Improving Novel Gene Discovery in High-Throughput Gene Expression Datasets

PhD. Thesis by

Bruce Rosa

In Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

Biotechnology Ph.D Program, Lakehead University
Thunder Bay, Ontario, Canada

May 25th, 2012

Abstract

High-throughput gene expression datasets (including RNA-seq and microarray datasets) can quantify the expression level of tens of thousands of genes in an organism, which allows for the identification of putative functions for previously unstudied genes involved in treatment/condition responses.

For static (single timepoint) high-throughput gene expression experiments, the most common first analysis step to discover novel genes is to filter out genes based on their degree of differential expression and the amount of inter-replicate noise. However, this filtering step may remove genes with very high baseline expression levels, and genes with important functional annotations in the experiment being studied. Chapter 2 presents a novel knowledge-based clustering approach for novel gene discovery, in which known functionally important genes as well as genes with very high expression levels (which would typically be removed by a strict fold change filter) are saved prior to filtering.

In stress-related experiments on plants (including *Arabidopsis*), novel gene discovery is complicated by stress-induced disruption of circadian rhythm pathways, leading to differential expression of many genes which are not involved in adaptive stress responses. Chapter 3 presents the PRIISM (Pattern Recomposition for the Isolation of Independent Signals in Microarray data) algorithm, which is a frequency-based method which is able to differentiate and isolate circadian-disruption signals, improving novel gene discovery in time-series stress-response datasets.

Another major factor limiting the effectiveness of novel gene discovery in time-series datasets is the experimenter’s choice of timepoints to sample; The identification of important novel treatment-response genes is strongly dependent on sampling the timepoints at which the most target response genes are the most significantly differentially expressed. Although there may be several other time-series datasets with similar treatments available, there is currently no approach in the literature for using the information in these datasets to guide timepoint selection in a new experiment. Chapter 4

presents a new machine-learning model called Optimal Timepoint Selection (OTS) to automatically design optimized sampling rates for microarray and RNA-seq experiments based on the expression data of known treatment-response genes in existing datasets.

Acknowledgements

I'd first like to thank my thesis supervisor Dr. Wensheng Qin for encouraging me to enroll in the biotechnology PhD program at Lakehead University, supporting my strong interest in bioinformatics work as much as possible, and arranging for me to visit Michigan State University to further my education in the field. I'd also like to especially thank Dr. Jin Chen, my supervisor at Michigan State University, for welcoming me into his lab as a visiting scholar, and for his time, guidance and expert instruction on these projects. I have learned a great deal from my time in Dr. Chen's lab, and I hope that I have been helpful to him and the other people in the lab.

I am grateful for the contributions of Dr. Beronda Montgomery and Dr. Sookyung Oh who provided biological expertise and opinions, as well as extensive help with revisions, for Chapters 2 and 3. I thank Dr. Yuhua Jiao from Jin Chen's lab for working closely with me for the Chapter 3 study, as well as for teaching me how to program using MATLab and assisting with other projects. Also, thanks to Dr. Ji Zhang for developing the core mathematical model for the Chapter 4 study, and for spending so much time meeting with us over Skype as we refined the study. As well, I thank Dr. Ian Major for overseeing the production of the RNA-seq dataset and for his biological expertise in the Chapter 4 study, Dr. Abe Koo and Lalita Patel for their roles in the production of the microarray pilot coronatine experiment, and Lingyong Wang for writing the first draft of the C# code for the OTS software, and for teaching me some programming skills. From Dr. Qin's lab at Lakehead University, I am thankful to Mehdi Dashtban, Miranda Maki and Amy Chen for their help and advice throughout my Ph.D.

I greatly appreciate the feedback and helpful advice for Chapters 2 and 3 from Dr. Michael Thomashow, Dr. Eva Farre, Dr. Gregg Howe and Dr. Sheng Yang He.

I'd like to thank Dr. Greg Pyle and Dr. Wely Floriano for giving their time to serve on my thesis committee and to review this relatively complicated thesis.

Finally, I thank my wife Paula for being patient and understanding about my spending so much of my time at home on my computer for the last few years.

Table of Contents

Table of Tables

Chapter 2

Chapter 3

Chapter 4

Table of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 1: Introduction

1.1 High-throughput gene expression measurement

Understanding the transcriptome (the complete quantification of transcripts in a biological sample) is essential for determining the functions of genes, inferring molecular pathways and ultimately for understanding development, maintenance and adaptive responses in an organism (Thilmony et al., 2006; Wang et al., 2009a). Quantitative real-time PCR (qPCR or RT-PCR) is the most accurate and widely used technology for measuring changes in the transcript levels of single genes over time or across different conditions or strains of an organism (Bustin et al., 2005). However, low-throughput technologies are not useful for discovering unstudied genes which are involved in a given pathway or adaptive response, because the target genes must be identified before sampling (Bustin et al., 2005). High-throughput gene expression technologies can quantify the expression level of the many thousands of genes in an organism (for eukaryotes, ranging from ~6,000 genes in yeast, up to ~50,000 genes in rice (Kurata et al., 2002)). This allows for the measurement and subsequent identification of previously unstudied genes which have large transcriptional changes in response to a given treatment or condition. The two most prevalent technologies for deducing and quantifying the transcriptome are hybridization-based microarray technology and sequence-based RNA-Seq technology (Allison et al., 2006; Wang et al., 2009a).

Microarrays are currently the predominant method for high-throughput gene expression measurement (Mueckstein et al., 2010). Microarray chips contain massive ordered arrays of unique PCR-amplified cDNA sequences (up to 390,000 in some experiments (Mueckstein et al., 2010)), comprising some or all of the known gene sequences in an organism (Schena et al., 1995). RNA samples are extracted, converted to cDNA using nucleotides tagged with a fluorescent dye (usually Cy3 or Cy5), and then are mixed with a hybridization buffer and hybridized to their complementary cDNA sequences on the array (Schulze and Downward, 2001). High-resolution confocal fluorescence scanners then read the fluorescence in each position of the array, which quantifies the amount of bound cDNA to each sequence. A relative measure for gene expression for each cDNA on the chip can be derived and matched to each gene (Schulze and Downward, 2001). Although microarrays are a relatively inexpensive way to measure global gene expression patterns, they have relatively high levels of noise (from both biological and technical sources (Tu et al., 2002)) and rely on existing knowledge to build the chips, which may be flawed or limited, particularly for poorly studied organisms (Okoniewski and Miller, 2006; Wang et al., 2009a).

RNA-sequencing (RNA-seq) technology is a recently developed revolutionary tool for transcriptomics which directly determines the cDNA sequences contained in a biological sample. For this technology, total RNA is converted to cDNA, broken into fragments with adapters added to one or both ends, and all of the fragments are sequenced using the adapters as a guide. The fragment length depends on the technology used, but typically varies from 30 to 400 base pairs (Okoniewski and Miller, 2006; Wang et al., 2009a). Once all of the RNA fragments are sequenced, they are either matched back up to a reference genome (if one is available), or are used for *de novo* assembly of a new genome (Wang et al., 2009a). The dynamic range of expression levels measured by RNA-seq is much larger than microarrays (greater than 9000-fold range in yeast (Nagalakshmi et al., 2008)). A correlation study showed that the two methods agree fairly well for genes with medium levels of expression, but the accuracy of RNA-seq is much higher at very low and high expression levels (Wang et al., 2009b). In addition, unlike microarrays, RNA-seq can provide information about alternative splicing of genes, and can measure the expression of genes which have not been previously identified as having open reading frames (Wang et al., 2009a). For these reasons, as the cost of sequencing continues to fall, RNA-seq is expected to replace microarrays for high-throughput gene expression measurement (Wang et al., 2009b).

High-throughput gene expression measurement is a rapidly growing field in biology, and new techniques are under development, which have the potential to significantly increase the accuracy of expression measurement while reducing costs. These new approaches (which are still at a proof-of-concept stage) will reduce noise by replacing fluorescence measurements with other signals to read bases, including using conductivity measurements as each base is added (Treffer and Deckert, 2010), tethering DNA to magnetic beads during hybridization in order to measure the pulling forces induced by the hybridization of each base (Ding et al., 2012), and utilizing a nanomechanical approach, where the stiffness of the hybridized strand is measured as each base is added (Husale et al., 2009). The high-throughput gene expression analysis methods outlined in this thesis are applicable to any of the platforms outlined here, and could also be applied to analyze high-throughput protein quantification datasets, which are expected to become increasingly available over the next few years (Stoevesandt et al., 2009).

1.2 Experimental design for high-throughput gene expression experiments

There are many ways to design a microarray or RNA-seq experiment, which should be carefully considered based on the experimental goals and the resources available to the researcher. For this thesis, the focus will be on novel gene discovery. Here, the definition of “novel gene discovery” is the identification of genes involved in a given treatment/condition response which have not been previously characterized as being involved in that response. Note that this definition does not imply that the gene has not been annotated with different functions separate from the given treatment/condition.

One can analyze global genetic responses at one timepoint to a given treatment (including abiotic treatments such as cold (Bieniawska et al., 2008) and biotic treatments such as bacterial infection (Tang et al., 2005)), or the genetic differences between different strains of the same organism (including analyzing the downstream effects of genetic knockout (Narusaka et al., 2003) or overexpression mutants (Osakabe et al., 2002)). In these “static” experiments, a snapshot of the expression levels of genes is measured, while in time-series experiments, a temporal process is examined (Bar-Joseph, 2004). These time-series experiments can include analyzing cell division cycles (Spellman et al., 1998), circadian rhythm patterns (Mockler et al., 2007), developmental processes (Arbeitman et al., 2002), or analyzing temporal treatment or mutation responses (e.g. analyzing genetic responses to cold treatment over time (Espinoza et al., 2010)). Approximately 30% of all existing high-throughput gene expression datasets are designed with a time series (Singh et al., 2005). For many experimental goals, time-series high-throughput gene expression data is more useful than static data because it can be used to infer signalling pathways and identify potential transcription factors, by analyzing the coexpression of genes over time (Filkov et al., 2002). However, while these experiments provide considerably more information, they are more costly due to the extra samples required at many timepoints, and so are typically ran with fewer replicates than static experiments, which limits the statistical approaches available for analyzing them.

1.3 Gene filtering using statistical measures

Both microarray and RNA-seq experiments provide gene expression measurements for tens of thousands of genes at a time. For high-throughput gene expression studies, a point-wise fold change measure is typically calculated (Allison et al., 2006), in which the expression level of a gene under a given treatment or condition is divided by the expression level of that same gene at the same time in the control sample. A two-fold difference (either upregulated or downregulated) is typically considered a worthwhile cut-off, and is a very common way of filtering out genes which have not been differentially expressed (Cui and Churchill, 2003). There are several drawbacks to reducing the dataset size simply by setting differential expression cut-offs. First, inter-replicate noise should be considered in order to filter out genes with strong fold changes that may simply appear to be differentially expressed due to noise (Verducci et al., 2006). Secondly, with a high fold change threshold, genes with relatively small but reproducible fold changes may be filtered out, which is a problem of particular importance for genes with high constitutive expression levels (Verducci et al., 2006).

With or without filtering, genes can be ranked by their fold change values, to determine which genes are the most strongly differentially expressed (relative to the control samples). However, high differential regulation values alone are not sufficient for a robust analysis, because there may be a great deal of inter-replicate variability (noise) for a given gene. Thus, if several replicates are available, statistical cut-offs based on the variability of the gene expression are typically applied (Cui and Churchill, 2003; McCarthy and Smyth, 2009). One of the simplest statistical measures for high-throughput gene expression experiments is the *t* test, which utilizes the degree of differential regulation and the inter-replicate variance to calculate a significance value for one gene, representing the probability that a gene is truly differentially regulated (Callow et al., 2000; Cui and Churchill, 2003). However, this approach may have low power because high-throughput gene expression experiments tend to have a low number of replicates due to the cost of running the experiments (Cui and Churchill, 2003). Because of the low number of replicates and the very high number of genes, several other statistical approaches are commonly applied to static high-throughput gene expression datasets, including:

(1) *Significance Analysis of Microarrays* (SAM, or *S*-test), which is a modified *t*-test approach that adds a constant value to each gene's variance to correct for false significances resulting from very small values (Tusher et al., 2001);

(2) False-discovery rate (FDR) tests, which use the proportion of truly differential regulated genes, the distribution of the true differences, the variability between replicates and the sample size to compute the expected number of false positive genes in a list (*i.e.,* the number of genes in a list identified as differentially expressed which are not truly differentially expressed) (Pawitan et al., 2005; Yang and Yang, 2006). False discovery rate (FDR) filtering is often preferred

to *p*-value filtering, because it allows the user to define the accepted proportion of false positives in the dataset (Pawitan et al., 2005; Yang and Yang, 2006);

(3) Analysis of Variance (ANOVA) tests, which are applied when analyzing the results from multiple conditions/treatments, and use the average expression of a gene across all conditions/treatments as an additional input (Ayroles and Gibson, 2006; Kerr et al., 2000). Of the many types of ANOVA analyses applied to multiple microarray datasets reviewed by Cui and Churchill (2003), the mixed-model approach which treats the condition and biological replication as random effects has shown the strongest performance.

All of these statistical approaches ultimately produce a significance value in addition to the fold change value for each gene; In many experiments, both a fold change cut-off (typically arbitrarily set at two-fold) and a significance cut-off based on variability in the data (typically set at 0.05 after population correction, for any of tests listed above) are used to determine which genes are significantly differentially regulated (Cui and Churchill, 2003). Often, this filtering is displayed using a “volcano plot”, which shows the Log_2 transformed fold change values plotted against the Log_{10} P values for each gene (Fig. 1.1) (Cui and Churchill, 2003). In simple novel gene discovery approaches, the genes closest to the top-left and top-right of the volcano plot (which have the largest differential expression with the lowest noise) can be considered for further study (Cui and Churchill, 2003).

1.4 Clustering and gene networks for novel gene discovery

After performing filtering, a commonly used approach for discovering novel genes in high-throughput datasets is to cluster the genes (based on one of many possible mathematical models), and analyze clusters containing known genes and novel genes, based on the assumption that genes with similar expression patterns across many timepoints or conditions are likely functionally similar in some way (Bar-Joseph et al., 2003a; Chiappetta et al., 2004; Dejean et al., 2007; Dembélé and Kastner, 2003;

Eisen et al., 1998; Ernst and Bar-Joseph, 2006; Ernst et al., 2005; Hand and Heard, 2005; Hestilow and Huang, 2009; Ji et al., 2006; Jiang et al., 2004; Koenig and Youn, 2011; Peddada et al., 2003; Syeda-Mahmood, 2003; Verducci et al., 2006; Wu, 2008). Similarly, gene coexpression networks (which link groups of genes in complex interaction networks rather than simply creating large clusters of related genes) can be used to identify putative functions of novel genes (Aoki et al., 2007; Hu et al., 2005; Mao et al., 2009; Stuart et al., 2003). Typically, detailed functional and structural annotations from the Gene Ontology project databases are used to determine clusters or portions of the networks which are interesting for further analysis (http://www.geneontology.org/).

One major drawback of all of the existing clustering and networking approaches is that genes which are already known to be involved in a specific treatment or condition being studied may be filtered prior to the analysis, due to low fold change values (sometimes resulting from high baseline expression levels) or high rates of noise. However, the filtering step is critical for reducing noise and for reducing the dataset size so that the results can be visualized and analyzed. Chapter 2 presents a novel knowledge-based clustering approach for novel gene discovery, in which known functionally important genes as well as genes with very high expression levels are saved prior to filtering.

1.5 Time-series dataset analyses

Time-series high-throughput gene expression datasets are more useful than static datasets for determining signalling pathways, identifying transcription factors, and discovering novel genes involved in specific pathways (Filkov et al., 2002), and are becoming increasingly available as the price of running microarrays continue to fall and online databases continue to collect them (Craigon et al., 2004; Hubble et al., 2009; Parkinson et al., 2009). Time-series datasets which analyze a condition or mutation in addition to a wild-type analysis are particularly useful for identifying novel genes involved in a biological process (Jiang et al., 2004). Current approaches for novel gene discovery in high-resolution time-series gene expression datasets include a statistic to measure whether each gene is significantly differentially expressed across the entire time series (Bar-Joseph et al., 2003b), transcription factor-target identification through time-shifting algorithms (Yu et al., 2003), principal and independent component analysis methods (Frigyesi et al., 2006; Kong et al., 2008; Raychaudhuri et al., 2000), and can also include the clustering and networking approaches outlined in the previous subsection.

1.6 Circadian clock disruption

Arabidopsis, like all other plants, relies on a molecular circadian clock which is used to influence gene expression to modify physiology and metabolism in preparation for predictable changes in light and temperature conditions in the environment (Adams and Carre, 2011). Plants with circadian clocks that are properly synchronized to their environments have been found to fix more carbon, grow larger and survive better than plants with clocks that are out of phase with their environments (Dodd et al., 2005). Several studies have shown that between 6% and 31% of the *Arabidopsis* genome is influenced by circadian clock genetic components (Edwards et al., 2006; Harmer et al., 2000; Michael et al., 2008), while another study suggests that there are significant baseline circadian oscillations for 100% of the genome (Ptitsyn, 2008). A number of approaches have been developed for analyzing circadian rhythms in time-series gene expression datasets (Lu et al., 2006; Michael et al., 2008; Mockler et al., 2007; Price et al., 2008; Wichert et al., 2004). However, recently, biotic and abiotic stress treatments have been shown to disrupt rhythmic clock patterns through amplitude changes or phase shifts (Bieniawska et al., 2008; Bilgin et al., 2010; Chaves et al., 2009; Espinoza et al., 2010; Michael et al., 2008; Nakamichi et al., 2009), resulting in significant fold changes for genes which are clock-influenced but are not involved in direct stress response. The use of constant light to avoid some of the gene expression changes caused by disruption of the clock is not feasible because clock genes continue to cycle even under constant light, and the unnatural conditions reduce the applicability of the results in such a study (Espinoza et al., 2010; Salome et al., 2008). Likewise, the use of genetic knockouts of clock components can reduce disruptions due to circadian input, but since many stress-response genes are regulated by clock components, the results of such a study are difficult to interpret (Dong et al., 2011; Espinoza et al., 2010).

All of the current high-throughput time-series gene expression models are unable to differentiate between gene expression patterns that are the result of disruption of circadian clock input and gene expression patterns from direct regulation as an adaptive response to treatment (which are much more interesting for novel gene discovery purposes). Chapter 3 presents the PRIISM (Pattern Recomposition for the Isolation of Independent Signals in Microarray data) algorithm, a frequency-based method capable of differentiating these signals and improving novel gene discovery in time-series stress-response datasets.

1.7 Sampling rates in time-series high-throughput gene expression datasets

Most time-series high-throughput gene expression datasets contain very few timepoints, primarily due to cost considerations; More than 75% of the time-series datasets in the Gene Expression Omnibus (GEO) database contain 5 or fewer timepoints (Edgar et al., 2002). The usefulness of these datasets in performing novel gene discovery is strongly dependent on sampling timepoints at which there are significant changes in important gene expression levels, but determining the *best* sampling timepoints for time-series gene expression experiments is a challenging optimization problem that is frequently discussed in the biological literature (Androulakis et al., 2007; Bar-Joseph, 2004; Luo et al., 2011; Peddada et al., 2003; Singh et al., 2005; Wang et al., 2008). Although the knowledge in existing relevant gene expression experiments has the potential to be extremely valuable in guiding timepoint selection in a new experiment, fully utilizing the power of the existing data is a difficult problem. To tackle the problem of learning the expression patterns in existing datasets, Chapter 4 presents a new machine-learning model called Optimal Timepoint Selection (OTS) to automatically design optimized sampling rates for microarray and RNA-seq experiments.

1.8 Summary

High-throughput gene expression datasets (including RNA-seq and microarray datasets) can quantify the expression level of tens of thousands of genes in an organism, which allows for the identification of putative functions for previously unstudied genes involved in treatment/condition responses. The studies presented in this thesis provide solutions to several limitations to the existing approaches for identifying these novel genes.

First, an approach for single-timepoint analyses is presented, in which important functional genes are retained through the typical filtering process, allowing researchers to integrate previous knowledge of gene functions into a cluster-based novel gene discovery approach (Rosa et al., 2010). Second, for time-series experiments performed on plants, a method is presented for isolating differential regulation patterns resulting from stress-induced disruption of circadian clock pathways from differential regulations patterns resulting from treatment-response pathways (Rosa et al., 2012a). Finally, an approach is presented for automatically identifying optimal timepoints for capturing the differential expression patterns of large groups of target genes in new high-throughput gene expression experiments, based on existing datasets (Rosa et al., 2012b). This method will help to generate datasets which are better able to identify novel target-responsive genes by ensuring that the timepoints with the strongest genetic response to a treatment are captured in new datasets.

1.9 Chapter 1 Figures

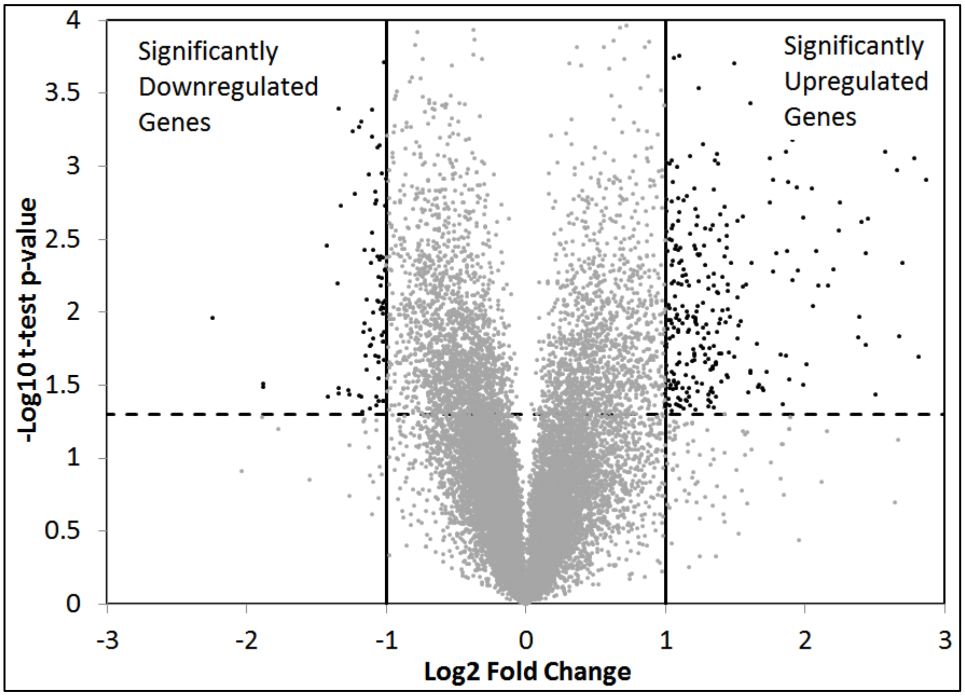


Figure 1.1: A volcano plot, used to filter non-significant genes (grey dots) from significantly differentially regulated genes (black dots) in a microarray experiment. Significantly upregulated and downregulated genes must be greater than a p-value cut-off (dashed black line) and either greater than or less than a fold change cut-off (vertical black lines). Typical cut-off values (2 fold up- or downregulation and p≤0.05) are shown. Data presented is from the "Whole plant" sample from the study in Chapter 2.

Chapter 2: Computing gene expression data with a knowledge-based gene clustering approach

This chapter has been published:

Rosa, B.A., Oh, S., Montgomery, B.L., Chen, J., Qin, W. (2010) Computing gene expression data with a knowledge-based gene clustering approach. International Journal of Biochemistry and Molecular Biology, 1(1), 51-68.

2.1 Abstract

Computational analysis methods for gene expression data gathered in microarray experiments can be used to identify the functions of previously unstudied genes. While obtaining the expression data is not a difficult task, interpreting and extracting the information from the datasets is challenging. In this study, a knowledge-based approach which identifies and saves important functional genes before filtering based on variability and fold change differences was utilized to study light-related gene regulation. Two clustering methods were used to cluster the filtered datasets, and clusters containing a key light regulatory gene were located. The common genes to both of these clusters were identified, and the genes in the common cluster were ranked based on their coexpression to the key gene. This process was repeated for 11 key genes in 3 treatment combinations. The initial filtering method reduced the dataset size from 22,814 probes to an average of 1134 genes, and the resulting common cluster lists contained an average of only 14 genes. These common cluster lists scored higher gene enrichment scores than two individual clustering methods. In addition, the filtering method increased the proportion of light responsive genes in the dataset from 1.8% to 15.2%, and the cluster lists increased this proportion to 18.4%. The relatively short length of these common cluster lists compared to gene groups generated through typical clustering methods or coexpression networks narrows the search for novel functional genes while increasing the likelihood that they are biologically relevant.

2.2 Introduction

There are a wide variety of approaches for data mining microarray datasets (including clustering and statistical coexpression networking), but knowledge-based approaches which integrate known gene information from various databases with statistical analysis methods are becoming increasingly common in the field of bioinformatics (Bellazzi and Zupan, 2007). For example, microarray gene expression data can be combined with genome sequence data to detect genetic regulatory elements controlling transcription (Mao et al., 2005), or it can be combined with known transcription factors and their targets to identify genetic regulation patterns useful for identifying novel transcription pathways (Yu et al., 2003). Another approach is to reduce the gene set to include only genes with known functions, and analyze their coexpression to find novel gene relationships in the dataset (Ma et al., 2007). More often, a variety of clustering methods are applied to gene coexpression data, and the common functions of genes within clusters are identified in order to attempt to discover the functions of other unstudied genes in the clusters (Hand and Heard, 2005; Hu et al., 2005; Stuart et al., 2003; Verducci et al., 2006). Almost all of these functional approaches utilize the Gene Ontology project databases, which characterize the functions of genes based on several categorizations (http://www.geneontology.org/).

The analysis of microarray datasets is complicated by the presence of biological noise (resulting from real changes in expression levels between different conditions and cell types) and experimental noise (resulting from differences in sample preparation, hybridization of cDNA to the probes, and reading of the wells on the gene chip) (Tu et al., 2002). Both of these sources of noise can limit the biological relevance of the results of microarray data mining (Tu et al., 2002).

One method employed for reducing noise in gene coexpression networks is to remove genes with a high-degree of inter-replicate variability. Typically, false discovery rate (FDR) filtering is used to remove highly variable genes, because it allows the user to define an acceptable proportion of false positives in the dataset (Pawitan et al., 2005; Yang and Yang, 2006). The calculation of FDR incorporates the proportion and distribution of truly differentially expressed genes, measurement variability and sample size, and thus is very useful to measuring variability in microarrays (Pawitan et al., 2005). FDR filtering is typically used to identify and remove highly variable genes before applying fold change filtering, but may be exclusively used when pooling together many datasets of various treatments, in which case the fold change statistic cannot be calculated (Ma et al., 2007).

Another method for reducing noise in gene coexpression networks is to remove genes exhibiting low fold changes between treatments. There is no generally accepted method for setting a threshold for this cut-off. Different approaches are to only keep genes exhibiting two-fold changes, to apply a t-statistic cut-off to keep only genes with significant changes, or to filter out genes which have fold changes below 30%, which is an estimate for the minimum biologically relevant change (Cui and Churchill, 2003; Filkov et al., 2002). There are several drawbacks to reducing the dataset size simply by setting differential expression cut-offs. First, inter-replicate noise should be considered in order to filter out genes with strong fold changes that may simply appear to be differentially expressed due to noise (Verducci et al., 2006). Secondly, with a high fold change threshold, genes with relatively small but reproducible fold changes may be filtered out, which is a problem of particular importance for genes with high constitutive expression levels (Verducci et al., 2006).

A common strategy for data mining microarray data using a knowledge-based clustering approach is to normalize the data, reduce the noise in the dataset by filtering out genes with a high degree of variability and low fold changes, cluster the genes, and then annotate the genes to analyze the biological functions of the clusters (Hu et al., 2005; Stuart et al., 2003; Verducci et al., 2006). This knowledge-based approach is capable of identifying upregulated clusters of functionally related genes (Mao et al., 2009), but when studying specific pathways, functionally important genes may be removed from the dataset by the filtering process.

The approach outlined in this paper overcomes this problem by annotating the probes first, saving genes important to the pathway being studied. After this step, genes with very high expression levels were also saved before filtering out highly variable genes and genes with low fold differences (Fig. 2.1). After filtering, all of the genes in the dataset were clustered using two different clustering approaches (K-means and Markov Clustering Algorithm [MCL]). K-means clustering was selected due to its common use in high-throughput gene expression clustering (Dembélé and Kastner, 2003; Hestilow and Huang, 2009; Macintyre, 2010; Schliep et al., 2004; Wu, 2008), and MCL clustering was chosen because it is an advanced clustering method which uses a "random walk" algorithm which is dissimmilar to the K-means algorithm, which makes it useful for cross-comparison (Dongen, 2008). At this point in the analysis, a novel key gene approach is used to group and compare clusters. Key genes of interest to the study were located in both the K-means and MCL clusters, and the common genes to both of these clusters were identified and ranked based on their Pearson correlations to the key gene.

Microarray expression datasets from three different strains of *Arabidopsis thaliana* plants were used for this experiment: A wild-type plant ("WT" treatment), one in which phytochromes have been inactivated in the leaf ("Leaf" treatment) and one in which phytochromes have been inactivated in the whole-plant ("Whole" treatment) (Warnasooriya and Montgomery, 2009). Phytochromes are proteins which are activated directly by light, then travel into the nucleus where they activate transcription factors in order to regulate light pathways (Montgomery, 2008). By degrading phytochromes, there will be a strong downregulation of light-regulated genes in the "Leaf" and "Whole" plants mutant phenotypes, so they will be used for studying light signalling pathways (Montgomery, 2008). All three potential combinations of treatments were compared in this study: WT-Leaf, WT-Whole and Leaf-Whole. The Leaf-Whole dataset was expected to have lower fold changes and more noise than either of the other dataset combinations because phytochromes have been inactivated in both treatments.

The analysis methods outlined in this paper creates small, ranked gene clusters developed for target light-regulation genes. These common clusters have a degree of coexpression in the common cluster lists, and many biologically relevant gene relationships, as well as many novel ones, were observed in the data.

2.3 Methods

2.3.1 Materials for microarray sample preparation

RNA was extracted and measured from seven-day old whole-*Arabidopsis* seedlings using standard methods for Affymetrix gene chips (The Arabidopsis ATH1 Genome Array; Affymetrix, Santa Clara, CA). The RNA extraction and microarray analysis was performed in triplicate for each plant using a procedure described by the manufacturer (Affymetrix).

2.3.2 Normalize the dataset using RMA normalization

First, RMA normalization was applied to the raw signal data, as this normalization method has been found to significantly reduce background noise, while still maintaining fold changes between up- and down-regulated genes (Bolstad et al., 2003; Irizarry et al., 2003). The “affylmGUI” package available as part of the *Bioconductor* software package for R was used to perform the normalization (Smyth, 2004).

2.3.3 Annotate the dataset using current gene descriptions

Annotation and gene ontology data was retrieved from *The Arabidopsis Information Resource* (TAIR) database (Genome release version 8, available for download from ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR8_genome_release/). TAIR is the largest collective database for *Arabidopsis* gene data, and consists of genes identified by a combination of manual and computational methods (Swarbreck et al., 2008). At this stage in the analysis, Affymetrix probes not matching any genes were removed from the dataset, and two or more probes that matched the same gene (due to probes hybridizing with different sequences on the same gene) were combined by averaging the values (183 probes total).

2.3.4 Identify and save important key genes involved in the target process

a) In this step, genes with important functions were identified, both by identifying specific genes involved in the target pathway and by searching all of the genes in the dataset for important functions by identifying "keywords" in both their "gene model description" (provided by TAIR) and their gene ontology functions. Because the gene model description contains a lot of detailed information about the gene function (including relationships to other genes and upstream and downstream effects on many pathways), far more functionally important genes are saved from filtering using this method than by relying on gene ontology alone.

Keywords must be chosen based on the goals of the research study, so the choice of relevant keywords requires some manual input by the researchers. Choosing broader keywords will save more genes based on function. After identifying the gene descriptions containing relevant keywords, the number of genes can be analyzed, and keywords can be added or removed to adjust the number of genes saved based on keywords.

b) At this point, there are still several probes in the dataset which have multiple genes (homologs) annotated to them. Sometimes, this occurs because there are several names assigned to the same gene in the literature, but often the sequences for several genes will be similar enough that they both bind to the same probe. There is no standard method for dealing with this problem. For this analysis, we have separated these genes into three categories:

i) Probes with multiple genes that contained one of the keywords were saved from further filtering, to be included in the final network.

ii) A "broad" keyword search using broader biological keywords which match many genes (e.g. "transcription") was applied on the probes with multiple genes. These probes were put through the rest of the filtering analysis (Fig. 2.1, steps 4 through 6).

iii) Probes with multiple genes that contained neither a keyword nor a “broad” keyword were discarded from the dataset.

2.3.5 Save genes with high expression levels

At this point in the analysis, filtering is performed based on signals, so the three treatment combinations had different sets of results. The high expression level cut-off step is applied because genes with high expression levels may have relatively low but consistent fold changes, which would otherwise be removed by fold change filtering. The frequency distribution of the expression levels for all of the genes in a microarray experiment is expected to fit a normal distribution for microarrays with sufficient background noise reduction (Konishi, 2004). Genes with a fold change higher than two standard deviations above the mean were saved from further filtering. This cut-off value was chosen because it is a parametric measure, and given a normal curve, this will result in only approximately 2.2% of the total dataset being saved due to high expression levels.

2.3.6 Remove highly variable genes using FDR

The “significance analysis of microarray” (SAM) excel plugin software package was used to calculate local and global FDR values for each of the datasets (Tusher et al., 2001). The first FDR value below 5% was used, to achieve datasets with a false discovery rate of less than 5%. This threshold cut-off of 5% or less is the most commonly used FDR cut-off for microarray datasets (Pawitan et al., 2005; Yang and Yang, 2006).

2.3.7 Remove genes with low fold changes

Genes with expression levels above two standard deviations above the mean and below two standard deviations below the mean were saved in the dataset, and the rest were discarded from the dataset. Similarly to the expression level filtering in step 4, this parametric cut-off was chosen because the fold difference distribution follows a normal distribution. Using two standard deviations above and below the mean as a cut-off results in approximately 4.4% of the remaining genes being saved in the dataset.

2.3.8 Calculate the Pearson correlation between each gene pair remaining in the dataset

Coexpression measurements for each gene pair are typically measured by calculating the Pearson correlation value between the gene signal intensity patterns (Mao et al., 2009; Stuart et al., 2003; Verducci et al., 2006). In this study, Pearson correlations were calculated for each gene pair using the normalized expression levels corresponding to each analysis (e.g., the three replicates for WT and the three replicates for Leaf were compared for each gene pair in the WT-Leaf analysis, resulting in 6 expression values being used to calculate the Pearson correlation values for each gene). The absolute values of the Pearson correlations were used throughout the rest of the analysis.

2.3.9 Rank the genes according to their Pearson correlation

The Pearson correlation values were converted to ranks, so that the most closely related genes to a target gene had the lowest rank values. Table 2.1 demonstrates this conversion, with the Pearson values in the left hand table, and the rank values in the right hand table. The ranks are calculated vertically, and do not apply horizontally across the table.

2.3.10 Apply K-means Clustering and MCL (Markov Clustering Algorithm) to the dataset

K-means clustering (the most popular clustering method for microarray analysis) clusters groups of related genes by defining cluster centers in the dataset and then arranging the genes such that their coexpression will be closest to the most correlated center (Wu, 2008). Another clustering method which is becoming popular for use in microarray analysis is the MCL clustering, which analyzes various random paths through the connections in a dataset, removes edges connecting clusters, and identifies related clusters with high precision (Mao et al., 2009; Pu et al., 2007; Van Dogen, 2000). In this study, both clustering methods were used to cluster genes based on the absolute Pearson correlation values. For K-means clustering, each of the datasets was split into 25 clusters. The R command “kmeans” was used to calculate these clusters. MCL clustering was performed using the freely available Linux package provided by the algorithm’s author (Van Dogen, 2000). The default inflation value was used, and the pre-inflation value was changed to adjust the number of clusters produced. The target number of 25 clusters resulted in one cluster being very large (usually 90% of the dataset), so this cluster was then split again into 25 clusters until each cluster contained less than 200 genes. This was done because the authors of this paper decided that clusters of greater than 200 genes were too large to be useful for novel gene identification.

For both the K-means and MCL clustering techniques, the number of clusters and the size of each cluster can be adjusted by the user (by either directly entering number for K-means clustering and by adjusting the pre-inflation parameters for MCL clustering). For this study, the approximately 1,000 genes in each dataset were separated into 25 clusters, resulting in clusters with an average of approximately 40 genes each. GSEA analysis showed that clusters smaller than these have significantly lower enrichment scores for both the MCL and K-means clustering methods.

2.3.11 Identify clusters containing key genes and rank clustered genes to key genes.

11 Key genes which are important for phytochrome-mediated light signalling were identified. For each of the 11 key genes in each treatment combination, four lists were generated: The ranked list of every gene in the dataset, the genes sharing the same K-means cluster, the genes sharing the same MCL cluster, and the common genes to both the K-means and the MCL clusters. All of the clusters have the gene ranks applied to them, so that the key gene is on the top of the list and the next most closely related gene is next.

2.3.12 GSEA and AmiGO analyses

Gene Set Enrichment Analysis (GSEA) is a useful algorithm and software suite that determines the degree to which a group of genes are significantly related by calculating an enrichment score for the gene group based on common biological functions, chromosomal locations, or regulatory patterns

(Subramanian et al., 2005). GSEA was employed to analyze each gene cluster derived for each of the 11 key genes (Subramanian et al., 2007).

AmiGO is a gene ontology analysis tool which can determine the percentage of genes in each gene group represented by several important light-related ontology categories (http://amigo.geneontology.org/cgi-bin/amigo/slimmer) (Carbon et al., 2009).

Currently, a fair amount of manual intervention is necessary to develop the cluster lists, as there is no program available to carry the dataset through each of the steps. When software packages were used for certain steps, they were indicated in the methods. Otherwise, the data manipulation was performed using Microsoft Excel 2007. The filtering steps have been standardized wherever possible, but the choice of keywords by the user will determine the amount of data that will be saved due to gene function.

2.4 Results

2.4.1 RMA normalization

The Affylmgui software was used to RMA normalize signal data for 22,810 probes (Fig. 2.1).

2.4.2 Gene annotation

Figure 2.1 indicates the number of genes removed from the dataset due to a lack of annotation in the TAIR database or due to probe duplication.

2.4.3 Identify and save important key genes involved in the target process (light regulation)

For this analysis, 11 key genes involved in phytochrome-mediated light signalling and 60 genes known to be related to them were saved. In addition, the keywords “Phytochrome, PhyA, PhyB, PhyC, PhyD, PhyE, CAB (Chlorophyll A/B-binding), chromophore, photomorphogenesis” were used to identify other genes involved in light response. The “broad keywords” used to identify probes with multiple genes which were to be saved were “Photo, light, chloroplast, transcription, kinase and auxin”.

2.4.4 Saving genes with high expression levels

The highest average intensity between two treatments was used to represent each gene for this step in the analysis. Figure 2.2 shows the distribution of the normalized gene intensities in the dataset. Despite a dip in the shape of the curve (at a signal intensity of approximately 5.5), the data fit a normal distribution with an R^2 of at least 0.88 for all three datasets (Table 2.2).

2.4.5 FDR filtering

The analyses containing the "Whole" dataset had less genes remaining after FDR filtering (834 and 825 compared to 5950 for the WT-Leaf dataset) (Table 2.3).

FDR values were calculated for the genes saved due to keyword and high expression level filtering. We found that, on average, 62% of important functional genes and 57% of genes with high intensity levels would have been removed due to FDR filtering, had they not been saved in the previous steps. The proportion of genes that would have been removed was much lower for the WT-Leaf dataset (33%) than for the WT-Whole (70%) and Leaf-Whole (72%) datasets.

2.4.6 Fold difference filtering

Figure 2.3 shows the distribution of all the genes in each of the three treatment combinations. Note that there are far less genes included in the WT-Leaf and Leaf-Whole datasets due to FDR filtering. The fit for these datasets to a normal distribution is lower (0.667 and 0.684) than for the WT-Leaf dataset (0.805) (Table 2.4).

2.4.7 Calculating Pearson correlations for each gene pair

The number of genes remaining in each dataset after FDR and fold difference filtering is summarized in Table 2.5. The frequency distribution of the absolute Pearson correlations can be seen in Figure 2.4, and the frequency distribution of the connectivity of the genes in the dataset is shown in Figure 2.5. This figure shows the number of significant correlations (>0.95) that each gene has in each dataset.

2.4.8 Gene ranking

The gene lists were ranked based on their correlation to each of the target genes, and these ranks were later used to order the genes in each cluster.

2.4.9 Clustering

MCL and K-means clustering was applied to the datasets for each of the three treatment combinations. The MCL and K-means clusters containing each of the 11 key genes were identified. A list of the genes common to both of these clusters was developed (See Tables 2.7 and 2.8 for an example of one of these lists). On average, 39% of the genes in the K-means and MCL clusters were found in both of the clusters. Table 2.5 shows the average number of genes in each gene cluster, as well as the average percentage of genes in each cluster that were included due to various steps in the filtering process.

2.4.10 GSEA and AmiGO analyses

Table 2.5 shows the average enrichment scores for each of the 11 key gene clusters based on GSEA analysis. Table 2.6 shows the results from a gene ontology analysis. Gene groups composed of 5 or fewer genes were not included in this analysis, as they tended to strongly bias the results, and the key genes used to identify the important clusters were removed from the analysis in order to avoid the bias that they introduce.

2.5 Discussion

The microarray analysis method outlined here combines the advantages of several microarray analysis techniques. It results in the development of short lists of genes most closely related to target genes by filtering out genes with high variability and low fold changes without removing functionally important genes, and then relies on two separate clustering methods to find genes closely co-expressed to light signalling genes. The use of two clustering methods which employ different algorithms narrows down a shorter list of important genes.

The removal of probes matching no genes, probes matching duplicate genes, and most of the probes matching multiple genes in steps 2 and 3 in the procedure reduces noise in the final clusters by removing 1,527 probes (~7% of the total dataset) with no meaningful gene matches before any steps of the filtering process (Fig. 2.1). This is one of the advantages of performing probe annotation before filtering the dataset.

The keyword search system outlined in the methods identified 383 genes potentially involved in phytochrome signalling and photomorphogenesis, which were the targets of this particular experiment. A search in the AmiGO gene ontology database using similar search terms (GO:0010017 "phytochrome signalling pathway" and GO:009640 "photomorphogenesis") identifies only 106 functionally important genes, indicating that the keyword search method outlined here identifies more than three times as many potential functionally important genes. The advantage of this method comes from searching the long TAIR gene descriptions, which often include putative functions as well as far up-stream or down-stream relationships between gene products. This keyword system also allows researchers to identify functionally important genes based on both specific as well as broad functions, and can be used to search for confirmed and putative relationships to other genes, which is not included in most ontology programs (Carbon et al., 2009; Wu et al., 2005).

By saving genes with high expression levels, genes with relatively small but consistent fold changes remained in the dataset. Also, by using the higher average intensity value between the treatments, we ensure that active photosynthesis genes (which are expected to be highly expressed in the WT condition in which light pathways are not disrupted) are kept in the dataset [22]. The close fits to a normal curve for these datasets (Table 2.2) justify the use of parametric cut-offs for filtering the dataset. The genes saved at this step in the analysis accounted for approximately half of the genes in the dataset.

As a result of FDR filtering, considerably more genes were removed from the datasets containing the “Whole” treatment. FDR values are affected by the amount of variability between replicates of genes, so this result indicates that there was a greater amount of inter-replicate variability in the “Whole” dataset (Table 2.3). This trend of significantly lower FDR values in the WT-Leaf dataset was also seen for the genes saved due to function or high intensity.

The number of genes filtered out of the datasets due to their fold change values varied considerably, primarily because of the much larger number of genes that were removed from FDR filtering in the WT-Whole and Leaf-Whole datasets (Table 2.4). The frequency distribution of the fold changes for all genes in the datasets is expected to fit a normal distribution (Konishi, 2004). The larger WT-Leaf dataset showed a strong fit to a normal distribution, justifying a parametric filtering cut-off for this step.

There was a much smaller proportion of genes with a low variability and high fold change for the Leaf-Whole dataset (4.8%) than for the WT-Leaf and WT-Whole datasets (18.6% and 11.8%), which was expected because the phenotype of these treatments was similar (phytochromes were deactivated in both, though in distinct tissues). Important genes and genes matching keywords accounted for an average of 6.4% and 28% of the genes in the datasets, respectively. In all of the datasets, genes saved because of high intensity values accounted for more than half of the genes in the dataset.

The frequency distribution of the absolute Pearson correlations (Fig. 2.4) shows that most of the gene correlations for the WT-Whole and WT-Leaf datasets were close to 1, indicating that the genes remaining in the dataset tend to be co-expressed with each other. The distribution for the Leaf-Whole dataset is much flatter, indicating that less of the genes were strongly co-expressed between these treatments. Figure 2.5, which shows the connectivity of the network, indicates that most genes in each dataset have relatively few strong correlations to other genes, and a few “hub” genes in each dataset have a lot of strong connections to other genes. This approximate power-law distribution is expected for all scale-free networks, including microarray datasets (Zhang and Horvath, 2005).

Table 2.5 shows the average proportion of each category of genes in the 11 key-gene clusters. There was a higher proportion of genes from the “Important Gene” and “Keyword” categories in each of the clusters than in the full filtered list, indicating that the genes identified in these categories were more closely co-expressed with the 11 key genes used to develop the cluster lists. The proportion of genes saved due to high intensity levels remained almost the same, while the proportion decreased considerably for genes saved due to low variability and high fold changes, indicating that less of these genes were involved in light regulation for this dataset. Table 2.5 also shows the GSEA enrichment scores for each of the cluster types. The “common” clusters (the common genes between the K-means and MCL analyses) had better average enrichment scores than the MCL clusters for all treatments, and better enrichment scores than the K-means clusters for two out of three of the treatments, indicating the presence of more closely co-expressed gene pairs in these clusters.

The gene ontology results indicate that the filtering process used in this study (prior to clustering) increases the proportion of known light-responsive genes in the dataset from 1.84% of all genes to 15.16% of all genes in the filtered lists and 18.43% in the common clusters (Table 2.6). Genes in the “photomorphogenesis” gene ontology category were expected to be strongly affected by the removal of phytochrome activity, and it was found that this term was enriched from 0.23% of all genes to 4.52% of all genes in the filtered list and 6.14% of genes in the clustered lists. Due to the keyword gene targeting process, it was expected that the proportion of functional genes would be increased considerably in the filtered lists, but the increase in the proportion of functional genes in the clusters indicates that the clustering steps are effectively identifying genes in target pathways. The GSEA and the GO results indicate that the common clusters better represent co-expressed light-related genes in the dataset than either of the clustering methods alone. The smaller size of these clusters also narrows down potential targets for further study, without reducing the importance of the gene relationships.

Table 2.7 shows one of the common cluster lists generated by the gene analysis methods outlined in this paper. The “Reason to Keep” column in this table shows the stage in the filtering system at which each gene was saved from filtering. Some of the genes (including *ATGRP8*, a known circadian-regulated gene) would have been removed by FDR filtering, but remained in the list because they were saved due to a high signal intensity (Carpenter et al., 1994).

Several genes known to be related to *SPA1* (*AT2G46340*, a repressor of PHYA activity) appear in the list in Table 2.7 (including two transducin genes as well as several uncharacterized genes that may be potential targets for further study). There are also genes encoding several WD-40 repeat family proteins (*AT3G49660*, *AT4G18900*, and *AT4G18905*) which are related to *SPA1* (Hoecker et al., 1999). Also in the list is *SHB1* (SHORT HYPOCOTYL UNDER BLUE1; *AT4G25350*) whose product can impact plant growth under blue, red and far-red light (Kang and Ni, 2006). To our knowledge, no previously published reports have made a connection between *SPA1* and *SHB1*, though both can presumably impact light signalling under far-red light. Notably also in this group is *CAM7*, a gene encoding a transcriptional regulator previously shown to interact with the promoter of light-induced genes and that impacts photomorphogenesis (Kushwaha et al., 2008). This gene has been previously associated with *HY5*, but not *SPA1* (Kushwaha et al., 2008). Also included is a gene encoding a circadian-regulated glycine rich protein (ATGRP8; AT4G39260), whose expression was previously shown to be impacted by far-red light (Zeidler et al., 2004) and another glycine rich protein (*AT4G29020*), which is likely a homolog of ATGRP5, involved in cellular elongation (Mangeon et al., 2010).

Table 2.8 shows the common cluster list for *HY5*, which is a positive regulatory factor that promotes photomorphogenesis. In this *HY5* common list is a gene encoding a largely uncharacterized protein SHW1, (SHORT HYPOCOTYL IN WHITE; *AT1G69935*) which serves as a negative regulator of photomorphogenesis (Bhatia et al., 2008). Also in this list is *PORB* (*AT4G27440*), whose product is known to be required for the light-dependent accumulation of chlorophyll (Armstrong et al., 1995).

2.6 Conclusion

By saving genes with important keywords and performing the clustering with unstudied genes, we can see new relationships between key genes that may otherwise be missed. Also, the possibility of new biological insight is represented by the identification of genes encoding proteins of unknown function (e.g., *AT2G25510* and *AT1G56660*).

The relatively short length of these lists (on average, 14 genes) makes them an ideal tool for researchers to narrow down potential genes for further study, compared to the large, interconnected networks typically generated by gene coexpression networks. In addition, the ranking system incorporated into lists gives them direction, and displays which genes in the common clusters had the highest correlation to the target gene.

The methods described here are the result of an attempt to combine knowledge-based and pure mathematical microarray analyses. Had a purely mathematical method been employed, an average of 61% of the important functional genes identified by the keyword analysis would have been removed from the dataset due to high FDR values. With a purely knowledge-based approach, unstudied genes with unknown functions would have been removed from the dataset, but instead are included due to the high intensity and FDR steps of this analysis. This study attempts to find a reasonable compromise between keeping important functional genes and removing highly variable genes, and the parameters used to filter the dataset at each step have been standardized as much as possible.

These cluster lists are a valuable tool for identifying novel gene targets, because they allow researchers to cluster key genes to important functional genes as well as to unstudied genes. By combining the K-means and MCL clusters, the size of the gene list is reduced by 60% (narrowing the search for novel genes), the gene enrichment scores are increased, and functionally important gene ontology categories are enriched, indicating a greater degree of coexpression in the common cluster lists.

2.7 Chapter 2 Tables

Table 2.1: The conversion of Pearson correlation values to ranks.

Absolute Pearson Correlations					Gene Ranks				
	Gene A	Gene B	Gene C	Gene D		Gene A	Gene B	Gene C	Gene D
Gene A	1	0.2	0.5	0.9	Gene A	1	4	3	2
Gene B	0.2	1	0.3	0.6	Gene B	4	1	4	4
Gene C	0.5	0.3	1	0.8	Gene C	3	3	1	3
Gene D	0.9	0.6	0.8	1	Gene D	2	2	2	1

Table 2.2: Parameters used for saving genes based on high expression levels.

Dataset	Cut-off (Signal Intensity)	Fit to Normal Curve	# Genes Saved	# Genes Remaining
WT - Leaf	11.24	0.880	712	20188
WT - Whole	11.33	0.886	573	20327
Leaf - Whole	11.35	0.880	541	20359

Table 2.3: Parameters used for filtering out highly variable genes in the datasets.

Dataset	Delta	FDR after filtering	# Genes Removed	# Genes Remaining
WT - Leaf	1.05	4.70%	14238	5950
WT - Whole	1.29	4.39%	19493	834
Leaf - Whole	1.28	4.21%	19534	825

Table 2.4: Parameters used for filtering out genes with low fold changes in each of the datasets.

	Cut-off Between				
Dataset	(Low)	(High)	Fit to Normal Distribution	# Genes Removed	# Genes Remaining
WT - Leaf	0.7019	1.316	0.805	5699	251
WT - Whole	0.759	1.342	0.667	704	130
Leaf - Whole	0.838	1.404	0.684	778	47

Table 2.5: Descriptions of gene groups used, including the total number of genes, the percentage of genes in each group representing the different filtering steps, and the enrichment scores calculated by GSEA.

Treat-ment	Cluster Type	Average Number of Genes	Reason for Keeping Gene in List (Average Percentage of Genes in Group)				Average GSEA Enrichment Score
			Import-ant Gene	Key-word	Intens-ity	Low Variability/ High Fold Change	
WT-Leaf	Entire Filtered List	1346	5.3%	23.2%	52.9%	18.6%	
	Ranked Lists (Top 100)	100	7.4%	24.5%	54.2%	13.9%	0.445
	K-Means	40.5	11.0%	28.7%	52.3%	8.0%	0.343
	MCL	25.3	8.3%	27.9%	54.9%	8.9%	0.310
	Common	9.5	13.5%	25.9%	52.7%	8.0%	0.335
WT-Whole	Entire Filtered List	1086	6.5%	28.8%	52.9%	11.8%	
	Ranked Lists (Top 100)	100	7.3%	28.6%	59.3%	4.9%	0.275
	K-Means	45.8	7.3%	30.8%	59.7%	2.2%	0.313
	MCL	34	7.8%	32.6%	56.7%	10.1%	0.344
	Common	16.3	6.1%	35.0%	54.5%	4.4%	0.428
Leaf-Whole	Entire Filtered List	971	7.3%	32.1%	55.7%	4.8%	
	Ranked Lists (Top 100)	100	9.0%	35.8%	54.5%	0.7%	0.255
	K-Means	33.6	9.9%	36.0%	54.1%	0.0%	0.321
	MCL	37.3	11.8%	32.3%	55.8%	0.0%	0.252
	Common	16.1	10.3%	33.7%	56.0%	0.0%	0.338
Average of All Treat-ments	Entire Filtered List	1134.3	6.4%	28.0%	53.8%	11.8%	
	Ranked Lists (Top 100)	100	7.9%	29.6%	56.0%	6.5%	0.325
	K-Means	40.0	9.4%	31.8%	55.4%	3.4%	0.326
	MCL	32.2	9.3%	30.9%	55.8%	6.3%	0.302
	Common	14.0	10.0%	31.5%	54.4%	4.1%	0.367

Table 2.6: Gene Ontology results for all gene groups**.**

		Percentage of Genes in Dataset with known GO functions					
Treatment	Cluster Type	Response to light stimulus	Response to red or far red light	Photomorphogenesis	Red or far red light signaling pathway	Response to blue light	Response to light intensity
WT-Leaf	Full List	11.60%	7.52%	3.58%	1.93%	1.86%	1.79%
	Ranked Lists (Top 100)	13.02%	8.60%	4.00%	2.04%	1.85%	1.33%
	K-Means	19.12%	15.00%	6.46%	4.85%	3.13%	2.07%
	MCL	15.58%	11.53%	5.70%	2.73%	2.81%	1.60%
	Common	18.72%	12.24%	7.36%	2.27%	2.48%	3.91%
WT-Whole	Full List	13.53%	9.11%	4.42%	2.39%	2.21%	2.03%
	Ranked Lists (Top 100)	13.98%	9.05%	3.67%	2.13%	2.98%	2.34%
	K-Means	15.99%	11.78%	4.82%	2.21%	3.97%	1.92%
	MCL	12.85%	10.02%	3.94%	2.62%	2.87%	2.23%
	Common	21.40%	18.03%	6.53%	7.22%	6.30%	2.37%
Leaf-Whole	Full List	13.43%	10.96%	5.01%	3.60%	1.79%	1.13%
	Ranked Lists (Top 100)	20.37%	12.96%	5.56%	4.63%	0.93%	0.93%
	K-Means	12.38%	10.04%	4.59%	3.46%	1.79%	1.13%
	MCL	14.68%	12.07%	6.33%	4.33%	2.07%	1.29%
	Common	13.29%	10.88%	6.08%	4.25%	1.58%	1.78%
Average of All Treatments	Full List	15.16%	9.86%	4.52%	2.98%	1.67%	1.58%
	Ranked Lists (Top 100)	13.13%	9.23%	4.09%	2.54%	2.21%	1.60%
	K-Means	16.59%	12.95%	5.87%	3.80%	3.05%	1.76%
	MCL	13.91%	10.81%	5.24%	3.20%	2.42%	1.87%
	Common	18.43%	13.38%	6.14%	4.16%	3.48%	2.62%
All *Arabidopsis* Genes		1.84%	0.65%	0.23%	0.16%	0.19%	0.32%

Table 2.7: A fully annotated gene list, showing the genes most closely related to the light regulated gene SPA1, based on both MCL and K-means clustering.

SPA1 K means and MCL Common List - WT-Leaf

10 genes

Gene Rank	Impor-tant Genes	Annotation	Gene Description	Gene Identifier	Keywords	Reason to Keep	Fold Change
1	SPA1	SPA1 (SUPPRESSOR OF PHYA-105 1)	Encodes a member of the SPA (suppressor of phyA-105) protein family (SPA1-SPA4). SPA1 is a PHYA signaling intermediate, putative regulator of PHYA signaling pathway. Light responsive repressor of photomorphogenesis...	AT2G46340	SPA1 (Gene Group), PhyA, Photomorphogenesis, Photo, Light, Kinase	Import-ant Gene	0.96
3		glycine-rich protein	glycine-rich protein; FUNCTIONS IN: molecular_function unknown; INVOLVED IN: biological_process unknown…	AT4G29020		Intensity	1.71
5		CAM7 (CALMODULIN 7); calcium ion binding	EF hand domain protein encodes a calmodulin. Can functionally complement a yeast CaM mutant.	AT3G43810	Photomorphogenesis, Photo	Keyword	1.18
10	Transducin	transducin family protein / WD-40 repeat family protein	transducin family protein / WD-40 repeat family protein; FUNCTIONS IN: nucleotide binding; INVOLVED IN: G-protein coupled receptor protein signaling pathway…	AT3G49660	SPA1 (Gene Group), Chloroplast	Import-ant Gene	1.16
13		ATOZI1 (ARABIDOPSIS THALIANA OZONE-INDUCED PROTEIN 1)	Putative pathogenesis-related protein whose transcript level is induced in response to ozone and pathogenic Pseudomonas strains.	AT4G00860		Intensity	1.14
14		EXGT-A1 (ENDOXYLOGLUCAN TRANSFERASE)	endoxyloglucan transferase (EXGT-A1) gene	AT2G06850	Light, Chloroplast, Auxin	Intensity	0.95
21		ATGRP8/GR-RBP8 (COLD, CIRCADIAN RHYTHM, AND	Encodes a glycine-rich protein with RNA binding domain at the N-terminus. Protein is structurally similar to proteins	AT4G39260	Chloroplast	Intensity	1.51

		RNA BINDING 1, GLYCINE-RICH PROTEIN 8)	induced by stress in other plants. Gene expression is induced by cold. Transcript undergoes circadian oscillations that are depressed by overexpression of AtGRP7...				
23		ATARF/ATARF1/ATARFA1A (ADP-RIBOSYLATION FACTOR 1)	Gene encoding ADP-ribosylation factor and similar to other ARFs and ARF-like proteins. A member of ARF GTPase family. Arabidopsis has 21 known members...	AT1G23490		Intensity	1.05
42	SPA-Like & Transducin	AT4G18900 and AT4G18905; transducin family proteins / WD-40 repeat family proteins	AT4G18900: transducin family protein / WD-40 repeat family protein; FUNCTIONS IN: nucleotide binding; INVOLVED IN: biological_process unknown...	AT4G18900;AT4G18905	SPA1 (Gene Group),	Important Gene	0.76
117		SHB1 (SHORT HYPOCOTYL UNDER BLUE1)	SHB1 encodes a nuclear and cytosolic protein that has motifs homologous with SYG1 protein family members...	AT4G25350	Phytochrome, Light	Keyword	1.53

Table 2.8: A fully annotated gene list, showing the genes most closely related to the light regulated gene HY5, based on both MCL and K-means clustering.

HY5 K means and MCL Common List - WT-Leaf							
17 genes							
Gene Rank	**Important Genes**	**Annotation**	**Gene Description**	**Gene Identifier**	**Keywords**	**Reason to Keep**	**Fold Change**
1	HY5	HY5 (ELONGATED HYPOCOTYL 5); DNA binding / TF	Basic leucine zipper (bZIP) transcription factor. Nuclear localization. Mutant studies showed that the gene product is involved in the positive regulation of the PHYA-mediated inhibition of hypocotyl elongation…	AT5G11260	HY5 (Gene Group), PhyA, Photomorphogenesis, Photo, …	Important Gene	2.616
3		60S ribosomal protein L30 (RPL30B)	60S ribosomal protein L30 (RPL30B); FUNCTIONS IN: structural constituent of ribosome…	AT1G77940		Intensity	0.602
4		PROPEP4 (Elicitor peptide 4 precursor)	Elicitor peptide 4 precursor (PROPEP4); FUNCTIONS IN: molecular_function unknown…	AT5G09980		Low variability, high fold change	4.25
5		LP1 (nonspecific lipid transfer protein 1)	Non-specific lipid transfer protein. Binds calmodulin in a Ca2+-independent manner. Localized to the cell wall…	AT2G38540		Intensity	0.484
13		29 kDa ribonucleoprotein, chloroplast, putative	Encodes a chloroplast RNA binding protein. A substrate of the type III effector HopU1 (mono-ADP-ribosyltransferase). Protein is tyrosine-phosphorylated…	AT2G37220	Chloroplast	Intensity	0.562
15		MT1C (metallothionein 1C)	One of the five metallothioneins genes identified in Arabidopsis. MTs are cysteine-rich proteins required for heavy metal tolerance.	AT1G07610		Intensity	1.446
32		60S ribosomal protein L15 (RPL15B)	60S ribosomal protein L15 (RPL15B); FUNCTIONS IN: structural constituent of ribosome; INVOLVED IN: translation; LOCATED IN: cytosolic large ribosomal subunit, ribosome, nucleolus, membrane…	AT4G17390		Intensity	0.594
46		60S ribosomal	60S ribosomal protein L18A	AT2G		Intens-	0.589

	protein L18A (RPL18aB)	(RPL18aB); FUNCTIONS IN: structural constituent of ribosome biogenesis…	34480		ity	
50	OST1/P44/SNRK2-6/SRK2E (OPEN STOMATA 1, SNF1-RELATED PROTEIN KINASE 2.6)	Encodes calcium-independent ABA-activated protein kinase, a member of SNF1-related protein kinases (SnRK2) whose activity is activated by ionic (salt) and non-ionic (mannitol) osmotic stress…	AT4G33950	Light, Signalling, Kinase	Keyword	2.424
52	ATRPL23A (RIBOSOMAL PROTEIN L23A)	Encodes a 60S ribosomal protein L23aA (AtrpL23aA). Paralog of RLPL23aB.	AT2G39460	Light	Intensity	0.527
96	similar to unknown protein (TAIR:AT4G33780.1)	Encodes a nuclear localized serine-arginine-aspartate-rich protein that acts as a negative regulator of photomorphogenesis.	AT1G69935	Photomorphogenesis, Photo	Keyword	0.818
153	unknown protein	unknown protein; FUNCTIONS IN: molecular_function unknown; INVOLVED IN: biological_process unknown…	AT2G25510		Low variability, high fold change	0.194
165	terpene synthase/cyclase family protein	terpene synthase/cyclase family protein; FUNCTIONS IN: lyase activity, magnesium ion binding; INVOLVED IN: metabolic process; LOCATED IN: chloroplast; EXPRESSED IN: flower, root;	AT5G48110	Chloroplast	Low variability, high fold change	4.824
181	unknown protein	unknown protein; FUNCTIONS IN: molecular_function unknown; INVOLVED IN: biological_process unknown…	AT1G56660		Low variability, high fold change	5.998
197	CWLP (CELL WALL-PLASMA MEMBRANE LINKER PROTEIN)	cell wall-plasma membrane linker protein homolog (CWLP)	AT3G22120		Intensity	0.314
253	PORB (PROTOCHLOROPHYLLIDE OXIDOREDUCTASE B)	light-dependent NADPH:protochlorophyllide oxidoreductase B	AT4G27440	Chlorophyll, Light, Chloroplast	Keyword	0.46
360	60S ribosomal protein L4/L1 (RPL4D)	60S ribosomal protein L4/L1 (RPL4D); FUNCTIONS IN: structural constituent of ribosome; INVOLVED IN: translation	AT5G02870	Chloroplast	Intensity	0.628

2.8 Chapter 2 Figures

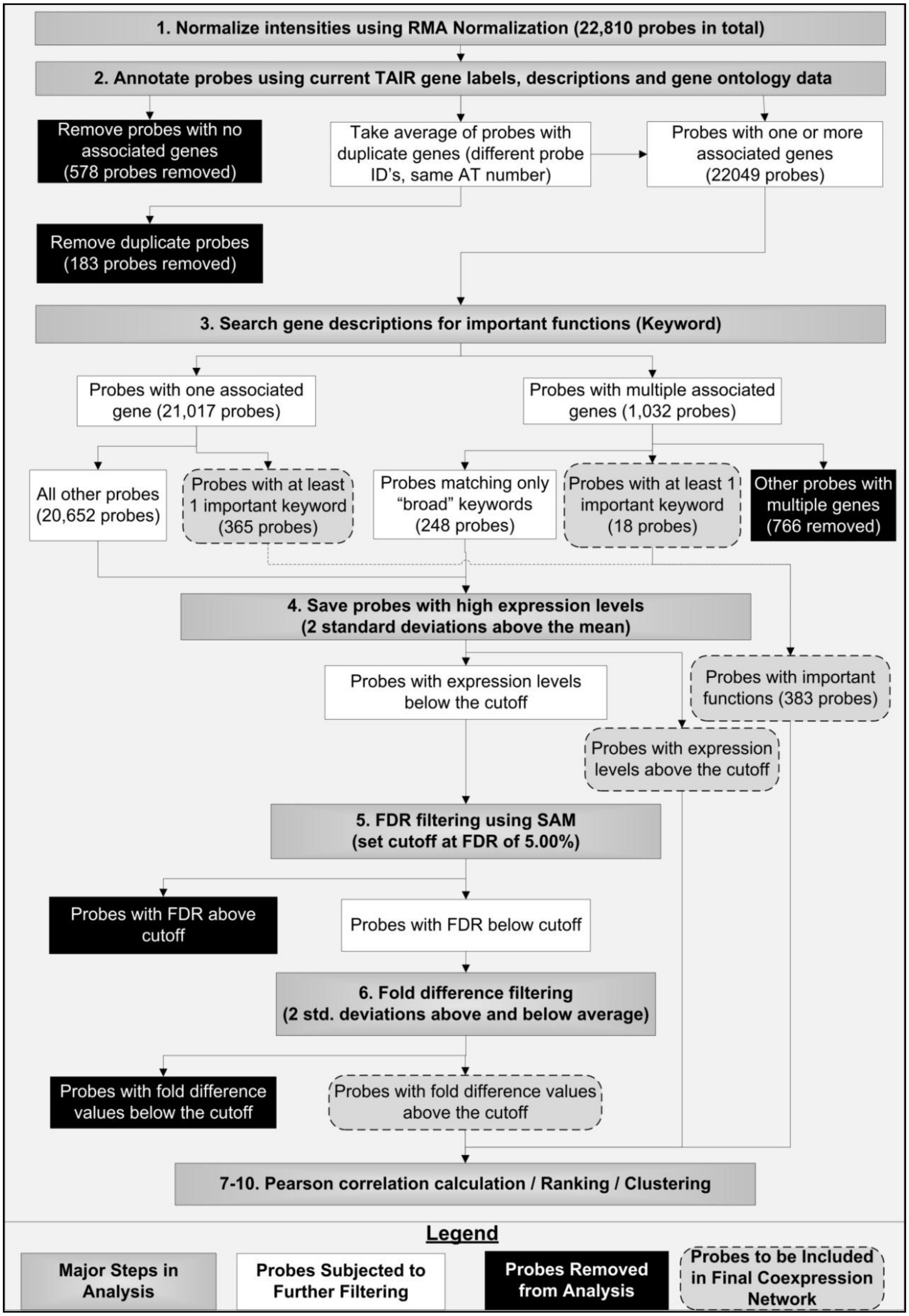

1. Normalize intensities using RMA Normalization (22,810 probes in total)
2. Annotate probes using current TAIR gene labels, descriptions and gene ontology data
Remove probes with no associated genes (578 probes removed)
Take average of probes with duplicate genes (different probe ID's, same AT number)
Probes with one or more associated genes (22049 probes)
Remove duplicate probes (183 probes removed)
3. Search gene descriptions for important functions (Keyword)
Probes with one associated gene (21,017 probes)
Probes with multiple associated genes (1,032 probes)
All other probes (20,652 probes)
Probes with at least 1 important keyword (365 probes)
Probes matching only "broad" keywords (248 probes)
Probes with at least 1 important keyword (18 probes)
Other probes with multiple genes (766 removed)
4. Save probes with high expression levels (2 standard deviations above the mean)
Probes with expression levels below the cutoff
Probes with important functions (383 probes)
Probes with expression levels above the cutoff
5. FDR filtering using SAM (set cutoff at FDR of 5.00%)
Probes with FDR above cutoff
Probes with FDR below cutoff
6. Fold difference filtering (2 std. deviations above and below average)
Probes with fold difference values below the cutoff
Probes with fold difference values above the cutoff
7-10. Pearson correlation calculation / Ranking / Clustering
Legend
Major Steps in Analysis
Probes Subjected to Further Filtering
Probes Removed from Analysis
Probes to be Included in Final Coexpression Network

Figure 2.1: A flowchart describing the network filtering procedure used. “AT numbers” are gene identifiers provided by TAIR. Grey rectangles (numbered) represent major steps in the analysis, and correspond to the steps in section 2.3. Black rectangles represent groups of probes or genes which were removed from the dataset. White rectangles represent groups of genes or probes which are retained in the dataset, but are subject to subsequent filtering steps. Dotted rectangles represent groups of genes which are retained in the final dataset and are not subject to further filtering.

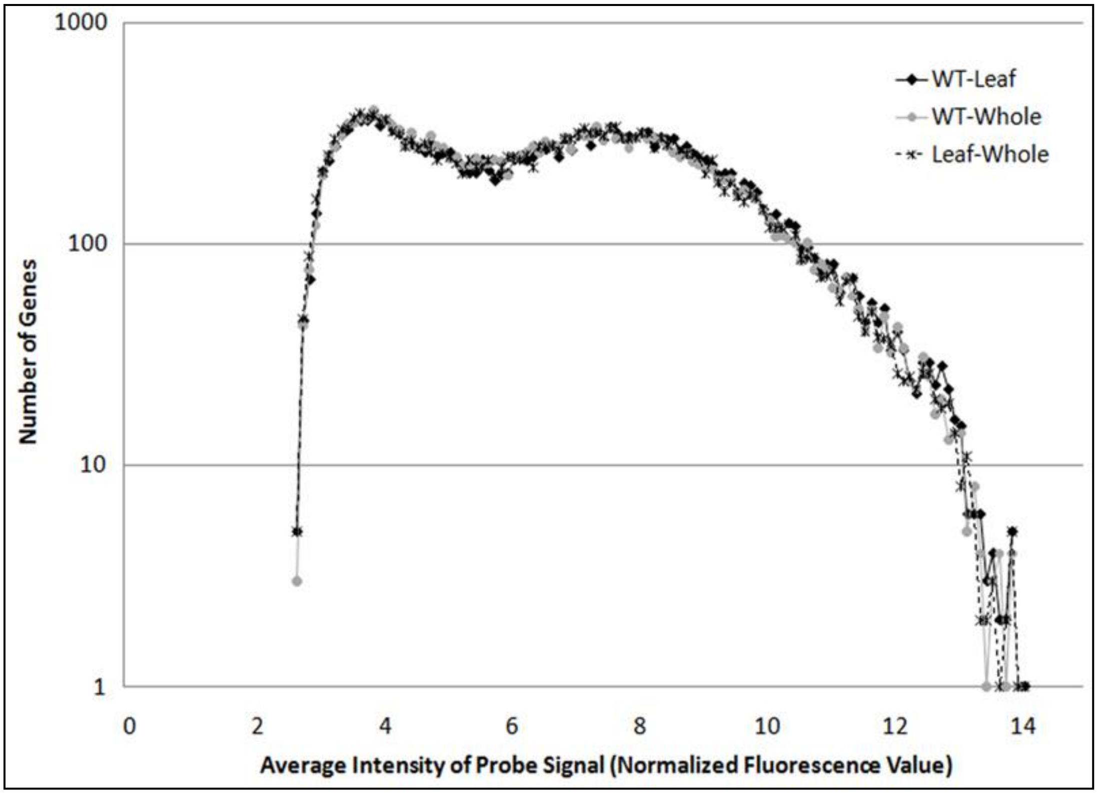


Figure 2.2: The frequency distribution of probe intensities for the three datasets. Black diamonds and the black line represent genes from the WT-Leaf dataset, grey circles and the grey line represent genes from the WT-Whole dataset, and stars and the dashed line represent genes from the Leaf-Whole dataset.

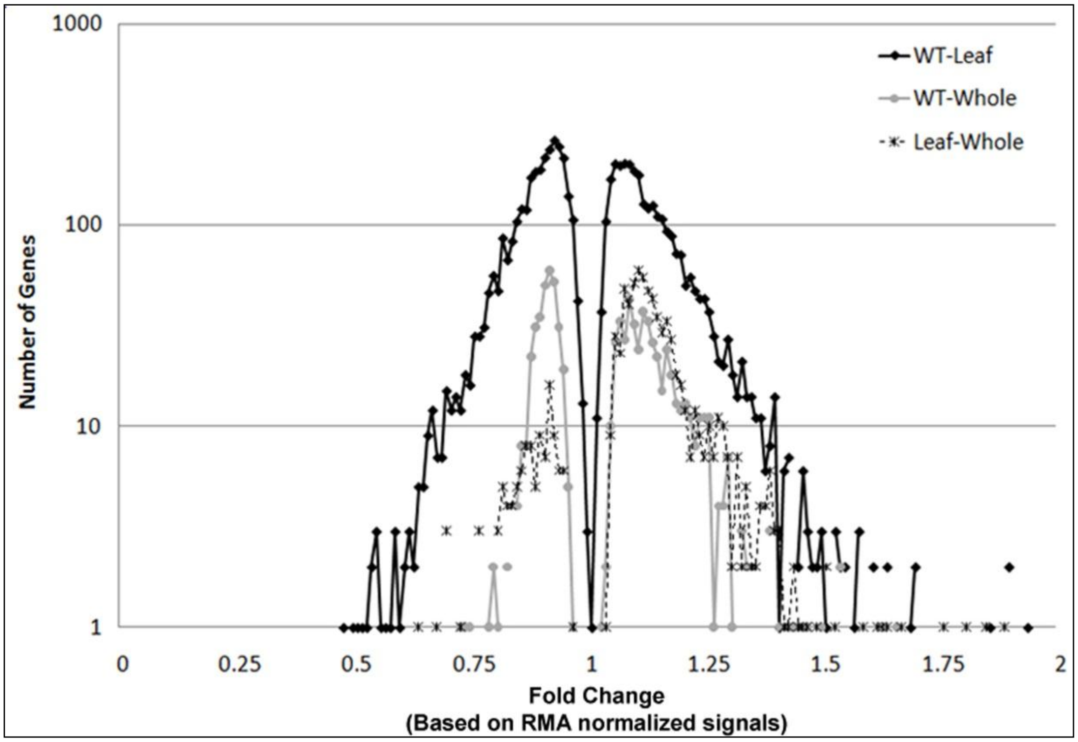


Figure 2.3: The frequency distribution of fold differences for the three datasets. Black diamonds and the black line represent genes from the WT-Leaf dataset, grey circles and the grey line represent genes from the WT-Whole dataset, and stars and the dashed line represent genes from the Leaf-Whole dataset.

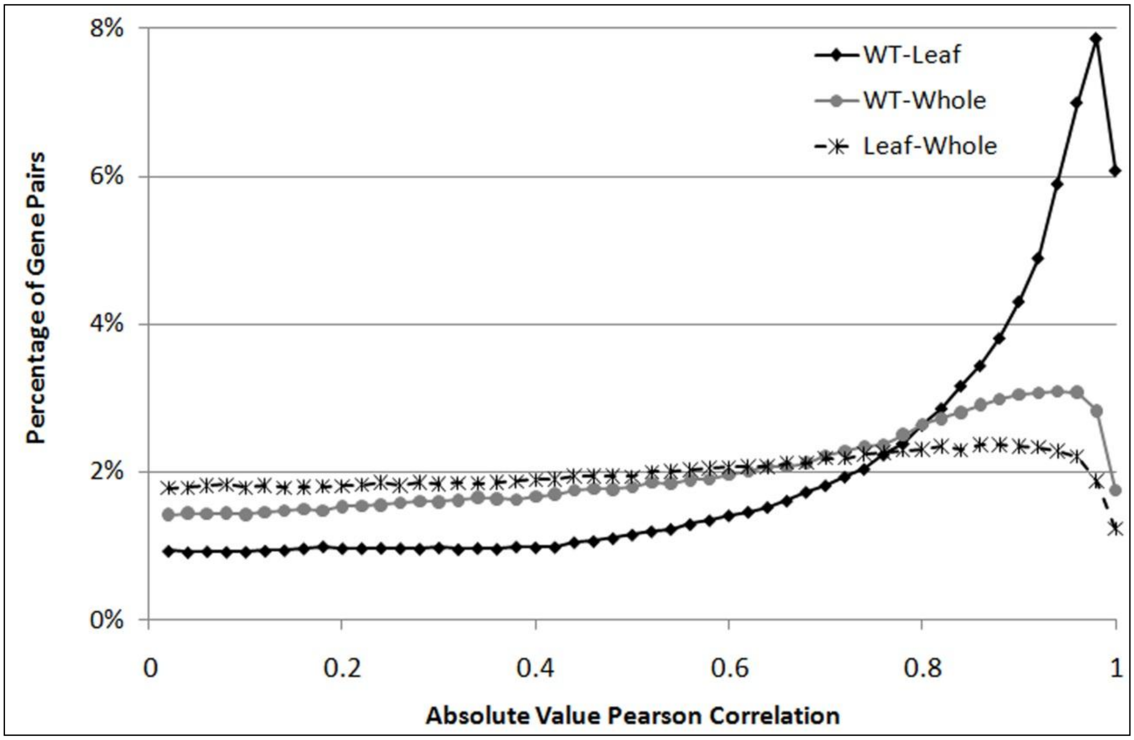


Figure 2.4: The frequency distribution of Pearson correlations between each gene pair in each dataset. Black diamonds and the black line represent genes from the WT-Leaf dataset, grey circles and the grey line represent genes from the WT-Whole dataset, and stars and the dashed line represent genes from

the Leaf-Whole dataset.

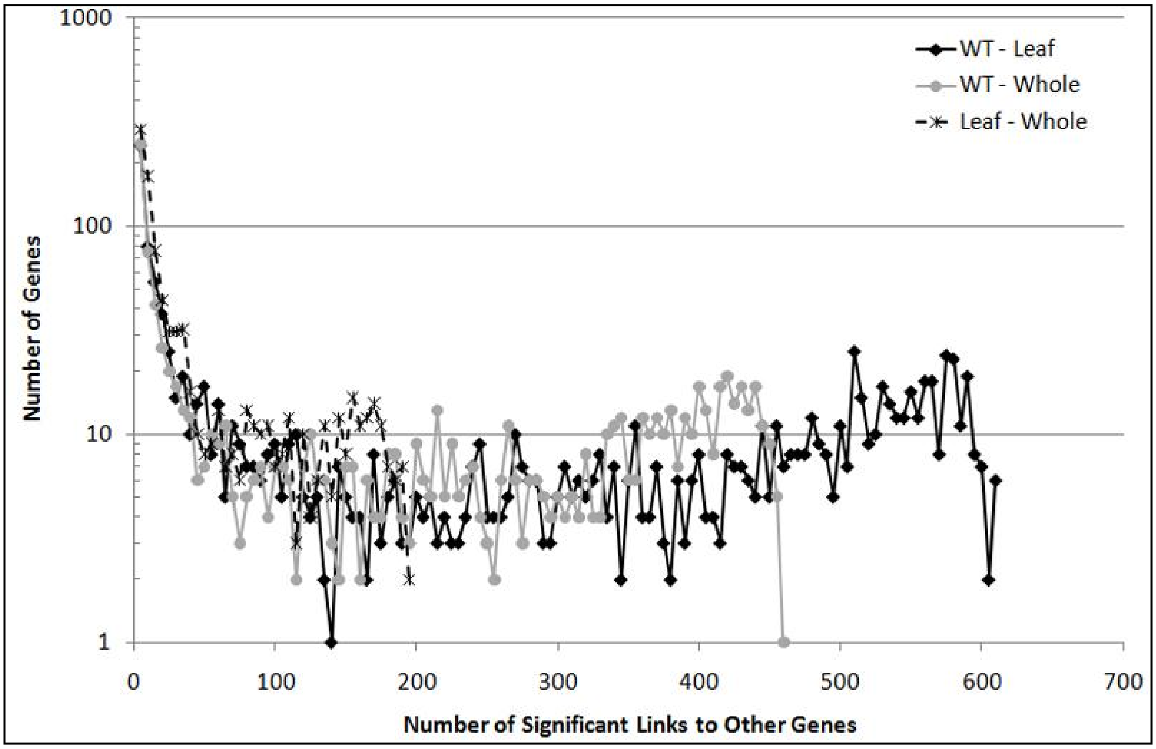


Figure 2.5: The frequency distribution of gene connectivity in each of the three datasets. Black diamonds and the black line represent genes from the WT-Leaf dataset, grey circles and the grey line represent genes from the WT-Whole dataset, and stars and the dashed line represent genes from the Leaf-Whole dataset.

Chapter 3: Frequency-based time-series gene expression recomposition using PRIISM

This chapter has been accepted for publication pending minor revisions, but has not yet been assigned an issue or page number:

Rosa, B.A., Jiao, Y., Oh, S., Montgomery, B.L., Qin, W., Chen, J. (2012a) Frequency-Based Time-Series Gene Expression Recomposition using PRIISM. BMC Systems Biology, In Press.

3.1 Abstract

Circadian rhythm pathways influence the expression patterns of as much as 31% of the *Arabidopsis* genome through complicated interaction pathways, and have been found to be significantly disrupted by biotic and abiotic stress treatments, complicating treatment-response gene discovery methods due to clock pattern mismatches in the fold change statistic. The PRIISM (Pattern Recomposition for the Isolation of Independent Signals in Microarray data) algorithm outlined in this paper is designed to separate pattern changes induced by different forces, including treatment-response pathways and circadian clock rhythm disruptions.

Using the Fourier transform, high-resolution time-series microarray data is projected to the frequency domain. By identifying the clock frequency range from the core circadian clock genes, we separate the frequency spectrum to different sections containing treatment-frequency (representing up- or down-regulation by an adaptive treatment response), clock-frequency (representing the circadian clock-disruption response) and noise-frequency components. Then, we project the components' spectra back to the expression domain to reconstruct isolated, independent gene expression patterns representing the effects of the different influences.

By applying PRIISM on a high-resolution time-series *Arabidopsis* microarray dataset under a cold treatment, we systematically evaluated our method using maximum fold change and principal component analyses. The results of this study showed that the ranked treatment-frequency fold change

results produce fewer false positives than the original methodology, and the 26-hour timepoint in our dataset was the best statistic for distinguishing the most known cold-response genes. In addition, six novel cold-response genes were discovered. PRIISM also provides gene expression data which represents only circadian clock influences, and may be useful for circadian clock analysis studies.

PRIISM is a novel approach for overcoming the problem of circadian disruptions from stress treatments on plants. PRIISM can be integrated with any existing analysis approach on gene expression data to separate circadian-influenced changes in gene expression, and it can be extended to apply to any organism with regular oscillations in gene expression patterns across a large portion of the genome.

3.2 Introduction

Differential gene expression studies typically use the fold change statistic (the ratio of mRNA quantities between two samples) as input, and have been used to discover genes involved in adaptive stress responses which have not been previously characterized (i.e., “novel genes”) (Cui and Churchill, 2003). Specifically, to correct for changes in gene expression induced by non-treatment related influences, fold-change values for time-series data are usually calculated using treatment and control data at every timepoint (Cui and Churchill, 2003). One of the major factors causing gene oscillations under control conditions are molecular circadian clock pathways, which influence physiology and metabolism in preparation for predictable changes in light and temperature (Adams and Carre, 2011). However, a wide range of biotic and abiotic stress treatments have been shown to disrupt rhythmic clock patterns through amplitude changes or phase shifts (Bieniawska et al., 2008; Bilgin et al., 2010; Chaves et al., 2009; Espinoza et al., 2010; Michael et al., 2008; Nakamichi et al., 2009), resulting in significant fold changes for genes which are clock-influenced but are not involved in direct stress response. Figure 3.1 demonstrates that genes can be differentially regulated due to direct stress responses (I), indirectly differentially regulated through disruption of clock pathways induced by the stress (II) or a combination of both (III). Additional complications in regulation patterns arise from the complexity of transcription factor pathways, in which targets may be regulated by clock components directly or through interactions with their transcription factors (Fig. 3.1). For this reason, novel treatment-response gene discovery methods are complicated by the disruption of synchronization of the circadian rhythm pathways, but this complexity is not reflected in existing methods including fold change studies, clustering analysis approaches, and more complex time-serial-based algorithms (Adams and Carre, 2011; Bieniawska et al., 2008; Chiappetta et al., 2004; Cui and Churchill, 2003; Dejean et al., 2007; Ernst et al., 2005; Espinoza et al., 2008; Hestilow and Huang, 2009; Koenig and Youn, 2011; Michael et al., 2008; Schliep et al., 2004; Syeda-Mahmood, 2003; Verducci et al., 2006).

Biological approaches such as the use of constant light and clock component genetic knockout mutants are applied in order to attempt to remove the influences of the circadian clock on target gene expression. However, constant light is an unnatural condition which reduces the applicability of the results, because natural biotic and abiotic genetic stress-response patterns depend on the time-of-day (the point in the light/dark cycle) at which the treatment is applied (Bieniawska et al., 2008; Morker and Roberts, 2011; Salome et al., 2008). Likewise, the use of genetic knockout mutants of circadian clock genes can reduce disruptions due to circadian input, but since stress response genes may be regulated by clock components, the results of such a study are also difficult to interpret (Dong et al., 2011; Espinoza et al., 2010; Morker and Roberts, 2011).

Most existing computational approaches for studying differential gene expression in microarray datasets involve clustering algorithms designed to group genes with similar expression profiles, with the goal of identifying potential annotations for unknown genes (Chiappetta et al., 2004; Dejean et al., 2007; Ernst et al., 2005; Hestilow and Huang, 2009; Koenig and Youn, 2011; Schliep et al., 2004; Syeda-Mahmood, 2003; Verducci et al., 2006). However, the gene distance measures used by all of these clustering methods are unable to distinguish adaptive-response gene expression patterns from circadian clock disruption gene expression patterns, and so may cluster genes with similar clock influences but very different treatment-response influences. Bar-Joseph et al's (2003) continuous representation model for finding differentially expressed genes in time series microarray datasets (which has been used to find more cell-cycle response genes in yeast than conventional clustering methods) is also unable to filter clock influences from treatment response influences on gene expression patterns (Bar-Joseph et al., 2003b).

Several studies have shown that between 6% and 31% of the *Arabidopsis* genome is influenced by circadian clock genetic components (Edwards et al., 2006; Harmer et al., 2000; Michael et al., 2008), while another study suggests that there are significant baseline circadian oscillations for 100% of the genome (Ptitsyn, 2008). A number of approaches have been developed for analyzing the circadian rhythms of genes in time-series datasets (Lu et al., 2006; Michael et al., 2008; Mockler et al., 2007; Price et al., 2008; Wichert et al., 2004). Fourier analysis (which can be used to identify dominant frequencies in time-series data) has been applied to successfully identify periodic genes by treating time-series microarray datasets as time-domain signals (Bozdech et al., 2003; Rustici et al., 2004; Spellman et al., 1998; Whitfield et al., 2002; Wichert et al., 2004). However, these Fourier analysis methods have not been widely used in differential gene expression study methods, because 1) in existing Fourier analysis applications (Bozdech et al., 2003; Rustici et al., 2004; Spellman et al., 1998; Whitfield et al., 2002; Wichert et al., 2004), a fixed frequency range was used as *a priori* knowledge to discover genes with similar oscillations, but novel genes may have totally different frequency patterns under different treatment conditions and; 2) to accurately capture oscillating rhythms, high resolution time course gene expression data is essential according to *Nyquist sampling theorem* (Marks II, 1991; Price et al., 2008), but such data have not been available until recently.

As the price of running microarrays and RNA-seq chips continues to fall, high-resolution time-series gene expression datasets that contain enough information to identify and characterize circadian-frequency rhythms for every gene are becoming available (Craigon et al., 2004; Hubble et al., 2009; Parkinson et al., 2009). Recently, Espinoza *et al* (2010) produced one such microarray dataset, which measured 16 timepoints covering a 58-hour time period with a cold treatment in *Arabidopsis* (Espinoza et al., 2010). Cold-stress genetic responses in *Arabidopsis* are particularly well-characterized, and have been shown to significantly dampen and phase-shift the oscillations of the core clock genes *CCA1* and *LHY*, which have regulatory influences over some cold-responsive transcription factors, including *CBF1*, *CBF2* and *CBF3* (Dong et al., 2011). Disruption of other major clock components due to cold treatment has also been reported, including constant overexpression of *CAB2* and *CCR2,* and constant underexpression of *CAT3* (Bieniawska et al., 2008; Espinoza et al., 2008). For these reasons, this is an ideal dataset to test whether the PRIISM algorithm is able to separate the strong circadian-clock influences on cold-response genes from treatment-response influences.

In this chapter, we present the PRIISM (Pattern Recomposition for the Isolation of Independent Signals in Microarray data) algorithm to perform novel stress-response gene discovery analyses which correct for differential gene expression patterns induced by the circadian clock. We observe that although core circadian clock gene patterns undergo significant changes in phase and amplitude as a result of stress, they maintain oscillating frequencies which remain similar to each other, and still remain close to the circadian pattern of one cycle per day (Bieniawska et al., 2008). We also observe that stress results in significantly increased average expression levels for stress-response genes (Bieniawska et al., 2008), which are reflected in the low-frequency signals (where one oscillation cycle occurs over the course of several days) for these genes. We assume that although circadian clock influences and adaptive stress-response influences can interact with each other (Fig. 3.1), they still cycle at very different rates from each other (and therefore maintain separate dominant frequency ranges) under stress conditions. Based on these observations, we have developed PRIISM to project gene expression data to the frequency domain using the Fourier Transform, isolate independent signals, and then project them back to the expression domain to reconstruct independent gene expression patterns representing the effects of different genetic influences. PRIISM is capable of separating one gene expression pattern into three distinct gene expression patterns: (1) The treatment-frequency gene expression pattern, which has much of the complicating circadian influences removed, and consequently can be used to more accurately identify differentially regulated genes which are involved in direct treatment response, (2) the clock-frequency gene expression pattern, representing rhythmic patterns with a period of approximately one cycle per day, and (3) the noise-frequency gene expression pattern (Fig. 3.2). By applying PRIISM on a cold-treatment dataset, we demonstrate that it can identify known treatment-response genes with a much lower false-positive rate than the existing methods, and can also identify important regulatory timepoints which are not obvious in the unprocessed data. In addition to improving performance when conducting novel treatment-response gene discovery, PRIISM also provides gene expression data which represent only circadian clock influences, and may be useful for circadian clock analysis studies.

3.3 Methods

A wide range of biotic and abiotic stress treatments have been shown to significantly disrupt the cyclic patterns of core circadian clock genes and their downstream target genes (Bieniawska et al., 2008; Bilgin et al., 2010; Chaves et al., 2009; Espinoza et al., 2010; Michael et al., 2008; Nakamichi et al., 2009). When a stress treatment is constantly applied, adaptive stress-response genes are expected to be differentially regulated, while influences from the circadian clock will cause oscillations in target gene expression patterns. In PRIISM, by projecting the gene expression data to the frequency domain using the Fourier transform (Oran Brigham, 1988), the resulting amplitude spectra peak at different frequencies, caused by these different influences. The Fourier transform is a mainstream signal processing technique that simplifies periodogram analysis by identifying the dominant frequencies in the amplitude spectrum. By distinguishing the clock frequency range from the core circadian clock genes in the frequency domain, we can separate the spectrum to different sections containing treatment-induced, clock-induced and noise-induced influences. Then, we project the amplitude spectra back to the expression domain to reconstruct isolated, independent gene expression patterns representing the effects of different frequency components. This method can be applied to any dataset which has sufficiently high resolution and length to measure frequencies of at least one cycle per day, and which uses a treatment that is applied at a frequency significantly different than the clock frequency.

PRIISM has four steps (Fig. 3.3). In the first step, gene expression data are pre-processed to fit the requirements of the Fourier transform (even timepoint spacing and zero average value (Oran Brigham, 1988)), after which the Fourier transform is performed to produce an amplitude spectrum for every gene (Fig. 3.3A & 3.3B). In the second step, a clock vector that defines the frequency range and the amplitudes of the core circadian clock genes is identified based on the spectra of core circadian clock genes (Fig. 3.3C; Section 3.3.2). In the third step, the clock vector is used to decompose every gene’s spectrum into three components (treatment, clock and noise; Fig. 3.3D). In the final step, the inverse Fourier transform is applied to project each spectrum component back to the expression domain, resulting in three independent expression patterns (Fig. 3.3E & Fig. 3.3F).

3.3.1 Pre-processing and Fourier analysis

Time series gene expression data are often unevenly sampled, and the disruption of clock patterns caused by the treatment varies over time. To be able to apply the Fourier transform (which requires steady and evenly sampled input), pre-processing is required. First, the whole time course is divided into overlapped frames. The size of these frames can be changed depending on the experiment;

If they are too long, then it may be difficult to capture changes over time, and if they are too short, then it is more difficult to capture the treatment-frequency patterns (particularly for low-resolution data). For this experiment, the first time frame is 26 hours long due to the two-hour light period at the start of the time period, and all the other time frames are 24 hours long, starting and ending at each light/dark transition (Fig. 3.4A). Second, within each time frame, the gene expression data is interpolated in order to make the time points evenly sampled, as required by the Fourier transform (Oran Brigham, 1988). After interpolation, the mean of the gene expression data for each gene is shifted to zero (refer to explanation of Equation 3.2). The Fourier transform is then applied on each overlapping time frame individually, and the final expression values for each timepoint are calculated using a weighted average for each time frame, where higher weights are used for expression values near the center of each time frame (Fig. 3.4B).

Fourier analysis is a signal processing technique (Oran Brigham, 1988) for the study of two processes: The Fourier transform (the process of decomposing a signal into a sum of components with different frequencies) and the inverse Fourier transform (the operation of reconstructing the signal from these components). Specifically, the discrete Fourier transform (DFT) and its inverse have been used to transform gene expression signals and to reconstruct the discrete signal, respectively (Tominaga, 2010). The Fourier coefficient of DFT (G_n) measures the contribution of the corresponding frequency component to the original signal and is given in Eq. 3.1 (Oran Brigham, 1988):

$$G_n = \sum_{k=0}^{K-1} g(kT) e^{-i2\pi \frac{n}{NT} k} \qquad n = 0, \ldots, N-1 \tag{3.1}$$

where, $g(kT)$ is the sampled signal of K samples with the sampling interval T; i is the imaginary unit. The frequency of the corresponding component n is denoted as f_n (*i.e.,* $\frac{n}{NT}$), where N is the number of frequency components. The DFT maps a time course signal into the frequency domain by producing a spectrum. An amplitude spectrum (plotted as the amplitude versus frequency) is a common frequency domain representation of the original signal. Fast Fourier transform (FFT) is an efficient algorithm to compute the DFT and its inverse (Oran Brigham, 1988). Because of its popularity, it has been built into most modern analysis tools including MATLAB and R.

The Fourier coefficient of the zero-frequency component (*G_0*), derived from Eq. 3.1 where f_n=0, is shown in Eq. 3.2 as given in (Oran Brigham, 1988):

$$G_0 = \sum_{k=0}^{K-1} g(kT)e^{-i2\pi k*0} = \sum_{k=0}^{K-1} g(kT) \tag{3.2}$$

Note that there is a dominant peak at zero frequency in the spectrum of the expression value, which may bias the identification of the true dominant peak to frequency zero. To avoid such bias, we shift the mean of the time course gene expression values for each gene to zero (and consequently $G_0=0$), leading to the removal of the peak at zero frequency. For example, the mean expression value for the gene shown in Figure 3.3A is reduced from 10.6 to 0, and will be added back proportionally to the reconstructed gene expression values during the inverse Fourier transform methods.

3.3.2 Identification of the circadian clock frequency range

The *Arabidopsis* circadian clock is composed of multiple feedback loops. Three genes, *Circadian Clock Associated 1* (*CCA1*), *Late Elongated Hypocotyl* (*LHY*) and *Timing of CAB Expression 1* (*TOC1*) compose the first and most important feedback loop controlling the circadian clock, while *Pseudo Response Regulators 7* and *9* (*PRR7* and *PRR9*) form a secondary feedback loop with *CCA1* and *LHY*, and a third feedback loop involving *TOC1* is regulated by unknown components (Harmer, 2009; Nakamichi, 2011). It has been found that through these feedback loops, eight core circadian rhythm genes (*CCA1*, *LHY*, *PRR7*, *PRR9*, *ELF4*, *GI*, *LUX* and *TOC1*) and their downstream gene targets regulate a wide range of downstream pathways, including germination, leaf development, organelle morphology, photosynthesis, and cell wall development (Adams and Carre, 2011; Li et al., 1994; Lu and Tobin, 2011; Mas, 2008; Salome et al., 2008; Thines and Harmon, 2011).

The Fourier transform is performed on these eight core circadian genes (Fig. 3Ci). The frequency components with relative amplitudes greater than 0.7 (corresponding to half of the maximum value in the spectra) are chosen as dominant frequencies (Sinclair and Dunton, 2007). We define the union of these eight sets of dominant frequencies as Circadian Clock Frequency Range (CCFR), noted as [f_{c_min}, f_{c_max}], where f_{c_min} is the lowest frequency, and f_{c_max} is the highest frequency (Fig. 3Ci). Note that in this example, the dominant clock frequency is significantly lower than one cycle per day, due to the stress-induced disruption of clock patterns. The weight of each frequency component in the CCFR is derived as:

$$w_n = \frac{\sum_{m=1}^{8} |G_{mn}|^2 - \min(\mathcal{G})}{\max(\mathcal{G}) - \min(\mathcal{G})} \quad n \in [c\_min, c\_max] \tag{3.3}$$

where $|G_{mn}|$ is the magnitude of the Fourier coefficient of the n_{th} frequency component for the m_{th} core circadian gene, $\mathcal{G} = \left\{ \sum_{m=1}^{8} |G_{mc\_min}|^2, \sum_{m=1}^{8} |G_{m(c\_min+1)}|^2, \ldots, \sum_{m=1}^{8} |G_{mc\_\max}|^2 \right\}$ is the set of the summed power of eight core clock genes present at each frequency component within the Circadian Clock Frequency Range (CCFR), and w_n is the weight for the frequency component at frequency f_n. The vector $\{w_{c\_min}, w_{c\_min+1}, \ldots, w_{c\_max}\}$ defines the gain-frequency response of a tapering bandpass filter within the CCFR.

3.3.3 Signal decomposition and recomposition

We apply Fourier analysis on each gene, producing the relative amplitude spectrum from which we identify three distinct sections defined according to the CCFR: Treatment-frequency, clock-frequency and noise-frequency components (Fig. 3.3Di). For the treatment-frequency decomposition, given a relatively narrow frequency band, we used a low pass filter with a steep cut-off frequency to gain the optimal balance between removing ringing artifact and approximating desired frequency response (Chatterjee et al., 2009). Fourier coefficients of the clock components of each gene are modulated by the weight of the corresponding frequency components, as given by Eq. 3.4:

$$\hat{G}_c = w_c G_c \quad c \in [c\_min, c\_max] \tag{3.4}$$

The tapering filtering results in clock-frequency expression patterns that are noise-reduced and contain less artifacts caused by a discontinuity in the filter function. The reconstructed high frequency expression pattern is considered to be noise, and it is not studied in this paper. Therefore, we simply applied an ideal high pass filter. The Fourier coefficients of the treatment-frequency components and the noise-frequency components are not modulated. The reweighted spectra used for the signal reconstruction of the three frequency components sections are shown in Figure 3.3Dii.

The inverse discrete Fourier transform (IDFT) is calculated according to Eq. 3.5 (Oran Brigham, 1988):

$$g(kT) = \frac{1}{N}\sum_{n=0}^{N-1} G_n e^{i2\pi \frac{n}{NT}k} \quad k = 0, \ldots, K-1 \tag{3.5}$$

The inverse Fourier transform is performed on the full spectrum, including the filtered spectra for each gene. Similar to using the clock vector as a tapering band-pass filter to remove noise, we added a course graining process to make sure there is no overlapping between any of the two frequency bands, which may increase the robustness of component selection. The mean of the original gene expression values (which was removed in the pre-processing step), is added back proportionally to each gene expression curve based on the amplitude distribution of each component in the spectra before shifting the mean (Fig. 3.3F), according to Eq. 3.6:

$$g'_L(kT) = g_L(kT) + \frac{\sum_{k=0}^{K-1} g(kT)}{K} \times \frac{\sum |G_L|^2}{\sum |G_n|^2} \quad k = 0, \ldots, K-1$$

(3.6)

where $g'_L(kT)$ is the treatment expression level at timepoint kT for a given gene, $g_L(kT)$ is result of inverse discrete Fourier transform (Eq. 3.5) on treatment frequency at timepoint kT, and G_L is the Fourier coefficient of treatment frequency component in frequency range $[0,f_{c\_min}\text{-}1]$. Similarly, we compute the clock expression level $g'_C(kT)$ and noise expression level $g'_N(kT)$.

Note that because the entire warm and cold gene expression datasets are mean-shifted based on their relative amplitudes in each component, the reconstructed time-zero fold change values may not necessarily be equal to zero (Fig. 3.2B).

3.4 Results / Discussion

This study analyzes an *Arabidopsis* Affymetrix ATH1 microarray dataset (containing 22,810 probes) generated by Espinoza *et al.* (2010), which consists of 16 timepoints collected over the course of 58 hours in both warm (20°C) and cold (4°C) conditions under a 16-hour light / 8-hour dark cycle starting at ZT14 (14 hours after dawn) (Espinoza et al., 2010). This dataset was chosen for the analysis because it has separate control and treatment arrays, it has sufficiently high resolution (sampled at 2 hours and every 4 hours after that), and cold is a well-studied treatment in *Arabidopsis* (Bieniawska et al., 2008; Dong et al., 2011; Espinoza et al., 2008; Espinoza et al., 2010; Fowler et al., 2005; Lee et al., 2005).

Gene expression data was RMA normalized using the “affylmgui” program available as part of the *Bioconductor* software package and annotated using annotation data available from TAIR (version 10, available ftp://ftp.arabidopsis.org/Genes/TAIR10_genome_release/). The gene expression data were interpolated to every 2 hours using B-spline regression, and were segmented into four overlapping gene expression time frames (from both the warm and cold treatments), which were combined using a weighted average (Fig. 3.4) (Bar-Joseph et al., 2003c; Smith and Craven, 2008). PRIISM was applied on this “original” dataset, resulting in three independent and isolated gene expression datasets (treatment-frequency, clock-frequency and noise-frequency).

3.4.1 Treatment-response gene discovery

In order to show the advantage of PRIISM, we compared the treatment-frequency dataset to the original dataset in terms of their ability to identify known cold-response genes using maximum fold changes and principal component analysis. Fold change values were calculated by subtracting the logged gene expression value in the warm from the logged gene expression value in the cold at every timepoint. Lists of *Arabidopsis* genes upregulated by cold treatment when grown on agar plates or grown in soil were collected from a previous study by Vogel et al (Vogel et al., 2005). The 302 cold-upregulated genes found in the intersection of these lists were used to define the set of “cold standard” (COS) upregulated genes. Receiver-Operator-Characteristic (ROC) curves (which have been shown to be an effective method for evaluating gene expression data (Parodi et al., 2003)) were generated for these COS-upregulated genes (Vogel et al., 2005) by distinguishing each ranked gene as either a true positive or a false positive (Fig. 3.5). A larger area under an ROC curve indicates that more COS-upregulated genes are identified. The line at which the number of true positives is equal to the number of false positives is indicated in Figure 3.5, and only the data above this line are considered biologically relevant. By ranking genes by their maximum fold change values in the treatment-frequency dataset, 52.6% (159/302) of known COS upregulated genes can be identified, compared to only 21.2% (64/302) in the original dataset (Table 3.1) (Vogel et al., 2005). This difference may be explained by the disruptions contributed by the clock-frequency influences and the noise-frequency influences, which are present in the original dataset. This shows that more COS-upregulated genes can be identified by ranking by the maximum fold change in the treatment-frequency dataset compared to the original dataset.

Principal component analysis (PCA) is a linear component composition method that has been applied to summarize different gene expression influences under different conditions, and consequently has been used for differential gene expression studies in microarray datasets (Raychaudhuri et al., 2000). PCA was performed on the original dataset (Fig. 3.6A), and the Euclidean distance from the bottom-left of the PCA plot of the first and second component was used to rank the genes, because the cold genes were biased towards lower values in these components (Fig. 3.6) allowing for the construction of an ROC curve based on this data (Fig. 3.5). These data show that only 13.9% (42/302) of the cold upregulated genes can be identified in the original PCA plot. The first PCA components of the treatment-frequency data and the clock-frequency data were also plotted (Fig. 3.6B) and ranking based on Euclidean distance from the bottom-right was able to identify 46.0% (139/302) of the COS-upregulated genes.

These results showed that, in both maximum fold change and PCA analyses, the ranked treatment-frequency fold change results produce fewer false positives than the original methodology by distinguishing more COS-upregulated genes (Table 3.1).

3.4.2 The identification of important gene regulation timepoints using PRIISM

In the previous section, it was shown that gene discovery in the treatment-frequency data produced by PRIISM constantly outperforms the same analyses on the original data. Although these approaches are useful for poorly studied treatment responses, a knowledge-based approach may be used to identify more treatment-response genes with a lower false positive rate.

Cold treatments have been shown to induce the expression of the transcription factors *C-repeat/DRE Binding Factor* genes *CBF1, CBF2* and *CBF3* (Gilmour et al., 1998), which are induced in parallel with the cold transcription factors *RAV1* and *ZAT12* (Fowler et al., 2005). Some of the important targets of *CBF* transcription factors include *Cold-Responsive* (*COR*) genes *COR15A, COR15B*, *COR47,* and *COR78* (Dong et al., 2011; Fowler and Thomashow, 2002; Fowler et al., 2005; Maruyama et al., 2004). All of the cold transcription factors and targets included in these lists have also been shown to be gated by the circadian clock, making them ideal for evaluating PRIISM's ability to remove clock-frequency influences (Bieniawska et al., 2008; Dong et al., 2011; Harmer et al., 2000).

In the treatment-frequency data, a peak in the fold change patterns can be observed in the well-studied cold response transcription factors and cold regulated (COR) response genes at the start of the first night (at approximately 26 hours) (Fig. 3.7C, 3.7D). The peaks of the transcription factors can be seen to occur before the peaks of their target genes, as is expected for a TF-target relationship. By contrast, these peaks are not apparent in the original fold change data (Fig. 3.7A, 3.7B). For this reason, an ROC curve was computed using the fold change value at 26 hours in the treatment-frequency fold change data (Fig. 3.5, Table 3.1). Table 3.1 shows that 194/302 (64.2%) of the true-positive COS-upregulated genes can be identified with a 50% false positive rate in the treatment-frequency 26-hour fold change data, compared to only 64 for the maximum fold change in original data and 42 for the PCA plot of the original data .

This data shows that the fold change value at 26 hours in the treatment-frequency data is the best predictor of whether a gene is involved in adaptive cold response. The top 25 ranked genes based on fold changes at 26 hours in the treatment-frequency dataset are shown in Table 3.2. Included in this table is the "Cold Upregulation Category" for each gene, which indicates whether a gene was upregulated in the cold when plants were grown in soil ("Soil"), on agar plates ("Plate"), on both growth mediums ("COS"), or on neither ("N/A") in Vogel et al's study (Vogel et al., 2005). In this table, 22/25 of the genes belonged to the COS group, two belonged to the "Plate" group, and one belonged to the "Soil" group, suggesting that the PRIISM method has successfully identified known cold-regulation genes (Vogel et al., 2005). As a comparison, Table 3.3 shows the top 25 ranked genes based on fold changes at 26 hours in the original data. Here, only 18/25 of the genes belongs to the COS group, two belonged to the "Plate" group, one belonged to the "soil" group, and four were in neither of the two groups. Also, the core clock genes *CCA1* and *LHY* are ranked 4th and 11th (respectively) in this table, showing that clock influences are strongly influencing the ranking in the original data.

Table 3.4 shows the top 25 ranked genes which were not part of the COS-upregulated gene list in Vogel et al (2005) (Vogel et al., 2005). 10/25 of the genes in this list belonged to the "Soil" group, 9 belonged to the "Plate" group, and 6 were novel genes not identified in Vogel et al's study (Vogel et al., 2005). All of the novel genes (and all but one of the 25 genes in this list) have been previously identified as being involved in cold response in other studies, suggesting that PRIISM has identified a list of very important cold-response genes (See "Comments" column in Table 3.3).

The results of a case study on *ATGolS3* (*AT1G09350*), the gene with the largest fold change in the treatment-frequency data at 26 hours are shown in Figure 3.8. The logged original gene expression curve under warm conditions has a minimum expression level of approximately 6, which is reflected by a flat treatment-frequency expression curve with a nearly constant value of 6 (Fig. 3.8A). The rhythmic pattern of the original data in warm conditions is captured in the clock-frequency gene expression curve, and the sharp peaks and sudden changes in slope are captured in the noise-frequency curve (Fig. 3.8A). The original gene expression data under cold conditions peaks quite strongly during the first night but retains some cyclical expression. The PRIISM-processed gene expression data shows that the treatment-frequency gene expression is constantly higher in the cold, with a peak at 26 hours, while the clock-frequency gene expression data is only marginally increased, but is increased more in the first day than in the second day (Fig. 3.8B and 3.8C). The fold change graph shown in Figure 3.8C indicates that most of the increase in gene expression is due to treatment-frequency influences for this gene, but the clock-frequency influences upregulate the gene more strongly early in the cold treatment. The noise-frequency fold change pattern matches many of the sharp peaks and valleys in the original fold change pattern, suggesting that much of the noise has indeed been removed (Fig. 3.8C).

To test the statistical significance of PRIISM's ability to discover treatment-response genes, P-values were calculated using a Z-test for both the maximum fold change from the original dataset and the fold change values at 26 hours in the treatment-frequency dataset. Figure 3.9 shows the number of genes that were found to be significant (P value<=0.05) in these tests, and how many belonged to the COS-upregulated gene list from Vogel et al (Vogel et al., 2005). Out of the 161 genes significant in the treatment-frequency data at 26 hours, 98 of them (60.9%) were COS upregulated genes, compared to 154 out of 379 (39.3%) for the original dataset.

3.4.3 Clock-frequency data analysis

The clock vectors calculated by Equation 3.3 under both warm and cold conditions for each of the time frames are shown in Figure 3.10. The difference between the length and the shape of the warm and cold vectors indicates the circadian rhythm disruption caused by the cold stress. Figure 3.10A shows drastically different frequency profiles for the warm and cold conditions, caused by an abrupt phase shift in the expression data. The clock genes continue to have disrupted frequencies in the second time frame (Fig 3.10B), but appear to return to normal oscillating frequencies, possibly with different phases, in time frames 3 and 4 (Fig. 3.10C, 3.10D).

To study whether the clock-frequency data produced by PRIISM successfully isolated cyclic clock influences from treatment-response influences, the clock-frequency gene expression patterns of eight well-studied cold response genes were matched with standard clock patterns according to the pattern-matching algorithm HAYSTACK (Mockler et al., 2007). This algorithm (the key component of The Diurnal Project) utilizes a model-based pattern matching algorithm to calculate the phase and cyclic pattern type for each gene in a dataset, and also calculates the correlation of each gene to the closest model, which can be used as an indication of how strong the clock influence is on the gene (Mockler et al., 2007). HAYSTACK provides T-test P-values indicating the probability that an input pattern matches a gene expression model, and provides several types of cyclic clock pattern models to use for comparison (Mockler et al., 2007). This analysis included the COR genes which have been shown to be under circadian clock control under warm conditions, but gated by cold transcription factors (including the CBF genes) under cold conditions (Dong et al., 2011). The results in Table 3.4 indicate that the P values for the clock-frequency gene expression data from PRIISM are substantially lower than the original data (under both warm and cold conditions), often by several orders of magnitude, demonstrating enrichment of clock-frequency gene expression in this data. This more cyclical data may be better for determining period, phase and amplitude characteristics and changes, leading to more informative circadian clock studies (Yang and Su, 2010).

Note that the lowest frequency band of the CCFR is simply discarded in PRIISM. In future work, it will be interesting to further test whether feeding it into the treatment-frequency component will construct more precise results.

3.5 Conclusion

Circadian rhythm pathways influence the expression patterns of as much as 31% of the *Arabidopsis* genome through complicated interaction pathways, and have been found to be significantly disrupted by biotic and abiotic stress treatments, complicating treatment-response gene discovery methods due to clock pattern mismatches in the fold change statistic. The PRIISM algorithm outlined in this paper is designed to separate pattern changes induced by different forces, including treatment pathways and circadian clock rhythm disruptions. By applying PRIISM on a cold-response dataset, we systematically evaluated our method using maximum fold change and PCA analyses. The results of this study showed that the ranked treatment-frequency fold change results produce fewer false positives than the original methodology, and the 26 hour timepoint in the PRIISM produced dataset was the best statistic for distinguishing the most known cold-response genes. In addition, PRIISM also provides gene expression data which represents only circadian clock influences, and may be useful for circadian clock analysis studies. In fact, any existing analysis approach on gene expression data can utilize PRIISM to separate circadian-influenced changes in gene expression. In conclusion, PRIISM is a novel approach for overcoming the problem of circadian disruptions from stress treatments on plants. PRIISM can be integrated with any existing analysis approach on gene expression data to separate circadian-influenced changes in gene expression, and it can be extended to apply to any organism with regular oscillations in gene expression patterns across a large portion of the genome. In future work, when higher resolution datasets become available, we will apply the discrete wavelet transforms (DWT) in order to further enhance the ability of PRIISM to distinguish circadian clock disruption influences from treatment-response pathway influences.

3.6 Chapter 3 Tables

Table 3.1: Summary of ROC analysis for genes upregulated by cold treatment

Statistic	Original Data		Treatment-Frequency Data		
	Maximum Fold Change	PCA Distance	Maximum Fold Change	PCA Distance	Fold Change at 26 Hours
Recall when true positives = false positives	21.2%	13.9%	52.6%	46.0%	64.2%
Number of true positives identified when true positives = false positives (Out of 302 true positives)	64	42	159	139	194

Table 3.2: Genes ranked based on their treatment-frequency fold change values at 26 hours. *"Cold Upregulation Category" indicates whether a gene was upregulated in the cold when plants were grown in soil ("Soil"), on agar plates ("Plate"), on both growth mediums ("COS"), or on neither ("N/A") in Vogel et al's study (Vogel et al., 2005).

Rank	**Annotation**	**P-Value (Treatment-Frequency Fold Change, 26 Hours)**	**Cold Upregulation Category* (Vogel et al., 2005)**
1	*AT1G09350*: Arabidopsis Thaliana Galactinol Synthase 3 (*AtGolS3*)	3.97E-31	COS
2	*AT4G14690*: Early Light-Inducible Protein 2 (*ELIP2*)	3.18E-28	COS
3	*AT4G12470*: Azelaic Acid Induced 1 (*AZI1*)	1.23E-25	COS
4	*AT3G50970*: Low Temperature-Induced 30 (*LTI30*)	3.62E-22	COS
5	*AT1G16850*: Unknown protein	7.38E-22	COS
6	*AT3G22840*: Early Light-Inducible Protein 1 (*ELIP1*)	8.66E-21	COS
7	*AT1G51090*: Heavy-metal-associated domain-containing	1.41E-16	COS
8	*AT3G55580*: Regulator of chromosome condensation (*RCC1*) family protein	3.62E-16	COS
9	*AT5G25110*: *CIPK25* (CBL-Interacting Protein Kinase 25)	1.13E-13	COS
10	*AT5G52310*: *COR78* (Cold Regulated 78)	3.25E-13	COS
11	*AT1G02820*: late embryogenesis abundant 3 family protein / LEA3 family protein	3.29E-13	Soil
12	*AT4G30830*: similar to unknown protein (*AT2G24140.1*)	5.57E-13	COS
13	*AT2G23910*: Cinnamoyl-CoA reductase-related	9.97E-13	Plate
14	*AT1G48100*: Glycoside hydrolase family 28 protein / polygalacturonase (pectinase) family protein	1.93E-12	COS
15	*AT5G17030*: UDP-Glucosyl Transferase 78D3 (*UGT78D3*)	3.07E-12	COS
16	*AT4G33070*: Pyruvate decarboxylase, putative	3.43E-12	COS
17	*AT3G17130*: Invertase/pectin methylesterase inhibitor family protein	4.21E-11	COS
18	*AT1G11210*: Similar to unknown protein (TAIR:*AT1G11220.1*)	1.06E-10	COS
19	*AT1G62570*: Flavin-monooxygenase Glucosinolates-Oxygenase 4 (*FMO GS-OX4*)	3.24E-10	COS
20	AT2G16890: UDP-glucoronosyl/UDP-glucosyl transferase family protein	5.52E-10	COS
21	*AT4G25480*: Dehydration response element B1A (*DREB1A*); C-Repeat Binding Factor 3 (*CBF3*)	6.1E-10	COS
22	*AT1G20440*: Cold-Regulated 47 (*COR47*); (*RD17*)	1.08E-09	COS
23	*AT1G61800*: Glucose-6-Phosphate/Phosphate Translocator 2 (*GPT2*)	1.37E-09	Plate
24	*AT4G17550*: Transporter-related	1.52E-09	COS
25	*AT1G62710*: Beta Vacuolar Processing Enzyme (*BETA-VPE*)	5.78E-09	COS

Table 3.3: Genes ranked based on their original fold change values at 26 hours. *"Cold Upregulation Category" indicates whether a gene was upregulated in the cold when plants were grown in soil ("Soil"), on agar plates ("Plate"), on both growth mediums ("COS"), or on neither ("N/A") in Vogel et al's study (Vogel et al., 2005).

Rank	**Annotation**	**P-Value (Treatment-Frequency Fold Change, 26 Hours)**	**Cold Upregula-tion Category* (Vogel et al., 2005)**
1	*AT1G09350*: *Arabidopsis thaliana* Galactinol Synthase 3 (*AtGolS3*)	1.31E-28	COS
2	*AT4G14690*: Early Light-Inducible Protein 2 (*ELIP2*)	9.41E-26	COS
3	*AT3G22840*: Early Light-Inducible Protein 1 (*ELIP1*)	7.48E-21	COS
4	*AT2G46830*: Circadian clock associated 1 (*CCA1*)	1.95E-18	COS
5	*AT1G02820*: late embryogenesis abundant 3 family protein / LEA3 family protein	1.53E-17	Soil
6	*AT2G16890*: UDP-glucoronosyl/UDP-glucosyl transferase family protein	8.19E-16	COS
7	*AT4G12470*: protease inhibitor/seed storage/lipid transfer protein (LTP) family protein	7.52E-16	COS
8	*AT4G33070*: pyruvate decarboxylase, putative	3.80E-15	COS
9	*AT3G51240*: Flavanone 3-hydroxylase (*F3H*)	4.99E-15	Plate
10	*AT1G10370*: Glutathione S-transferase 30 (*GST30*)	2.57E-12	Plate
11	*AT1G01060*: Late elongated hypocotyl (*LHY*)	2.60E-12	**N/A**
12	*AT1G32900*: starch synthase, putative	4.74E-12	**N/A**
13	AT1G62570: flavin-monooxygenase glucosinolate s-oxygenase 4 (*FMO GS-OX4*)	8.51E-12	COS
14	*AT1G48100*: glycoside hydrolase family 28 protein	3.01E-11	COS
15	*AT5G62210*: embryo-specific protein-related	6.17E-11	COS
16	*AT2G16890*: UDP-glucoronosyl/UDP-glucosyl transferase family protein	8.69E-11	COS
17	*AT4G30830*: Unknown Protein	1.31E-10	COS
18	*AT5G06980*: Unknown Protein	1.33E-10	COS
19	AT5G25110: CBL-interacting protein kinase 25 (*CIPK25*)	1.61E-10	COS
20	*AT1G07180*: Alternative NAD(P)H Dehydrogenase 1 (*ATNDI1*)	2.00E-10	**N/A**
21	*AT3G50970*: Low Temperature-Induced 30 (*LTI30*)	6.33E-10	COS
22	*AT1G73480*: hydrolase, alpha/beta fold family protein	1.17E-09	COS
23	*AT3G55580*: regulator of chromosome condensation (RCC1) family protein	2.45E-09	COS
24	*ATCG00270*: PSII D2 protein	2.39E-09	**N/A**
25	*AT1G51090*: heavy-metal-associated domain-containing protein	3.71E-09	COS

Table 3.4: The top 25 ranked non-COS genes based on treatment-frequency fold change values at 26 hours.

Rank	Annotation	Comments	P Value (Treatment-Frequency Fold Change, 26 Hours)	Cold Upregul-ation Category* (Vogel et al., 2005)
11	*AT1G02820*: Late embryogenesis abundant 3 family protein / LEA3 family protein	LEA family proteins are associated with dehydration stress (and therefore cold) and general environmental stress in plants, and desiccation tolerance in other organisms including bacteria (Hundertmark and Hincha, 2008). Cold response genes *COR15A*, *COR15B* and *COR47* are classified as *LEA* genes. Although not to the same degree as the *COR* genes, expression of this gene was upregulated by cold according to quantitative RT-PCR (Hundertmark and Hincha, 2008)	3.29E-13	Soil
13	*AT2G23910*: Cinnamoyl-CoA reductase-related	Implicated in the biosynthesis of phenylpropanoids (Boerjan et al., 2003; Lacombe et al., 1997), which contribute to many different plant responses to biotic and abiotic stress/challenge (Solecka, 1997)	9.97E-13	Plate
23	*AT1G61800*: *GPT2* (Glucose-6-Phosphate Translocator 2)	A *gpt2* mutant shows an impairment in photosynthetic acclimation in response to shifts to high irradiance light, which can be exacerbated under cold conditions (Athanasiou et al., 2010)	1.37E-09	Plate
29	*AT5G06760*: Late embryogenesis abundant group 1 (LEA group 1) domain-containing protein	Similar to other *LEA* above (*AT1G02820*), expression of this gene is upregulated by cold according to quantitative RT-PCR (Hundertmark and Hincha, 2008)	1.86E-08	Soil
37	*AT3G51240*: *F3H*; *TT6* (Flavanone 3-Hydroxylase; Transparent Testa 6)	Implicated in freezing stress response (Hannah et al., 2006)	7.07E-08	Plate
50	*AT1G60190*: Armadillo/beta-catenin repeat family / U-box domain-containing		1.92E-06	Soil
53	*AT5G24120*: *SIGE/SIG5* (RNA polymerase sigma subunit E); DNA binding / DNA-directed RNA polymerase/ sigma/ transcription factor	Regulated in blue light by cryptochromes and involved in light-dependent regulation of the photosynthetic apparatus (Onda et al., 2008). In a separate study shown to be essential for Arabidopsis (Yao et al., 2003)	3.24E-06	Soil
55	*AT1G10370*: *ATGSTU17/ERD9/* (Early-Responsive to Dehydration 9)	Dehydration responsive (Swarbreck et al., 2008)	4.43E-06	Plate

57	*AT1G32900*: Starch synthase, putative	Identified in a study on light/cold interactions (Soitamo et al., 2008). Upregulated by cold generally, but upregulated more under cold/light conditions than cold/dark	6.47E-06	**Novel**
60	*AT4G33905*: Peroxisomal membrane protein 22 kDa, putative	Upregulated by stress, including cold treatment (Ma and Bohnert, 2007)	7.71E-06	**Novel**
61	*AT1G01520*: Myb family transcription factor	Upregulated in mutant that has improved freezing tolerance (i.e. *esk1* mutant) (Xin et al., 2007)	1.63E-05	**Novel**
63	*AT5G57760*: Unknown		1.95E-05	Plate
70	*AT5G14760*: AO (L-aspartate oxidase)	Involved in the synthesis of NAD (Katoh et al., 2006), which is phosphorylated by cold in other plants (Ruiz et al., 2002)	4.96E-05	**Novel**
71	*AT1G10585*: Transcription factor	Upregulated under conditions associated with oxidative stress/high light (Vanderauwera et al., 2005)	5.66E-05	Soil
72	*AT5G07010*: Sulfotransferase family	Jasmonate responsive (Swarbreck et al., 2008)	5.95E-05	Soil
75	*AT2G22590*: Glycosyltransferase family protein	In the same gene family as *UGT91A1*, (a target of a TF that regulates flavonol synthesis), and is thus proposed to impact flavonol biosynthesis, which is a product associated with cold response (Korn et al., 2008; Stracke et al., 2007)	6.55E-05	Plate
76	*AT3G17609*: *HYH* (HY5-Homolog); DNA binding / transcription factor	Involved in phyB signaling (Jonassen et al., 2008); Required for low temperature-induced anthocyanin accumulation (Zhang et al., 2011)	6.78E-05	**Novel**
81	*AT1G17170*: *ATGSTU24* (*Arabidopsis thaliana* Glutathione S-Transferase (*TAU*) 24)	Member of the Glutathione S-transferase family (involved in flavonoid synthesis and general abiotic stress response) (Sappl et al., 2004)	0.000123	Soil
82	*AT5G07990*: *TT7* (Transparent Testa 7); flavonoid 3'-monooxygenase	Flavonoid biosynthesis protein, which is a product associated with cold response (Korn et al., 2008; Swarbreck et al., 2008)	0.000125	Plate
83	*AT3G55940*: Phosphoinositide-specific phospholipase C, putative	Phospholipase C genes, to which this is related, have been associated with responses to stress in Arabidopsis (Lin et al., 2004)	0.000142	Plate
84	*AT3G21560*: *UGT84A2*; UDP-glycosyltransferase/ sinapate 1-glucosyltransferase	Upregulated by cold via the phospholipase D-dependent phosphatidic acid production (Vergnolle et al., 2005)	0.000145	Plate
85	*AT5G49480*: *ATCP1* (CA2+-Binding Protein 1)	A "cold regulated signaling gene" that is altered in an ice1 mutant background (*ICE1* is a cold/freezing related TF) (Lee et al., 2005). Regulation altered under drought conditions (Huang et al., 2008). Also (like *UGT84A2*, above) upregulated by cold via phospholipase D-dependent phosphatidic acid production	0.000168	Soil

		(Vergnolle et al., 2005)		
86	*AT5G44110*: *POP1*	Shown to be upregulated by cold in supplemental table of (Kreps et al., 2002). Response to Red and Far-Red light via phyA (Tepperman et al., 2006). Also a target of *HY5* (Lee et al., 2007) , which is a transcription factor in light signaling/responsiveness, but also shown to be important for cold dependent anthocyanin accumulation together with *HYH* (above) (Zhang et al., 2011)	0.00017	Soil
87	*AT5G36910*: *THI2.2* (Thionin 2.2); toxin receptor binding	Downregulated under high temperature stress (Larkindale and Vierling, 2008), associated with jasmonic acid/salicylic acid signalling (Li et al., 2004) and target of *FAR1* and *FHY3*, which function in phyA signaling (Hudson et al., 2003)	0.000174	**Novel**
88	*AT2G31380*: *STH1* (salt tolerance homologue); transcription factor/ zinc ion binding, also previously denoted *ZF3*	Like *POP1* above, shown to be upregulated by cold in supplemental table of (Kreps et al., 2002). Circadian-controlled zinc finger gene with role in light signaling (Kumagai et al., 2008). Additional evidence for role in light signaling and regulation by phytochrome (Khanna et al., 2006; Tepperman et al., 2004), and like *THI2* (above), target of *FAR1* and *FHY3*, which function in phyA signaling (Hudson et al., 2003)	0.000176	Soil

*”Cold Upregulation Category” indicates whether a gene was upregulated in the cold when plants were grown in soil (“Soil”), on agar plates (“Plate”), on both growth mediums (“COS”), or on neither (“Novel”) in Vogel et al’s study (Vogel et al., 2005).

Table 3.5: A comparison of the clock patterns between PRIISM-processed and original gene expression data. P-values (calculated using the T-test HAYSTACK function) indicate the correlation of the gene expression patterns of well-studied cold-responsive genes to pre-defined cyclic clock patterns.

Gene Name	AGI Number	**P-values for Warm Gene Expression Data**		**P-values for Cold Gene Expression Data**	
		Original	Clock-Frequency (PRIISM)	Original	Clock-Frequency (PRIISM)
COR15A	*AT2G42540*	5.6E-17	0	0.125	0.039
COR15B	*AT2G42530*	0	0	0.038	4.3E-03
COR47	*AT1G20440*	4.7E-09	1.8E-13	0.012	3.8E-03
COR78	*AT5G52310*	0	0	0.013	2.3E-03
CBF1	*AT4G25490*	5.0E-07	7.2E-08	4.5E-04	3.8E-05
CBF2	*AT4G25470*	3.9E-06	1.6E-13	3.2E-08	2.2E-09
CBF3	*AT4G25480*	5.5E-14	5.6E-17	2.6E-07	5.0E-10
RAV1	*AT1G13260*	1.8E-06	3.0E-10	7.4E-05	2.8E-04
ZAT12	*AT5G59820*	3.4E-03	4.9E-05	1.4E-04	1.9E-05

3.7 Chapter 3 Figures

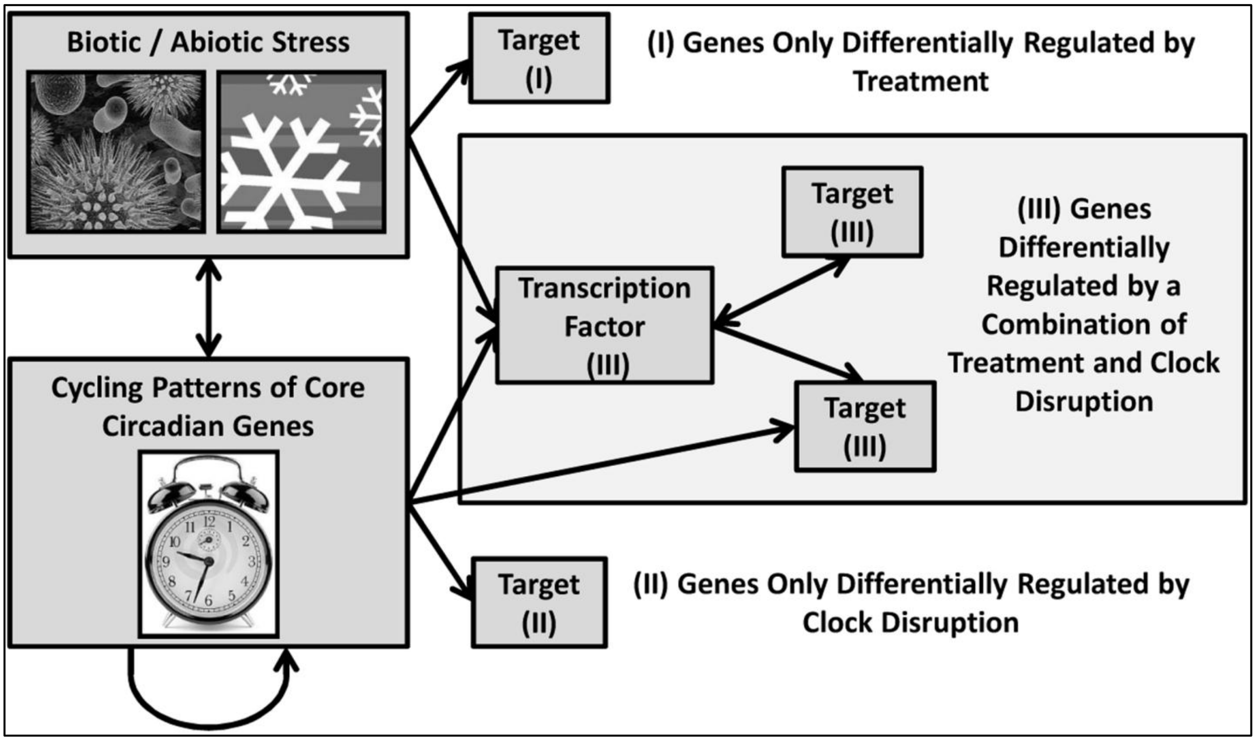

Biotic / Abiotic Stress
Target (I)
(I) Genes Only Differentially Regulated by Treatment
Target (III)
(III) Genes Differentially Regulated by a Combination of Treatment and Clock Disruption
Transcription Factor (III)
Target (III)
Cycling Patterns of Core Circadian Genes
Target (II)
(II) Genes Only Differentially Regulated by Clock Disruption

Figure 3.1: Biotic and abiotic stresses both directly and indirectly influence target gene expression patterns. Genes found to be differentially expressed may be influenced by (I) only direct treatment influences, (II) only indirect circadian-clock disruption influences, or (III) both direct treatment response and indirect clock influences.

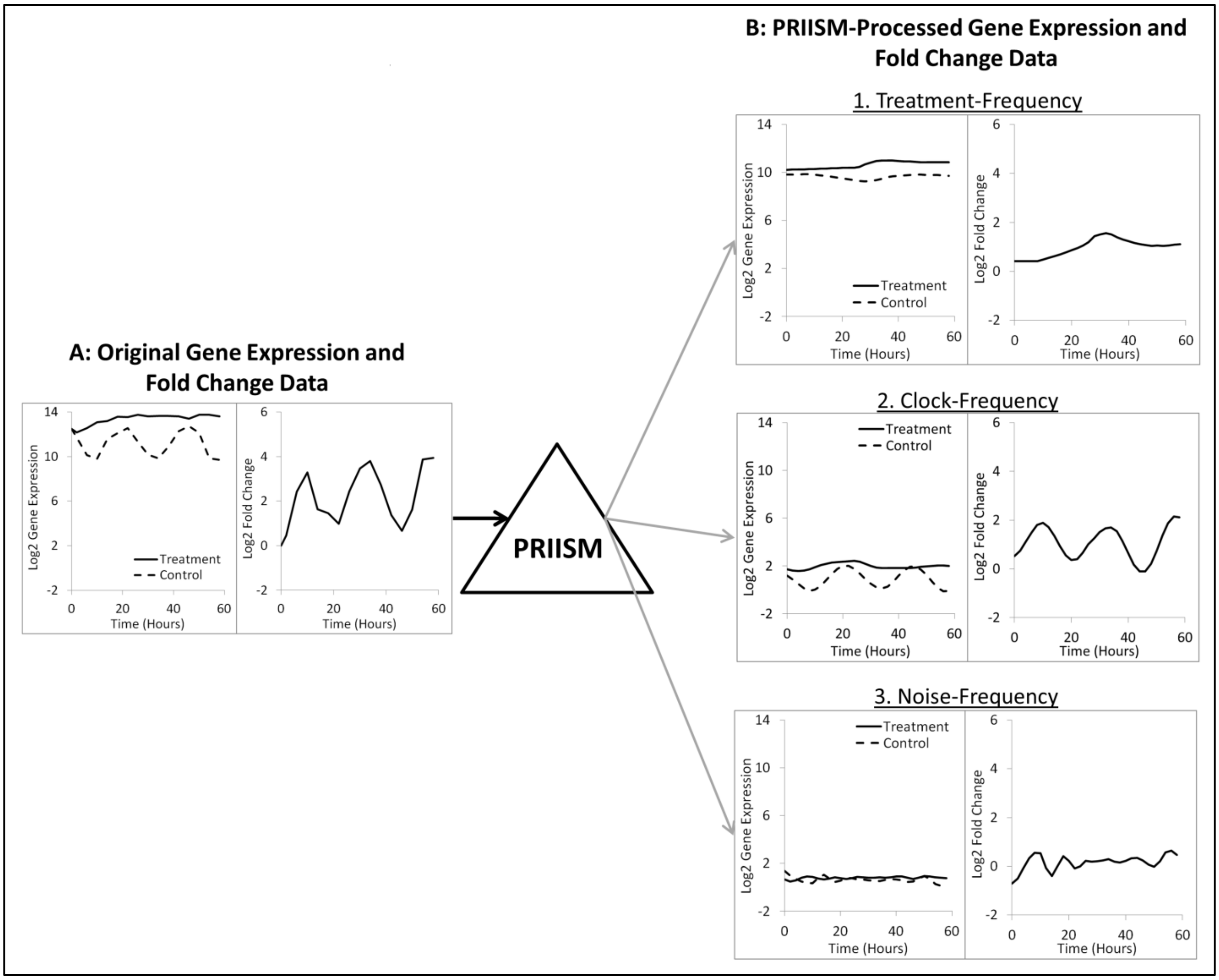


Figure 3.2: PRIISM separates gene expression data into three independent gene expression datasets. PRIISM separates (A) the original gene expression patterns under control and treatment conditions (used to calculate the fold change pattern) into (B) treatment-frequency, clock-frequency and noise-frequency gene expression patterns. The cold-induced gene *COR15A* (*AT2G42540*) is shown as an example.

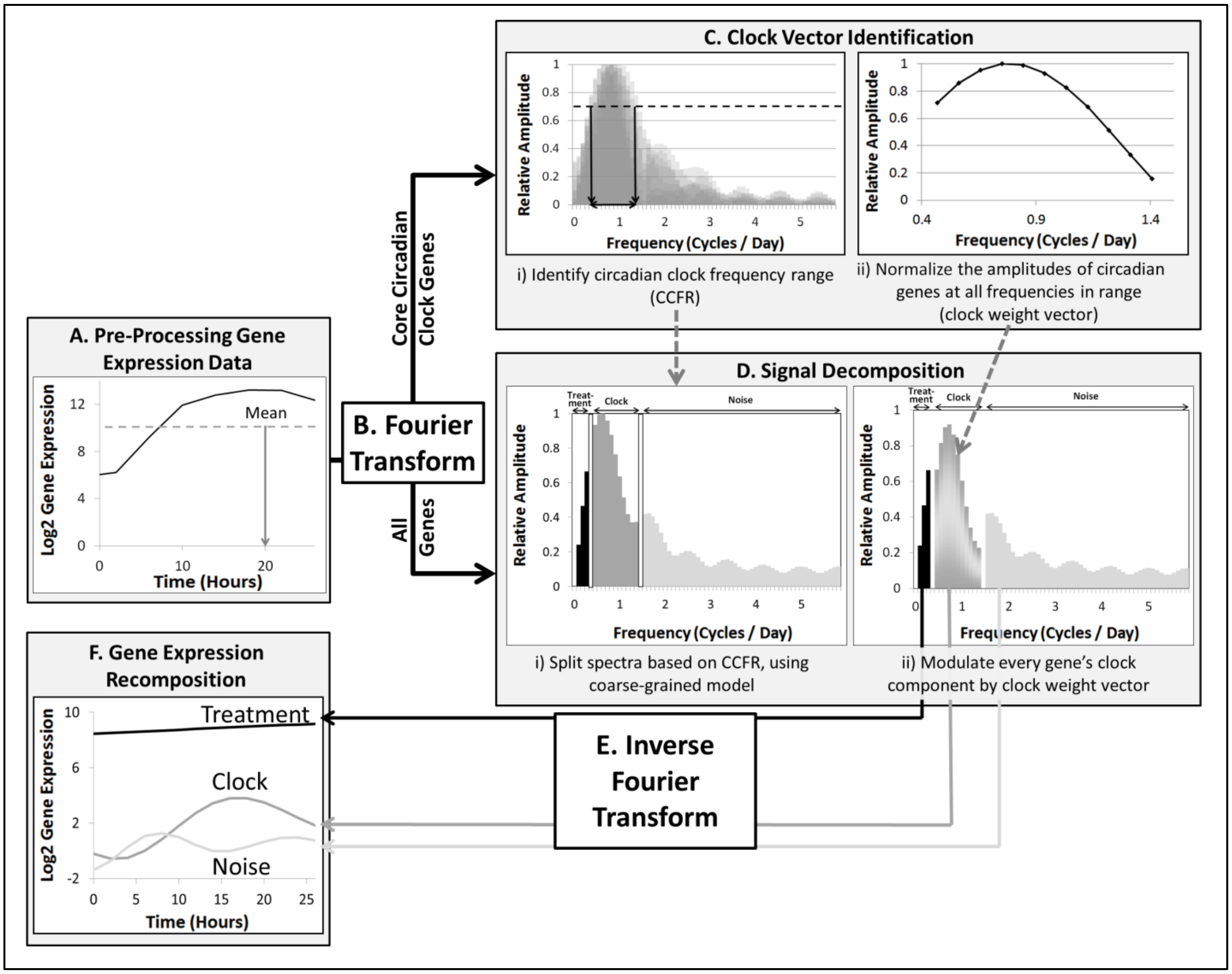

A. Pre-Processing Gene Expression Data
Log2 Gene Expression
Mean
Time (Hours)
B. Fourier Transform
Core Circadian Clock Genes
All Genes
C. Clock Vector Identification
Relative Amplitude
Frequency (Cycles / Day)
i) Identify circadian clock frequency range (CCFR)
ii) Normalize the amplitudes of circadian genes at all frequencies in range (clock weight vector)
D. Signal Decomposition
Treat-ment
Clock
Noise
i) Split spectra based on CCFR, using coarse-grained model
ii) Modulate every gene's clock component by clock weight vector
E. Inverse Fourier Transform
F. Gene Expression Recomposition
Treatment
Clock
Noise
Time (Hours)

Figure 3.3: Workflow of the PRIISM algorithm. The 0 to 26 hour time-frame in the cold for *AtgolS3* (*AT1G09350*) is used as an example.

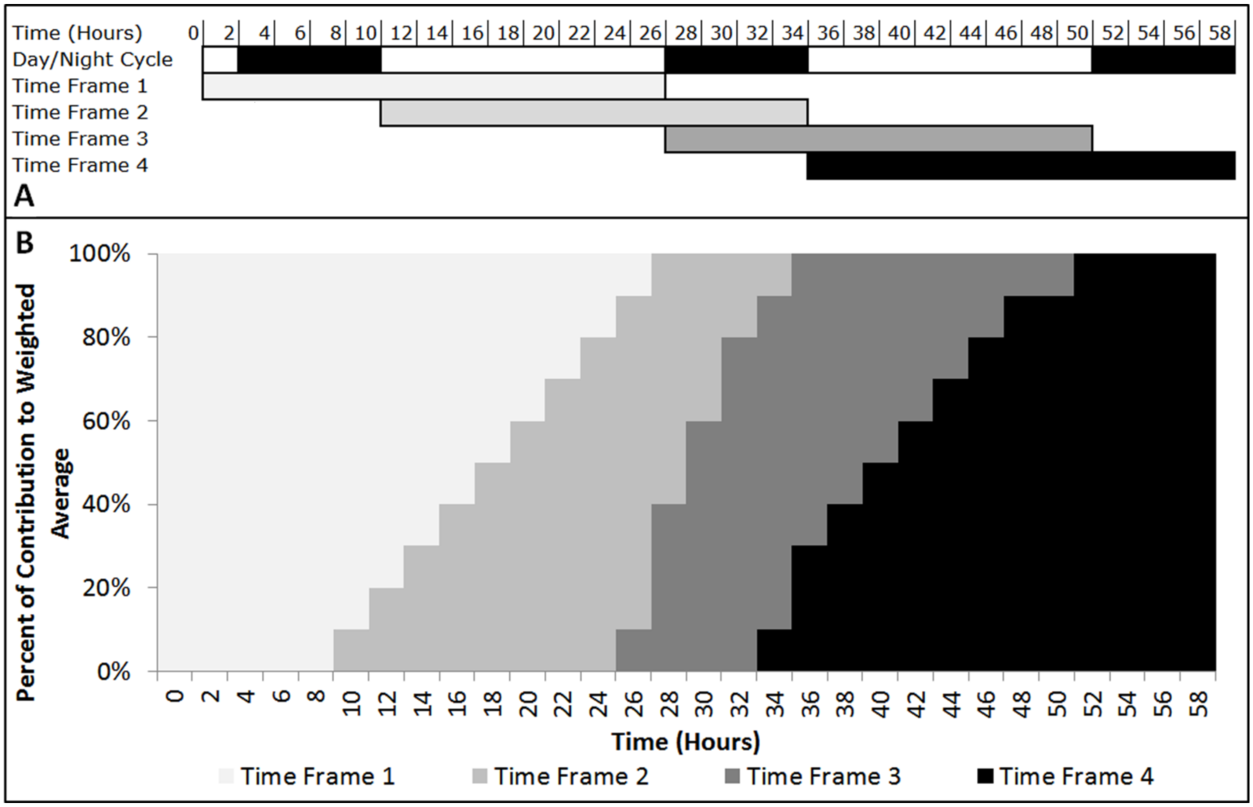


Figure 3.4: Time frames used to generate FFT results**.** Frame sizes and positions are shown in (A) and the contribution of each frame to the weighted average at each timepoint is shown in (B).

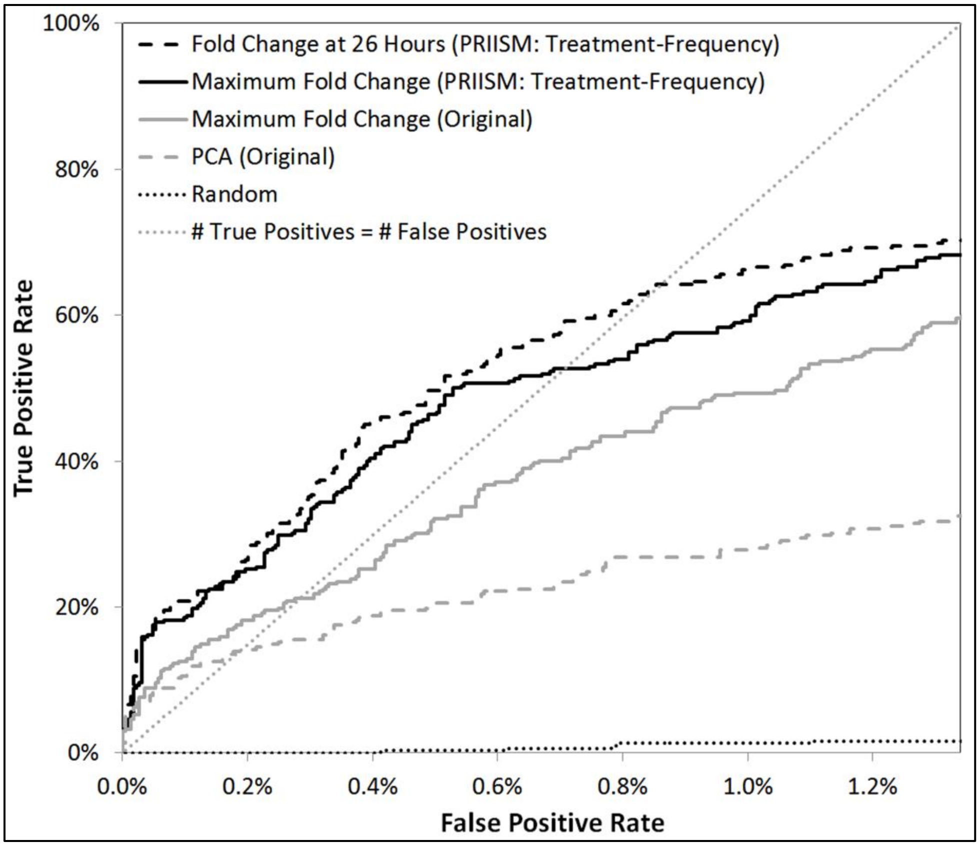


Figure 3.5: ROC curves for COS-upregulated genes. ROC Curves for the 26-hour treatment-frequency fold change (dashed black line), the treatment-frequency maximum fold change (solid black line), the original maximum fold change (solid grey line), and original PCA plot distance data (dashed grey line) are shown. The point at which the number of false positives is equal to the number of true positives (dotted grey line) and random gene selection (dotted black line) are also shown.

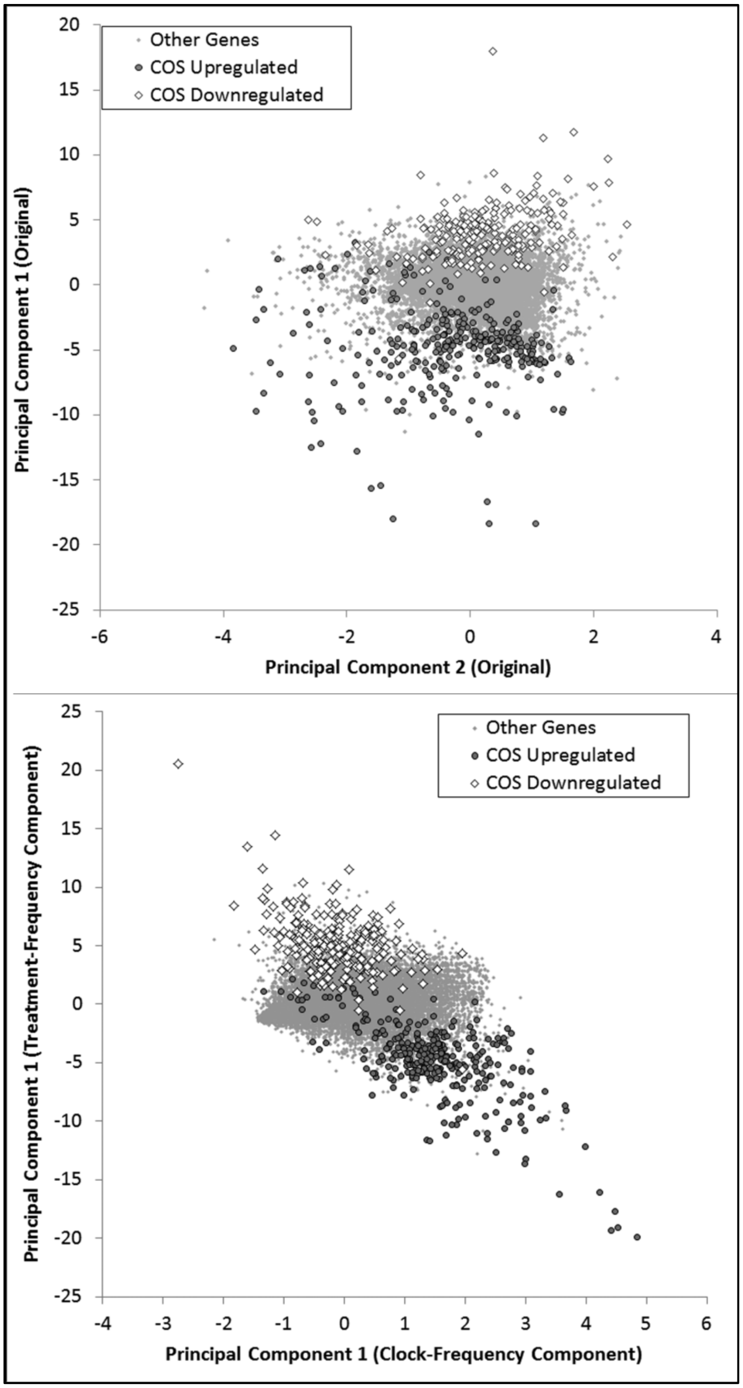


Figure 3.6: Principal Component Analysis (PCA) Plots. Principal component analysis (PCA) plots for the original data (A) and the first components of the clock-frequency and treatment-frequency data (B) are shown. COS-upregulated genes are shown in black circles, COS-downregulated genes (which are not analyzed in detail here) are shown in white diamonds and all other genes are shown as grey dots.

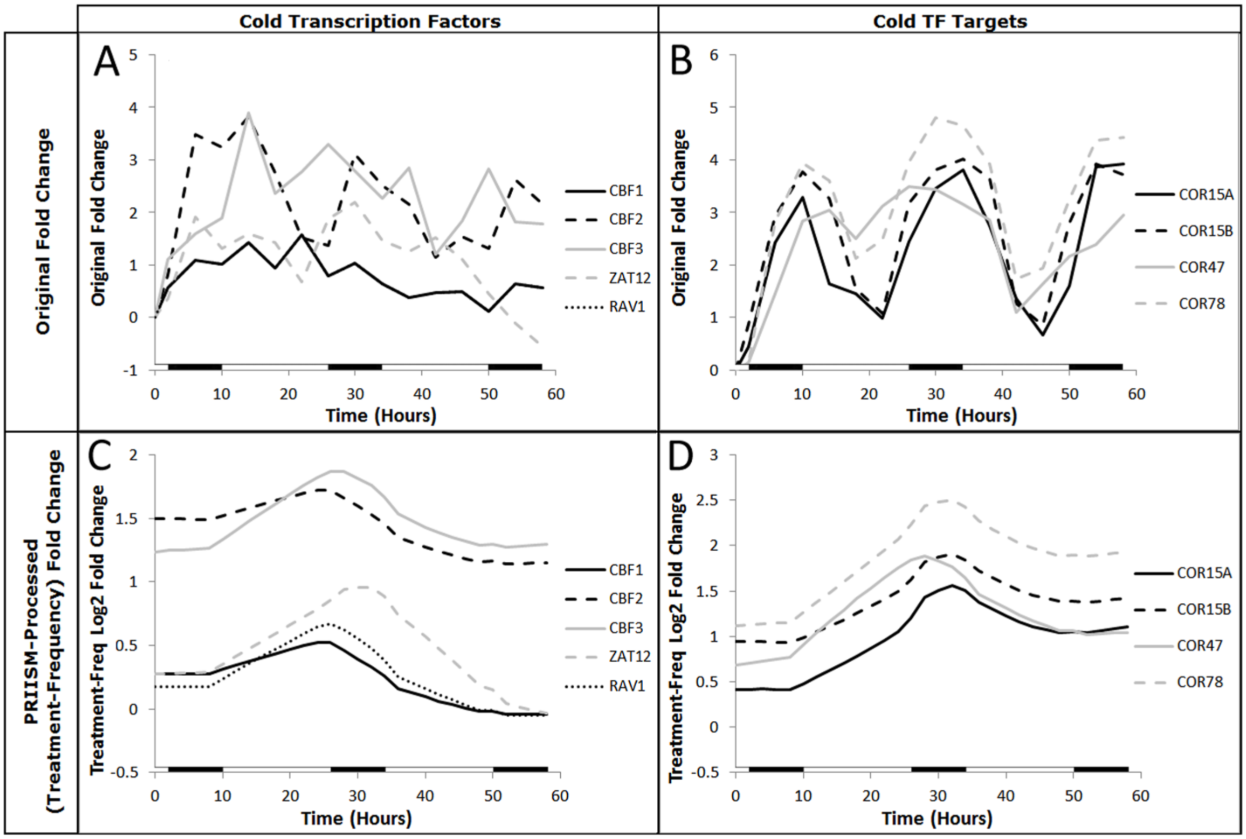


Figure 3.7: Fold change patterns of cold transcription factors and target genes before and after PRIISM processing. The original fold change patterns for important cold transcription factors (A) and some of their important target (COR) genes (B) are shown, along with their the treatment-frequency fold change patterns for the same genes (C & D).

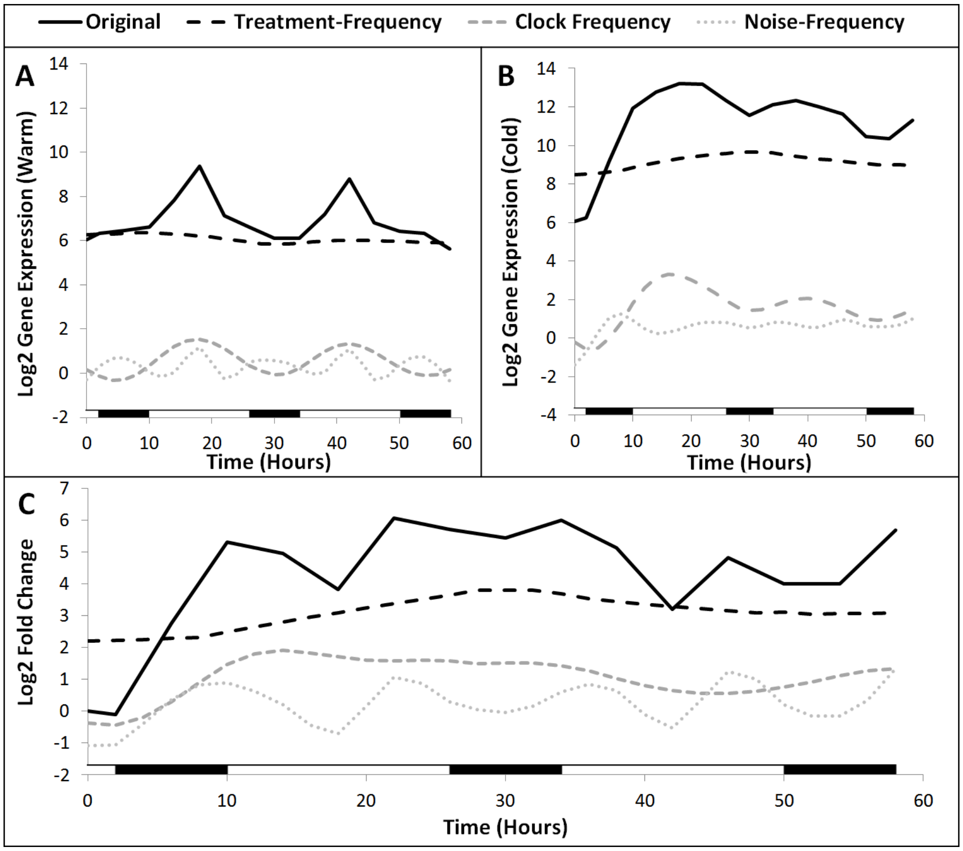


Figure 3.8: A case study examining PRIISM output gene expression and fold change data. The fold change patterns (A), warm gene expression patterns (B) and cold gene expression patterns (C) for the original and PRIISM-processed data for *AtgolS3* (*AT1G09350*), the most highly upregulated gene in response to cold at 26 hours in the treatment-frequency data.

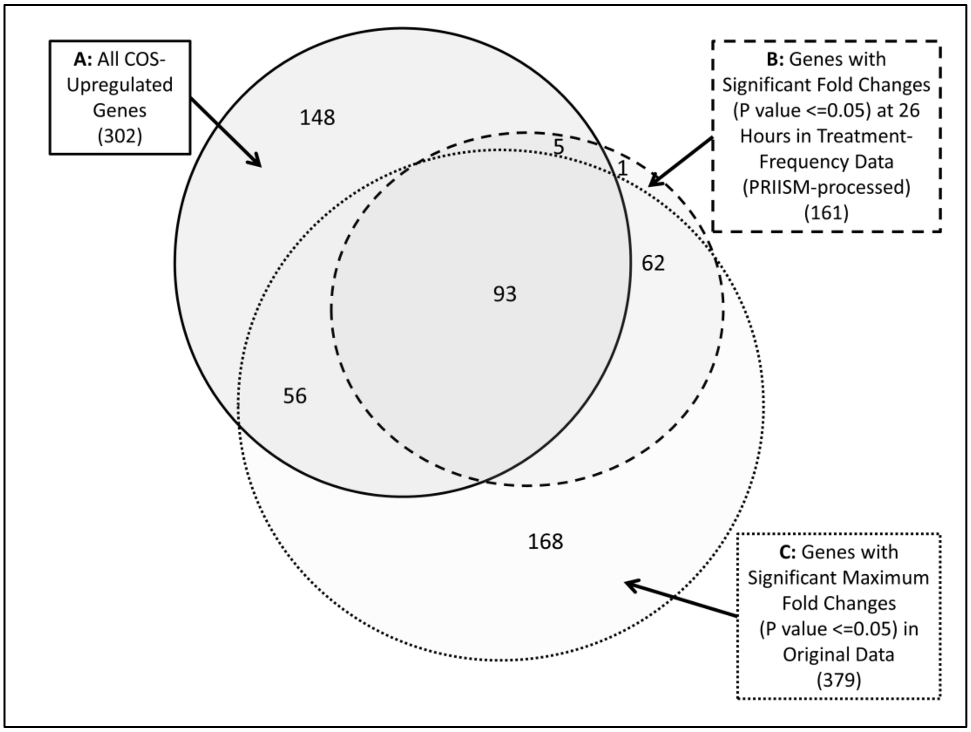

A: All COS-Upregulated Genes (302)
B: Genes with Significant Fold Changes (P value <=0.05) at 26 Hours in Treatment-Frequency Data (PRIISM-processed) (161)
C: Genes with Significant Maximum Fold Changes (P value <=0.05) in Original Data (379)
148
5
1
62
93
56
168

Figure 3.9: Venn diagram showing COS-upregulated genes in original and PRIISM-processed significant gene lists. The number of genes in the overlaps between COS-upregulated genes (A) and the significant genes (P value ≤ 0.05) in both the maximum fold change in the original dataset (B) and the fold change at 26 hours in the treatment-frequency dataset (C) are shown.

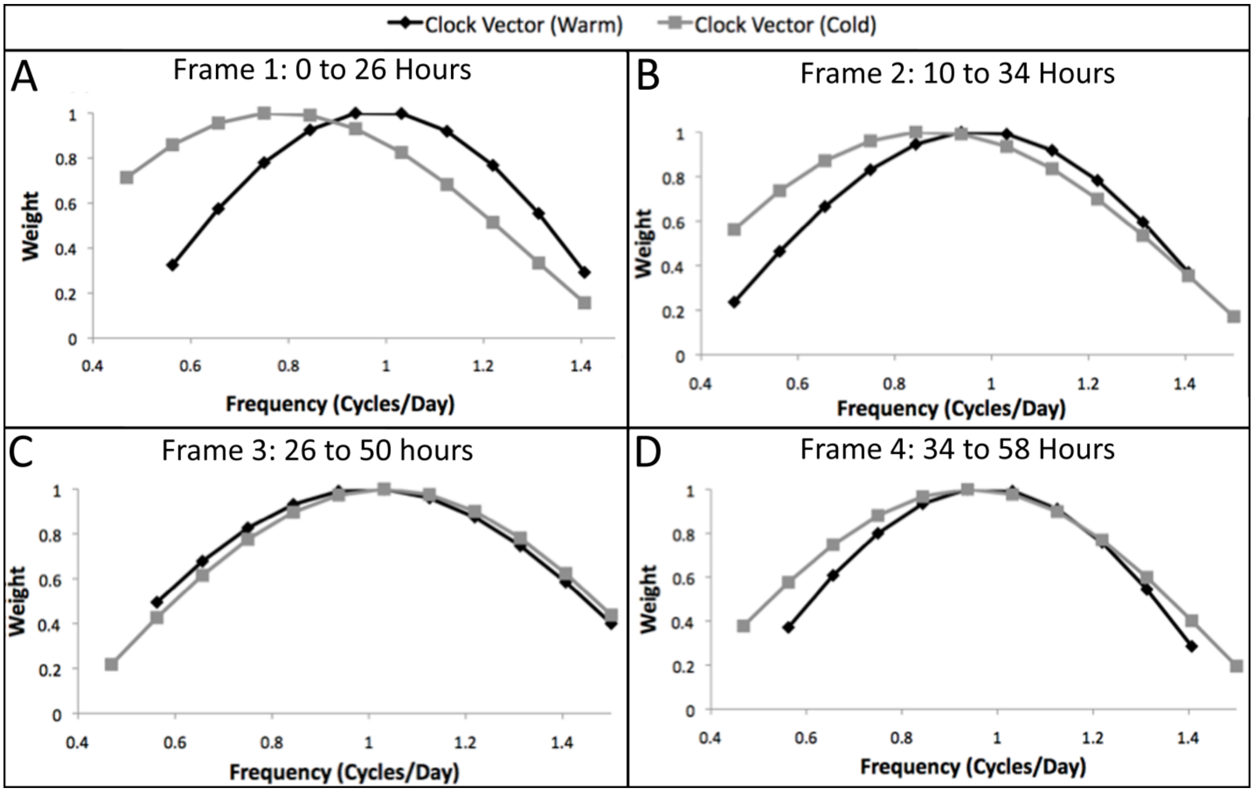


Figure 3.10: Clock vectors under warm and cold conditions

Chapter 4: Knowledge-based optimal timepoint sampling in high-throughput temporal experiments

This chapter has very recently been submitted for publication review:

Rosa, B.A., Zhang, J., Major, I.T., Qin, W., Chen, J. (2012b) Knowledge-based optimal timepoint sampling in high-throughput temporal experiments. Manuscript Submitted for Publication to Bioinformatics.

4.1 Abstract

Determining the *best* sampling rates (which maximize information yield and minimize cost) for time-series high-throughput gene expression experiments is a challenging optimization problem. Existing approaches infer timepoints using low-throughput technology or the uncertainty in the sparsely-sampled expression curves. The knowledge in existing relevant gene expression experiments is extremely valuable, but it is a difficult problem to utilize it for timepoint selection. Here, we present a new data-integrative model, Optimal Timepoint Selection (OTS), to address the sampling rate problem by integrating existing datasets with a novel optimization approach, and identifying the maximal disagreement between the current dataset (to which timepoints will be added) and the integrated dataset. Two experimental settings were used to test the performance of OTS. In "iterative-online" sampling, timepoints were added iteratively, starting from the first and the last timepoints. OTS selected early timepoints containing the strongest upregulation peaks of the target genes, reducing the interpolation error rate by 50% (or 72%) after adding one (or seven) suggested timepoint(s) to an *Arabidopsis* coronatine-response dataset, compared with 32% when adding one uniformly-distributed timepoint (or 69% for seven timepoints with "active learning", the most recent approach). In a similarly designed experiment studying yeast cell-cycle genes, OTS reduced the error rate by 35% (or 56%) after adding one (or seven) timepoints, compared with uniform distribution (5%) and active learning (51%). In the "top-up" sampling test, several timepoints were added at once to a uniformly-sampled time series.

OTS reduced the error rate by 26% (14% in active learning) in the same yeast dataset by capturing early (G_1 phase) timepoints with strong differential regulation events. In conclusion, OTS performed consistently better across these experiments, showing that by integrating existing datasets, it can optimize the distribution of a limited number of timepoints, leading to better biological insights about the regulation patterns of treatment/condition-response genes.

4.2 Introduction

High-throughput gene expression experiments are capable of measuring the expression levels of tens of thousands of genes in a biological sample. Time-series high-throughput gene expression datasets in particular can provide dynamic information about gene regulation patterns, and can be used to construct regulatory networks and infer regulatory relationships among genes (Wang et al., 2008). As of February 2012, there are 644 time-series datasets (representing approximately 23% of the total datasets) on the Gene Expression Omnibus (GEO) online database (Edgar et al., 2002). We build our approach based on three observations: 1) Gene expression experiments can be sampled in an online fashion, as outlined in Singh *et al* (Singh et al., 2005); *i.e.,* biological samples can be treated and collected at a very high rate and then stored at a relatively low cost, and researchers can measure gene expression in particular samples at a later time after deciding which timepoint will be optimal (Singh et al., 2005). 2) A researcher is usually interested in capturing the expression patterns of a subset of genes (which may be grouped into several clusters with similar expression patterns) associated with a given treatment/condition; and 3) Treatment-response gene expression patterns from different experiments performed under similar treatments/conditions can provide differential regulation information that could be valuable for defining an optimal timepoint for sampling, even if the sampling rates are different from the new experiment.

Based on these observations, a straightforward approach to choosing the best timepoint to add to a sparsely sampled (current) dataset is to learn information from other existing (training) datasets, and then find unsampled timepoints at which there are significant upregulation or downregulation events for the genes of interest in the training datasets. This approach is based on the assumption that the differential expression patterns for the genes of interest in the training data are similar to each other and are similar to the dataset to which a timepoint will be added. However, in practice, this assumption may be violated in most of cases due to 1) large differences in the dynamic ranges between platforms (e.g., RNA-seq technology has a dynamic range several orders of magnitude higher than microarray technology (Marioni et al., 2008)), 2) inconsistency among different datasets, either due to slightly different growing conditions, different treatments, or "lab signatures", which result in differences in gene expression patterns among different labs even after attempts to reproduce conditions exactly (Massonnet et al., 2010), 3) high noise rates in expression values (particularly for microarray datasets (Marioni et al., 2008)), and 4) sparse sampling rates in existing data.

To tackle the data-integration challenges mentioned above, we have developed OTS to address the sampling-rate problem by utilizing a novel method of combining differential gene expression data from training datasets based on their similarity to the experiment to which timepoints will be added (the "current" experiment). OTS then identifies the timepoints with the maximal difference between the current and the integrated training data, leading to the identification of the timepoint(s) that may result in the most significant information gain in the current dataset. Comparing to the existing algorithms, OTS is novel in the following ways:

- **Gene expression values are projected to threshold space.** In contrast to existing gene expression prediction algorithms (Chikina et al., 2009; Gustafsson and Hornquist, 2010; Ruan, 2010), the goal of our method is to predict the best timepoints to add to a high-throughput experiment. Therefore, rather than focusing on specific expression patterns, we are instead interested in how many genes are significantly differentially expressed at each timepoint, and how significant the overall expression values are (in a categorized fashion). Consequently, we project the gene expression values to threshold space to better capture important regulatory timepoints (explained in section 4.3.5).

- **A novel integrative measure is utilized for timepoint selection.** After projecting gene expression data to threshold space, we look for the optimal timepoint by developing a novel integrative measure. Instead of averaging or pooling all of the training data together, we first weight each training data's contribution to the overall result based on their similarity to the current dataset. Then, we adjust the weighted-average values with a shifting function for local fitting (explained in section 4.3.6).
- **Optimal timepoints are selected using multi-objective optimization.** By clustering all of the genes based on their expression patterns, we are able to select the best timepoint for each cluster of target genes by using the integrative measure. Then, we adopt a multi-objective optimization (MOO) model to select the overall optimal timepoint for all of the clusters (Coello, 1999). MOO is superior to the sampling voting method used in (Singh et al., 2005) because timepoints chosen by MOO benefit all (or the majority) of clusters, while the sampling voting method may be biased to one or a few clusters (explained in section 4.3.7).

The overall experimental approach of the sampling rate design with OTS is shown in Figure 4.2. First, a biological experiment is performed, and samples are preserved (typically by freezing) at dense timepoints. A subset of timepoints (including at least the last timepoint in the range of interest and one other timepoint) is sampled in order to input an initial dataset into OTS. Then, time-series training datasets are collected. Note that it is not necessary for the training datasets to be collected using the same technology (*i.e.*, PCR, microarray or RNA-seq experiments), but they should use treatments or conditions that are expected to affect target treatment-response genes in the same way as in the current dataset. The timepoints in these training datasets can suggest differential expression patterns at the unsampled timepoints in the current dataset. Finally, OTS produces a ranked list of the optimal timepoints to be selected next. The optimal timepoint(s) can then be sampled and added to the current dataset for the identification of the next optimal timepoint. This process can be continued iteratively until the maximum number of samples is reached *i.e.*, all of the samples or all of the resources available for sampling are used up.

We have applied OTS to select timepoints in high-throughput time-series experiments for two different organisms (yeast and *Arabidopsis*) using different platforms (microarray and RNA-seq), and used noisy as well as sparse and poorly-matching training data to suggest the optimal timepoints. In both datasets, our method clearly outperforms active learning (Singh et al., 2005).

4.3 Methods

The goal of this paper is to develop a computational algorithm to design the sampling rate of time series gene expression experiments such that the real gene expression patterns for genes of interest are captured as accurately as possible. Specifically, our approach is to generate an estimate dataset by integrating training data, and to identify the timepoint at which the estimate dataset is the most different from the current dataset, which may result in the identification of the most significant differential regulation events in the current dataset (see Figure 4.2 for the overall experimental design of OTS). Mathematically, given training datasets R_1, R_2, …, R_m, a current dataset U (with gene expression values available at timepoint set T_S and unmeasured biological samples available at timepoint set T_A), identify the optimal timepoint t_O which minimizes the difference between the interpolated and real gene expression curves for all genes of interest G.

The Methods section is organized as follows: we will review the existing approaches for determining optimal sampling rates in section 4.3.1, introduce a case study for the better understanding of OTS in section 4.2.2, and discuss the individual steps in OTS in sections 4.3.3 to 4.3.7 (as outlined in Algorithm 4.1).

4.3.1 Existing approaches

Before introducing OTS, we will review the existing approaches for determining optimal sampling rates. One common approach involves observing the expression patterns of a small number of genes of interest in a separate experiment, sampled using a different platform. For example, real-time PCR experiments can produce high-resolution time-series gene expression data, and bioluminescent and fluorescent tags can be added to the genetic sequence of a gene so that protein levels can be measured over time (Kalir et al., 2001; Vance et al., 2002; Yuan et al., 2006). If there is a time range in which strong differential expression occurs in a preselected small set of genes, then this may suggest timepoints that should be sampled in the high-throughput experiment. However, these approaches are very costly and time consuming, and there is no approach to determine which timepoints to select if the target genes being studied have very different expression patterns from each other. Also, this approach can only use a very small number of genes relative to the number of genes typically affected by a given treatment or condition.

In the field of signal processing, determining effective sampling rates is a well-studied problem (Orfanidis, 1995). However, most approaches for determining a sampling rate can only be applied to high-resolution signals (according to the Nyquist sampling theorem (Marks II, 1991)), while in the case of high-throughput gene expression datasets, the temporal resolution is usually much lower than the traditional signals (Edgar et al., 2002). Also, because many genes may be regulated by common transcription factors or are localized in the same biological pathway, gene expression patterns of biologically related genes are correlated and dependent on each other, which further complicates traditional signal processing methods (Singh et al., 2005).

An active learning algorithm was recently developed for iteratively choosing timepoints to sample, using the uncertainty in the interpolation of the currently estimated time-dependent curve as the objective function (Singh et al., 2005). This algorithm used local cross-validation to enable effective sampling from non-uniform locations along a time-series, and suggests the optimal timepoint to add based on the existing gene expression patterns in one dataset (Singh et al., 2005). The performance evaluation in this study (on a yeast dataset) showed that this algorithm can find optimal timepoints such that majority cycling yeast genes in the dataset can be identified using just 18 out of 24 of the timepoints in the original dataset (Singh et al., 2005). However, to precisely capture the gene expression patterns, the interpolation step in this algorithm requires a minimum of five timepoints to start (according to the source code available from http://theory.csail.mit.edu/tsample), so it would not have been applicable for 75% of the existing datasets in GEO, and would have only been able to predict very few timepoints in almost all of the existing datasets (Edgar et al., 2002; Singh et al., 2005). Furthermore, the timepoint selection is based only on the gene expression in the dataset to which a new timepoint will be added, and existing time-series gene expression datasets using similar treatments (which may be high-resolution and contain useful gene expression information) cannot be applied in the algorithm.

4.3.2 Case study

For demonstration purposes, we used a novel *Arabidopsis* coronatine-treatment dataset as a case study (Fig. 4.3A). This dataset was produced in a separate study to determine the effect of the phytotoxin coronatine (a molecular mimic of the plant hormone jasmonate) on global gene expression in Arabidopsis (manuscript in preparation). In our current study, we used this densely-sampled (21 timepoint) RNA-seq dataset as a mock “current” biological experiment (Step 2 in Figure 4.2), and a number of existing microarray datasets involving coronatine/jasmonate treatment as training data (Step 4 in Figure 4.2) (Chung et al., 2008; Wierstra and Kloppstech, 2000). A total of 195 genes of interest were selected based on gene ontology categories related to these biological treatments (see section 4.4.1 for details). In order to best illustrate the OTS algorithm, this case study starts with six timepoints, at 0.25 hours (the first), 24 hours (the last), and 1, 2, 3 and 5 hours (selected iteratively by the first four rounds of OTS selection; Figure 4.2), and the fifth round of timepoint selection will be outlined in detail here.

We define the "current mock" (first row in Figure 4.3A) as the current dataset (*U*) and the rest of the rows as training datasets (*R*). Note that the "current mock" in Figure 4.3A shows all of the possible timepoints for sampling, but only a small subset of the timepoints will be selected to simulate the online design fashion.

4.3.3 Gene expression clustering

We observe that a researcher is usually interested in capturing the differential expression patterns of a subset of genes associated with a given treatment/condition, rather than the whole genome. This observation prompted us to focus on capturing the expression patterns of a subset of the entire gene set, and since many of these gene expression patterns are correlated, they can be separated into clusters (*C*). A few genes of interest with unique differential expression patterns will be classified in their own cluster, and because each cluster contributes equally to the results regardless of how many genes are in it, this allows unique genes to exert a stronger influence over the results compared to if they were pooled equally with the rest of the genes of interest.

As the first step of OTS (line 1 in Algorithm 4.1), all of the genes of interest (*G*) are clustered based on their differential expression values in the training datasets *R*. In this step, we applied K-means clustering (Dembélé and Kastner, 2003) (implemented in Cluster 3.0 (Eisen et al., 1998), available for download from rana.lbl.gov/EisenSoftware.htm), a common method for clustering genes based on Log2 fold change values. OTS can be easily extended to use any other clustering algorithms and the Log2 fold change values can be conveniently replaced by other differential expression measurements, depending on user’s needs. In the case study, the 195 coronatine-responsive genes were separated into ten clusters; the Log2 fold change curves for one of the clusters containing 13 genes in this dataset are shown in Figure 4.4A.

4.3.4 Data interpolation

To be able to estimate the differential gene expression patterns at all timepoints, OTS linearly interpolates the differential expression measurements from the current (*U*) and the training (*R*) datasets to every available timepoint in T_A (Algorithm 4.1 line 4-5) (Meijering, 2002). Linear interpolation was used in order to minimize the inference of false peaks and valleys in the expression data by directly connecting data points and not calculating expression levels above or below the two points used in the calculation (Benesty et al., 2004). Linear interpolation also avoids over-smoothing unevenly spaced timepoints, which occurs on sparsely-sampled datasets when using other common interpolation methods such as B-spline (Meijering, 2002). Given a gene *g*, its estimated differential gene expression value *e* at time point *t* is:

$$e = e_i + (t - t_i)\frac{e_{i+1} - e_i}{t_{i+1} - t_i} \tag{4.1}$$

where e_i is the differential gene expression value at sampled timepoint t_i ($t_i<t<t_{i+1}$), and t_i and t_{i+1} are the closest sampled timepoints to *t*. OTS can be easily extended to use any other interpolation algorithms depending on user’s needs.

4.3.5 Projection to threshold space

Rather than focusing on specific expression patterns, we capture important regulatory timepoints by measuring how many genes are significantly differentially expressed at each timepoint, and how significant the overall expression values are (in a categorized fashion). Consequently, unlike the existing approaches of sampling rate design (which focus on the inference of values of differential gene expressions (Chikina et al., 2009; Gustafsson and Hornquist, 2010; Ruan, 2010)), we project the gene expression data of each cluster into threshold space, where the values for a given timepoint are determined based on how many genes have differential gene expression values higher (or lower) than a series of differential regulation thresholds (Fig. 4B) (Algorithm 4.1 line 6-10). This thresholding process reduces noise in the comparison among datasets by ignoring small fluctuations in differential gene expression value patterns and only considering the relative magnitude of the signals.

In order to avoid the bias introduced by setting only one significant regulation threshold value, multiple evenly-spaced positive and negative differential regulation threshold values are defined to determine the degree to which a cluster of genes is differentially regulated at a given timepoint in a given experiment, according to Equation 4.2. Basically, given a user-defined threshold number *H*, we divide the threshold space (three standard deviations above and below the average differential gene expression value) evenly into sections of (1/*H*). For example in Figure 4.4B, an *H* value of 6 has been used, and the dotted lines represent the evenly spaced positive and negative thresholds.

Mathematically, for a gene *g*'s expression value at time *i* in dataset *j*, we compute its *differential regulation count* (DRC) by counting how many thresholds it is higher (or lower) than if it is up (or down) regulated. The sum of these count values for all the genes in cluster *c* (defined in Equation 4.2) represents the DRC for timepoint *i* in cluster *c* in dataset *j*. Higher DRC numbers indicate stronger differential regulation, regardless of whether the genes are upregulated or downregulated.

$$D_c^{ij} = \sum_{g \in G_c} \sum_{h=1}^{H} \left[e_g^{ij} - \left(\frac{(\mu + 3\sigma)(h-1)}{H} \right) > 0 \right] + \left[e_g^{ij} - \left(\frac{(\mu - 3\sigma)(h-1)}{H} \right) < 0 \right] \qquad (4.2)$$

where D_c^{ij} is the *differential regulation count* (DRC) for timepoint *i* in cluster *c* in dataset *j*, *H* is the user-defined threshold (*H*>0), e_g^{ij} is the differential expression measurement for gene *g* (out of the set of genes G_c in cluster *c*), and *μ* and *σ* are the average and the standard deviation values (respectively) for the differential gene expression values across all timepoints and all genes of interest *G* in all the training datasets. Operator [*x*] returns 1 if *x* is true, otherwise it returns 0.

For example, in Figure 4.4C, one gene crosses the top upregulation threshold (at 4.11), and three genes cross the next upregulation threshold (3.29). These counts are made for each regulation threshold and summed up as shown in the Figure. DRC curves for all the training and current datasets for cluster 2 in this case study are shown in Figure 4.4D. Although these DRC curves have a similar upregulation trend, they are very different from each other at a detailed level, mainly because of differential sampling rates, but also because of slightly different growing conditions, different (but relevant) treatments, “lab signatures” (Massonnet et al., 2010), and high noise rates in microarray datasets (Marioni et al., 2008).

4.3.6 Curve matching

As described above, we require a timepoint selection algorithm that is capable of integrating heterogeneous training data, and that allows for efficient computation. To achieve this goal, we save the resulting DRC values in a cluster-time-experiment (CTE) table. Table 4.1 shows the layout of the CTE table, which includes DRC values (D^{ij}) for every timepoint *i* ($1 \le i \le n$) in every training data *j* ($1 \le j \le m$) in cluster *c*. Additional columns are added to the CTE table for the current dataset ($\overset{\square}{D}$), and for the estimate dataset ($\bar{D}$) (explained below).

For the CTE table, our approach is to generate an estimate curve for each cluster by combining the training DRC curves, and then identifying the timepoint at which the estimate DRC curve is the most different from the current DRC curve. Adding this timepoint to the current dataset may result in the most significant information gain. However, combining the training datasets into an estimate curve is a difficult problem because the training datasets may not be similar to each other and may not be similar to the current dataset (see Figure 4.4D for an example). Although there are numerous ways to normalize and scale the datasets (such as least-squares estimation), the challenge is that the difference between the training and current datasets will not converge to 0 even if numerous timepoints are added (because of biological differences among experimental conditions, different sampling rates and other factors that may affect the training datasets and not the current dataset), leading to biased estimations of the gene expression patterns. To tackle this problem, we utilize a novel two-step (global matching and local fitting) normalization and scaling approach for the curve-matching problem (Algorithm 4.1, line 11).

In the first step (global matching), rather than pooling all of the training datasets together, we first weight each training data's contribution to the overall result based on their similarity to the current DRC curve in each cluster using non-negative least-squares (NNLS) regression (Chen et al., 2010; Lawson and Hanson, 1995) and then save the results in the estimate dataset. Mathematically, given a $n \times m$ CTE table of DRC values derived from the training data (D^{ij}) and an n x 1 vector of DRC values derived from the current data ($\widehat{D}$), find a non-negative $m \times 1$ weight vector w that minimizes the difference between weighted training and current datasets (*i.e.*, $\min_w f(w) = \frac{1}{2}\left\|Dw - \widehat{D}\right\|^2$), where the weight vector w is then used to calculate a weighted-sum estimate DRC curve for cluster c. An example of an NNLS-regression estimate for one of the clusters in the case study is shown in Figure 4.4E.

If there is a training dataset that poorly matches the current dataset for a majority of the target genes, then the NNLS will assign low weights, resulting in that training dataset having a low contribution to the estimation. This allows OTS to be robust against poorly-matched training datasets. By forcing all of the weight values to be non-negative, it also avoids a problem introduced by standard LSE regression, wherein negative weights can "flip" the patterns, changing peaks to valleys and providing false information in the estimation.

In the second step (local fitting), in order to correct the estimate fit, DRC values resulting from the NNLS weighted-sum are shifted for each timepoint, such that the estimate DRC values are equal to the current DRC values at every sampled timepoint (indicated by vertical dashed grey lines in Figure 4.4F). The rest of the timepoints in the NNLS-weighted estimate DRC curve are shifted by an amount suggested by the sampled timepoints, and modulated by their distance from the sampled timepoints according to a sigmoid weight distributed according to Equations 4.3 and 4.4 (Chen and Mangasarian, 1995; Marler et al., 2006). The estimate value at timepoint i ($\bar{D}^i$) is defined as:

$$\bar{D}^i = \begin{cases} \widehat{D}^i & \text{if } i \in Ts \\ \sum_{j=1}^{m} w_j D^{ij} + \dfrac{2\left(\widehat{D}^t - \sum_{j=1}^{m} w_j D^{tj}\right)}{1+e^{\frac{5|t-i|}{n'}}} & \text{otherwise} \end{cases} \qquad (4.3)$$

and *t* is defined as $t = \underset{t \in Ts}{\arg\max} \left| \widehat{D}^t - \sum_{j=1}^{m} w_j D^{tj} \right|$ (4.4)

where *i* is a timepoint in the interpolated current dataset ($T_A \cup T_S$), $\widehat{D}_t$ is the DRC value for timepoint *t* in the current dataset, D^{ij} is the DRC value for timepoint *i* in training data *j*, w_j is the weight assigned by NNLS for training data *j*, and *n′* is the number of timepoint in the current dataset.

The curve difference score ($Q^i = \left|\widehat{D}^i - \bar{D}^i\right|$) is the difference between the estimate and current dataset curves at timepoint *i*. Figure 4.4F shows that for the cluster outlined in the case study, 12 hours is the optimal timepoint, which is in agreement with the actual DRC value at 12 hours for this cluster. Figure 4.5A shows the curve difference score table for all of the clusters in the case study experiment, in which each timepoint is associated with one curve difference score for each cluster.

4.3.7 Timepoint selection with multi-objective optimization

By clustering all the genes based on their expression patterns and comparing the estimate and current curves in each cluster, we are able to select the best timepoint for each cluster using the maximum curve difference scores. But if the best timepoint for each cluster is different, a cross-cluster ranking method is needed to calculate one timepoint for the entire dataset. Instead of applying a sampling voting method (used in (Singh et al., 2005)), OTS applies a *multi-objective optimization* (MOO) model to rank and select optimal timepoints for all the clusters (Algorithm 4.1, lines 13-14) (Coello, 1999). MOO is better than the sampling voting method because timepoints chosen by MOO benefit all (or the majority) of clusters, while the sampling voting method may be biased towards the optimal timepoints in one or a few clusters.

Mathematically, MOO computes a *λ-score* for each timepoint (to indicate how optimal that timepoint is) in two steps. First, *λ -dominance* is determined for each timepoint pair as follows: We say timepoint t_1 λ *-dominates* timepoint t_2 (denoted as $t_1 \underset{\lambda}{\succ} t_2$) if Q^{t1} *i*s larger than Q^{t2} in λ clusters (1 ≤ λ≤|*C*|):

$$t_1 \underset{\lambda}{\succ} t_2 \text{ iff } \left|\{c_t \mid Q_{ct}^{t_1} > Q_{ct}^{t_2}, c_t \in C\}\right| = \lambda \qquad (4.5)$$

For example, in Figure 4.5A, all of the values in the 12-hour column are larger than all ten of the values in the 6-hour column, so the 12-hour timepoint *λ-dominates* the 6-hour timepoint at λ =10.

Second, the *λ-score* of a timepoint *t* is defined as the number of timepoints other than *t* that are *λ-dominated* by *t*, which is mathematically defined as:

$$\lambda - score(j,\lambda) = \left|\{j' \mid j \neq j', j \in T_A, j \underset{\lambda}{\succ} j'\}\right| \qquad (4.6)$$

For example, Figure 4.5B shows that timepoint 12 hours has a λ-score of 2 at λ =10.

Optimal timepoints are selected by the ranking generated based on λ *-score* values of timepoints. Initially, λ is set to |*C*| (the number of clusters; 10 in the case study), but if two or more timepoints share the same λ*-score* (such as Timepoints 1.5, 8, 10 and 14 in the case study) then they are compared at $\lambda = |C|-1$ (where timepoint 1.5 outranks the others to get a second-place overall rank). If there remains a tie, then they are compared at $\lambda = |C|-2$, and the process is repeated until each timepoint is ranked. The λ*-scores* for the case study experiment, presented in Figure 4.5B, show that 12 hours is the optimal timepoint for selection in the next round of sampling because it has the largest λ*-score* value when $\lambda = 10$. The researcher can then decide the number of selected timepoints to be analyzed. For example, if a microarray experiment is being conducted, then the researcher may only want to choose the top-ranked timepoint, and run two chips (with control and treatment, each in duplicate). However, the high capacity of many next generation sequencing platforms offers the possibility to multiplex several samples per run, so several highly-ranked optimal timepoints could be accommodated per sequencing run (as in the top-up sampling experimental design outlined in section 4.4.2) (Islam et al., 2011).

4.4 Results / Discussion

Three experiments were used to verify the performance of OTS, and the performance was compared with uniform distribution and with active learning timepoint selection (where applicable). The OTS software has been implemented with Microsoft C#, and its executable code, manual, instructions for installation, as well as the datasets used in this manuscript are available for download from http://flash.lakeheadu.ca/~wqin/OTS/index.htm

4.4.1 Dataset description

We tested the performance of OTS on gene expression datasets from two different organisms. First, we selected *Arabidopsis*, for which certain gene functions are well studied but dense time-series gene expression datasets are difficult to find. Here, coronatine/jasmonate-associated microarray datasets are used as training data for a RNA-seq experiment, to demonstrate the effectiveness of OTS across platforms (Fig. 4.3A). This experiment also demonstrates the effectiveness of OTS when using relatively sparsely sampled training datasets, as well as using different but related biological treatments in the training datasets. Because coronatine is a toxin produced by *Pseudomonas syringae* pv. tomato DC3000 (*Pst.* DC3000), and is a molecular mimic of the jasmonate hormone, which mediates would response in *Arabidopsis* (Thilmony et al., 2006), several different datasets utilizing coronatine, *Pst.* DC3000 and wounding treatments were used as training input into the algorithm. In addition, a pilot microarray experiment using a coronatine treatment with three timepoints is also incorporated as a training dataset.

The second organism tested was yeast, and utilized two very high resolution microarray experiments (from Pramila *et al.* (2006) which had 25 timepoints (every 5 minutes from 0 minutes to 120 minutes), making them two of the most densely sampled time-series experiments ever conducted (Edgar et al., 2002; Parkinson et al., 2009). One of these datasets was used as the current mock dataset, and the other was used as one of four training datasets (Fig. 4.3B). In the Pramila *et al* and Spellman *et al* experiments, α-factor synchronization is used to synchronize the cell cycles of the yeast cells to the G_1 phase (Pramila et al., 2006; Spellman et al., 1998), while in the Cho *et al* experiment, temperature changes were also used to synchronize cell cycles to the G_1 phase (Cho et al., 1998).

4.4.2 Experiment design

Three different OTS test experiments were performed, in order to demonstrate its performance using different datasets and different sampling strategies. The first iterative-online sampling experiment uses the *Arabidopsis* coronatine-response datasets. Because the treatments in the current and training datasets activate jasmonate-responsive genes (Chung et al., 2008) and because jasmonate responses are linked to the circadian clock (Goodspeed et al., 2012), the GO-SLIM categories “involved in response to jasmonic acid synthesis” (GO:0009753, 139 genes) and “involved in circadian rhythm” (GO: 0007623, 76 genes) were used, for a total of 195 genes after removing duplicate genes, genes not expressed (zero counts) in the RNA-seq dataset and genes not available in the microarray platform. In this experiment, only the first and last timepoint from the current dataset were used as input, with five additional timepoints added one-at-a-time, to simulate iterative-online sampling on an initially very sparse dataset.

Another iterative-online sampling experiment was run with the yeast cell-cycle datasets. In this experiment, we have chosen all of the genes available in every training and current dataset in the GO category “mitosis” (*i.e.*, cell division; GO:0007067, 90 genes total) (Ashburner et al., 2000) as the set of target genes. This yeast experiment demonstrates the ability of OTS to work despite a great deal of noise; cell cycle patterns are only weakly reproducible (even between replicates), α-factor synchronization and temperature treatments may elicit stress-responses in addition to cell cycle differences (Cooper and Shedden, 2003), and a very diverse set of genes with very different functions is responsible for mitosis (Cho et al., 1998), all of which add considerably to the overall noise in the experiment.

Besides the “iterative-online” sampling experiment, a different sampling strategy was also used with the same yeast cell-cycle datasets. In this experiment, we start with five evenly distributed timepoints (at 5, 30, 60, 90 and 120 minutes), and then add two more timepoints as a batch to “top-up” the timepoints sampled, simulating the situation of choosing extra timepoints after conducting a pilot sampling which is determined by researcher’s knowledge/intuition.

In the *Arabidopsis* experiment, very large differential gene expression values were expected for jasmonic-acid response, based on the literature (Chung et al., 2008; Wierstra and Kloppstech, 2000), and very low levels of noise are expected in the RNA-seq dataset (Marioni et al., 2008), so a threshold number (*H*) of 6 was selected in the OTS to preferentially capture these larger changes in expression. In contrast, for yeast experiments, a threshold number (*H*) of 3 was used in order to reduce the high expected noise in the datasets (Cooper and Shedden, 2003), by ignoring the small fluctuations in gene expression measurements. Ten clusters were used in the *Arabidopsis* dataset, and 8 clusters were used in the yeast dataset.

As a comparison, Singh et al's (Singh et al., 2005) active learning algorithm was also used to choose optimal timepoints, using the same number of clusters as OTS. However, the active learning algorithm requires at least five timepoints as initial input, so in the iterative-online experiments (which start with only two timepoints) the first three selected timepoints were chosen using a uniform distribution across the time series, and the last two timepoints were chosen using the active learning algorithm (Singh et al., 2005). For the iterative-online experiments, random timepoint selection was also performed, where timepoints were randomly selected within the time range of each experiment 250 times (for each number of timepoints selected).

4.4.3 Performance measurement

For the performance-testing experiments, given the gene expression data at a subset of all the timepoints, the differential gene expression values at every unsampled timepoint were estimated with linear interpolation (see Figure 4.4A for an example in which 6 out of 20 timepoints have been used for interpolation). By comparing the actual and the interpolated values, we are able to evaluate the performance of the timepoint selection. Figure 4.6A shows a scatterplot of the Interpolated vs. actual Log2 fold change values (for all genes at all available timepoints) in the *Arabidopsis* experiment at the start of the iterative-online experiment, when the interpolated dataset only uses the first (0.25 hours) and last (24 hours) timepoints. A measure of error between the interpolated and actual Log2 differential gene expression values was derived, such that larger errors result from measurements with (a) poor agreement between the actual and interpolated values and (b) large actual differential expression. For example, in Figure 4.6A, one gene at one timepoint has an actual Log2 fold change value of 8.4 and an interpolated Log2 fold change value of 3.1, giving an error area of 22.2 according to Equation 4.7. At this timepoint, this gene has a large error area (highlighted in white) because it has a poor agreement between the actual and interpolated values (far from the diagonal line), and it has large actual differential gene expression value (far from the center point of the x axis).

These error areas are summed for all genes and timepoints to calculate the Sum Error Area (E_{TA}) for each set of timepoints used for the interpolation, according to Equation 4.7. To measure performance, these values are presented as a percentage of the initial sum error area (the sum error area at the start of each experiment).

$$E_{T_A} = \frac{1}{2}\sum_{g \in G}\sum_{i=1}^{|T_A|}\left|(\hat{e}_g^i + \phi)\cdot(\hat{e}_g^i - e_g^i)\right| \qquad (4.7)$$

where E_{TA} is the Sum Error Area for interpolation based on a subset of the full dataset timepoints (T_S), $\hat{e}_g^i$ (or e_g^i) is the actual (or interpolated) differential gene expression value for gene *g* in the set of genes of interest *G* at timepoint *i* ($1 \le I \le |T_A|$), ϕ is a small constant with the same sign as $\hat{e}_g^i$.

4.4.4 Test experiment results

Figure 4.7 shows that in the iterative-online experiments, OTS consistently chooses timepoints which model the differential gene expression patterns of target genes more effectively than random selection and uniform distribution followed by the active learning algorithm. The increase in performance of OTS is particularly clear for the early timepoints selected. In the *Arabidopsis* experiment (Fig. 4.7A), the addition of the first OTS timepoint (at 5 hours) reduces the error in the interpolation by more than 50%, compared with uniform distribution, which had only a 32% reduction in the initial error. Random timepoint selection (for one timepoint, the average of all of the timepoints) also had a 32% reduction in the initial error. Figure 4.6B shows that a large portion of the error in the uniform distribution selection results from measurements that are predicted to be low or negative fold change values, but are actually very high (circles near the bottom right), but OTS successfully identifies most of these upregulation events. The second timepoint added by OTS (at 2 hours) further reduced the error to

35% of the initial error, compared with uniform distribution, which still has 66% of the initial error rate after the addition of two timepoints and still misses many of the very strong upregulation events (Fig. 4.6C). Random timepoint selection still had 46.5% of the initial error after the addition of two timepoints, suggesting that it outperforms uniform distribution selection, but is not as optimal as OTS selection. Overall, OTS has a lower error for every number of timepoints added in this experiment, including the first two timepoints for which active learning could be applied (Fig. 4.7A).

The timepoint selection distribution for every round in the iterative-online *Arabidopsis* experiment is shown in Figure 4.8A, which shows an early-timepoint selection bias by OTS. Figures 4.9A and 4.9B show a case study example of the interpolated differential gene expression pattern after the addition of five OTS timepoints for a bHLH transcription factor called *MYC3* (*AT5G46760*), which interacts with Jasmonate ZIM-domain proteins to mediate the jasmonate response (Fernández-Calvo et al., 2011; Xie et al., 1998). Unlike the interpolated differential gene expression pattern using uniform distribution followed by active learning (Fig. 4.9B), the strong upregulation peak from 2 to 5 hours is well-defined by OTS interpolation (Fig. 4.9A), with the correct peak time and strength accurately captured. The early timepoint at 0.5 hours that was suggested by active learning fails to capture this peak, as the gene expression level does not significantly increase until an hour after treatment (Fig. 4.9B). By suggesting more early timepoints, OTS is able to more effectively define the early peaks of coronatine-induced genes, and after four rounds of selecting early timepoints, OTS chooses 12 hours as the fifth timepoint, leading to good overall coverage of the time range.

In the similarly designed iterative-online yeast experiment, despite large amount of noise among datasets and large differences among genes (see section 4.4.2), OTS outperformed uniformly-distributed timepoint selection (Fig. 4.7B), reducing the error by 35% after the addition of just one timepoint (compared with only 5% in uniform distribution, which selected the center timepoint at 60 minutes, and 16% in random selection). The error plot in Figure 4.6E shows that the strongest upregulation and downregulation events (close to the bottom left and bottom right of the plot) are much more accurately defined by the timepoint selection in OTS, compared with active learning. This trend continues in the plot in Figure 4.6F; after the addition of two timepoints, OTS selection reduces the initial sum error area by 47%, compared with just 14% in active learning and 24.1% in random selection (Fig. 4.7B). At the end of the experiment (five timepoints added, for a total of seven timepoints), the initial error is reduced by 56.1% using OTS timepoints, compared with 51.3% using active learning timepoints and 50.3% using random selection. Like in the *Arabidopsis* experiment, OTS outperforms active learning at every number of timepoints tested, and in yeast there also appears to be a bias towards early timepoints (Fig. 4.8B), probably due to stronger and more co-ordinated cyclic gene responses immediately after synchronization (Cho et al., 1998; Spellman et al., 1998), and because many yeast cell cycle genes peak in late G_1, the point at which the cell needs to "decide" whether to divide or to continue to grow (Rodriguez-Sanchez et al., 2011).

Figures 4.9C and 4.9D show the interpolated differential gene expression curves for OTS and uniform distribution/active learning (respectively) for *Mitotic Arrest Deficient 1* (*MAD1; YGL086W*), which encodes a protein that is critical for regulating the transition into anaphase, by modulating the activity of the mitotic spindle (Chen et al., 1999). At the end of the iterative/online experiment, OTS-timepoint selection has accurately defined the peak expression of this cell-cycle gene (occurring at 40 minutes, Figure 4.9C), whereas uniform distribution/active learning-timepoint selection has missed the true peak expression of this gene, and incorrectly identified the peak time at 60 minutes (Fig. 4.9D).

In the top-up yeast experiment, OTS also outperformed uniformly-distributed timepoint selection, reducing the error by 25.9% after the addition two timepoints to the initial five timepoints, compared with 14.4% for the two active-learning timepoints. The error plot in Figure 4.6H shows that the strongest upregulation events (close to the bottom-right side of the plot) are much more accurately defined by the timepoint selection in OTS, compared with active learning. The timepoint selection distribution in Figure 4.8C shows that active learning adds timepoints far apart at 25 and 95 minutes (very close to existing timepoints in the dataset), while OTS has an early bias (as in the iterative-online experiment), suggesting 10 and 20 minutes as the optimal timepoints.

Figures 4.9E and 4.9F show the differential gene expression interpolation results in the top-up experiment for *APC/C^{Cdh1} modulator 1* (*ACM1*; *YPL267W*), which encodes a protein that inhibits key cell cycle proteins, and is cell-cycle regulated (appearing late in G_1, and disappearing in late M phase) (Martinez et al., 2006). The initial interpolation based on 5 evenly distributed timepoints (every 30 minutes; grey line, Figure 4.9E) suggests that this gene has a peak Log2 fold change of 0.22, occurring at 30 minutes. The 25 minute timepoint added by active learning (Fig. 4.9F) has a Log2 fold change value of 0.36, but still misses the true peak of 0.69 (occurring at 20 minutes), which is captured by the OTS timepoint. Furthermore, compared with active learning, OTS generally captured the expression peaks very well for the many yeast cell cycle genes that peak in late G_1. Overall, this yeast top-up study demonstrates that OTS timepoint selection has nearly twice the performance of active learning selection, by identifying strong upregulation and downregulation events which more accurately model true gene expression patterns for more cell cycle genes, providing more biologically interesting information.

4.5 Conclusion

Here, we have demonstrated that OTS can out-perform existing algorithms in finding optimal timepoints for defining differential gene expression patterns for large groups of target genes, by utilizing training data as a guide. We have demonstrated that the algorithm is robust to noise and to sparsely-sampled, poorly matched, and cross-platform training data. Because it relies on training data as well as the existing expression pattern in the current dataset, OTS can be applied on datasets containing as few as two timepoints and still produce strong results, in contrast to the best existing algorithm which requires a minimum of five timepoints as input (Singh et al., 2005). Although OTS was tested here on differential gene expression values, it can potentially also be used on other types of data, including raw transcript number counts, relative protein quantities, or any other type of measurement that can be sampled in an online fashion. Overall, OTS can be used to significantly improve the results from biological experiments, by allowing researchers to optimize the distribution of timepoints when there is a limit on the number of samples that can be measured across a time-series dataset, and to our knowledge is the first to utilize existing datasets in timepoint suggestion.

4.6 Chapter 4 Tables

Table 4.1: Cluster-Time-Experiment (CTE) Table storing Differential Regulation Count (DRC) values for cluster c. Each row represents a timepoint available for sampling in the current dataset (T_A), each column represents a training dataset (R_j), the current dataset (U), or the estimate dataset (E) and each value in the table is a DRC value for timepoint i in training data j (D^{ij}), current dataset ($\hat{D}^{i}$) or estimate dataset ($\bar{D}^{i}$).

Timepoint	Training Datasets				Current Dataset	Estimate
	R_1	R_2	…	R_m	$\hat{D}$	$\bar{D}$
t_1	D^{11}	D^{12}	…	D^{1m}	$\hat{D}^{1}$	$\bar{D}^{1}$
t_2	D^{21}	D^{22}	…	D^{2m}	$\hat{D}^{2}$	$\bar{D}^{2}$
…	…	…	D^{ij}	…	…	…
t_n	D^{n1}	D^{n2}	…	D^{nm}	$\hat{D}^{n}$	$\bar{D}^{n}$

4.7 Chapter 4 Figures

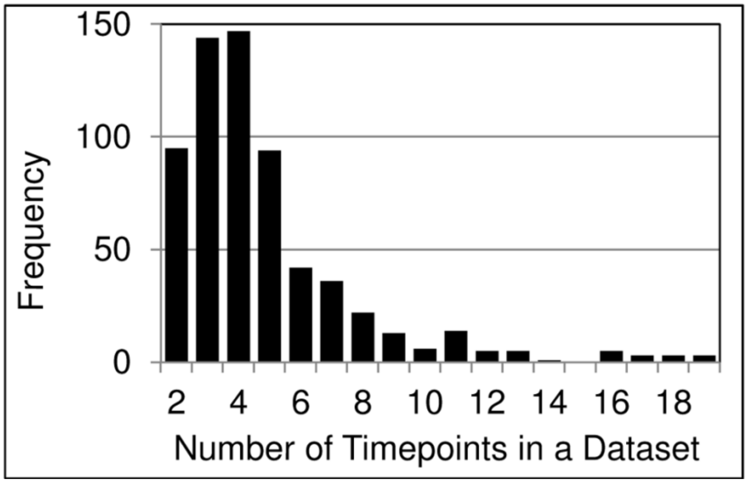


Figure 4.1: Histogram of the number of timepoints in each time-series high-throughput gene expression dataset in the GEO database.

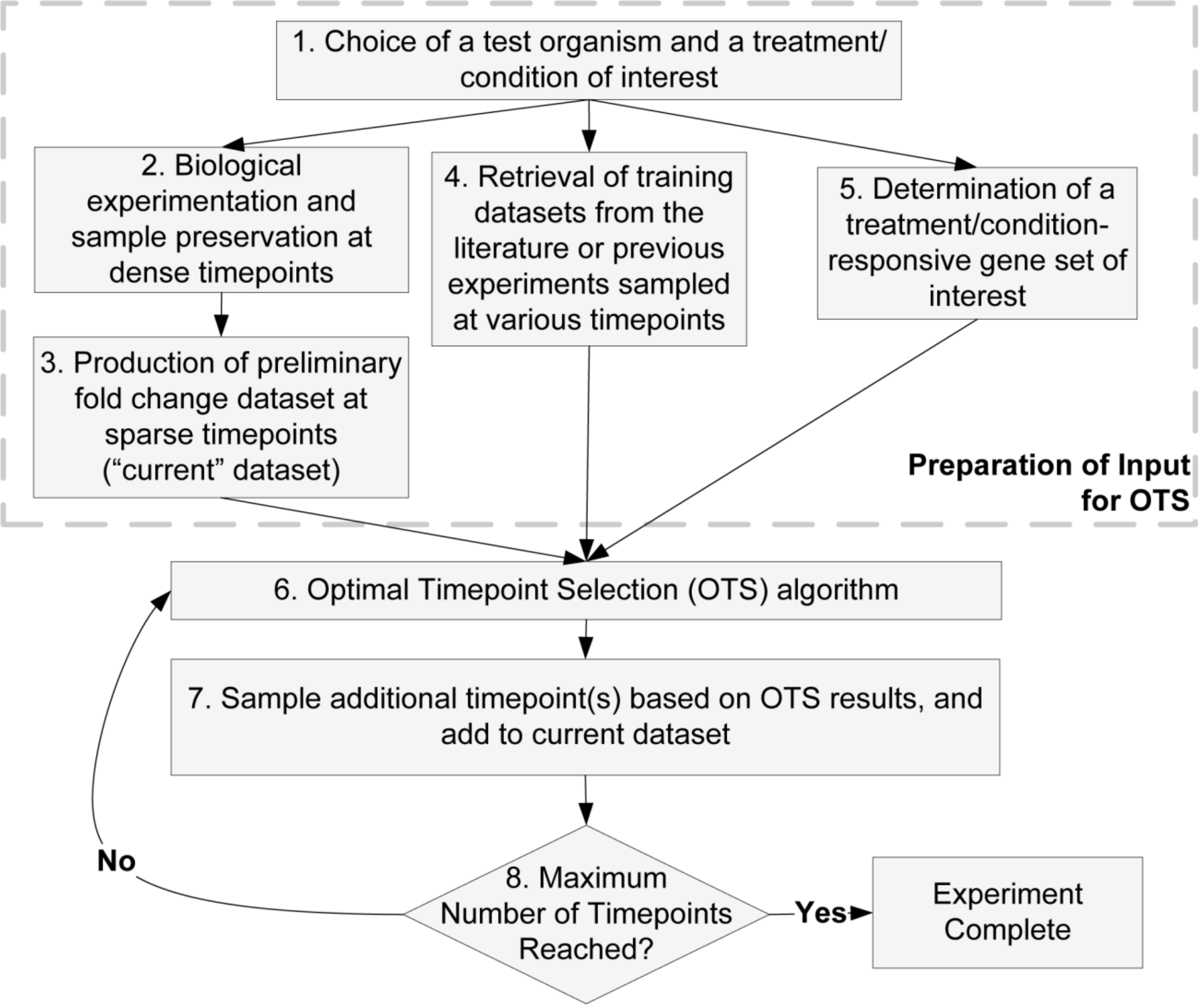
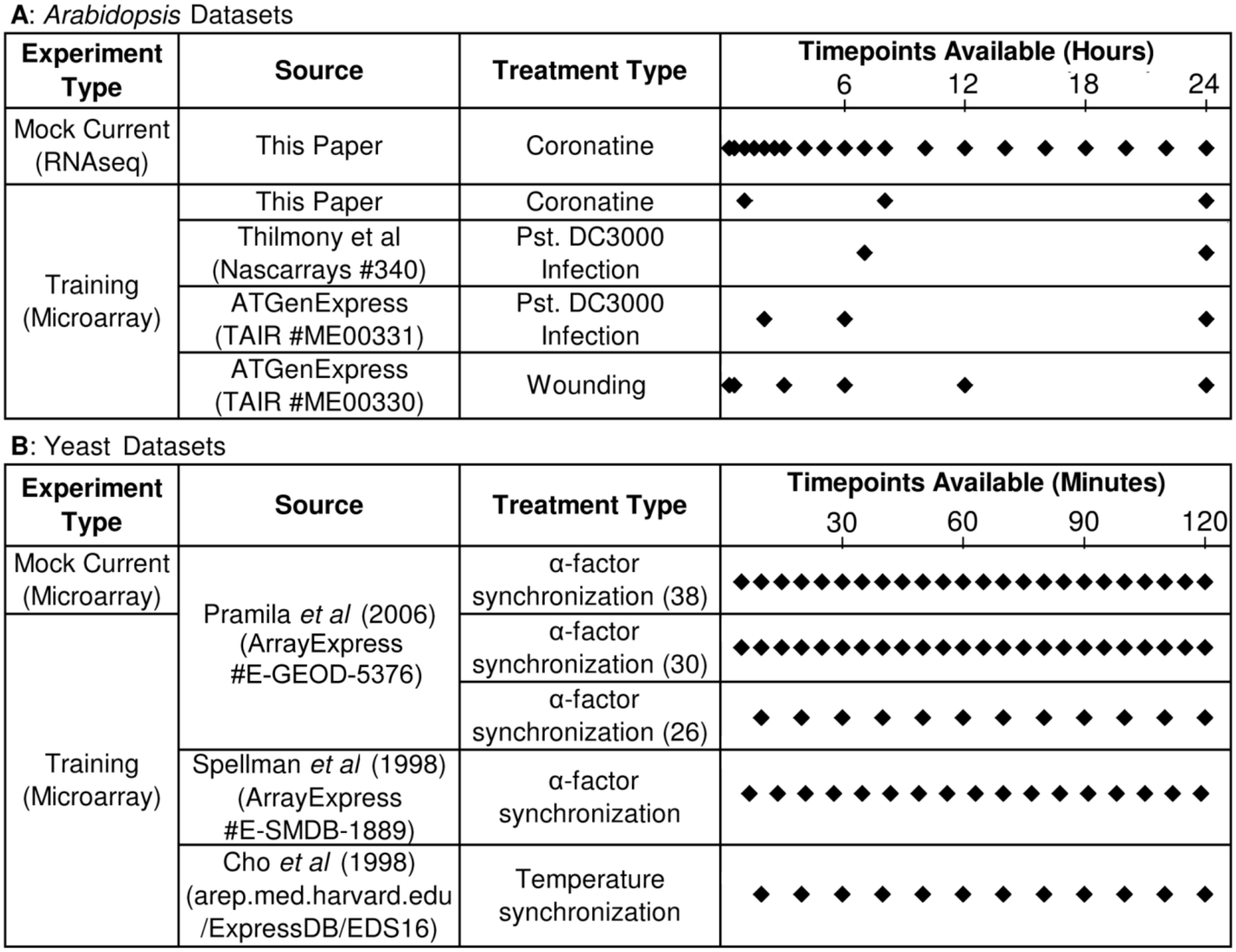

1. Choice of a test organism and a treatment/ condition of interest
2. Biological experimentation and sample preservation at dense timepoints
3. Production of preliminary fold change dataset at sparse timepoints ("current" dataset)
4. Retrieval of training datasets from the literature or previous experiments sampled at various timepoints
5. Determination of a treatment/condition-responsive gene set of interest
Preparation of Input for OTS
6. Optimal Timepoint Selection (OTS) algorithm
7. Sample additional timepoint(s) based on OTS results, and add to current dataset
8. Maximum Number of Timepoints Reached?
No
Yes
Experiment Complete

Figure 4.2: Flowchart depicting the overall experimental approach for utilizing OTS



A: *Arabidopsis* Datasets

Experiment Type	Source	Treatment Type	Timepoints Available (Hours)
Mock Current (RNAseq)	This Paper	Coronatine	◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆
Training (Microarray)	This Paper	Coronatine	◆ ◆ ◆
	Thilmony et al (Nascarrays #340)	Pst. DC3000 Infection	◆ ◆
	ATGenExpress (TAIR #ME00331)	Pst. DC3000 Infection	◆ ◆ ◆
	ATGenExpress (TAIR #ME00330)	Wounding	◆◆ ◆ ◆ ◆ ◆

B: Yeast Datasets

Experiment Type	Source	Treatment Type	Timepoints Available (Minutes)
Mock Current (Microarray)	Pramila *et al* (2006) (ArrayExpress #E-GEOD-5376)	α-factor synchronization (38)	◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆
Training (Microarray)		α-factor synchronization (30)	◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆◆
		α-factor synchronization (26)	◆ ◆ ◆ ◆ ◆ ◆ ◆ ◆ ◆ ◆ ◆ ◆
	Spellman *et al* (1998) (ArrayExpress #E-SMDB-1889)	α-factor synchronization	◆ ◆ ◆ ◆ ◆ ◆ ◆ ◆ ◆ ◆ ◆ ◆ ◆ ◆ ◆ ◆ ◆
	Cho *et al* (1998) (arep.med.harvard.edu /ExpressDB/EDS16)	Temperature synchronization	◆ ◆ ◆ ◆ ◆ ◆ ◆ ◆ ◆ ◆ ◆ ◆

Figure 4.3: Description of current and training datasets used for (A) the Arabidopsis testing experiments and (B) the yeast experiment.

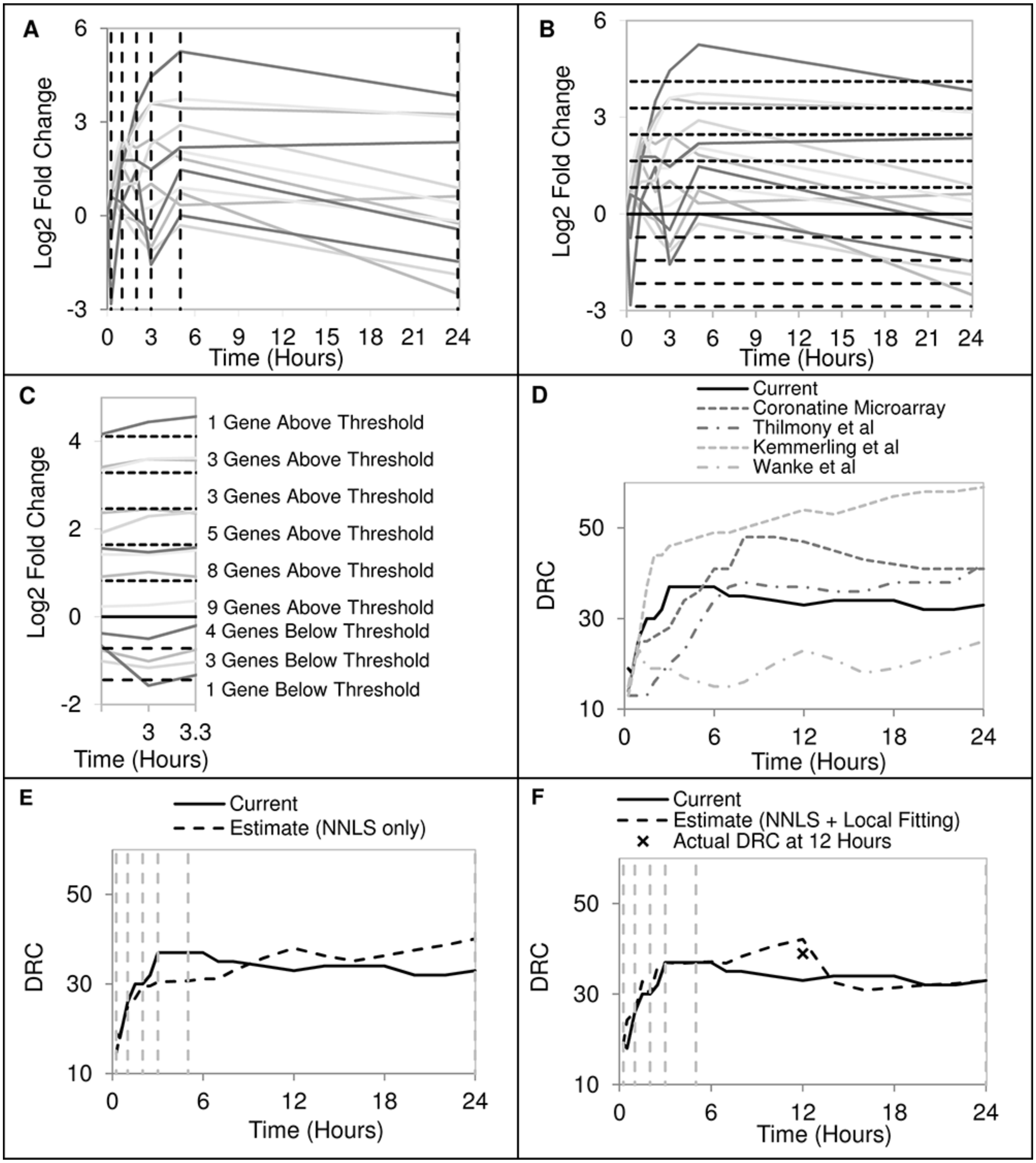
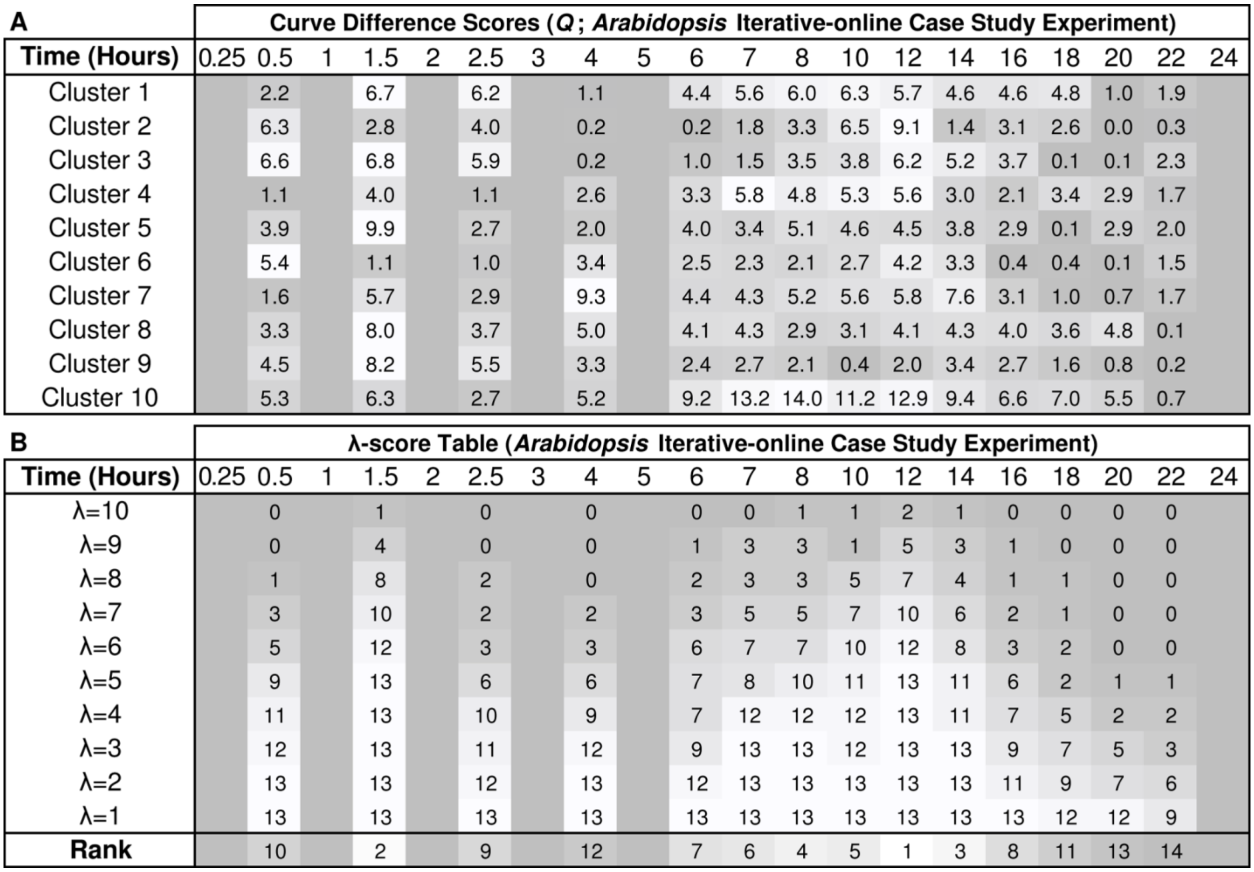


Figure 4.4: The conversion of differential gene expression value curves in one cluster to a differential regulation count (DRC) curve. One of the clusters in the "current" dataset of the Arabidopsis experiment is shown after the addition of four OTS timepoints. (A) Differential gene expression values from the current dataset are used to interpolate differential gene expression values to every available preserved

timepoint in the biological experiment. Vertical dashed black lines indicate actual timepoints used as input and the lines of various shades of grey represent the differential gene expression curves for individual genes. (B) Upregulation and downregulation thresholds are determined according to the average and standard deviation of the uninterpolated training datasets (horizontal black dashed lines). (C) The number of genes crossing each threshold value are counted at each timepoint. Counting is shown at 3 hours, indicated with vertical dotted black line ($\widehat{D}^{3}$ = 1+3+3+5+8+9+4+3+1=37; Equation 4.2). (D) The DRC curves for one cluster in the current dataset (black line) and the four training datasets (grey lines) are shown. (E) The current DRC curve (4black line), and the estimate DRC curve suggested by NNLS regression alone (dashed line). Actual sampled timepoints are indicated with vertical dashed grey lines (F) The current DRC curve (black line), the final estimate DRC curve (shifted according to the Equation 4.3; dashed line), and the actual DRC value for the optimal timepoint at 12 hours (black x mark; hidden from OTS, but shown for demonstration). Actual sampled timepoints are indicated with vertical dashed grey lines



A	Curve Difference Scores (*Q*; *Arabidopsis* Iterative-online Case Study Experiment)																				
Time (Hours)	0.25	0.5	1	1.5	2	2.5	3	4	5	6	7	8	10	12	14	16	18	20	22	24
Cluster 1		2.2		6.7		6.2		1.1		4.4	5.6	6.0	6.3	5.7	4.6	4.6	4.8	1.0	1.9	
Cluster 2		6.3		2.8		4.0		0.2		0.2	1.8	3.3	6.5	9.1	1.4	3.1	2.6	0.0	0.3	
Cluster 3		6.6		6.8		5.9		0.2		1.0	1.5	3.5	3.8	6.2	5.2	3.7	0.1	0.1	2.3	
Cluster 4		1.1		4.0		1.1		2.6		3.3	5.8	4.8	5.3	5.6	3.0	2.1	3.4	2.9	1.7	
Cluster 5		3.9		9.9		2.7		2.0		4.0	3.4	5.1	4.6	4.5	3.8	2.9	0.1	2.9	2.0	
Cluster 6		5.4		1.1		1.0		3.4		2.5	2.3	2.1	2.7	4.2	3.3	0.4	0.4	0.1	1.5	
Cluster 7		1.6		5.7		2.9		9.3		4.4	4.3	5.2	5.6	5.8	7.6	3.1	1.0	0.7	1.7	
Cluster 8		3.3		8.0		3.7		5.0		4.1	4.3	2.9	3.1	4.1	4.3	4.0	3.6	4.8	0.1	
Cluster 9		4.5		8.2		5.5		3.3		2.4	2.7	2.1	0.4	2.0	3.4	2.7	1.6	0.8	0.2	
Cluster 10		5.3		6.3		2.7		5.2		9.2	13.2	14.0	11.2	12.9	9.4	6.6	7.0	5.5	0.7	

B	λ-score Table (*Arabidopsis* Iterative-online Case Study Experiment)																				
Time (Hours)	0.25	0.5	1	1.5	2	2.5	3	4	5	6	7	8	10	12	14	16	18	20	22	24
λ=10		0		1		0		0		0	0	1	1	2	1	0	0	0	0	
λ=9		0		4		0		0		1	3	3	1	5	3	1	0	0	0	
λ=8		1		8		2		0		2	3	3	5	7	4	1	1	0	0	
λ=7		3		10		2		2		3	5	5	7	10	6	2	1	0	0	
λ=6		5		12		3		3		6	7	7	10	12	8	3	2	0	0	
λ=5		9		13		6		6		7	8	10	11	13	11	6	2	1	1	
λ=4		11		13		10		9		7	12	12	12	13	11	7	5	2	2	
λ=3		12		13		11		12		9	13	13	12	13	13	9	7	5	3	
λ=2		13		13		12		13		12	13	13	13	13	13	11	9	7	6	
λ=1		13		13		13		13		13	13	13	13	13	13	13	12	12	9	
Rank		10		2		9		12		7	6	4	5	1	3	8	11	13	14	

Figure 4.5: Curve difference score (Q) and λ -score tables used to determine the optimal timepoint for all the selected genes in the case study experiment. (A) Curve difference scores are calculated at every timepoint and every cluster according to Equation 4.4. The highest curves difference scores in each cluster (row) are indicated with the brightest shades, and empty columns represent timepoints already sampled (TS) (B) λ -scores are calculated to determine the optimal timepoint by comparing curve difference scores according to Equation 4.5. Brighter shading indicates higher λ-score values and lower (better) ranks. The optimal timepoint selected here is at 12 hours because it has the highest λ-score at λ=10. The second-best ranked timepoint is 1.5 hours because although its λ-score is tied with several timepoints at λ=10, it has a higher λ-score than these tied timepoints at λ=9.

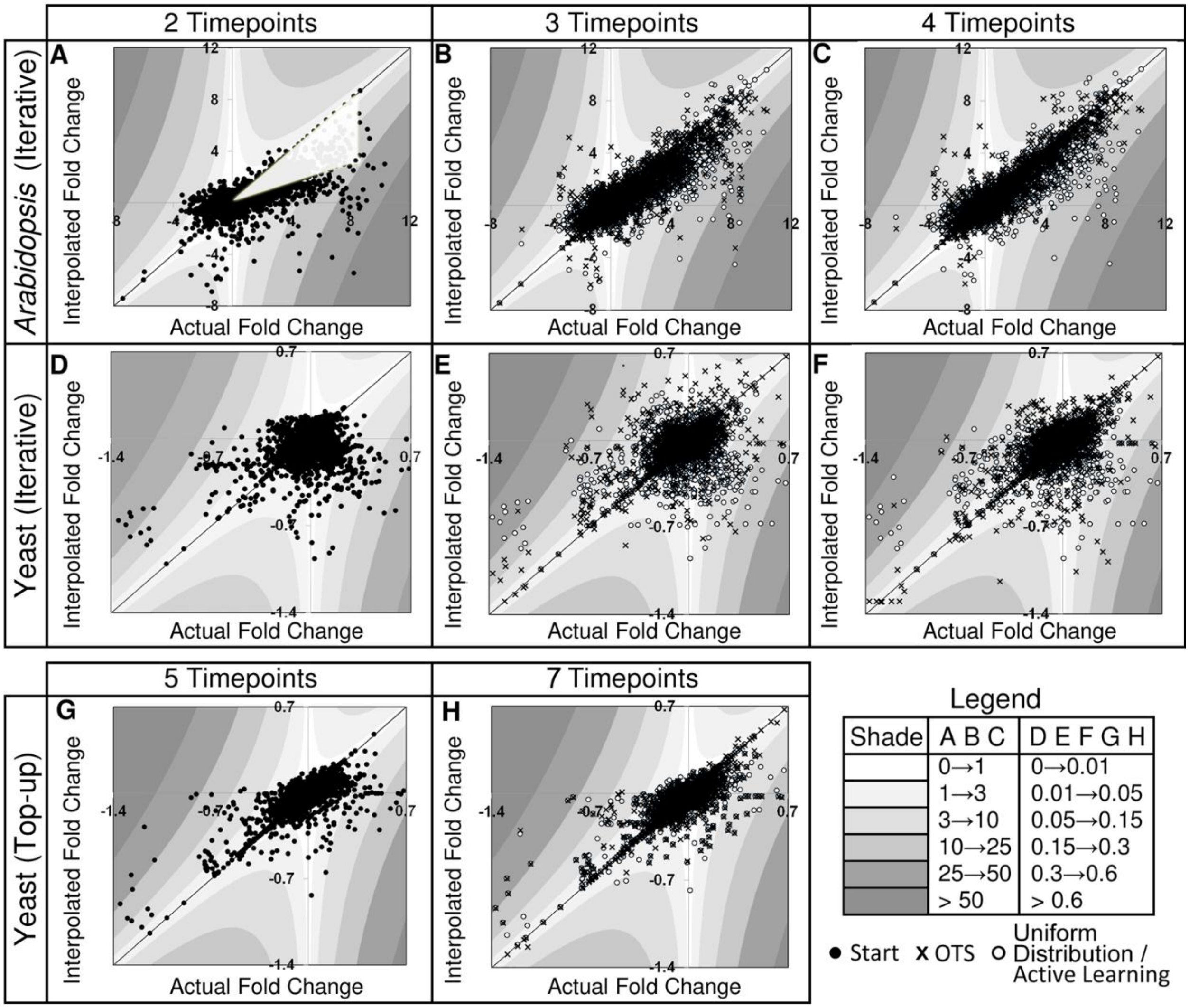


Figure 4.6: The interpolated Log2 fold change values plotted against the actual Log2 fold change values for every gene at every timepoint in the dataset. Shaded grey areas indicate the error area (Equation 4.7) for a given region of each plot, with darker shades indicating higher error area values. Black circles represent the values at the start of each experiment, black X marks represent values based on OTS-selected timepoints, and hollow circles represent values based on uniform distribution/active-learning selected timepoints. (A) Interpolation based on the first two timepoints (0.25 hours and 24 hours) in the iterative-online Arabidopsis experiment. The highlighted triangular area indicates the the “error area” for one gene at one timepoint, where the actual value is 8.4 and the interpolated value is 3.1, giving an

error area of 22.2 according to Equation 4.7. (B) Interpolation after the addition of one OTS and one active learning timepoint in the iterative-online Arabidopsis experiment. (C) Interpolation after the addition of two OTS and two active learning timepoints in the iterative-online Arabidopsis experiment. (D) Interpolation based on the first two timepoints (5 minutes and 120 minutes) in the iterative-online Yeast experiment. (E) Interpolation after the addition of one OTS and one active learning timepoint in the iterative-online Yeast experiment. (F) Interpolation after the addition of two OTS and two active learning timepoints in the iterative-online yeast experiment. (G) Interpolation based on the first five timepoints (every 30 minutes) in the top-up yeast experiment. (H) Interpolation after the addition of two OTS and two active learning timepoints in the top-up yeast experiment.

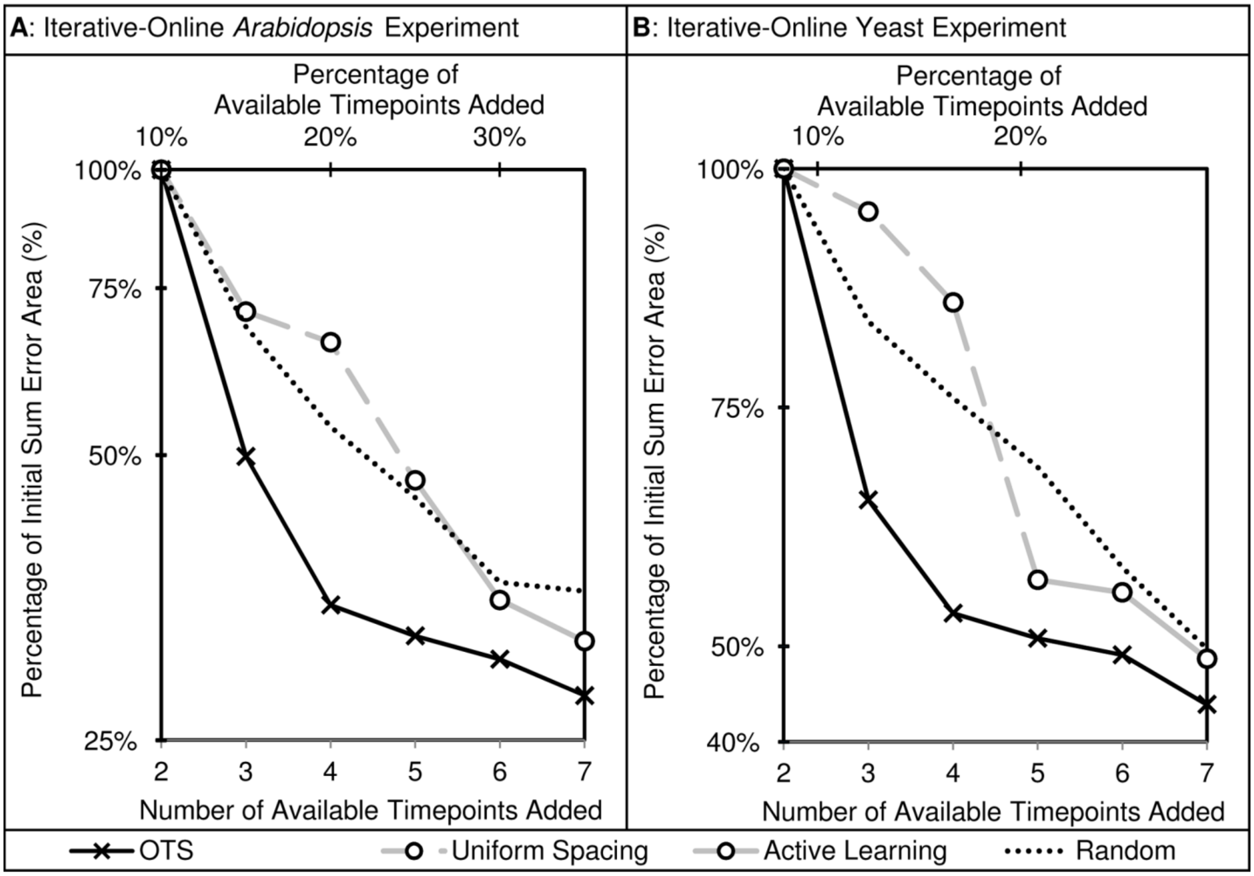


Figure 4.7: Sum Error Area values for the Arabidopsis and yeast iterative-online experiments. The results for OTS (black line, black X marks), uniform distribution (dashed grey line, white circles) followed by active learning (solid grey line, white circles) and random selection (dotted line) are shown for (A) the iterative-online Arabidopsis experiment (195 genes) and (B) the iterative-online yeast experiment (90 genes). OTS outperformed uniform distribution/active learning and random selection at every number of timepoints added.

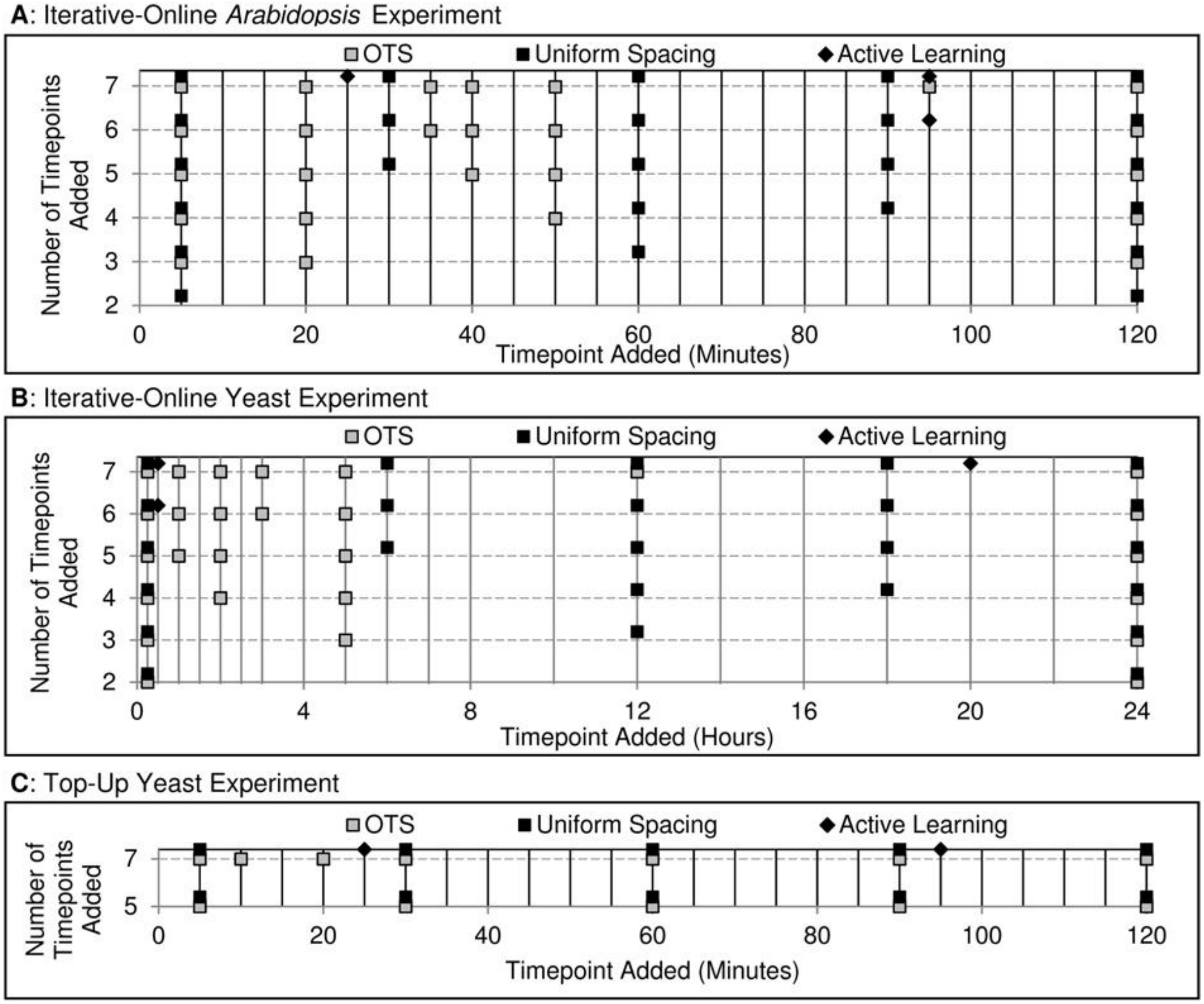


Figure 4.8: The timepoint selection distribution for every round of timepoint addition in the iterative-online Arabidopsis experiment and the top-up Yeast experiment. (A) Timepoints selected in the iterative-online *Arabidopsis* experiment and (B) Timepoints selected in the iterative-online Yeast experiment (C) Timepoints selected in the top-up Yeast experiment. For all experiments, grey squares represent timepoints initially inputted or or selected by OTS, black squares represent timepoints selected by uniform distribution, and black diamonds represent timepoints selected by active learning.

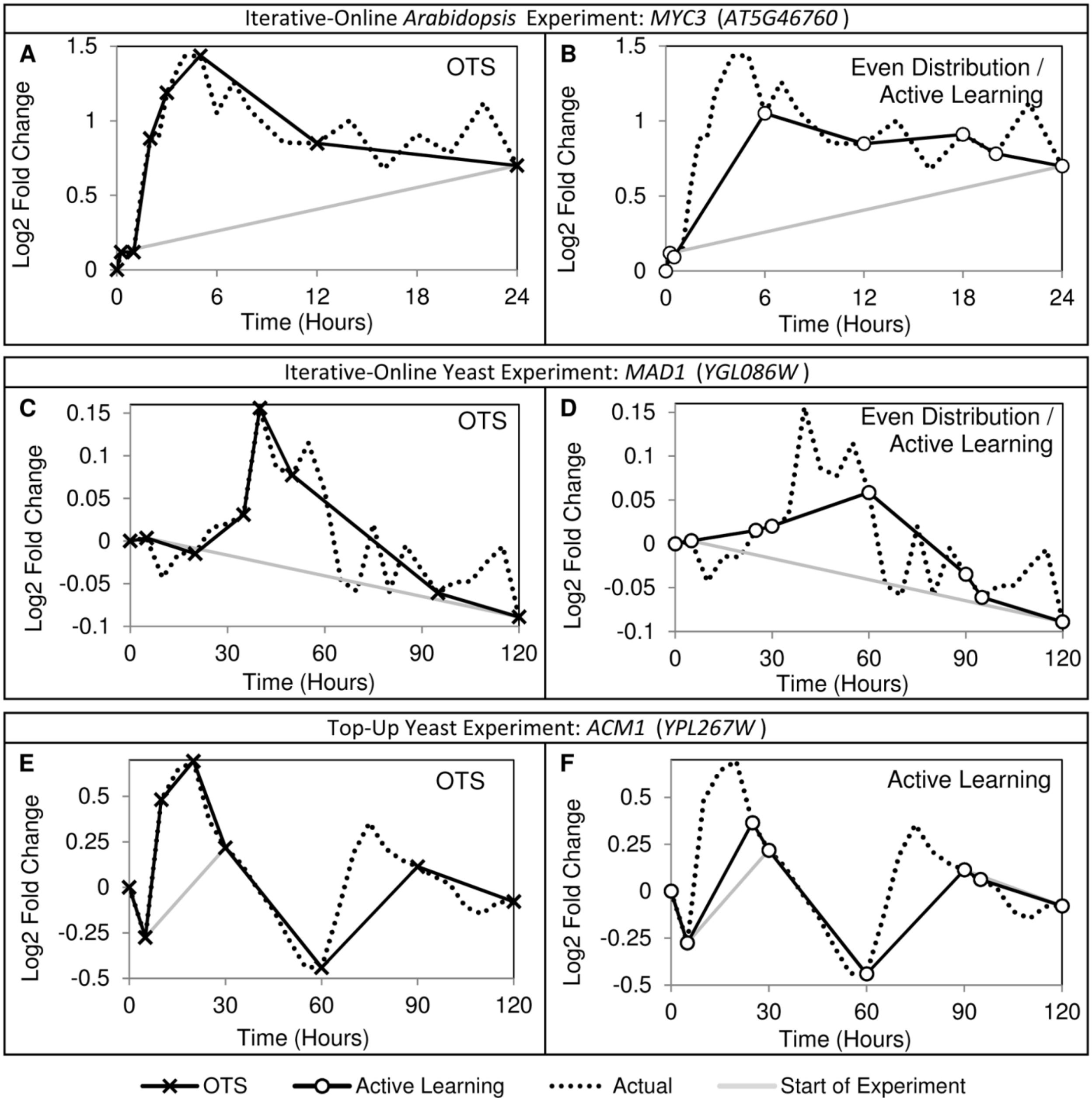
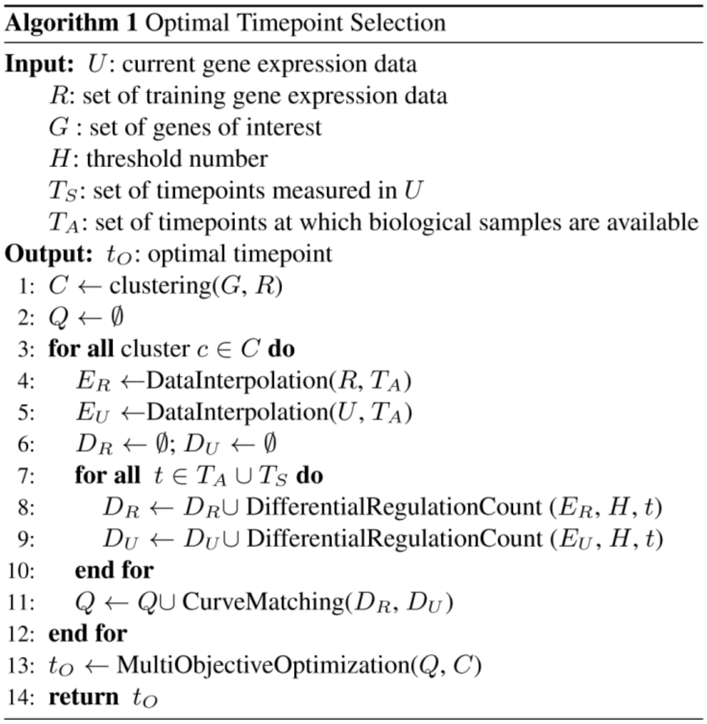


Figure 4.9: Case studies for one gene in each of the three timepoint selection experiments. For the iterative-online *Arabidopsis* experiment, the interpolated differential gene expression patterns of the jasmonate-responsive bHLH transcription factor *MYC3* (*AT5G46760*) are shown based on (A) the addition of five OTS timepoints and (B) based on the addition of three uniformly distributed and two active learning timepoints. For the iterative-online yeast experiment, the interpolated differential gene expression patterns of *Mitotic Arrest Deficient 1* (*MAD1*; *YGL086W*) are shown based on (C) the addition

of five OTS timepoints and (D) based on the addition of three uniformly distributed and two active learning timepoints. For the top-up yeast experiment, the interpolated differential gene expression patterns of *APC/C*^{Cdh1} *modulator 1* (*ACM1; YPL267W*) are shown based on (E) the addition of two OTS timepoints and (F) the addition of two active learning timepoints.

4.8 Chapter 4 Algorithms



Algorithm 1 Optimal Timepoint Selection

```
Input: $U$: current gene expression data
    $R$: set of training gene expression data
    $G$ : set of genes of interest
    $H$: threshold number
    $T_S$: set of timepoints measured in $U$
    $T_A$: set of timepoints at which biological samples are available
Output: $t_O$: optimal timepoint
$C \leftarrow$ clustering($G$, $R$)
$Q \leftarrow \emptyset$
for all cluster $c \in C$ do
    $E_R \leftarrow$DataInterpolation($R$, $T_A$)
    $E_U \leftarrow$DataInterpolation($U$, $T_A$)
    $D_R \leftarrow \emptyset$; $D_U \leftarrow \emptyset$
    for all $t \in T_A \cup T_S$ do
        $D_R \leftarrow D_R \cup$ DifferentialRegulationCount ($E_R$, $H$, $t$)
        $D_U \leftarrow D_U \cup$ DifferentialRegulationCount ($E_U$, $H$, $t$)
    end for
    $Q \leftarrow Q \cup$ CurveMatching($D_R$, $D_U$)
end for
$t_O \leftarrow$ MultiObjectiveOptimization($Q$, $C$)
return $t_O$
```

Algorithm 4.1: The pseudocode for Optimal Timepoint Selection (OTS). Refer to section 4.3.3 to 4.3.7 for the introduction of Clustering, DataInterpolation, DifferentialRegulationCount, CurveMatching and MultiObjectiveOptimization respectively.

Chapter 5: Conclusion

High-throughput gene expression datasets allow researchers to sample tens of thousands of genes at once, providing a wealth of information both about known treatment/condition-response genes, as well as previously uncharacterized genes which may be very important targets for further biological study. However, given the varying levels of annotations of the genes, the high rates of noise inherent to these systems, and the complexities of genetic responses to various treatments and conditions, performing accurate novel gene discovery (*i.e.* identifying previously uncharacterized genes which are actually involved in a biological response to a given treatment/condition) is a difficult challenge. In this thesis, three different approaches for improving novel gene discovery in high-throughput gene expression datasets have been presented.

First, a novel knowledge-based approach which identifies and saves important functional genes before filtering based on variability and fold change differences was utilized to study light regulation. When combined with a novel clustering approach, it was shown that this experiment produced relatively short cluster lists compared to gene groups generated through typical clustering methods or coexpression networks, which narrowed the search for novel functional genes while increasing the likelihood that they are biologically relevant. This method can be applied to any single-timepoint experiment in which there is some level of background knowledge about the expected genetic response to the treatment or condition used. In the future, the knowledge-based filtering step presented here can be applied to other analysis approaches in order to prevent the removal of important, and other clustering algorithms could be combined in a similar way as in this study. In addition, the biological results from this study (which suggest that several novel genes may be involved in light regulation) could be verified by biological experiments.

Second, the PRIISM algorithm was developed in order to significantly reduce the complications of circadian clock pathway disruptions in plant novel gene discovery experiments. PRIISM was applied on a high-resolution time-series *Arabidopsis* microarray dataset under a cold treatment, and the results of this study showed that the ranked treatment-frequency fold change results contain fewer false positives than the original methodology. This experiment also showed that many known target response genes had very strong differential regulation through treatment-response pathways at 26 hours, which was not obvious in the original dataset due to the noise resulting from circadian rhythm pathways, showing that in addition to novel gene discovery, this approach may be able to better characterize treatment responses for known genes. In addition, PRIISM also provides gene expression data which represents only circadian clock influences, and may be useful for circadian clock analysis studies. In this study, six strong candidates for novel cold response genes were discovered which could be verified biologically in future experiments. As higher-resolution time series gene expression datasets become available, it would be useful to modify PRIISM to support wavelet analyses, where instead of using a sliding window based on light/dark cycles, every possible sliding window of a given length would be split.

Third, a computational approach was developed to design the timepoint selection in time-series high-throughput gene expression experiments, such that the strong expression changes in known target genes are the most accurately captured, based on expression data from training datasets. By finding the timepoints at which most of the known target genes are differentially expressed, the potential to discover other related novel treatment/condition response genes is significantly increased. OTS was tested using several sampling approaches on several datasets, and consistently outperformed other existing timepoint sampling strategies in terms of defining the true expression patterns of known target genes. In future versions, OTS may be modified to include more advanced interpolation algorithms, different clustering algorithms, and may be tested on much higher-resolution training and current datasets as they become available.

Overall, the computational projects outlined in this thesis have contributed significantly to the field of novel gene discovery in high-throughput gene expression datasets. Methods have been developed to decrease noise (and increase the accuracy of novel gene discovery) in both static as well as time-series gene expression datasets, and an approach has been developed to design experiments such that important regulatory timepoints can be selected in new experiments, leading to more accurate treatment/condition-response novel gene discovery in the dataset produced.

References

Adams, S. and Carre, I.A. (2011) Downstream of the plant circadian clock: output pathways for the control of physiology and development. Essays Biochem, **49**(1), 53-69.

Allison, D.B., Cui, X., Page, G.P., Sabripour, M. (2006) Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet, **7**(1), 55-65.

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

Aoki, K., Ogata, Y., Shibata, D. (2007) Approaches for extracting practical information from gene co-expression networks in plant biology. Plant Cell Physiol, **48**(3), 381-390.

Arbeitman, M.N., Furlong, E.E.M., Imam, F., Johnson, E., Null, B.H., Baker, B.S., Krasnow, M.A., Scott, M.P., Davis, R.W., White, K.P. (2002) Gene Expression During the Life Cycle of Drosophila melanogaster. Science, **297**(5590), 2270-2275.

Armstrong, G.A., Runge, S., Frick, G., Sperling, U., Apel, K. (1995) Identification of NADPH:protochlorophyllide oxidoreductases A and B: a branched pathway for light-dependent chlorophyll biosynthesis in Arabidopsis thaliana. Plant Physiol, **108**(4), 1505-1517.

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

Athanasiou, K., Dyson, B.C., Webster, R.E., Johnson, G.N. (2010) Dynamic acclimation of photosynthesis increases plant fitness in changing environments. Plant Physiol, **152**(1), 366-373.

Ayroles, J.F. and Gibson, G. (2006) Analysis of variance of microarray data. Methods Enzymol, **411**, 214-233.

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

Bar-Joseph, Z., Demaine, E.D., Gifford, D.K., Srebro, N., Hamel, A.M., Jaakkola, T.S. (2003a) K-ary clustering with optimal leaf ordering for gene expression data. Bioinformatics, **19**(9), 1070-1078.

Bar-Joseph, Z., Gerber, G., Simon, I., Gifford, D.K., Jaakkola, T.S. (2003b) Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. Proc Natl Acad Sci U S A, **100**(18), 10146-10151.

Bar-Joseph, Z., Gerber, G.K., Gifford, D.K., Jaakkola, T.S., Simon, I. (2003c) Continuous representations of time-series gene expression data. Journal of Computational Biology, **10**(3-4), 341-356.

Bellazzi, R. and Zupan, B. (2007) Towards knowledge-based gene expression data mining. J Biomed Inform, **40**(6), 787-802.

Benesty, J., Jingdong, C., Yiteng, H. (2004) Time-delay estimation via linear interpolation and cross correlation. Speech and Audio Processing, IEEE Transactions on, **12**(5), 509-519.

Bhatia, S., Gangappa, S.N., Kushwaha, R., Kundu, S., Chattopadhyay, S. (2008) SHORT HYPOCOTYL IN WHITE LIGHT1, a serine-arginine-aspartate-rich protein in Arabidopsis, acts as a negative regulator of photomorphogenic growth. Plant Physiol, **147**(1), 169-178.

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

Bilgin, D.D., Zavala, J.A., Zhu, J., Clough, S.J., Ort, D.R., DeLucia, E.H. (2010) Biotic stress globally downregulates photosynthesis genes. Plant Cell and Environment, **33**(10), 1597-1613.

Boerjan, W., Ralph, J., Baucher, M. (2003) Lignin biosynthesis. Annu Rev Plant Biol, **54**, 519-546.

Bolstad, B.M., Irizarry, R.A., Astrand, M., Speed, T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, **19**(2), 185-193.

Bozdech, Z., Llinás, M., Pulliam, B.L., Wong, E.D., Zhu, J., DeRisi, J.L. (2003) The Transcriptome of the Intraerythrocytic Developmental Cycle of Plasmodium falciparum. PLoS Biol, **1**(1), e5.

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

Callow, M.J., Dudoit, S., Gong, E.L., Speed, T.P., Rubin, E.M. (2000) Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. Genome Res, **10**(12), 2022-2029.

Carbon, S., Ireland, A., Mungall, C.J., Shu, S., Marshall, B., Lewis, S. (2009) AmiGO: online access to ontology and annotation data. Bioinformatics, **25**(2), 288-289.

Carpenter, C.D., Kreps, J.A., Simon, A.E. (1994) Genes encoding glycine-rich Arabidopsis thaliana proteins with RNA-binding motifs are influenced by cold treatment and an endogenous circadian rhythm. Plant Physiol, **104**(3), 1015-1025.

Chatterjee, P., Mukherjee, S., Chaudhuri, S., Seetharaman, G. (2009) Application Of PapoulisGerchberg Method In Image Super-Resolution and Inpainting. Computer Journal, **52**(1), 80-89.

Chaves, M.M., Flexas, J., Pinheiro, C. (2009) Photosynthesis under drought and salt stress: regulation mechanisms from whole plant to cell. Ann Bot, **103**(4), 551-560.

Chen, C. and Mangasarian, O.L. (1995) Smoothing methods for convex inequalities and linear complementarity problems. Mathematical Programming, **71**(1), 51-69.

Chen, R.H., Brady, D.M., Smith, D., Murray, A.W., Hardwick, K.G. (1999) The spindle checkpoint of budding yeast depends on a tight complex between the Mad1 and Mad2 proteins. Mol Biol Cell, **10**(8), 2607-2618.

Chen, Z., Luan, D., Riofrio, L., Ma, A. (2010) A study on the focusing power of dynamic photon painting. The 52nd Annual Meeting of American Association of Physicists in Medicine (AAPM), AbstractID: 12676.

Chiappetta, P., Roubaud, M.C., Torresani, B. (2004) Blind source separation and the analysis of microarray data. J Comput Biol, **11**(6), 1090-1109.

Chikina, M.D., Huttenhower, C., Murphy, C.T., Troyanskaya, O.G. (2009) Global prediction of tissue-specific gene expression and context-dependent gene networks in Caenorhabditis elegans. PLoS Comput Biol, **5**(6), e1000417.

Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., Davis, R.W. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell, **2**(1), 65-73.

Chung, H.S., Koo, A.J., Gao, X., Jayanty, S., Thines, B., Jones, A.D., Howe, G.A. (2008) Regulation and function of Arabidopsis JASMONATE ZIM-domain genes in response to wounding and herbivory. Plant Physiol, **146**(3), 952-964.

Coello, C.A. (1999) A Comprehensive Survey of Evolutionary-Based Multi-objective Optimization Techniques. Knowledge and Information Systems, **1**(3), 129-156.

Cooper, S. and Shedden, K. (2003) Microarray analysis of gene expression during the cell cycle. Cell Chromosome, **2**(1), 1.

Craigon, D.J., James, N., Okyere, J., Higgins, J., Jotham, J., May, S. (2004) NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. Nucleic Acids Res, **32**(Database issue), D575-577.

Cui, X. and Churchill, G.A. (2003) Statistical tests for differential expression in cDNA microarray experiments. Genome Biol, **4**(4), 210.

Dejean, S., Martin, P.G., Baccini, A., Besse, P. (2007) Clustering time-series gene expression data using smoothing spline derivatives. EURASIP J Bioinform Syst Biol, 70561.

Dembélé, D. and Kastner, P. (2003) Fuzzy C-means method for clustering microarray data. Bioinformatics, **19**(8), 973-980.

Ding, F., Manosas, M., Spiering, M.M., Benkovic, S.J., Bensimon, D., Allemand, J.-F., Croquette, V. (2012) Single-molecule mechanical identification and sequencing. Nat Meth, **advance online publication**.

Dodd, A.N., Salathia, N., Hall, A., Kevei, E., Toth, R., Nagy, F., Hibberd, J.M., Millar, A.J., Webb, A.A. (2005) Plant circadian clocks increase photosynthesis, growth, survival, and competitive advantage. Science, **309**(5734), 630-633.

Dong, M.A., Farre, E.M., Thomashow, M.F. (2011) CIRCADIAN CLOCK-ASSOCIATED 1 and LATE ELONGATED HYPOCOTYL regulate expression of the C-REPEAT BINDING FACTOR (CBF) pathway in Arabidopsis. Proc Natl Acad Sci U S A, **108**(17), 7241-7246.

Dongen, S.V. (2008) Graph Clustering Via a Discrete Uncoupling Process. SIAM J. Matrix Anal. Appl., **30**(1), 121-141.

Edgar, R., Domrachev, M., Lash, A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res, **30**(1), 207-210.

Edwards, K.D., Anderson, P.E., Hall, A., Salathia, N.S., Locke, J.C.W., Lynn, J.R., Straume, M., Smith, J.Q., Millar, A.J. (2006) FLOWERING LOCUS C Mediates Natural Variation in the High-Temperature Response of the Arabidopsis Circadian Clock. The Plant Cell Online, **18**(3), 639-650.

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences, **95**(25), 14863-14868.

Ernst, J. and Bar-Joseph, Z. (2006) STEM: a tool for the analysis of short time series gene expression data. BMC Bioinformatics, **7**, 191.

Ernst, J., Nau, G.J., Bar-Joseph, Z. (2005) Clustering short time series gene expression data. Bioinformatics, **21 Suppl 1**, i159-168.

Espinoza, C., Bieniawska, Z., Hincha, D.K., Hannah, M.A. (2008) Interactions between the circadian clock and cold-response in Arabidopsis. Plant Signal Behav, **3**(8), 593-594.

Espinoza, C., Degenkolbe, T., Caldana, C., Zuther, E., Leisse, A., Willmitzer, L., Hincha, D.K., Hannah, M.A. (2010) Interaction with diurnal and circadian regulation results in dynamic metabolic and transcriptional changes during cold acclimation in Arabidopsis. PLoS One, **5**(11), e14101.

Fernández-Calvo, P., Chini, A., Fernández-Barbero, G., Chico, J.-M., Gimenez-Ibanez, S., Geerinck, J., Eeckhout, D., Schweizer, F., Godoy, M., Franco-Zorrilla, J.M., Pauwels, L., Witters, E., Puga, M.I., Paz-Ares, J., Goossens, A., Reymond, P., De Jaeger, G., Solano, R. (2011) The Arabidopsis bHLH Transcription Factors MYC3 and MYC4 Are Targets of JAZ Repressors and Act Additively with MYC2 in the Activation of Jasmonate Responses. The Plant Cell Online, **23**(2), 701-715.

Filkov, V., Skiena, S., Zhi, J. (2002) Analysis techniques for microarray time-series data. J Comput Biol, **9**(2), 317-330.

Fowler, S. and Thomashow, M.F. (2002) Arabidopsis Transcriptome Profiling Indicates That Multiple Regulatory Pathways Are Activated during Cold Acclimation in Addition to the CBF Cold Response Pathway. The Plant Cell Online, **14**(8), 1675-1690.

Fowler, S.G., Cook, D., Thomashow, M.F. (2005) Low temperature induction of Arabidopsis CBF1, 2, and 3 is gated by the circadian clock. Plant Physiol, **137**(3), 961-968.

Frigyesi, A., Veerla, S., Lindgren, D., Hoglund, M. (2006) Independent component analysis reveals new and biologically significant structures in micro array data. BMC Bioinformatics, **7**, 290.

Gilmour, S.J., Zarka, D.G., Stockinger, E.J., Salazar, M.P., Houghton, J.M., Thomashow, M.F. (1998) Low temperature regulation of the Arabidopsis CBF family of AP2 transcriptional activators as an early step in cold-induced COR gene expression. Plant J, **16**(4), 433-442.

Goodspeed, D., Chehab, E.W., Min-Venditti, A., Braam, J., Covington, M.F. (2012) Arabidopsis synchronizes jasmonate-mediated defense with insect circadian behavior. Proc Natl Acad Sci U S A.

Gustafsson, M. and Hornquist, M. (2010) Gene expression prediction by soft integration and the elastic net-best performance of the DREAM3 gene expression challenge. PLoS One, **5**(2), e9134.

Hand, D.J. and Heard, N.A. (2005) Finding groups in gene expression data. J Biomed Biotechnol, **2005**(2), 215-225.

Hannah, M.A., Wiese, D., Freund, S., Fiehn, O., Heyer, A.G., Hincha, D.K. (2006) Natural genetic variation of freezing tolerance in Arabidopsis. Plant Physiol, **142**(1), 98-112.

Harmer, S.L. (2009) The Circadian System in Higher Plants. In Annu Rev Plant Biol, Annual Reviews, Palo Alto: pp 357-377.

Harmer, S.L., Hogenesch, J.B., Straume, M., Chang, H.S., Han, B., Zhu, T., Wang, X., Kreps, J.A., Kay, S.A. (2000) Orchestrated transcription of key pathways in Arabidopsis by the circadian clock. Science, **290**(5499), 2110-2113.

Hestilow, T.J. and Huang, Y. (2009) Clustering of gene expression data based on shape similarity. EURASIP J Bioinform Syst Biol, 195712.

Hoecker, U., Tepperman, J.M., Quail, P.H. (1999) SPA1, a WD-repeat protein specific to phytochrome A signal transduction. Science, **284**(5413), 496-499.

Hu, H., Yan, X., Huang, Y., Han, J., Zhou, X.J. (2005) Mining coherent dense subgraphs across massive biological networks for functional discovery. Bioinformatics, **21 Suppl 1**, i213-221.

Huang, D., Wu, W., Abrams, S.R., Cutler, A.J. (2008) The relationship of drought-related gene expression in Arabidopsis thaliana to hormonal and environmental factors. J Exp Bot, **59**(11), 2991-3007.

Hubble, J., Demeter, J., Jin, H., Mao, M., Nitzberg, M., Reddy, T.B., Wymore, F., Zachariah, Z.K., Sherlock, G., Ball, C.A. (2009) Implementation of GenePattern within the Stanford Microarray Database. Nucleic Acids Res, **37**(Database issue), D898-901.

Hudson, M.E., Lisch, D.R., Quail, P.H. (2003) The FHY3 and FAR1 genes encode transposase-related proteins involved in regulation of gene expression by the phytochrome A-signaling pathway. The Plant Journal, **34**(4), 453-471.

Hundertmark, M. and Hincha, D.K. (2008) LEA (late embryogenesis abundant) proteins and their encoding genes in Arabidopsis thaliana. BMC Genomics, **9**, 118.

Husale, S., Persson, H.H.J., Sahin, O. (2009) DNA nanomechanics allows direct digital detection of complementary DNA and microRNA targets. Nature, **462**(7276), 1075-1078.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., Speed, T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, **4**(2), 249-264.

Islam, S., Kjallquist, U., Moliner, A., Zajac, P., Fan, J.B., Lonnerberg, P., Linnarsson, S. (2011) Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res, **21**(7), 1160-1167.

Ji, L., Mock, K.W., Tan, K.L. (2006) Quick Hierarchical Biclustering on Microarray Gene Expression Data. Sixth IEEE Symposium on BioInformatics and BioEngineering, 110–120.

Jiang, D., Pei, J., Ramanathan, M., Tang, C., Zhang, A. (2004) Mining coherent gene clusters from gene-sample-time microarray data. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, Seattle, WA, USA: pp 430-439.

Jonassen, E., Lea, U., Lillo, C. (2008) HY5 & HYH are positive regulators of nitrate reductase in seedlings and rosette stage plants. Planta, **227**(3), 559-564.

Kalir, S., McClure, J., Pabbaraju, K., Southward, C., Ronen, M., Leibler, S., Surette, M.G., Alon, U. (2001) Ordering Genes in a Flagella Pathway by Analysis of Expression Kinetics from Living Bacteria. Science, **292**(5524), 2080-2083.

Kang, X. and Ni, M. (2006) Arabidopsis SHORT HYPOCOTYL UNDER BLUE1 contains SPX and EXS domains and acts in cryptochrome signaling. Plant Cell, **18**(4), 921-934.

Katoh, A., Uenohara, K., Akita, M., Hashimoto, T. (2006) Early steps in the biosynthesis of NAD in Arabidopsis start with aspartate and occur in the plastid. Plant Physiol, **141**(3), 851-857.

Kerr, M.K., Martin, M., Churchill, G.A. (2000) Analysis of variance for gene expression microarray data. J Comput Biol, **7**(6), 819-837.

Khanna, R., Shen, Y., Toledo-Ortiz, G., Kikis, E.A., Johannesson, H., Hwang, Y.-S., Quail, P.H. (2006) Functional Profiling Reveals That Only a Small Number of Phytochrome-Regulated Early-Response Genes in Arabidopsis Are Necessary for Optimal Deetiolation. The Plant Cell Online, **18**(9), 2157-2171.

Koenig, L. and Youn, E. (2011) Hierarchical Signature Clustering for Time Series Microarray Data: Software Tools and Algorithms for Biological Systems, Springer, New York: pp 57-65.

Kong, W., Vanderburg, C.R., Gunshin, H., Rogers, J.T., Huang, X. (2008) A review of independent component analysis application to microarray gene expression data. Biotechniques, **45**(5), 501-520.

Konishi, T. (2004) Three-parameter lognormal distribution ubiquitously found in cDNA microarray data and its application to parametric data treatment. BMC Bioinformatics, **5**, 5.

Korn, M., Peterek, S., Mock, H.P., Heyer, A.G., Hincha, D.K. (2008) Heterosis in the freezing tolerance, and sugar and flavonoid contents of crosses between Arabidopsis thaliana accessions of widely varying freezing tolerance. Plant Cell Environ, **31**(6), 813-827.

Kreps, J.A., Wu, Y., Chang, H.-S., Zhu, T., Wang, X., Harper, J.F. (2002) Transcriptome Changes for Arabidopsis in Response to Salt, Osmotic, and Cold Stress. Plant Physiol, **130**(4), 2129-2141.

Kumagai, T., Ito, S., Nakamichi, N., Niwa, Y., Murakami, M., Yamashino, T., Mizuno, T. (2008) The common function of a novel subfamily of B-Box zinc finger proteins with reference to circadian-associated events in Arabidopsis thaliana. Biosci Biotechnol Biochem, **72**(6), 1539-1549.

Kurata, N., Nonomura, K., Harushima, Y. (2002) Rice genome organization: the centromere and genome interactions. Ann Bot, **90**(4), 427-435.

Kushwaha, R., Singh, A., Chattopadhyay, S. (2008) Calmodulin7 plays an important role as transcriptional regulator in Arabidopsis seedling development. Plant Cell, **20**(7), 1747-1759.

Lacombe, E., Hawkins, S., Van Doorsselaere, J., Piquemal, J., Goffner, D., Poeydomenge, O., Boudet, A.M., Grima-Pettenati, J. (1997) Cinnamoyl CoA reductase, the first committed enzyme of the lignin branch biosynthetic pathway: cloning, expression and phylogenetic relationships. Plant J, **11**(3), 429-441.

Larkindale, J. and Vierling, E. (2008) Core Genome Responses Involved in Acclimation to High Temperature. Plant Physiol, **146**(2), 748-761.

Lawson, C.L. and Hanson, R.J. (1995) *Solving Least Squares Problems* SIAM, Philidelphia

Lee, B.-h., Henderson, D.A., Zhu, J.-K. (2005) The Arabidopsis Cold-Responsive Transcriptome and Its Regulation by ICE1. The Plant Cell Online, **17**(11), 3155-3175.

Lee, J., He, K., Stolc, V., Lee, H., Figueroa, P., Gao, Y., Tongprasit, W., Zhao, H., Lee, I., Deng, X.W. (2007) Analysis of Transcription Factor HY5 Genomic Binding Sites Revealed Its Hierarchical Role in Light Regulation of Development. The Plant Cell Online, **19**(3), 731-749.

Li, H.M., Altschmied, L., Chory, J. (1994) Arabidopsis mutants define downstream branches in the phototransduction pathway. Genes Dev, **8**(3), 339-349.

Li, J., Brader, G., Palva, E.T. (2004) The WRKY70 Transcription Factor: A Node of Convergence for Jasmonate-Mediated and Salicylate-Mediated Signals in Plant Defense. The Plant Cell Online, **16**(2), 319-331.

Lin, W.H., Ye, R., Ma, H., Xu, Z.H., Xue, H.W. (2004) DNA chip-based expression profile analysis indicates involvement of the phosphatidylinositol signaling pathway in multiple plant responses to hormone and abiotic treatments. Cell Res, **14**(1), 34-45.

Lu, S.X. and Tobin, E.M. (2011) Chromatin remodeling and the circadian clock: Jumonji C-domain containing proteins. Plant Signal Behav, **6**(6).

Lu, Y., Rosenfeld, R., Bar-Joseph, Z. (2006) Identifying cycling genes by combining sequence homology and expression data. Bioinformatics, **22**(14), e314-322.

Luo, W., Friedman, M., Hankenson, K., Woolf, P. (2011) Time series gene expression profiling and temporal regulatory pathway analysis of BMP6 induced osteoblast differentiation and mineralization. BMC Systems Biology, **5**(1), 82.

Ma, S. and Bohnert, H. (2007) Integration of Arabidopsis thaliana stress-related transcript profiles, promoter structures, and cell-specific expression. Genome Biol, **8**(4), R49.

Ma, S., Gong, Q., Bohnert, H.J. (2007) An Arabidopsis gene network based on the graphical Gaussian model. Genome Res, **17**(11), 1614-1625.

Macintyre, G., Bailey, J., Gustaffson, D., Haviv, I., Kowalczyk, A. (2010) Using Gene Ontology annotations in exploratory microarray clustering to understand cancer etiology Pattern Recognition Letters, **31**(14), 2138-2146.

Mangeon, A., Magioli, C., Tarre, E., Cardeal, V., Araujo, C., Falkenbach, E., Rocha, C.A., Rangel-Lima, C., Sachetto-Martins, G. (2010) The tissue expression pattern of the AtGRP5 regulatory region is controlled by a combination of positive and negative elements. Plant Cell Rep, **29**(5), 461-471.

Mao, L., Mackenzie, C., Roh, J.H., Eraso, J.M., Kaplan, S., Resat, H. (2005) Combining microarray and genomic data to predict DNA binding motifs. Microbiology, **151**(Pt 10), 3197-3213.

Mao, L., Van Hemert, J.L., Dash, S., Dickerson, J.A. (2009) Arabidopsis gene co-expression network and its functional modules. BMC Bioinformatics, **10**, 346.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., Gilad, Y. (2008) RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Res, **18**(9), 1509-1517.

Marks II, R.J. (1991) *Introduction to Shannon Sampling and Interpolation Theory* Springer-Verlag, New York, USA.

Marler, M.R., Gehrman, P., Martin, J.L., Ancoli-Israel, S. (2006) The sigmoidally transformed cosine curve: a mathematical model for circadian rhythms with symmetric non-sinusoidal shapes. Stat Med, **25**(22), 3893-3904.

Martinez, J.S., Jeong, D.E., Choi, E., Billings, B.M., Hall, M.C. (2006) Acm1 is a negative regulator of the CDH1-dependent anaphase-promoting complex/cyclosome in budding yeast. Mol Cell Biol, **26**(24), 9162-9176.

Maruyama, K., Sakuma, Y., Kasuga, M., Ito, Y., Seki, M., Goda, H., Shimada, Y., Yoshida, S., Shinozaki, K., Yamaguchi-Shinozaki, K. (2004) Identification of cold-inducible downstream genes of the Arabidopsis DREB1A/CBF3 transcriptional factor using two microarray systems. Plant J, **38**(6), 982-993.

Mas, P. (2008) Circadian clock function in Arabidopsis thaliana: time beyond transcription. Trends Cell Biol, **18**(6), 273-281.

Massonnet, C., Vile, D., Fabre, J., Hannah, M.A., Caldana, C., Lisec, J., Beemster, G.T., Meyer, R.C., Messerli, G., Gronlund, J.T., Perkovic, J., Wigmore, E., May, S., Bevan, M.W., Meyer, C., Rubio-Diaz, S., Weigel, D., Micol, J.L., Buchanan-Wollaston, V., Fiorani, F., Walsh, S., Rinn, B., Gruissem, W., Hilson, P., Hennig, L., Willmitzer, L., Granier, C. (2010) Probing the reproducibility of leaf growth and molecular phenotypes: a comparison of three Arabidopsis accessions cultivated in ten laboratories. Plant Physiol, **152**(4), 2142-2157.

McCarthy, D.J. and Smyth, G.K. (2009) Testing significance relative to a fold-change threshold is a TREAT. Bioinformatics, **25**(6), 765-771.

Meijering, E. (2002) A Chronology of Interpolation: From Ancient Astronomy to Modern Signal and Image Processing. Proceedings of the IEEE, **90**(3), 319-342.

Michael, T.P., Mockler, T.C., Breton, G., McEntee, C., Byer, A., Trout, J.D., Hazen, S.P., Shen, R., Priest, H.D., Sullivan, C.M., Givan, S.A., Yanovsky, M., Hong, F., Kay, S.A., Chory, J. (2008) Network discovery pipeline elucidates conserved time-of-day-specific cis-regulatory modules. PLoS Genet, **4**(2), e14.

Mockler, T.C., Michael, T.P., Priest, H.D., Shen, R., Sullivan, C.M., Givan, S.A., McEntee, C., Kay, S.A., Chory, J. (2007) The DIURNAL project: DIURNAL and circadian expression profiling, model-based pattern matching, and promoter analysis. Cold Spring Harb Symp Quant Biol, **72**, 353-363.

Montgomery, B.L. (2008) Right place, right time: Spatiotemporal light regulation of plant growth and development. Plant Signal Behav, **3**(12), 1053-1060.

Morker, K.H. and Roberts, M.R. (2011) Light exerts multiple levels of influence on the Arabidopsis wound response. Plant Cell Environ, **34**(5), 717-728.

Mueckstein, U., Leparc, G.G., Posekany, A., Hofacker, I., Kreil, D.P. (2010) Hybridization thermodynamics of NimbleGen microarrays. BMC Bioinformatics, **11**, 35.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., Snyder, M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science, **320**(5881), 1344-1349.

Nakamichi, N. (2011) Molecular Mechanisms Underlying the Arabidopsis Circadian Clock. Plant and Cell Physiology.

Nakamichi, N., Kusano, M., Fukushima, A., Kita, M., Ito, S., Yamashino, T., Saito, K., Sakakibara, H., Mizuno, T. (2009) Transcript profiling of an Arabidopsis PSEUDO RESPONSE REGULATOR arrhythmic triple mutant reveals a role for the circadian clock in cold stress response. Plant Cell Physiol, **50**(3), 447-462.

Narusaka, Y., Narusaka, M., Seki, M., Ishida, J., Nakashima, M., Kamiya, A., Enju, A., Sakurai, T., Satoh, M., Kobayashi, M., Tosa, Y., Park, P., Shinozaki, K. (2003) The cDNA Microarray Analysis Using an Arabidopsis pad3 Mutant Reveals the Expression Profiles and Classification of Genes Induced by Alternaria brassicicola Attack. Plant and Cell Physiology, **44**(4), 377-387.

Okoniewski, M.J. and Miller, C.J. (2006) Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. BMC Bioinformatics, **7**, 276.

Onda, Y., Yagi, Y., Saito, Y., Takenaka, N., Toyoshima, Y. (2008) Light induction of Arabidopsis SIG1 and SIG5 transcripts in mature leaves: differential roles of cryptochrome 1 and cryptochrome 2 and dual function of SIG5 in the recognition of plastid promoters. The Plant Journal, **55**(6), 968-978.

Oran Brigham, E. (1988) *The fast Fourier transform and its applications* Prentice-Hall, Inc., Upper Saddle River, NJ, USA: pp.448.

Orfanidis, S. (1995) Introduction to signal processing. Prentice Hall.

Osakabe, Y., Miyata, S., Urao, T., Seki, M., Shinozaki, K., Yamaguchi-Shinozaki, K. (2002) Overexpression of Arabidopsis response regulators, ARR4/ATRR1/IBC7 and ARR8/ATRR3, alters cytokinin responses differentially in the shoot and in callus formation. Biochem Biophys Res Commun, **293**(2), 806-815.

Parkinson, H., Kapushesky, M., Kolesnikov, N., Rustici, G., Shojatalab, M., Abeygunawardena, N., Berube, H., Dylag, M., Emam, I., Farne, A., Holloway, E., Lukk, M., Malone, J., Mani, R., Pilicheva, E., Rayner, T.F., Rezwan, F., Sharma, A., Williams, E., Bradley, X.Z., Adamusiak, T., Brandizi, M., Burdett, T., Coulson, R., Krestyaninova, M., Kurnosov, P., Maguire, E., Neogi, S.G., Rocca-Serra, P., Sansone, S.-A., Sklyar, N., Zhao, M., Sarkans, U., Brazma, A. (2009) ArrayExpress update—from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Res, **37**(suppl 1), D868-D872.

Parodi, S., Muselli, M., Fontana, V., Bonassi, S. (2003) ROC curves are a suitable and flexible tool for the analysis of gene expression profiles. Cytogenet Genome Res, **101**(1), 90-91.

Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A., Ploner, A. (2005) False discovery rate, sensitivity and sample size for microarray studies. Bioinformatics, **21**(13), 3017-3024.

Peddada, S.D., Lobenhofer, E.K., Li, L., Afshari, C.A., Weinberg, C.R., Umbach, D.M. (2003) Gene selection and clustering for time-course and dose–response microarray experiments using order-restricted inference. Bioinformatics, **19**(7), 834-841.

Pramila, T., Wu, W., Miles, S., Noble, W.S., Breeden, L.L. (2006) The Forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S-phase gap in the transcriptional circuitry of the cell cycle. Genes Dev, **20**(16), 2266-2278.

Price, T.S., Baggs, J.E., Curtis, A.M., Fitzgerald, G.A., Hogenesch, J.B. (2008) WAVECLOCK: wavelet analysis of circadian oscillation. Bioinformatics, **24**(23), 2794-2795.

Ptitsyn, A. (2008) Comprehensive analysis of circadian periodic pattern in plant transcriptome. BMC Bioinformatics, **9**(Suppl 9), S18.

Pu, S., Vlasblom, J., Emili, A., Greenblatt, J., Wodak, S.J. (2007) Identifying functional modules in the physical interactome of Saccharomyces cerevisiae. Proteomics, **7**(6), 944-960.

Raychaudhuri, S., Stuart, J.M., Altman, R.B. (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac Symp Biocomput, 455-466.

Rodriguez-Sanchez, L., Rodriguez-Lopez, M., Garcia, Z., Tenorio-Gomez, M., Schvartzman, J.B., Krimer, D.B., Hernandez, P. (2011) The fission yeast rDNA-binding protein Reb1 regulates G1 phase under nutritional stress. J Cell Sci, **124**(Pt 1), 25-34.

Rosa, B.A., Jiao, Y., Oh, S., Montgomery, B.L., Qin, W., Chen, J. (2012a) Frequency-based time-series gene expression recomposition using PRIISM. BMC Systems Biology, **In Press**.

Rosa, B.A., Oh, S., Montgomery, B.L., Chen, J., Qin, W. (2010) Computing gene expression data with a knowledge-based gene clustering approach. International Journal of Biochemistry and Molecular Biology, **1**(1), 51-68.

Rosa, B.A., Zhang, J., Major, I.T., Qin, W., Chen, J. (2012b) Knowledge-based optimal timepoint sampling in high-throughput temporal experiments. **Submitted**.

Ruan, J. (2010) A top-performing algorithm for the DREAM3 gene expression prediction challenge. PLoS One, **5**(2), e8944.

Ruiz, J.M., Sanchez, E., Garcia, P.C., Lopez-Lefebre, L.R., Rivero, R.M., Romero, L. (2002) Proline metabolism and NAD kinase activity in greenbean plants subjected to cold-shock. Phytochemistry, **59**(5), 473-478.

Rustici, G., Mata, J., Kivinen, K., Lio, P., Penkett, C.J., Burns, G., Hayles, J., Brazma, A., Nurse, P., Bahler, J. (2004) Periodic gene expression program of the fission yeast cell cycle. Nat Genet, **36**(8), 809-817.

Salome, P.A., Xie, Q., McClung, C.R. (2008) Circadian timekeeping during early Arabidopsis development. Plant Physiol, **147**(3), 1110-1125.

Sappl, P.G., Onate-Sanchez, L., Singh, K.B., Millar, A.H. (2004) Proteomic analysis of glutathione S - transferases of Arabidopsis thaliana reveals differential salicylic acid-induced expression of the plant-specific phi and tau classes. Plant Mol Biol, **54**(2), 205-219.

Schena, M., Shalon, D., Davis, R.W., Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, **270**(5235), 467-470.

Schliep, A., Steinhoff, C., Schonhuth, A. (2004) Robust inference of groups in gene expression time-courses using mixtures of HMMs. Bioinformatics, **20 Suppl 1**, i283-289.

Schulze, A. and Downward, J. (2001) Navigating gene expression using microarrays--a technology review. Nat Cell Biol, **3**(8), E190-195.

Sinclair, I. and Dunton, J. (2007) *Electronic and Electrical Servicing: Consumer and commercial electronics* Elsevier, Burlington, MA: pp.322.

Singh, R., Palmer, N., Gifford, D., Berger, B., Bar-Joseph, Z. (2005) Active learning for sampling in time-series experiments with application to gene expression analysis. Proceedings of the 22nd international conference on Machine learning, 832 - 839.

Smith, A.A. and Craven, M. (2008) Fast multisegment alignments for temporal expression profiles. Comput Syst Bioinformatics Conf, **7**, 315-326.

Smyth, G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol, **3**, Article3.

Soitamo, A., Piippo, M., Allahverdiyeva, Y., Battchikova, N., Aro, E.-M. (2008) Light has a specific role in modulating Arabidopsis gene expression at low temperature. BMC Plant Biology, **8**(1), 13.

Solecka, D. (1997) Role of phenylpropanoid compounds in plant responses to different stress factors. Acta Physiologiae Plantarum, **19**(3), 257-268.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell, **9**(12), 3273-3297.

Stoevesandt, O., Taussig, M.J., He, M. (2009) Protein microarrays: high-throughput tools for proteomics. Expert Review of Proteomics, **6**(2), 145-157.

Stracke, R., Ishihara, H., Huep, G., Barsch, A., Mehrtens, F., Niehaus, K., Weisshaar, B. (2007) Differential regulation of closely related R2R3-MYB transcription factors controls flavonol accumulation in different parts of the Arabidopsis thaliana seedling. Plant J, **50**(4), 660-677.

Stuart, J.M., Segal, E., Koller, D., Kim, S.K. (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science, **302**(5643), 249-255.

Subramanian, A., Kuehn, H., Gould, J., Tamayo, P., Mesirov, J.P. (2007) GSEA-P: a desktop application for Gene Set Enrichment Analysis. Bioinformatics, **23**(23), 3251-3253.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, **102**(43), 15545-15550.

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T.Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A., Singh, S., Swing, V., Tissier, C., Zhang, P., Huala, E. (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res, **36**(Database issue), D1009-1014.

Syeda-Mahmood, T. (2003) Clustering time-varying gene expression profiles using scale-space signals. Proc IEEE Comput Soc Bioinform Conf, **2**, 48-56.

Tang, B.S.F., Chan, K.-h., Cheng, V.C.C., Woo, P.C.Y., Lau, S.K.P., Lam, C.C.K., Chan, T.-l., Wu, A.K.L., Hung, I.F.N., Leung, S.-y., Yuen, K.-y. (2005) Comparative Host Gene Transcription by Microarray Analysis Early after Infection of the Huh7 Cell Line by Severe Acute Respiratory Syndrome Coronavirus and Human Coronavirus 229E. Journal of Virology, **79**(10), 6180-6193.

Tepperman, J.M., Hudson, M.E., Khanna, R., Zhu, T., Chang, S.H., Wang, X., Quail, P.H. (2004) Expression profiling of phyB mutant demonstrates substantial contribution of other phytochromes to red-light-regulated gene expression during seedling de-etiolation. The Plant Journal, **38**(5), 725-739.

Tepperman, J.M., Hwang, Y.-S., Quail, P.H. (2006) phyA dominates in transduction of red-light signals to rapidly responding genes at the initiation of Arabidopsis seedling de-etiolation. The Plant Journal, **48**(5), 728-742.

Thilmony, R., Underwood, W., He, S.Y. (2006) Genome-wide transcriptional analysis of the Arabidopsis thaliana interaction with the plant pathogen Pseudomonas syringae pv. tomato DC3000 and the human pathogen Escherichia coli O157:H7. Plant J, **46**(1), 34-53.

Thines, B. and Harmon, F.G. (2011) Four easy pieces: mechanisms underlying circadian regulation of growth and development. Curr Opin Plant Biol, **14**(1), 31-37.

Tominaga, D. (2010) Periodicity detection method for small-sample time series datasets. Bioinform Biol Insights, **4**, 127-136.

Treffer, R. and Deckert, V. (2010) Recent advances in single-molecule sequencing. Current Opinion in Biotechnology, **21**(1), 4-11.

Tu, Y., Stolovitzky, G., Klein, U. (2002) Quantitative noise analysis for gene expression microarray experiments. Proc Natl Acad Sci U S A, **99**(22), 14031-14036.

Tusher, V.G., Tibshirani, R., Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A, **98**(9), 5116-5121.

Van Dogen, S. (2000) Graph Clustering by Flow Simulation. In PhD Thesis, University of Utrecht.

Vance, W., Arkin, A., Ross, J. (2002) Determination of causal connectivities of species in reaction networks. Proc Natl Acad Sci U S A, **99**(9), 5816-5821.

Vanderauwera, S., Zimmermann, P., Rombauts, S., Vandenabeele, S., Langebartels, C., Gruissem, W., Inzé, D., Van Breusegem, F. (2005) Genome-Wide Analysis of Hydrogen Peroxide-Regulated Gene Expression in Arabidopsis Reveals a High Light-Induced Transcriptional Cluster Involved in Anthocyanin Biosynthesis. Plant Physiol, **139**(2), 806-821.

Verducci, J.S., Melfi, V.F., Lin, S., Wang, Z., Roy, S., Sen, C.K. (2006) Microarray analysis of gene expression: considerations in data mining and statistical treatment. Physiol Genomics, **25**(3), 355-363.

Vergnolle, C., Vaultier, M.-N., Taconnat, L., Renou, J.-P., Kader, J.-C., Zachowski, A., Ruelland, E. (2005) The Cold-Induced Early Activation of Phospholipase C and D Pathways Determines the Response of Two Distinct Clusters of Genes in Arabidopsis Cell Suspensions. Plant Physiol, **139**(3), 1217-1233.

Vogel, J.T., Zarka, D.G., Van Buskirk, H.A., Fowler, S.G., Thomashow, M.F. (2005) Roles of the CBF2 and ZAT12 transcription factors in configuring the low temperature transcriptome of Arabidopsis. Plant J, **41**(2), 195-211.

Wang, X., Wu, M., Li, Z., Chan, C. (2008) Short time-series microarray analysis: methods and challenges. BMC Syst Biol, **2**, 58.

Wang, Z., Gerstein, M., Snyder, M. (2009a) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet, **10**(1), 57-63.

Wang, Z., Gerstein, M., Snyder, M. (2009b) RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, **10**(1), 57-63.

Warnasooriya, S.N. and Montgomery, B.L. (2009) Detection of spatial-specific phytochrome responses using targeted expression of biliverdin reductase in Arabidopsis. Plant Physiol, **149**(1), 424-433.

Whitfield, M.L., Sherlock, G., Saldanha, A.J., Murray, J.I., Ball, C.A., Alexander, K.E., Matese, J.C., Perou, C.M., Hurt, M.M., Brown, P.O., Botstein, D. (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol Biol Cell, **13**(6), 1977-2000.

Wichert, S., Fokianos, K., Strimmer, K. (2004) Identifying periodically expressed transcripts in microarray time series data. Bioinformatics, **20**(1), 5-20.

Wierstra, I. and Kloppstech, K. (2000) Differential effects of methyl jasmonate on the expression of the early light-inducible proteins and other light-regulated genes in barley. Plant Physiol, **124**(2), 833-844.

Wu, F.X. (2008) Genetic weighted k-means algorithm for clustering large-scale gene expression data. BMC Bioinformatics, **9 Suppl 6**, S12.

Wu, H., Su, Z., Mao, F., Olman, V., Xu, Y. (2005) Prediction of functional modules based on comparative genome analysis and Gene Ontology application. Nucleic Acids Res, **33**(9), 2822-2837.

Xie, D.X., Feys, B.F., James, S., Nieto-Rostro, M., Turner, J.G. (1998) COI1: an Arabidopsis gene required for jasmonate-regulated defense and fertility. Science, **280**(5366), 1091-1094.

Xin, Z., Mandaokar, A., Chen, J., Last, R.L., Browse, J. (2007) Arabidopsis ESK1 encodes a novel regulator of freezing tolerance. The Plant Journal, **49**(5), 786-799.

Yang, J.J. and Yang, M.C. (2006) An improved procedure for gene selection from microarray experiments using false discovery rate criterion. BMC Bioinformatics, **7**, 15.

Yang, R. and Su, Z. (2010) Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation. Bioinformatics, **26**(12), i168-174.

Yao, J., Roy-Chowdhury, S., Allison, L.A. (2003) AtSig5 Is an Essential Nucleus-Encoded Arabidopsis σ-Like Factor. Plant Physiol, **132**(2), 739-747.

Yu, H., Luscombe, N.M., Qian, J., Gerstein, M. (2003) Genomic analysis of gene expression relationships in transcriptional regulatory networks. Trends Genet, **19**(8), 422-427.

Yuan, J.S., Reed, A., Chen, F., Stewart, C.N., Jr. (2006) Statistical analysis of real-time PCR data. BMC Bioinformatics, **7**, 85.

Zeidler, M., Zhou, Q., Sarda, X., Yau, C.P., Chua, N.H. (2004) The nuclear localization signal and the C-terminal region of FHY1 are required for transmission of phytochrome A signals. Plant J, **40**(3), 355-365.

Zhang, B. and Horvath, S. (2005) A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol, **4**, Article17.

Zhang, Y., Zheng, S., Liu, Z., Wang, L., Bi, Y. (2011) Both HY5 and HYH are necessary regulators for low temperature-induced anthocyanin accumulation in Arabidopsis seedlings. Journal of Plant Physiology, **168**(4), 367-374.