	D
40	Cy3	88	11 pm	11	pm	D
89	Cy5	88	11 pm	11	pm	D
176	Cy5	88	11 pm	11	pm	D

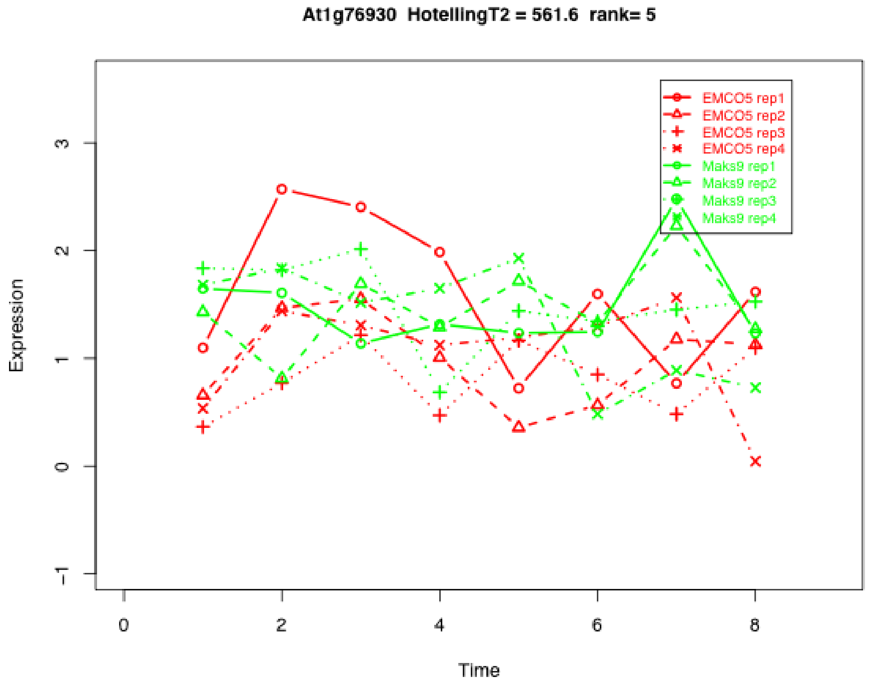
* Time of Day

Appendix H. A set of line graphs representing the top 4 most differentially expressed genes between *H. arabidopsidis* isolates *Maks9* and *Emco5* using Timecourse.

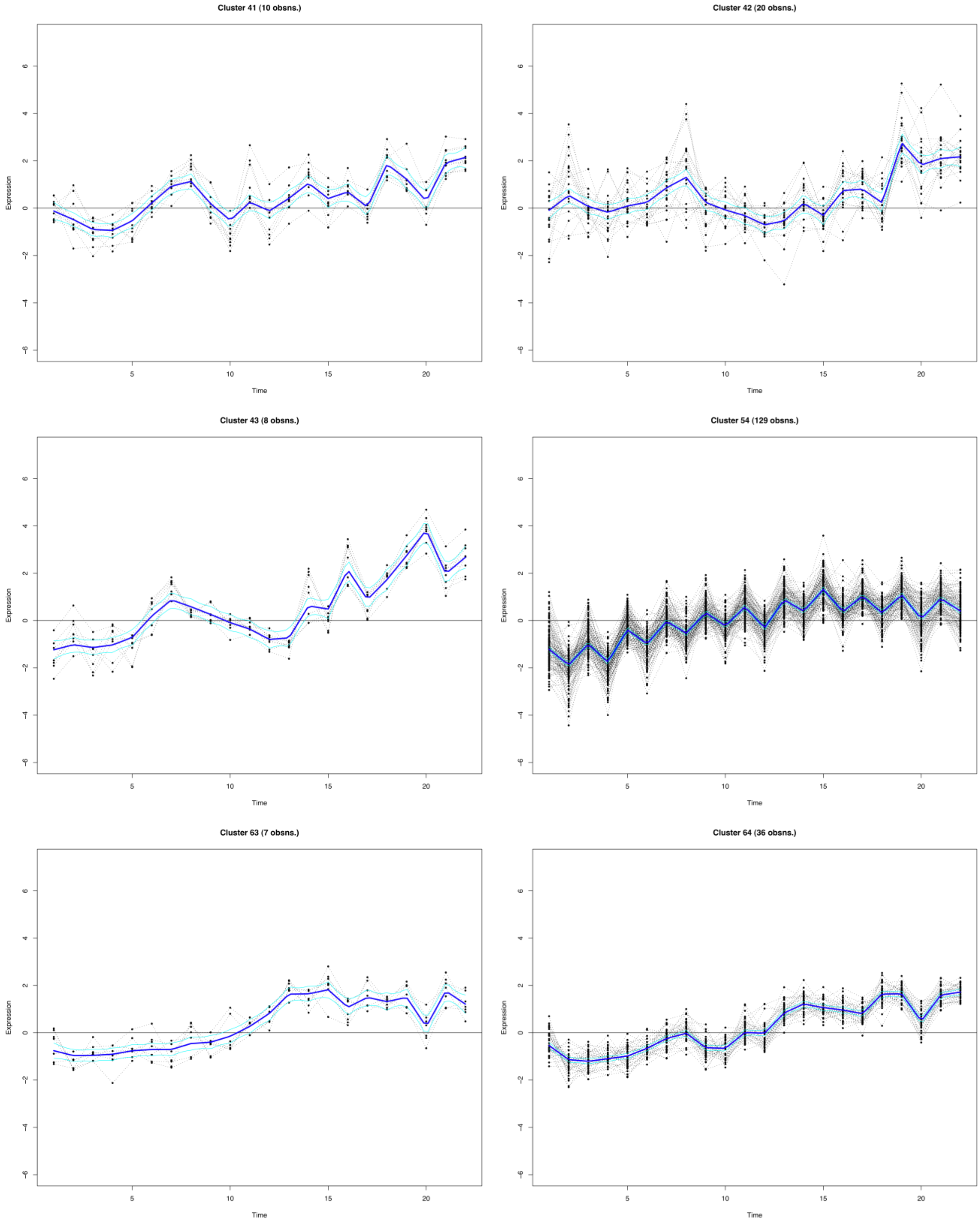




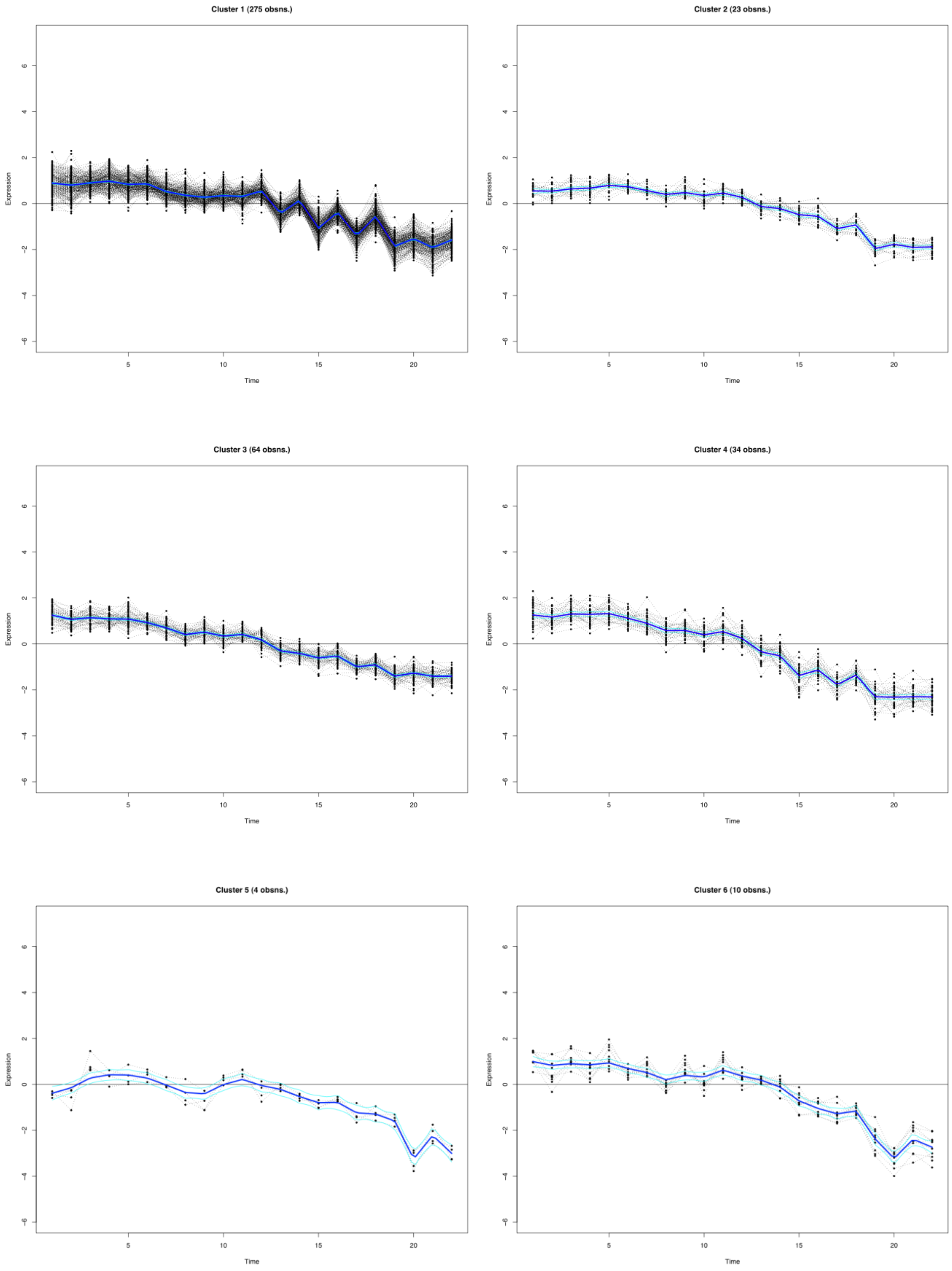

At5g23240 HotellingT2 = 714.7 rank= 4
Expression
Time
0
2
4
6
8
−2
−1
0
1
2
3
4
EMCO5 rep1
EMCO5 rep2
EMCO5 rep3
EMCO5 rep4
Maks9 rep1
Maks9 rep2
Maks9 rep3
Maks9 rep4


At1g76930 HotellingT2 = 561.6 rank= 5
Expression
Time
0
2
4
6
8
−1
0
1
2
3
EMCO5 rep1
EMCO5 rep2
EMCO5 rep3
EMCO5 rep4
Maks9 rep1
Maks9 rep2
Maks9 rep3
Maks9 rep4

Appendix. I. A set of lines graphs representing the expression profiles clusters 41, 42, 43, 54, and 63,64. The horizontal axis represents each of the 22 time points and the vertical axis represents the fold change in expression of the $\log_2$ signal intensities.



Appendix. J. A set of lines graphs representing the expression profiles clusters 1 to 6. The horizontal axis represents each of the 22 time points and the vertical axis represents the fold change in expression of the log$_2$ signal intensities.



Cluster 1 (275 obsns.)

Cluster 2 (23 obsns.)

Cluster 3 (64 obsns.)

Cluster 4 (34 obsns.)

Cluster 5 (4 obsns.)

Cluster 6 (10 obsns.)

Appendix. K. A table containing the gene identifiers and functions of the genes selected for modelling. The gene annotations obtained from the TAIR7 database.

Gene Number	Gene Identifier	Gene Function
1	*At1g19210*	AP2 domain-containing transcription factor putative
2	*At1g72520*	Lipoxygenase putative
3	*At1g74930*	AP2 domain-containing transcription factor putative
4	*At1g80840*	WRKY family transcription factor, similar to WRKY transcription factor GB:BAA87058 GI:6472585 from (Nicotiana tabacum)
5	*At2g44840*	Ethylene-responsive element-binding protein putative
6	*At3g23250*	Myb family transcription factor (*At*MYB15)
7	*At4g23800*	High mobility group (HMG1/2) family protein
8	*At4g23810*	WRKY family transcription factor
9	*At4g34410*	AP2 domain-containing transcription factor putative
10	*At5g21960*	Encodes a member of the DREB subfamily A-5 of ERF/AP2 transcription factor family. The protein contains one AP2 domain. There are 15 members in this subfamily including RAP2.1, RAP2.9 and RAP2.10.
11	*At1g34020*	Transporter-related
12	*At1g43160*	AP2 domain-containing protein RAP2.6 (RAP2.6)
13	*At1g49900*	Zinc finger (C2H2 type) family protein
14	*At2g02990*	Ribonuclease 1 (RNS1)
15	*At2g38240*	Oxidoreductase 2OG-Fe(II) oxygenase family protein
16	*At2g38380*	Peroxidase 22 (PER22) (P22) (PRXEA) / basic peroxidase E, identical to SP:P24102 Peroxidase 22 precursor (EC 1.11.1.7) (Atperox P22) (ATPEa) (Basic peroxidase E); identical to cDNA class III peroxidase ATPEa, GI:17530569
17	*At2g43870*	Polygalacturonase, putative / pectinase, putative, similar to SP:P48979 Polygalacturonase precursor (EC 3.2.1.15) (PG) (Pectinase) {Prunus persica}; contains PF00295: Glycosyl hydrolases family 28 (polygalacturonases)
18	*At3g11480*	S-adenosyl-L-methionine:carboxyl methyltransferase family protein
19	*At3g48520*	Cytochrome P450 family protein
20	*At4g21830*	Methionine sulfoxide reductase domain-containing protein / SeIR domain-containing protein
21	*At4g21850*	Methionine sulfoxide reductase domain-containing protein / SeIR domain-containing protein
22	*At4g22470*	Protease inhibitor/seed storage/lipid transfer protein (LTP) family protein
23	*At4g35160*	O-methyltransferase family 2 protein
24	*At4g36950*	Pseudogene similar to OSJNBa0042L16.2

25	*At5g02490*	Heat shock cognate 70 kDa protein 2 (HSC70-2) (HSP70-2)
26	*At5g05270*	Chalcone-flavanone isomerase family protein
27	*At5g13220*	Expressed protein
28	*At5g13930*	Chalcone synthase / naringenin-chalcone synthase
29	*At5g28237*	Tryptophan synthase beta subunit putative
30	*At5g28238*	Tryptophan synthase beta subunit putative
31	*At1g56650*	Myb family transcription factor (*At*MYB75)
32	*At1g66390*	Myb family transcription factor putative / production of anthocyanin pigment 2 protein (PAP2)
33	*At4g22880*	Myb family transcription factor, putative / production of anthocyanin pigment 2 protein (PAP2), contains Pfam profile: PF00249 myb-like DNA-binding domain; similar to GB:AAF66727 from (Petunia x hybrida) (Plant Cell 11 (8), 1433-1444 (1999)); identical to cDNA production of anthocyanin pigment 2 protein (PAP2) GI:11935172
34	*At5g07990*	Leucoanthocyanidin dioxygenase putative / anthocyanidin synthase putative
35	*At5g17220*	Flavonoid 3'-monooxygenase / flavonoid 3'-hydroxylase (F3'H) / cytochrome P450 75B1 (CYP75B1) / transparent testa 7 protein (TT7)
36	*At5g42800*	Glutathione S-transferase putative
37	*At5g54060*	Dihydroflavonol 4-reductase (dihydrokaempferol 4-reductase) (DFR)
38	*At2g18680*	Expressed protein

Appendix. L. A table showing the predicted interactions between the 38 genes chosen for modelling. The interaction strength is determined from the CBDZ score generated by the model, which is calculated in part from the standard deviation. Interactions are considered significant if their CBDZ scores are at least +/- 1.69 deviations from a standard normal distribution of 0, representing no interaction between the genes.

Source Gene	Gene Identifier	Target Gene	Gene Identifier	Standard Deviation (sds)
17	*At2g43870*	23	*At4g35160*	1.69017
18	*At3g11480*	13	*At1g49900*	2.29644
20	*At4g21830*	19	*At3g48520*	-1.75917
20	*At4g21830*	32	*At1g66390*	2.08421
22	*At4g22470*	11	*At1g34020*	2.28156
22	*At4g22470*	26	*At5g05270*	2.20862
22	*At4g22470*	28	*At5g13930*	1.77606
22	*At4g22470*	31	*At1g56650*	2.64274
22	*At4g22470*	32	*At1g66390*	2.69653
22	*At4g22470*	33	*At4g22880*	2.68096
22	*At4g22470*	34	*At5g07990*	2.22393
22	*At4g22470*	35	*At5g17220*	2.36572
22	*At4g22470*	36	*At5g42800*	3.07881
22	*At4g22470*	37	*At5g54060*	2.59433
32	*At1g66390*	13	*At1g49900*	1.94772
32	*At1g66390*	15	*At2g38240*	1.98133
32	*At1g66390*	16	*At2g38380*	1.96192
32	*At1g66390*	18	*At3g11480*	1.77558
32	*At1g66390*	19	*At3g48520*	2.31819
32	*At1g66390*	22	*At4g22470*	2.13633
32	*At1g66390*	26	*At5g05270*	-2.03374
32	*At1g66390*	27	*At5g13220*	1.69428
32	*At1g66390*	29	*At5g28237*	2.18901
32	*At1g66390*	30	*At5g28238*	2.00162
32	*At1g66390*	31	*At1g56650*	-1.69108
32	*At1g66390*	32	*At1g66390*	-2.30739
32	*At1g66390*	33	*At4g22880*	-2.17769
32	*At1g66390*	34	*At5g07990*	-2.26241
32	*At1g66390*	36	*At5g17220*	-1.96536
32	*At1g66390*	37	*At5g42800*	-1.80495

Appendix. M. A Table showing the At numbers of the genes and their respective gene annotations for cluster 10 of the Arabidopsis long day experiment after clustering with the program SplineCluster.

Gene Identifier	Gene Ontology
At1g06680	photosystem II oxygen-evolving complex 23 (OEC23)
At1g09390	GDSL-motif lipase/hydrolase family protein
At1g10900	phosphatidylinositol-4-phosphate 5-kinase family protein
At1g14280	phytochrome kinase putative
At1g15810	ribosomal protein S15 family protein
At1g16720	expressed protein
At1g18060	expressed protein
At1g18360	hydrolase alpha/beta fold family protein
At1g19450	integral membrane protein putative / sugar transporter family protein
At1g31330	photosystem I reaction centre subunit III family protein
At1g35420	dienelactone hydrolase family protein
At1g49130	zinc finger (B-box type) family protein
At1g51400	photosystem II 5 kD protein
At1g51940	protein kinase family protein / peptidoglycan-binding LysM domain-containing protein
At1g52220	expressed protein
At1g52230	photosystem I reaction centre subunit VI chloroplast putative / PSI-H putative (PSAH2)
At1g52240	expressed protein
At1g55670	photosystem I reaction centre subunit V chloroplast putative / PSI-G putative (PSAG)
At1g57770	amine oxidase family
At1g58290	glutamyl-tRNA reductase 1 / GluTR (HEMA1)
At1g62510	protease inhibitor/seed storage/lipid transfer protein (LTP) family protein
At1g64720	expressed protein
At1g66150	leucine-rich repeat protein kinase putative (TMK1)
At1g69530	expansin putative (EXP1)
At1g70820	phosphoglucomutase putative / glucose phosphomutase putative
At1g71430	expressed protein
At1g72540	protein kinase putative
At1g73600	phosphoethanolamine N-methyltransferase 3 putative (NMT3)
At1g73870	zinc finger (B-box type) family protein
At1g74470	geranylgeranyl reductase
At1g74960	3-ketoacyl-ACP synthase putative
At1g76800	nodulin putative
AT1G78830	curculin-like (mannose-binding) lectin family protein, similar to S glycoprotein (Brassica rapa) GI:2351186; contains Pfam profile PF01453: Lectin (probable mannose binding)
At1g15820	chlorophyll A-B binding protein, chloroplast (LHCB6), nearly identical to Lhcb6 protein (Arabidopsis thaliana) GI:4741960; contains Pfam profile PF00504: Chlorophyll A-B binding protein
At1g29930	chlorophyll A-B binding protein 2, chloroplast / LHCII type I CAB-2 / CAB-140 (CAB2B), identical to SP:P04778 Chlorophyll A-B binding protein 2, chloroplast precursor (LHCII type I CAB-2) (CAB-140) (LHCP) {Arabidopsis thaliana}
At1g44446	chlorophyll a oxygenase (CAO) / chlorophyll b synthase, identical to chlorophyll a oxygenase GI:5853117 from (Arabidopsis thaliana); contains Pfam PF00355 Rieske (2Fe-2S) domain
At1g54780	thylakoid lumen 18.3 kDa protein, SP:Q9ZVL6
At2g07000	expressed protein
No match	unknown protein
At2g18300	basic helix-loop-helix (bHLH) family protein
At2g20260	photosystem I reaction centre subunit IV chloroplast putative / PSI-E putative (PSAE2)
At2g22230	beta-hydroxyacyl-ACP dehydratase putative
At2g29650	inorganic phosphate transporter putative
At2g30790	photosystem II oxygen-evolving complex 23 putative
At2g32100	ovate protein-related
At2g33400	expressed protein
At2g34430	chlorophyll A-B binding protein / LHCII type I (LHB1B1)
At2g34460	flavin reductase-related
At2g39940	coronatine-insensitive 1 / COI1 (FBL2)
At2g42320	nucleolar protein gar2-related
At2g42580	tetratricopeptide repeat (TPR)-containing protein
At2g45180	protease inhibitor/seed storage/lipid transfer protein (LTP) family protein
At2g45560	cytochrome P450 family protein
At2g22990	sinapoylglucose:malate sinapoyltransferase (SNG1), similar to serine carboxypeptidase I precursor (SP:P37890) (Oryza sativa); contains Pfam profile PF00450: Serine carboxypeptidase; identical to cDNA sinapoylglucose:malate sinapoyltransferase (SNG1) GI:8699618
At2g32880	meprin and TRAF similarity domain-containing protein / MATH domain-containing protein, low similarity to ubiquitin-specific protease 12 (Arabidopsis thaliana) GI:11993471; contains Pfam profile PF00917: MATH domain
EUG4#2#04330	no match

At3g02380	zinc finger protein CONSTANS-LIKE 2 (COL2)
At3g05900	neurofilament protein-related
At3g06070	expressed protein
At3g08940	chlorophyll A-B binding protein (LHCB4.2)
At3g15570	phototropic-responsive NPH3 family protein
At3g16140	photosystem I reaction centre subunit VI chloroplast putative / PSI-H putative (PSAH1)
At3g17040	tetratricopeptide repeat (TPR)-containing protein
At3g17350	expressed protein
At3g18080	glycosyl hydrolase family 1 protein
At3g18773	zinc finger (C3HC4-type RING finger) family protein
At3g19850	phototropic-responsive NPH3 family protein
No match	unknown protein
At3g28130	nodulin MtN21 family protein
At3g44450	expressed protein
At3g52840	beta-galactosidase putative / lactase putative
At3g53800	armadillo/beta-catenin repeat family protein
At3g54890	chlorophyll A-B binding protein / LHCI type I (CAB)
At3g56940	dicarboxylate diiron protein putative (Crd1)
At3g59400	expressed protein
At3g60290	oxidoreductase 2OG-Fe(II) oxygenase family protein
At3g62630	expressed protein
At3g61470	chlorophyll A-B binding protein (LHCA2)
At3g47470	chlorophyll A-B binding protein 4, chloroplast / LHCI type III CAB-4 (CAB4), identical to SP:P27521 Chlorophyll A-B binding protein 4, chloroplast precursor (LHCI type III CAB-4) (LHCP) {Arabidopsis thaliana}
At3g52840	beta-galactosidase, putative / lactase, putative, similar to beta-galactosidase precursor GI:3869280 from (Carica papaya)
At4g03400	auxin-responsive GH3 family protein
At4g10340	chlorophyll A-B binding protein CP26 chloroplast / light-harvesting complex II protein 5 / LHCIIc (LHCB5)
At4g15160	protease inhibitor/seed storage/lipid transfer protein (LTP) family protein
At4g15920	nodulin MtN3 family protein
At4g21870	26.5 kDa class P-related heat shock protein (HSP26.5-P)
At4g23290	protein kinase family protein
At4g23400	major intrinsic family protein / MIP family protein
At4g24670	alliinase family protein
At4g25050	acyl carrier family protein / ACP family protein
At4g26555	immunophilin / FKBP-type peptidyl-prolyl cis-trans isomerase family protein
At4g27440	protochlorophyllide reductase B chloroplast / PCR B / NADPH-protochlorophyllide oxidoreductase B (PORB)
At4g28750	photosystem I reaction centre subunit IV chloroplast putative / PSI-E putative (PSAE1)
At4g29905 (gene only)	Expressed protein
At4g30110	ATPase E1-E2 type family protein / haloacid dehalogenase-like hydrolase family protein / heavy-metal-associated domain-containing protein
At4g32280	auxin-responsive AUX/IAA family protein
At4g34220	leucine-rich repeat transmembrane protein kinase putative
At4g34730	ribosome-binding factor A family protein
At4g34770	auxin-responsive family protein
At4g35250	vestitone reductase-related
At4g38860	auxin-responsive protein putative
At4g39350	cellulose synthase catalytic subunit (Ath-A)
At4g12900	gamma interferon responsive lysosomal thiol reductase family protein / GILT family protein, similar to SP:P13284 Gamma-interferon inducible lysosomal thiol reductase precursor {Homo sapiens}; contains Pfam profile PF03227: Gamma interferon inducible lysosomal thiol reductase (GILT)
At4g27440	protochlorophyllide reductase B, chloroplast / PCR B / NADPH-protochlorophyllide oxidoreductase B (PORB), identical to SP:P21218 protochlorophyllide reductase B, chloroplast precursor (EC 1.3.1.33) (PCR B) (NADPH-protochlorophyllide oxidoreductase B) (POR B) (Arabidopsis thaliana)
At4g29905	expressed protein
At4g38820	expressed protein
At4g38850	auxin-responsive protein / small auxin up RNA (SAUR-AC1), identical to GP:546362 small auxin up RNA {Arabidopsis thaliana}; belongs to auxin-induced (indole-3-acetic acid induced) protein family
At5g08330	TCP family transcription factor putative
At5g16030	expressed protein
At5g16400	thioredoxin putative
At5g19220	glucose-1-phosphate adenylyltransferase large subunit 1 (APL1) / ADP-glucose pyrophosphorylase (ADG2)
At5g19950	expressed protein
At5g24490	30S ribosomal protein putative
At5g25610	dehydration-responsive protein (RD22)
At5g28635	copia-like retrotransposon family
At5g39080	transferase family protein
At5g39530	expressed protein
At5g44190	myb family transcription factor (GLK2)
At5g44530	subtilase family protein

At5g44770	DC1 domain-containing protein
At5g46110	phosphate/triose-phosphate translocator putative
At5g49730	ferric reductase-like transmembrane component family protein
At5g54270	chlorophyll A-B binding protein / LHCII type III (LHCB3)
At5g57345	expressed protein
At5g65310	homeobox-leucine zipper protein 5 (HB-5) / HD-ZIP transcription factor 5
At5g24150	squalene monooxygenase 1,1 / squalene epoxidase 1,1 (SQP1,1), identical to SP:O65404
At5g24490	30S ribosomal protein, putative, similar to SP:P19954 Plastid-specific 30S ribosomal protein 1, chloroplast precursor (CS-S5) (CS5) (S22) (Ribosomal protein 1) (PSRP-1) {Spinacia oleracea}; contains Pfam profile PF02482: Sigma 54 modulation protein / S30EA ribosomal protein
At5g35490	expressed protein (MRU1), contains Pfam domain, PF04827: Protein of unknown function (DUF635)
At5g66570	Encodes a protein which is an extrinsic subunit of photosystem II and which has been proposed to play a central role in stabilization of the catalytic manganese cluster. In <iArabidopsis thaliana</i the PsbO proteins are encoded by two genes: <ipsbO1</i and <ipsbO2</i. PsbO1 is the major isoform in the wild-type.
At1g06680	photosystem II oxygen-evolving complex 23 (OEC23)
At1g09390	GDSL-motif lipase/hydrolase family protein
At1g10900	phosphatidylinositol-4-phosphate 5-kinase family protein
At1g14280	phytochrome kinase putative
At1g15810	ribosomal protein S15 family protein
At1g16720	expressed protein
At1g18060	expressed protein
At1g18360	hydrolase alpha/beta fold family protein
At1g19450	integral membrane protein putative / sugar transporter family protein
At1g31330	photosystem I reaction centre subunit III family protein
At1g35420	dienelactone hydrolase family protein
At1g49130	zinc finger (B-box type) family protein
At1g51400	photosystem II 5 kD protein
At1g51940	protein kinase family protein / peptidoglycan-binding LysM domain-containing protein
At1g52220	expressed protein
At1g52230	photosystem I reaction centre subunit VI chloroplast putative / PSI-H putative (PSAH2)
At1g52240	expressed protein
At1g55670	photosystem I reaction centre subunit V chloroplast putative / PSI-G putative (PSAG)
At1g57770	amine oxidase family
At1g58290	glutamyl-tRNA reductase 1 / GluTR (HEMA1)
At1g62510	protease inhibitor/seed storage/lipid transfer protein (LTP) family protein
At1g64720	expressed protein
At1g66150	leucine-rich repeat protein kinase putative (TMK1)
At1g69530	expansin putative (EXP1)
At1g70820	phosphoglucomutase putative / glucose phosphomutase putative
At1g71430	expressed protein
At1g72540	protein kinase putative
At1g73600	phosphoethanolamine N-methyltransferase 3 putative (NMT3)
At1g73870	zinc finger (B-box type) family protein
At1g74470	geranylgeranyl reductase
At1g74960	3-ketoacyl-ACP synthase putative
At1g76800	nodulin putative
AT1G78830	curculin-like (mannose-binding) lectin family protein, similar to S glycoprotein (Brassica rapa) GI:2351186; contains Pfam profile PF01453: Lectin (probable mannose binding)
At1g15820	chlorophyll A-B binding protein, chloroplast (LHCB6), nearly identical to Lhcb6 protein (Arabidopsis thaliana) GI:4741960; contains Pfam profile PF00504: Chlorophyll A-B binding protein
At1g29930	chlorophyll A-B binding protein 2, chloroplast / LHCII type I CAB-2 / CAB-140 (CAB2B), identical to SP:P04778 Chlorophyll A-B binding protein 2, chloroplast precursor (LHCII type I CAB-2) (CAB-140) (LHCP) {Arabidopsis thaliana}
At1g44446	chlorophyll a oxygenase (CAO) / chlorophyll b synthase, identical to chlorophyll a oxygenase GI:5853117 from (Arabidopsis thaliana); contains Pfam PF00355 Rieske (2Fe-2S) domain
At1g54780	thylakoid lumen 18.3 kDa protein, SP:Q9ZVL6
At2g07000	expressed protein
No match	unknown protein
At2g18300	basic helix-loop-helix (bHLH) family protein
At2g20260	photosystem I reaction centre subunit IV chloroplast putative / PSI-E putative (PSAE2)
At2g22230	beta-hydroxyacyl-ACP dehydratase putative
At2g29650	inorganic phosphate transporter putative
At2g30790	photosystem II oxygen-evolving complex 23 putative
At2g32100	ovate protein-related
At2g33400	expressed protein
At2g34430	chlorophyll A-B binding protein / LHCII type I (LHB1B1)
At2g34460	flavin reductase-related
At2g39940	coronatine-insensitive 1 / COI1 (FBL2)
At2g42320	nucleolar protein gar2-related
At2g42580	tetratricopeptide repeat (TPR)-containing protein
At2g45180	protease inhibitor/seed storage/lipid transfer protein (LTP) family protein
At2g45560	cytochrome P450 family protein
At2g22990	sinapoylglucose:malate sinapoyltransferase (SNG1), similar to serine carboxypeptidase I precursor (SP:P37890) (Oryza sativa); contains Pfam profile PF00450: Serine carboxypeptidase; identical to cDNA

	sinapoylglucose:malate sinapoyltransferase (SNG1) GI:8699618
At2g32880	meprin and TRAF similarity domain-containing protein / MATH domain-containing protein, low similarity to ubiquitin-specific protease 12 (Arabidopsis thaliana) GI:11993471; contains Pfam profile PF00917: MATH domain
EUG4#2#04330	no match
At3g02380	zinc finger protein CONSTANS-LIKE 2 (COL2)
At3g05900	neurofilament protein-related
At3g06070	expressed protein
At3g08940	chlorophyll A-B binding protein (LHCB4.2)
At3g15570	phototropic-responsive NPH3 family protein
At3g16140	photosystem I reaction centre subunit VI chloroplast putative / PSI-H putative (PSAH1)
At3g17040	tetratricopeptide repeat (TPR)-containing protein
At3g17350	expressed protein
At3g18080	glycosyl hydrolase family 1 protein
At3g18773	zinc finger (C3HC4-type RING finger) family protein
At3g19850	phototropic-responsive NPH3 family protein
No match	unknown protein
At3g28130	nodulin MtN21 family protein
At3g44450	expressed protein
At3g52840	beta-galactosidase putative / lactase putative
At3g53800	armadillo/beta-catenin repeat family protein

Appendix. N. A Table showing the At numbers of the genes and their respective gene annotations for cluster 65 of the Arabidopsis long day experiment after clustering with the program SplineCluster.

Gene Identifier	Gene Ontology
At1g05710	ethylene-responsive protein putative
At1g08310	esterase/lipase/thioesterase family protein
At1g08320	bZIP family transcription factor
At1g12640	membrane bound O-acyl transferase (MBOAT) family protein
At1g13340	expressed protein
At1g18860	WRKY family transcription factor
At1g19180	expressed protein
At1g24620	polcalcin putative / calcium-binding pollen allergen putative
At1g51890	leucine-rich repeat protein kinase putative
At1g62300	WRKY family transcription factor
At1g64610	WD-40 repeat family protein
At1g72700	haloacid dehalogenase-like hydrolase family protein
At1g75000	GNS1/SUR4 membrane family protein
At1g76070	expressed protein
At1g76980	expressed protein
At1g14870	expressed protein, similar to PGPS/D12 (Petunia x hybrida) GI:4105794; contains Pfam profile PF04749: Protein of unknown function, DUF614
At1g32940	subtilase family protein, contains similarity to subtilase; SP1 GI:9957714 from (Oryza sativa)
At1g32960	subtilase family protein, contains similarity to subtilase; SP1 GI:9957714 (Oryza sativa)
At1g77450	no apical meristem (NAM) family protein, contains Pfam PF02365: No apical meristem (NAM) domain; similar to GRAB1 protein GB:CAA09371, a novel member of the NAC domain family
At2g21780	expressed protein
At2g23150	NRAMP metal ion transporter 3 (NRAMP3)
At2g40340	AP2 domain-containing transcription factor putative (DRE2B)
At2g44500	expressed protein
At2g23150	NRAMP metal ion transporter 3 (NRAMP3)
At2g04050	MATE efflux family protein, similar to ripening regulated protein DDTFR18 (Lycopersicon esculentum) GI:12231296; contains Pfam profile: PF01554 uncharacterized membrane protein family
At2g18680	expressed protein
At2g22470	arabinogalactan-protein (AGP2), identical to gi:3883122:gb:AAC77824; supported by cDNA gi:3883121:gb:AF082299
At2g32210	expressed protein
At2g41905	expressed protein
AT2G39725	complex 1 family protein / LVR family protein, contains Pfam PF05347: Complex 1 protein (LYR family)
At3g03440	armadillo/beta-catenin repeat family protein
At3g10320	expressed protein
At3g10450	serine carboxypeptidase S10 family protein
At3g12580	heat shock protein 70 putative / HSP70 putative
At3g16860	phytochelatin synthetase-related
At3g29250	short-chain dehydrogenase/reductase (SDR) family protein
At3g55900	F-box family protein
At3g57520	alkaline alpha galactosidase putative
At3g57550	guanylate kinase 2 (GK-2)
At3g54680	proteophosphoglycan-related, contains similarity to proteophosphoglycan (Leishmania major) gi:5420389:emb:CAB46680
At4g02280	sucrose synthase putative / sucrose-UDP glucosyltransferase putative
At4g18280	glycine-rich cell wall protein-related
At4g25000	alpha-amylase putative / 14-alpha-D-glucan glucanohydrolase putative
At4g28490	leucine-rich repeat transmembrane protein kinase putative
At4g30470	cinnamoyl-CoA reductase-related
At4g37530	peroxidase putative
At4g12735	expressed protein
At5g01540	lectin protein kinase putative
At5g02580	expressed protein
At5g16910	cellulose synthase family protein
At5g22540	expressed protein
At5g25820	exostosin family protein
At5g39050	transferase family protein
At5g46050	proton-dependent oligopeptide transport (POT) family protein
At5g64230	expressed protein
At5g64310	arabinogalactan-protein (AGP1)
At5g67160	transferase family protein
At5g67310	cytochrome P450 family protein
At5g07440	glutamate dehydrogenase 2 (GDH2), identical to glutamate dehydrogenase 2 (GDH 2) (Arabidopsis thaliana) SWISS-PROT:Q38946

At5g27360	sugar-porter family protein 2 (SFP2), identical to sugar-porter family protein 2 (Arabidopsis thaliana) GI:14585701
At5g62520	Encodes a protein with similarity to RCD1 but without the WWE domain. The protein does have a PARP signature upstream of the C-terminal protein interaction domain. The PARP signature may bind NAD+ and attach the ADP-ribose-moiety from NAD+ to the target molecule. Its presence suggests a role for the protein in ADP ribosylation.
At1g05710	ethylene-responsive protein putative
At1g08310	esterase/lipase/thioesterase family protein
At1g08320	bZIP family transcription factor
At1g12640	membrane bound O-acyl transferase (MBOAT) family protein
At1g13340	expressed protein
At1g18860	WRKY family transcription factor
At1g19180	expressed protein
At1g24620	polcalcin putative / calcium-binding pollen allergen putative
At1g51890	leucine-rich repeat protein kinase putative
At1g62300	WRKY family transcription factor
At1g64610	WD-40 repeat family protein
At1g72700	haloacid dehalogenase-like hydrolase family protein
At1g75000	GNS1/SUR4 membrane family protein
At1g76070	expressed protein
At1g76980	expressed protein
At1g14870	expressed protein, similar to PGPS/D12 (Petunia x hybrida) GI:4105794; contains Pfam profile PF04749: Protein of unknown function, DUF614
At1g32940	subtilase family protein, contains similarity to subtilase; SP1 GI:9957714 from (Oryza sativa)
At1g32960	subtilase family protein, contains similarity to subtilase; SP1 GI:9957714 (Oryza sativa)
At1g77450	no apical meristem (NAM) family protein, contains Pfam PF02365: No apical meristem (NAM) domain; similar to GRAB1 protein GB:CAA09371, a novel member of the NAC domain family
At2g21780	expressed protein
At2g23150	NRAMP metal ion transporter 3 (NRAMP3)
At2g40340	AP2 domain-containing transcription factor putative (DRE2B)
At2g44500	expressed protein
At2g23150	NRAMP metal ion transporter 3 (NRAMP3)
At2g04050	MATE efflux family protein, similar to ripening regulated protein DDTFR18 (Lycopersicon esculentum) GI:12231296; contains Pfam profile: PF01554 uncharacterized membrane protein family
At2g18680	expressed protein
At2g22470	arabinogalactan-protein (AGP2), identical to gi:3883122:gb:AAC77824; supported by cDNA gi:3883121:gb:AF082299
At2g32210	expressed protein
At2g41905	expressed protein
AT2G39725	complex 1 family protein / LVR family protein, contains Pfam PF05347: Complex 1 protein (LYR family)
At3g03440	armadillo/beta-catenin repeat family protein
At3g10320	expressed protein
At3g10450	serine carboxypeptidase S10 family protein
At3g12580	heat shock protein 70 putative / HSP70 putative
At3g16860	phytochelatin synthetase-related
At3g29250	short-chain dehydrogenase/reductase (SDR) family protein
At3g55900	F-box family protein
At3g57520	alkaline alpha galactosidase putative
At3g57550	guanylate kinase 2 (GK-2)
At3g54680	proteophosphoglycan-related, contains similarity to proteophosphoglycan (Leishmania major) gi:5420389:emb:CAB46680
At4g02280	sucrose synthase putative / sucrose-UDP glucosyltransferase putative
At4g18280	glycine-rich cell wall protein-related
At4g25000	alpha-amylase putative / 14-alpha-D-glucan glucanohydrolase putative
At4g28490	leucine-rich repeat transmembrane protein kinase putative
At4g30470	cinnamoyl-CoA reductase-related
At4g37530	peroxidase putative
At4g12735	expressed protein
At5g01540	lectin protein kinase putative
At5g02580	expressed protein
At5g16910	cellulose synthase family protein
At5g22540	expressed protein
At5g25820	exostosin family protein
At5g39050	transferase family protein
At5g46050	proton-dependent oligopeptide transport (POT) family protein
At5g64230	expressed protein
At5g64310	arabinogalactan-protein (AGP1)
At5g67160	transferase family protein
At5g67310	cytochrome P450 family protein
At5g07440	glutamate dehydrogenase 2 (GDH2), identical to glutamate dehydrogenase 2 (GDH 2) (Arabidopsis thaliana) SWISS-PROT:Q38946
At5g27360	sugar-porter family protein 2 (SFP2), identical to sugar-porter family protein 2 (Arabidopsis thaliana) GI:14585701
At5g62520	Encodes a protein with similarity to RCD1 but without the WWE domain. The protein does have a PARP signature upstream of the C-terminal protein interaction domain. The PARP signature may bind NAD+ and

	attach the ADP-ribose-moiety from NAD+ to the target molecule. Its presence suggests a role for the protein in ADP ribosylation.
At1g05710	ethylene-responsive protein putative
At1g08310	esterase/lipase/thioesterase family protein
At1g08320	bZIP family transcription factor
At1g12640	membrane bound O-acyl transferase (MBOAT) family protein
At1g13340	expressed protein
At1g18860	WRKY family transcription factor
At1g19180	expressed protein
At1g24620	polcalcin putative / calcium-binding pollen allergen putative
At1g51890	leucine-rich repeat protein kinase putative
At1g62300	WRKY family transcription factor
At1g64610	WD-40 repeat family protein
At1g72700	haloacid dehalogenase-like hydrolase family protein
At1g75000	GNS1/SUR4 membrane family protein
At1g76070	expressed protein
At1g76980	expressed protein
At1g14870	expressed protein, similar to PGPS/D12 (Petunia x hybrida) GI:4105794; contains Pfam profile PF04749: Protein of unknown function, DUF614
At1g32940	subtilase family protein, contains similarity to subtilase; SP1 GI:9957714 from (Oryza sativa)
At1g32960	subtilase family protein, contains similarity to subtilase; SP1 GI:9957714 (Oryza sativa)
At1g77450	no apical meristem (NAM) family protein, contains Pfam PF02365: No apical meristem (NAM) domain; similar to GRAB1 protein GB:CAA09371, a novel member of the NAC domain family
At2g21780	expressed protein
At2g23150	NRAMP metal ion transporter 3 (NRAMP3)
At2g40340	AP2 domain-containing transcription factor putative (DRE2B)
At2g44500	expressed protein
At2g23150	NRAMP metal ion transporter 3 (NRAMP3)
At2g04050	MATE efflux family protein, similar to ripening regulated protein DDTFR18 (Lycopersicon esculentum) GI:12231296; contains Pfam profile: PF01554 uncharacterized membrane protein family
At2g18680	expressed protein
At2g22470	arabinogalactan-protein (AGP2), identical to gi:3883122:gb:AAC77824; supported by cDNA gi:3883121:gb:AF082299
At2g32210	expressed protein
At2g41905	expressed protein
AT2G39725	complex 1 family protein / LVR family protein, contains Pfam PF05347: Complex 1 protein (LYR family)
At3g03440	armadillo/beta-catenin repeat family protein
At3g10320	expressed protein
At3g10450	serine carboxypeptidase S10 family protein
At3g12580	heat shock protein 70 putative / HSP70 putative
At3g16860	phytochelatin synthetase-related
At3g29250	short-chain dehydrogenase/reductase (SDR) family protein
At3g55900	F-box family protein
At3g57520	alkaline alpha galactosidase putative
At3g57550	guanylate kinase 2 (GK-2)
At3g54680	proteophosphoglycan-related, contains similarity to proteophosphoglycan (Leishmania major) gi:5420389:emb:CAB46680
At4g02280	sucrose synthase putative / sucrose-UDP glucosyltransferase putative
At4g18280	glycine-rich cell wall protein-related
At4g25000	alpha-amylase putative / 14-alpha-D-glucan glucanohydrolase putative
At4g28490	leucine-rich repeat transmembrane protein kinase putative
At4g30470	cinnamoyl-CoA reductase-related
At4g37530	peroxidase putative
At4g12735	expressed protein
At5g01540	lectin protein kinase putative
At5g02580	expressed protein
At5g16910	cellulose synthase family protein
At5g22540	expressed protein
At5g25820	exostosin family protein
At5g39050	transferase family protein
At5g46050	proton-dependent oligopeptide transport (POT) family protein
At5g64230	expressed protein
At5g64310	arabinogalactan-protein (AGP1)
At5g67160	transferase family protein
At5g67310	cytochrome P450 family protein
At5g07440	glutamate dehydrogenase 2 (GDH2), identical to glutamate dehydrogenase 2 (GDH 2) (Arabidopsis thaliana) SWISS-PROT:Q38946
At5g27360	sugar-porter family protein 2 (SFP2), identical to sugar-porter family protein 2 (Arabidopsis thaliana) GI:14585701
At5g62520	Encodes a protein with similarity to RCD1 but without the WWE domain. The protein does have a PARP signature upstream of the C-terminal protein interaction domain. The PARP signature may bind NAD+ and attach the ADP-ribose-moiety from NAD+ to the target molecule. Its presence suggests a role for the protein in ADP ribosylation.
At1g05710	ethylene-responsive protein putative
At1g08310	esterase/lipase/thioesterase family protein

At1g08320	bZIP family transcription factor
At1g12640	membrane bound O-acyl transferase (MBOAT) family protein
At1g13340	expressed protein
At1g18860	WRKY family transcription factor
At1g19180	expressed protein
At1g24620	polcalcin putative / calcium-binding pollen allergen putative
At1g51890	leucine-rich repeat protein kinase putative
At1g62300	WRKY family transcription factor
At1g64610	WD-40 repeat family protein
At1g72700	haloacid dehalogenase-like hydrolase family protein
At1g75000	GNS1/SUR4 membrane family protein
At1g76070	expressed protein
At1g76980	expressed protein
At1g14870	expressed protein, similar to PGPS/D12 (Petunia x hybrida) GI:4105794; contains Pfam profile PF04749: Protein of unknown function, DUF614
At1g32940	subtilase family protein, contains similarity to subtilase; SP1 GI:9957714 from (Oryza sativa)